

Methods in  
Molecular Biology 1386

Springer Protocols

Ulf Schmitz  
Olaf Wolkenhauer *Editors*

# Systems Medicine

 Humana Press

# METHODS IN MOLECULAR BIOLOGY

*Series Editor*  
**John M. Walker**  
School of Life and Medical Sciences  
University of Hertfordshire  
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:  
<http://www.springer.com/series/7651>



# **Systems Medicine**

Edited by

**Ulf Schmitz**

*Department of Systems Biology & Bioinformatics, University of Rostock, Rostock, Germany*

**Olaf Wolkenhauer**

*Department of Systems Biology & Bioinformatics, University of Rostock, Rostock, Germany*



*Editors*

Ulf Schmitz  
Department of Systems Biology & Bioinformatics  
University of Rostock  
Rostock, Germany

Olaf Wolkenhauer  
Department of Systems Biology & Bioinformatics  
University of Rostock  
Rostock, Germany

ISSN 1064-3745                      ISSN 1940-6029 (electronic)  
Methods in Molecular Biology  
ISBN 978-1-4939-3282-5            ISBN 978-1-4939-3283-2 (eBook)  
DOI 10.1007/978-1-4939-3283-2

Library of Congress Control Number: 2015956342

Springer New York Heidelberg Dordrecht London  
© Springer Science+Business Media New York 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Humana Press is a brand of Springer  
Springer Science+Business Media LLC New York is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

---

## Preface

It is now more than a decade ago since systems biology was celebrated as a new promising approach that can predict the behavior of cellular systems through mathematical modeling and simulation. It was the hope of many that, once systems biology helped deciphering mechanisms underlying disease emergence and progression, it is only a short path toward the design of novel, more successful therapeutic approaches.

Although technological advancements allow the generation of ever more detailed snapshots of life across multiple levels of temporal and spatial scales, and despite a wealth of new insights on how life is organized, it is a long way before we are able to translate this improved understanding to achieve a sustainable impact on clinical practice.

To guide this path, there is a need to survey the diverse approaches, the multitude of methodologies, and the myriad of tools that should and will be integrated into customized protocols and workflows for the reliable prognoses of disease outbreak and course, for the identification of therapeutic targets, the development of targeted therapies for individual patients, and for monitoring therapy success and patient well-being.

This book presents trends, initiatives, and recent developments in this emerging field called systems medicine, which has the goal of finding solutions to the challenges described above. We are glad that leading experts have contributed to this first book on systems medicine and provided their insights into the state of the art in the field.

## Structure of the Book

This book is structured in four parts. The first part, “A Road Map Toward Systems Medicine,” consists of six chapters that outline the field of systems medicine by defining the terminology and describing how established computational methods from bioinformatics and systems biology can be taken forward to an integrative systems medicine approach. One chapter describes the necessity for redefining training curricula for medical and computational students, and in two chapters the impact is discussed that a systems medicine approach possibly has on handling diseases and patients and on the pharmaceutical industry.

The second part of the book, “Opinions and Perspectives,” provides an outlook on the role that systems medicine may or should play in various medical fields like oncology, neurology, the study of lung diseases, immune-related diseases and therapies, and infectious diseases.

In Part III case studies are presented that demonstrate different facets of the systems medicine approach in action to study, e.g., the human metabolism, chronic obstructive pulmonary disease, transcriptomics, and regenerative stem cell medicine. These chapters nicely illustrate the interdisciplinary combination of computational methods with wet lab experiments.

The fourth part of the book, “Tools and Methodologies,” contains four chapters that introduce tools, resources, and methodologies from bioinformatics and systems biology and shows how to apply these in a systems medicine project.

*Rostock, Germany*

*Ulf Schmitz  
Olaf Wolkenhauer*



---

## **Acknowledgements**

We thank all authors for finding the time in their busy schedules and making the effort contributing to this book with their highly engaging chapters.



---

# Contents

<i>Preface</i> . . . . .	<i>v</i>
<i>Contributors</i> . . . . .	<i>xi</i>
PART I A ROAD MAP TOWARD SYSTEMS MEDICINE	
1 Systems Medicine: Sketching the Landscape . . . . . <i>Marc Kirschner</i>	3
2 Taking Bioinformatics to Systems Medicine . . . . . <i>Antoine H.C. van Kampen and Perry D. Moerland</i>	17
3 Systems Medicine: The Future of Medical Genomics, Healthcare, and Wellness . . . . . <i>Mansoor Saqi, Johann Pellet, Irina Roznovat, Alexander Mazein, Stéphane Ballereau, Bertrand De Meulder, and Charles Auffray</i>	43
4 Next-Generation Pathology . . . . . <i>Peter D. Caie and David J. Harrison</i>	61
5 Training in Systems Approaches for the Next Generation of Life Scientists and Medical Doctors . . . . . <i>Damjana Rozman, Jure Acimovic, and Bernd Schmeck</i>	73
6 Systems Medicine in Pharmaceutical Research and Development . . . . . <i>Lars Kuepfer and Andreas A. Schuppert</i>	87
PART II OPINIONS AND PERSEPECTIVES	
7 Systems Medicine and Infection . . . . . <i>Ruth Bowness</i>	107
8 Systems Medicine for Lung Diseases: Phenotypes and Precision Medicine in Cancer, Infection, and Allergy . . . . . <i>Bernd Schmeck, Wilhelm Bertrams, Xin Lai, and Julio Vera</i>	119
9 Third-Kind Encounters in Biomedicine: Immunology Meets Mathematics and Informatics to Become Quantitative and Predictive . . . . . <i>Martin Eberhardt, Xin Lai, Namrata Tomar, Shailendra Gupta, Bernd Schmeck, Alexander Steinkasserer, Gerold Schuler, and Julio Vera</i>	135
10 Systems Medicine in Oncology: Signaling Network Modeling and New-Generation Decision-Support Systems . . . . . <i>Silvio Parodi, Giuseppe Riccardi, Nicoletta Castagnino, Lorenzo Tortolina, Massimo Maffei, Gabriele Zoppoli, Alessio Nencioni, Alberto Ballestrero, and Franco Patrone</i>	181
11 Neurological Diseases from a Systems Medicine Point of View . . . . . <i>Marek Ostaszewski, Alexander Skupin, and Rudi Balling</i>	221

PART III SYSTEMS MEDICINE PROJECTS AND CASE STUDIES

12 Computational Modeling of Human Metabolism and Its Application  
to Systems Biomedicine. . . . . 253  
*Maïke K. Aurich and Ines Thiele*

13 From Systems Understanding to Personalized Medicine:  
Lessons and Recommendations Based on a Multidisciplinary  
and Translational Analysis of COPD . . . . . 283  
*Josep Roca, Isaac Cano, David Gomez-Cabrero, and Jesper Tegnér*

14 RNA Systems Biology for Cancer: From Diagnosis to Therapy . . . . . 305  
*Rabeleh Amirkhah, Ali Farazmand, Olaf Wolkenhauer, and Ulf Schmitz*

15 Mathematical Models of Pluripotent Stem Cells: At the Dawn  
of Predictive Regenerative Medicine . . . . . 331  
*Pinar Pir and Nicolas Le Novère*

PART IV TOOLS AND METHODOLOGIES

16 Network-Assisted Disease Classification and Biomarker Discovery . . . . . 353  
*Sonja Strunz, Olaf Wolkenhauer, and Alberto de la Fuente*

17 Anatomy and Physiology of Multiscale Modeling and Simulation  
in Systems Medicine . . . . . 375  
*Alexandru Mizeranschi, Derek Groen, Joris Borgdorff,  
Alfons G. Hoekstra, Bastien Chopard, and Werner Dubitzky*

18 Mathematical and Statistical Techniques for Systems Medicine:  
The Wnt Signaling Pathway as a Case Study . . . . . 405  
*Adam L. MacLean, Heather A. Harrington, Michael P.H. Stumpf,  
and Helen M. Byrne*

19 Modeling and Simulation Tools: From Systems Biology  
to Systems Medicine . . . . . 441  
*Brett G. Olivier, Maciej J. Swat, and Martijn J. Moné*

*Index* . . . . . 465

---

## Contributors

- JURE ACIMOVIC • *Centre for Functional Genomics and Bio-Chips, Institute of Biochemistry, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia*
- RAHELEH AMIRKHAH • *Department of Cell and Molecular Biology, School of Biology, College of Science, University of Tehran, Tehran, Iran*
- CHARLES AUFFRAY • *European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL, Université de Lyon, Lyon, France; Université Claude Bernard, Lyon, Cedex 07, France*
- MAIKE K. AURICH • *Luxembourg Center for Systems Biomedicine, University of Luxembourg, Luxembourg, UK*
- STÉPHANE BALLEREAU • *European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL, Université de Lyon, Lyon, France*
- ALBERTO BALLESTRERO • *Department of Internal Medicine (DIMI), Genoa University, Genoa, Italy*
- RUDI BALLING • *Luxembourg Centre for Systems Biomedicine (LCSB), Université du Luxembourg, Luxembourg, UK*
- WILHELM BERTRAMS • *Systems Biology Platform, Institute for Lung Research/iLung, German Center for Lung Research, Universities of Giessen and Marburg Lung Centre, Philipps-University Marburg, Marburg, Germany*
- JORIS BORGENDORFF • *Netherlands eScience Center, Amsterdam, The Netherlands*
- RUTH BOWNESS • *School of Medicine, University of St Andrews, St Andrews, UK*
- HELEN M. BYRNE • *Department of Life Sciences, Imperial College London, London, UK*
- PETER D. CAIE • *Quantitative and systems pathology, University of St Andrews, St Andrews, UK*
- ISAAC CANO • *Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Bunyola, Balearic Islands; IDIBAPS, Hospital Clínic, CIBERES, Universitat de Barcelona, Barcelona, Spain*
- NICOLETTA CASTAGNINO • *Department of Internal Medicine (DIMI), Genoa University, Genoa, Italy*
- BASTIEN CHOPARD • *Computer Science Department, University of Geneva, Carouge, Switzerland*
- BERTRAND DE MEULDER • *European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL, Université de Lyon, Lyon, France*
- WERNER DUBITZKY • *Biomedical Sciences Research Institute, University of Ulster, Londonderry, UK; School of Biomedical Sciences, University of Ulster, Londonderry, UK*
- MARTIN EBERHARDT • *Laboratory of Systems Tumor Immunology, Department of Dermatology, University Hospital Erlangen and Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany*
- ALI FARAZMAND • *Department of Cell and Molecular Biology, School of Biology, College of Science, University of Tehran, Tehran, Iran*
- ALBERTO DE LA FUENTE • *Biomathematics and Bioinformatics Unit, Leibniz-Institute for Farm Animal Biology Biomathematics and Bioinformatics Unit, Dummerstorf, Germany*



- DAVID GOMEZ-CABRERO • *Unit of Computational Medicine, Department of Medicine Solna, Center for Molecular Medicine, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden*
- DEREK GROEN • *Chemistry Department, Centre for Computational Science, University College London, London, UK*
- SHAILENDRA GUPTA • *Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany*
- HEATHER A. HARRINGTON • *Department of Life Sciences, Imperial College London, London, UK*
- DAVID J. HARRISON • *Quantitative and systems pathology, University of St Andrews, St Andrews, UK*
- ALFONS G. HOEKSTRA • *Computational Science Lab, Faculty of Science, Institute for Informatics, University of Amsterdam, Amsterdam, The Netherlands; Advanced Computing Lab, ITMO University, St. Petersburg, Russia*
- ANTOINE H.C. VAN KAMPEN • *Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center (AMC), University of Amsterdam, Amsterdam, The Netherlands; Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands*
- MARC KIRSCHNER • *Forschungszentrum Jülich GmbH, Projektträger Jülich, Molekulare Lebenswissenschaften, Jülich, Germany*
- LARS KUEPFER • *Computational Systems Biology, Bayer Technology Services GmbH, Leverkusen, Germany; Institute of Applied Microbiology, RWTH Aachen University, Aachen, Germany*
- XIN LAI • *Laboratory of Systems Tumor Immunology, Department of Dermatology, Faculty of Medicine, University of Erlangen-Nurnberg, Erlangen, Germany University Hospital Erlangen and Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany*
- NICOLAS LE NOVÈRE • *Babraham Institute, Babraham Research Campus, Cambridge, UK*
- ADAM L. MACLEAN • *Mathematical Institute, University of Oxford, Oxford, UK; Department of Life Sciences, Imperial College London, London, UK*
- MASSIMO MAFFEI • *Department of Internal Medicine (DIMI), Genoa University, Genoa, Italy*
- ALEXANDER MAZEIN • *European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL, Université de Lyon, Lyon, France*
- ALEXANDRU MIZERANSCHI • *Biomedical Sciences Research Institute, University of Ulster, Londonderry, UK*
- PERRY D. MOERLAND • *Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center (AMC), University of Amsterdam, Amsterdam, The Netherlands; Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands*
- MARTIJN J. MONÉ • *Molecular Cell Physiology, VU University Amsterdam, Amsterdam, The Netherlands; Systems and Synthetic Biology, Wageningen University, Wageningen, The Netherlands*
- ALESSIO NENCIONI • *Department of Internal Medicine (DIMI), Genoa University, Genoa, Italy*

- BRETT G. OLIVIER • *Systems Bioinformatics, VU University Amsterdam, Amsterdam, The Netherlands*
- MAREK OSTASZEWSKI • *Luxembourg Centre for Systems Biomedicine (LCSB), Université du Luxembourg, Luxembourg, UK*
- SILVIO PARODI • *Department of Internal Medicine (DIMI), Genoa University, Genoa, Italy*
- FRANCO PATRONE • *Department of Internal Medicine (DIMI), Genoa University, Genoa, Italy*
- JOHANN PELLET • *European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL, Université de Lyon, Lyon, France*
- PINAR PIR • *Babraham Institute, Babraham Research Campus, Cambridge, UK*
- GIUSEPPE RICCARDI • *Signals and Interactive Systems lab, Department of Engineering and Information Science, Trento University, Trento, Italy*
- JOSEP ROCA • *Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Bunyola, Balearic Islands; IDIBAPS, Hospital Clínic, CIBERES, Universitat de Barcelona, Barcelona, Spain*
- DAMJANA ROZMAN • *Centre for Functional Genomics and Bio-Chips, Institute of Biochemistry, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia*
- IRINA ROZNOVAT • *European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL, Université de Lyon, Lyon, France*
- MANSOOR SAQI • *European Institute for Systems Biology and Medicine, CNRS-ENS-UCBL, Université de Lyon, Lyon, France*
- BERND SCHMECK • *Department of Medicine, Pulmonary and Critical Care Medicine, University Medical Center, Philipps-University Marburg, Marburg, Germany; Systems Biology Platform, Institute for Lung Research/iLung, German Center for Lung Research, Universities of Giessen and Marburg Lung Centre, Philipps-University, Marburg, Marburg, Germany*
- ULF SCHMITZ • *Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany*
- GEROLD SCHULER • *Department of Dermatology, University Hospital Erlangen and Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany*
- ANDREAS SCHUPPERT • *Lehrstuhl für datenbasierte Modellierung in CES, Joint Research Center for Computational Biomedicine, AICES RWTH Aachen University, Aachen, Germany*
- ALEXANDER SKUPIN • *Luxembourg Centre for Systems Biomedicine (LCSB), Université du Luxembourg, Luxembourg, UK*
- ALEXANDER STEINKASSERER • *Department of Immune Modulation at the Department of Dermatology, University Hospital Erlangen, Erlangen, Germany*
- SONJA STRUNZ • *Biomathematics and Bioinformatics Unit, Leibniz-Institute for Farm Animal Biology (FBN), Institute of Genetics and Biometry, Dummerstorf, Germany*
- MICHAEL P.H. STUMPF • *Mathematical Institute, University of Oxford, Oxford, UK*
- MACIEJ J. SWAT • *EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridgeshire, UK*
- JESPER TEGNÉR • *Unit of Computational Medicine, Department of Medicine Solna, Center for Molecular Medicine, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden*
- INES THIELE • *Luxembourg Center for Systems Biomedicine, University of Luxembourg, Luxembourg, UK*

NAMRATA TOMAR • *Laboratory of Systems Tumor Immunology, Department of Dermatology, Faculty of Medicine, University of Erlangen-Nurnberg, Erlangen, Germany; University Hospital Erlangen and Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany*

LORENZO TORTOLINA • *Department of Internal Medicine (DIMI), Genoa University, Genoa, Italy*

JULIO VERA • *Laboratory of Systems Tumor Immunology, Department of Dermatology, Faculty of Medicine, University of Erlangen-Nurnberg, Erlangen, Germany*

OLAF WOLKENHAUER • *Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany; Stellenbosch Institute for Advanced Study (STIAS), Wallenberg Research Centre at Stellenbosch University, Stellenbosch, South Africa*

GABRIELE ZOPPOLI • *Department of Internal Medicine (DIMI), Genoa University, Genoa, Italy*

# **Part I**

## **A Road Map Toward Systems Medicine**

# Chapter 1

## Systems Medicine: Sketching the Landscape

Marc Kirschner

### Abstract

To understand the meaning of the term Systems Medicine and to distinguish it from seemingly related other expressions currently in use, such as precision, personalized, -omics, or big data medicine, its underlying history and development into present time needs to be highlighted. Having this development in mind, it becomes evident that Systems Medicine is a genuine concept as well as a novel way of tackling the manifold complexity that occurs in nowadays clinical medicine—and not just a rebranding of what has previously been done in the past. So looking back it seems clear to many in the field that Systems Medicine has its origin in an integrative method to unravel biocomplexity, namely, Systems Biology. Here scientist by now gained useful experience that is on the verge toward implementation in clinical research and practice.

Systems Medicine and Systems Biology have the same underlying theoretical principle in systems-based thinking—a methodology to understand complexity that can be traced back to ancient Greece. During the last decade, however, and due to a rapid methodological development in the life sciences and computing/IT technologies, Systems Biology has evolved from a scientific concept into an independent discipline most competent to tackle key questions of biocomplexity—with the potential to transform medicine and how it will be practiced in the future. To understand this process in more detail, the following section will thus give a short summary of the foundation of systems-based thinking and the different developmental stages including systems theory, the development of modern Systems Biology, and its transition into clinical practice. These are the components to pave the way toward Systems Medicine.

**Key words** Biocomplexity, Systems-based thinking, Systems Biology, Systems Medicine, 4P medicine, Clinical translation, Multidisciplinarity, Integration

---

### 1 What Is Systems Medicine?

Defining Systems Medicine is a necessary task in order to sell the concept and to define its scope and limits. However, a definition, especially for such a young discipline, will be adopted over time according to the specific needs and shouldn't be too explicit or narrow to allow for enough developmental space. Based on the actual experiences, several working definitions have been used in publications, projects, and presentations that describe Systems Medicine and clinically relevant topics. Table 1 will thus focus on some selected definitions and give a short overview of their key points.

**Table 1**  
**Current working definitions of Systems Medicine**

Scientific perspectives
<p><i>Systems Medicine is the implementation of Systems Biology approaches in medical concepts, research, and practice. This involves iterative and reciprocal feedback between clinical investigations and practice with computational, statistical, and mathematical multiscale analysis and modeling of pathogenetic mechanisms, disease progression and remission, disease spread and cure, treatment responses and adverse events, as well as disease prevention both at the epidemiological and individual patient level. As an outcome Systems Medicine aims at a measurable improvement of patient health through systems-based approaches and practice (CASyM—Coordinating Action Systems Medicine; <a href="http://www.casym.eu">www.casym.eu</a>) [40]</i></p>
<p><i>An integrative approach to medical needs taking advantage and emphasizing information and tools made available by the greatest possible spectrum of scientific disciplines aimed at improving risk prediction and individual treatment respecting ethical and legal requirements. This approach should improve medical practice by standardization, information, integration, monitoring, and personalization [41]</i></p>
<p><i>With the availability of increasingly powerful high-throughput technologies, computational tools, and integrated knowledge bases, it has become possible to establish new links between genes, biological functions, and a wide range of human diseases. This is providing signatures of pathological biology and links to clinical research and drug discovery. These are the hallmarks of Systems Medicine as it is emerging from the initial, more targeted efforts of medical genomics. In addition to genomics and Systems Biology, the key components that will ensure the successful development of Systems Medicine are the modeling of physiopathology in a clinical-practice context, imaging, and biobanking that complies with strictly enforced ethical regulations [42]</i></p>
Funder perspectives
<p><i>Systems Medicine is the application of Systems Biology approaches to medical research and medical practice. Its objective is to integrate a variety of biological/medical data at all relevant levels of cellular organization using the power of computational and mathematical modeling, to enable understanding of the pathophysiological mechanisms, prognosis, diagnosis, and treatment of disease [38]</i></p>
<p><i>Systems Medicine uses systems-oriented approaches, in both research and clinical care, to illuminate complex physiological and pathological processes and, thereby, to create a basis for development of innovative therapies and preventive measures [39]</i></p>

By taking these selected definitions into account, it becomes possible to derive a strategic and scientific principle of Systems Medicine (O. Wolkenhauer, personal communication, modified):

*Strategic definition*—Systems Medicine is an interdisciplinary approach to improve diagnosis, prognosis, and therapy, through the integration of multiple data and the integration of expertise from all relevant disciplines including biostatistics, bioinformatics, health informatics, and Systems Biology into clinical research. *Scientific definition*—Systems Medicine is the science that studies how physiological functions emerge from the interactions between cells and tissues and how this influences the behavior of these components in the human body.

In essence, Systems Medicine is a way of thinking, a conceptual framework, focused on outcome and impact, rather than a theoretical discipline; it aims at a measurable improvement of patient health through systems approaches [1].

---

## 2 Development of a Systems Theory of Life to Tackle Biocomplexity

A systems-oriented strategy toward understanding complexity is not a modern time invention. Instead, it has a long history, and parts of its theoretical foundations can already be found in a theorem postulated by Aristotle: *What is composed of parts so that it forms a uniform whole, not just a bunch, but rather like [...] a syllable, which obviously more than the sum of its parts* [2]. In essence, the whole is more than the sum of its parts.

This simple statement makes clear that, when applied to biological problems and the underlying organization of biocomplexity, it is simply not enough to study all single components of a system—biology itself cannot be subdivided into ever smaller, and thus more understandable, units without losing crucial information about the whole. So, it took until the twentieth century that modern biology was able to push the limits and transform systems thinking into a practical discipline.

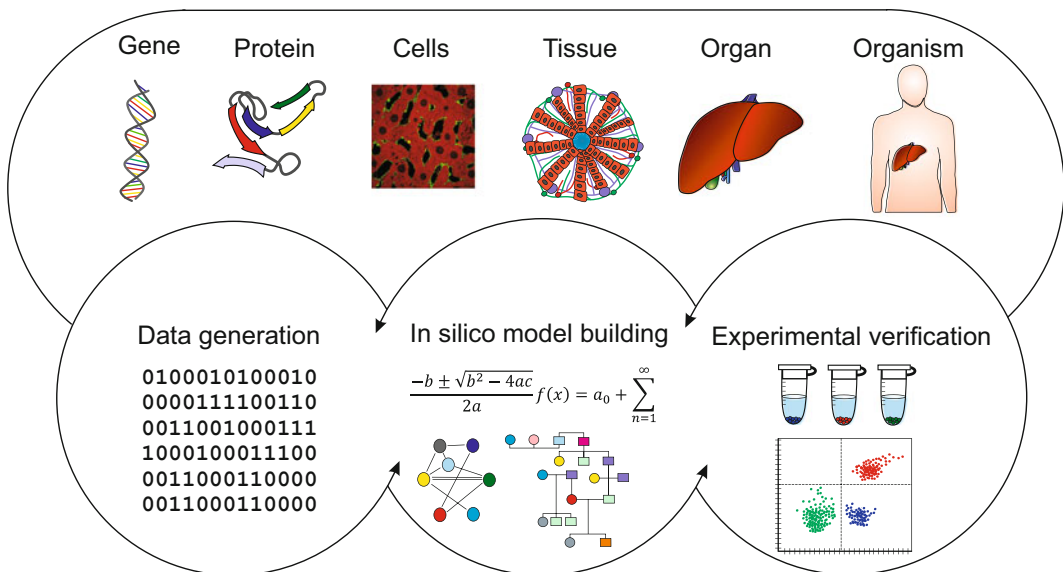
One of the theoretical foundations of Systems Biology is tightly associated with the “Modern Systems Theory,” which was significantly influenced by two Austrian scientists, the biologist *Paul A. Weiss* and the philosopher and theoretical biologist *Ludwig von Bertalanffy*. Toward the end of the 1960s, Bertalanffy published the concept of a general systems theory [3], where one of the key statements can be summarized as the following: *Biology is autonomous and life itself cannot be reduced to disciplines in physics or chemistry or to physico-mechanistic relationships. As consequence, biology needs to be described from a differentiated and methodological point of view, where integrative and holistic aspects play the key role. As a consequence of this view, Bertalanffy postulated the organismic biology, describing the entity of biological units as a main feature that, in its whole, is more than just the sum of all (single) units. This leads to the systems theory (of life), describing a biological organism as a system of complex interactions of different elements, that can be described with the help of mathematical models (summarized in [4]).*

---

## 3 Toward Modern Systems Biology

Due to their dynamic and multilevel features, biological systems cannot be compared to engineered machines, such as airplanes or cars, where all parts are exactly known and an existing blueprint

allows for a precise assembly (as well as disassembly and repair) process that gives rise to the end product. In contrast, biocomplexity needs to be tackled from a different angle—beyond linear dependencies with an integrative strategy that takes dynamic, spatiotemporal features into account. Approximately a decade ago life sciences made a step change toward such an integrative understanding of biology that was fueled by modern analytical technologies including -omics sciences and a rapid development in computer performance and storage capacity. It was this development that gave life scientists the proper computational as well as methodological tools at hand to produce, store, and analyze the large amounts of biological (genomic, proteomic, metabolomic, or physiomic) data that modern biology was able to produce. Only this combination gave rise to the development of mathematical models and computer simulations, based on truly large sets of biological/experimental data, that were able to describe the underlying dynamics of life processes (Fig. 1). Based on the principles of the systems theory of the last century, Systems Biology evolved into an independent and worldwide recognized discipline over the last decade. With an initial focus on basic science, Systems Biology is now fully established in the life science communities and has evolved into a discipline that has many potential applications, especially for medical sciences and clinical practice. Systems



**Fig. 1** Understanding biocomplexity through an integrative systems approach. Modern Systems Biology tackles biocomplexity through a unique approach that integrates biological data, mathematical modeling, and experimental verification in iterative feedback cycles. *Upper half*: Different levels of biocomplexity ranging from a gene over the organ to the organizational context of an organism. *Lower half*: Iterative cycles of biological data generation, mathematical modeling, and its experimental verification. Picture source of “cells”: With friendly permission from Prof. Gabriele Pradel



Medicine in this context is regarded as the next logical step toward the translation of the Systems Biology approach into clinically relevant settings [5].

---

## 4 Systems Biology to Understand Medical Complexity

The human body and the diseases it develops are highly complex involving many different factors that interact through multifaceted networks over time and space. These systems are also influenced by external and environmental factors on yet many unknown levels of interaction. The classical way to challenge this intrinsic complexity has been a reductionistic one by breaking it down to ever smaller, simpler, and more tractable units (from larger to smaller) [5–8]. As a consequence, this led to a great amount of fragmented and highly specialized disciplines that allowed biological as well as medical sciences to discover and precisely dissect molecular mechanisms, underlying gene functions [5], and cellular processes. In clinical medicine such a reductionistic way is very useful when a single factor is responsible for a given disease. If the origin of disease can be traced back to a single cause, e.g., bacteria in urinary tract infections, inflammation of the appendix (appendicitis), or a broken bone, a reductionist approach to diagnosis makes a lot of sense [7]. However, the diagnostic power of such a reductionism to trace back the causative origin(s) is greatly challenged as soon as multiple patho-phenotypes come into play—features of many clinically relevant complex diseases.

In a medical context reductionism usually lacks an integrative concept toward the biological systems as a whole—its interacting networks, its spatiotemporal dependencies, as well as its (disease) perturbations [9, 10]. Complex diseases such as cancer, asthma, and chronic or noncommunicable diseases are heterogeneous with different comorbidities and multiple phenotypes [7, 8, 11]. Here, a reductionistic way to prevention and treatment is limited due to the nonlinear dynamical nature of the underlying disease-relevant biological interactions and cellular networks that involve many factors, such as regulatory proteins, signal transduction pathways, or transmitter molecules, on different scales and dimensions. To reveal the underlying mechanisms and functional interplays of the human body and the diseases it develops, a more far-reaching approach is needed that explores medicine beyond linear relationships and single parameters [7]. This started with the dawn of modern Systems Biology and is currently paving its way into medical research and practice.

---

## 5 Computational Models to Understand Biocomplexity

In the 1990s the quantitative description of functional biological modules, such as interaction of complex metabolic reactions or signal transduction pathways, gained a key role in biology [12–14].

Through the successful interplay of computer sciences, mathematics, and life sciences, it was now possible to quantitatively describe complex biological systems (Fig. 1). With the help of computational models, complex biological processes could be tackled in a rational and focused way, whereas computer simulations allowed it to make predictions toward their behavior under different conditions. A pioneering work was the virtual heart by Denis Noble [15], who integrated complex quantitative information on different organizational levels to model the complete organ. Other large-scale projects that include such modeling methodologies are the German *Virtual Liver Network* [16], the European Commission (EC) flagship project the *Human Brain Project* [17], and the *Virtual Physiological Human* [18]. These projects wish to understand biocomplexity on many scales and dimensions, whereas it is believed that this knowledge can be used in a medical context not only to make the best possible diagnosis and precise treatments for each individual but also to make predictions about the onset and continuation of disease as well as to apply possible prevention strategies—core features of a modern proactive and personalized medicine.

---

## 6 Transition into Clinical Medicine

Clinicians have always integrated information about their patients on many scales from family history, to lifestyle, to actual diagnostic findings as well as individual doctor-patient experience/relationships gained over a long-term period. This is a proven and valid concept, which has evolved over time in classical medicine. A general concept of this medicine is to treat disease when symptoms occur; it can thus be described as reactive or evidence-based medicine [19, 20]. However, this is dramatically challenged by the sheer amount of patient-relevant information and data modern life science technologies are able to produce. For instance, a metabolomics blood sample analysis produces hundreds to thousands of different parameters simultaneously that have to be interpreted by statistics to provide a meaningful output [21]. But modern diagnostics also produces imaging, genomic sequencing, and additional -omics data that all need to be interpreted and integrated into the context of an individual patient. This simply can no longer be achieved by a classical diagnosis of a single doctor. The sheer complexity of this amount of data produces a gap that needs to be bridged in order to make sense out of this information—for the benefit of the patient. This gap is what the computational modeling part of Systems Biology can bridge. Translated into clinical research and practice, the principles of Systems Biology offer an approach that is scalable to the individual needs of the doctor-patient relationship and helps to explain diagnostic findings (e.g.,

via decision-aid or expert systems) and justifications of therapeutic decisions and offer novel paths to prognosis of diseases. By using a Systems Medicine concept, clinicians will be able to integrate complex dynamic information and will be properly guided to make optimal decisions for diagnosis, prevention, and therapy.

Many clinicians by now started to appreciate that tackling complex diseases requires such an integrative strategy that involves multiple, spatiotemporal parameters to achieve a true far-reaching perspective of an individual patient, including disease history, lifestyle, as well as individual molecular/genetic profiles. It is most likely that this will overcome the challenges complex diseases possess to clinical medicine, but it is also a promising concept for a proactive medicine focused on better diagnosis, more precise treatments, and timely prevention measures. On a small scale this is already happening and has been proven as a successful concept in clinical medicine for some diseases such as chronic pulmonary infections [22, 23] or several cancer entities [24–28]. These examples demonstrate that medicine is undergoing a change in mind-set and that the classical, reactive approach is now flanked by a proactive Systems Medicine way of thinking. The comprehensive appraisal and analysis of individual patient data that Systems Medicine enables will thus lay the grounds of a more predictive, preventive, personalized, and participatory (P4) medicine [19, 20].

Major advances in biotechnology, harnessed by Systems Medicine, will enable medical practice to manage a person's health, instead of managing a patient's disease. By applying new computational and mathematical tools to medical practice, medicine can move from a largely reactive mode to a more socioeconomically compatible one that is focused on a patient's well-being utilizing rationally designed (systems) strategies instead of disease-driven reactions. The broad goals of going along this novel road are easy to define within the overarching health and wealth, whereas P4 medicine is a systematic framework that Systems Medicine is most likely to deliver.

---

## 7 What Does Systems Medicine Promise and Where Are the Challenges?

Systems Medicine offers a novel and innovative concept that is based on integration. It is patient centered and has the potential to evolve into a proactive P4 medicine supported by mathematical/computational modeling approaches and modern technological developments such as knowledge management informatics and clinical decision support systems. Especially for multicomponent types of diseases that include different comorbidities and pathophenotypes, a differentiated and well-managed care network is crucial that is able to (1) integrate diagnosis, therapy, and preven-

tion measures as well as (2) cooperate between patients, general practitioners, hospitals, insurers, and policy makers. Success stories that demonstrate the feasibility of Systems Medicine in clinical practice already began to unfold [1], and it is probably only a matter of time until sustained socioeconomic benefits can be demonstrated. However, there are many hurdles that need to be taken. Besides the immanent scientific challenges to unravel biocomplexity and disease, Systems Medicine also faces practical difficulties associated with a far-reaching implementation strategy. Some of these challenges are summarized below:

*Multidisciplinarity*—Because of the manifold requirements necessary to decipher complex diseases, Systems Medicine requires multidisciplinary thinking within different educational backgrounds [29, 30], whereas understanding each other's position represents a major challenge for developing a common vision: Systems Biology researchers need to think in terms of Systems Medicine and vice versa for clinicians and medical doctors, who need to learn to think in terms of Systems Biology. Due to this heterogeneity it takes time to develop a coherent Systems Medicine community. Such a community should be able to integrate all relevant stakeholders, from research, the clinic, industry, regulators, and policy makers to develop a common vision, a strategy, and the willingness for a change (with regard to the classical/reactive medicine paradigm).

*Technological development*—A key feature of Systems Medicine and its clinical application is the development of mathematical/computational models that are able to integrate a whole range of different patient-relevant information and knowledge [31], such as -omics data, family history, lifestyle, and other diagnostic findings. Models, bioinformatics tools, and decision-aid systems (expert systems) need to be user-friendly and adaptable/interoperable and need to make the daily clinical routine easier as well as more precise. However, the quality and reproducibility of this information are essential as well as its validation in clinical research and practice—including proper management, utilization/sharing, and handling procedures for data derived from electronic medical records, connected health devices, and integrated information and communication technology (ICT)-based systems [1]. This requires a proper technological infrastructure that robustly and securely supports management of information, knowledge, and data. It is key to Systems Medicine that information is put into the context of each individual patient and that doctors are properly supported to make sense out of this, for routine diagnosis, treatment of disease, and proper prevention measures.

*Education*—Especially in the current clinical practice, there is a lack of embedded training and educational programs that show the versatility of Systems Medicine for the daily routine. Even though Systems Medicine has the potential to provide individualized deci-

sion support as well as expert systems for diagnosis, therapy, and prevention, the average clinician (including the medical curriculum) is simply too busy to integrate additional items into the average workflow. Thus, lack of encouragement of research in spite of high clinical workload may be a major threat to clinically oriented research including Systems Medicine. Any lack of opportunities for research could therefore challenge efforts to implement Systems Medicine in the clinical practice.

---

## 8 What Is Already There and What Is Missing?

As described in earlier paragraphs, a wide acceptance by all involved stakeholders is crucial for a sustained implementation of Systems Medicine. The practical feasibility including cost-effectiveness and socioeconomic benefits is important to prove. In some complex diseases, this is currently under way, whereas the field of personalized therapy—as one of the Ps of P4 medicine—is advancing well. Examples exist in cancer therapy that allow for a more focused therapy (lessening the therapy burden) with fewer side effects. Prominent is gene expression profiling that is used to better stratify patients with breast cancer in order to improve decision-making with regard to adjuvant chemotherapy [32]. In another approach, focus is put on personalization of chronotherapy to optimize cancer treatment and minimize treatment burden [33]. Especially in chronic diseases such as asthma and chronic obstructive pulmonary disease (COPD), it is recognized that the classical medicine is limited in solving challenges associated with different susceptibility states, various patho-phenotypes, or pre-clinical disease manifestations [34] and that Systems Medicine might offer an alternative, more promising solution. For many clinically relevant diseases, systems-based approaches are under way that take into account the disease-specific challenges, such as multiple symptoms, the patient's social environment, genetics as well as -omics data/information, and cost-effectiveness of treatment [32, 34–36] to develop a proactive P4 medicine. This is in part supported by a general reduction in assay costs that allow for inexpensive genetics/-omics testing. In addition, many researchers realize that the development of novel drugs nowadays is a truly large-scale undertaking that can only be managed in an integrated effort-combining infrastructures, such as full molecular screening programs, freely available information-sharing platforms, and biobanks, including support by industry as well as regulators [37].

In order to use the full potential of Systems Medicine, this broad applicability including many stakeholders from different disciplines is needed. This needs to be backed up by well-defined studies flanked by specific key performance indicators that clearly identify the socio-

economic benefits, a lowered treatment/disease burden, and transparent diagnose, therapy, and prevention systems.

---

## 9 Systems Medicine in Europe

Very early the EC recognized the potential of Systems Medicine and supported research teams with a competitive edge for this innovative approach. From 2004 to 2010, under the sixth/seventh framework program, more than 400 M € was committed by the EC for research, training, as well as infrastructure projects to push Systems Medicine forward ([www.ec.europa.eu](http://www.ec.europa.eu)). This led to many successful projects with focus on medical aspects and the potential for a tangible applications in clinical research and practice [1].

In order to further develop this field and to provide a sound strategic foundation as well as a large-scale community building process for Systems Medicine, in 2012 the EC funded the coordinating action for the implementation of Systems Medicine across Europe—CASyM ([www.casym.eu](http://www.casym.eu)) with the aim to strengthen the European Systems Medicine community and to build a vision and practical implementation strategy (road map) for this approach. This road map identifies four core priority actions and asserts that (1) investment in proof of concept and demonstrator projects is needed to help to precipitate a paradigm shift in the way medicine is practiced. This shift will be supported by (2) a strong Systems Medicine community, (3) new multidisciplinary training programs, and (4) the development of new European-wide practices in clinical data access, sharing, and standardization. These actions are outlined along with ten crosscutting key areas and specific recommendations over a period of 2, 5, and 10 years. In addition, and based on the recommendations of the CASyM road map, the EC launched the first Systems Medicine-oriented ERA-NET under Horizon 2020—a consortium of 15 European funding bodies with support by the EC (co-fund scheme) that agreed on a common research agenda. The ERA-NET “ERACoSysMed-Systems Medicine to address clinical needs” started in January 2015 with the aim to specifically fund demonstrator projects that identify areas where a systems approach addresses a clinical question and provides solution strategies to clinical problems.

Keeping the above-described development “From Systems Biology to Systems Medicine” [5, 38] in mind, these activities clearly foster a joint European Systems Medicine approach. This is also reflected in the current Horizon 2020 work program 2014/2015 “Health demographic change and well-being” of the EC, where Systems Medicine and Personalized Medicine are specific topics for research and innovation calls, demonstrating again the overarching importance of this approach.

---

## 10 Systems Medicine on National Levels in Europe

Across Europe on the national level, there are only a few countries specifically developing a coherent funding agenda for research and technological development in Systems Medicine. Countries like the Netherlands, Israel, Spain, Slovenia, Belgium, and Norway do not (yet) invest specifically in Systems Medicine initiatives. However, most countries do have invested quite substantially in Systems Biology or its translational aspect into clinically relevant fields. For instance, the Netherlands has a long tradition in Systems Biology funding and recently launched three Systems Biology centers of excellence ranging from cancer to bioenergetics to metabolism as well as aging [1]. Additional Systems Medicine funding across EU countries does exist but focuses more on specific projects as a component of more or less broad research calls.

Countries with the most well-structured Systems Medicine programs are Germany and Luxembourg: With the 2012 published novel concept *Paving the Way for Systems Medicine—The e:Med research and funding concept* [39], the Federal Ministry of Education and Research (BMBF), Germany, is launching a highly structured, coherent, and far-reaching national funding program that puts a clear focus on translational research toward clinical practice and the patient. This initiative will lead to funding Systems Medicine in five different modules and will amount to a financial investment of 200 M € within a time period of 8 years [1]. This concept can be regarded as a logical consequence of a long history of Systems Biology funding with a commitment of more than 430 M € that included important initiatives such as *The Virtual Liver Network* (<http://network.virtual-liver.de/en/vision/>); *CancerSys*, *MedSys*, or *GerontoSys* (<https://www.ptj.de/systembiologie>); as well as the *e:Bio—innovation competition Systems Biology* program (<https://www.ptj.de/e-bio>). Along with this is the Fonds National de la Recherche, Luxembourg, who recently invested a total budget of 12 M € into the NCER program (<http://www.fnrlu/calls2/ncer-programme>), a pilot call on Systems Medicine—National Centre of Excellence in Research on the topic of early diagnosis and stratification of Parkinson’s disease.

These examples demonstrate that Systems Medicine indeed has a broad foundation in many European countries with regard to governmental support and funding but that there are only very few dedicated national programs pushing forward the development of this approach. However, with the newly established Horizon 2020 ERA-NET 14 European funding organizations joined forces and actually agreed on a common research and innovative agenda toward the practical implementation of Systems Medicine via so-called demonstrator projects that show feasibility and socio-economic benefits.



## References

1. The CASyM Consortium (2014) The CASyM roadmap – implementation of systems medicine across Europe. <https://www.casym.eu/index.php?index=90>
2. Aristoteles. *Metaphysik*, e-artnow (2014); ISBN 978-80-268-1784-0
3. von Bertalanffy L (1968) *General system theory: foundations, development, applications*. G. Braziller, New York
4. Drack M, Apfalter W, Pouvreau D (2007) On the making of a system theory of life: Paul A Weiss and Ludwig von Bertalanffy's conceptual connection. *Q Rev Biol* 82:349–373
5. Wolkenhauer O, Auffray C, Jaster R et al (2013) The road from systems biology to systems medicine. *Pediatr Res* 73:502–507. doi:10.1038/pr.2013.4
6. Vandamme D, Fitzmaurice W, Kholodenko B, Kolch W (2013) Systems medicine: helping us understand the complexity of disease. *QJM Mon J Assoc Phys* 106:891–895. doi:10.1093/qjmed/hct163
7. Ahn AC, Tewari M, Poon C-S, Phillips RS (2006) The clinical applications of a systems approach. *PLoS Med* 3:e209. doi:10.1371/journal.pmed.0030209
8. Ahn AC, Tewari M, Poon C-S, Phillips RS (2006) The limits of reductionism in medicine: could systems biology offer an alternative? *PLoS Med* 3:e208. doi:10.1371/journal.pmed.0030208
9. Hood L, Balling R, Auffray C (2012) Revolutionizing medicine in the 21st century through systems approaches. *Biotechnol J* 7:992–1001. doi:10.1002/biot.201100306
10. Del Sol A, Balling R, Hood L, Galas D (2010) Diseases as network perturbations. *Curr Opin Biotechnol* 21:566–571. doi:10.1016/j.copbio.2010.07.010
11. Bousquet J, Anto JM, Sterk PJ et al (2011) Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Med* 3:43. doi:10.1186/gm259
12. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* 402:C47–C52. doi:10.1038/35011540
13. Endy D, Brent R (2001) Modelling cellular behaviour. *Nature* 409:391–395. doi:10.1038/35053181
14. Lauffenburger DA (2000) Cell signaling pathways as control modules: complexity for simplicity? *Proc Natl Acad Sci* 97:5031–5033. doi:10.1073/pnas.97.10.5031
15. Noble D (2002) Modeling the heart – from genes to cells to the whole organ. *Science* 295:1678–1682. doi:10.1126/science.1069881
16. Holzhütter H-G, Drasdo D, Preusser T et al (2012) The virtual liver: a multidisciplinary, multilevel challenge for systems biology. *Wiley Interdiscip Rev Syst Biol Med* 4:221–235. doi:10.1002/wsbm.1158
17. Markram H (2012) The human brain project. *Sci Am* 306:50–55. doi:10.1038/scientificamerican0612-50
18. Fenner J, Brook B, Clapworthy G et al (2008) The EuroPhysiome, STEP and a roadmap for the virtual physiological human. *Philos Trans R Soc Math Phys Eng Sci* 366:2979–2999. doi:10.1098/rsta.2008.0089
19. Sobradillo P, Pozo F, Agustí Á (2011) P4 medicine: the future around the corner. *Arch Bronconeumol Engl Ed* 47:35–40. doi:10.1016/S1579-2129(11)70006-4
20. Hood L, Flores M (2012) A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnol* 29:613–624. doi:10.1016/j.nbt.2012.03.004
21. Spratlin JL, Serkova NJ, Eckhardt SG (2009) Clinical applications of metabolomics in oncology: a review. *Clin Cancer Res* 15:431–440. doi:10.1158/1078-0432.CCR-08-1059
22. Bangs A (2005) Predictive biosimulation and virtual patients in pharmaceutical R and D. *Stud Health Technol Inform* 111:37–42
23. Roca J, Rodríguez DA, Falciani F et al. Systems medicine in complex chronic diseases: chronic obstructive pulmonary disease (COPD) as a use case. C98. Of men and mice: better understanding of chronic obstructive pulmonary disease. *American Thoracic Society*, pp A5132–A5132
24. Wiens AL, Martin SE, Bertsch EC et al (2014) Luminal subtypes predict improved survival following central nervous system metastasis in patients with surgically managed metastatic breast carcinoma. *Arch Pathol Lab Med* 138:175–181. doi:10.5858/arpa.2012-0541-OA
25. Levi F (2008) The circadian timing system: a coordinator of life processes [chronobiological investigations]. *IEEE Eng Med Biol Mag* 27:17–19. doi:10.1109/MEMB.2007.907361
26. Faratian D, Goltsov A, Lebedeva G et al (2009) Systems biology reveals new strategies for personalizing cancer medicine and confirms the role of PTEN in resistance to trastuzumab. *Cancer Res* 69:6713–6720. doi:10.1158/0008-5472.CAN-09-0777
27. Neal ML, Trister AD, Ahn S et al (2013) Response classification based on a minimal model of glioblastoma growth is prognostic for clinical outcomes and distinguishes progression from pseudoprogression. *Cancer Res*



- 73:2976–2986. doi:[10.1158/0008-5472.CAN-12-3588](https://doi.org/10.1158/0008-5472.CAN-12-3588)
28. Taylor IW, Linding R, Warde-Farley D et al (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* 27:199–204. doi:[10.1038/nbt.1522](https://doi.org/10.1038/nbt.1522)
  29. Hood L (2003) Leroy hood expounds the principles, practice and future of systems biology. *Drug Discov Today* 8:436–438
  30. Michener WK, Baerwald TJ, Firth P et al (2001) Defining and unraveling biocomplexity. *BioScience* 51:1018–1023
  31. Wolkenhauer O, Auffray C, Brass O et al (2014) Enabling multiscale modeling in systems medicine. *Genome Med* 6:21. doi:[10.1186/gm538](https://doi.org/10.1186/gm538)
  32. Ward S, Scope A, Rafia R et al (2013) Gene expression profiling and expanded immunohistochemistry tests to guide the use of adjuvant chemotherapy in breast cancer management: a systematic review and cost-effectiveness analysis. *Health Technol Assess Winch Engl* 17:1–302. doi:[10.3310/hta17440](https://doi.org/10.3310/hta17440)
  33. Innominato PF, Roche VP, Palesh OG et al (2014) The circadian timing system in clinical oncology. *Ann Med* 46:191–207. doi:[10.3109/07853890.2014.916990](https://doi.org/10.3109/07853890.2014.916990)
  34. Vanfleteren LEGW, Kocks JWH, Stone IS et al (2014) Moving from the Oslerian paradigm to the post-genomic era: are asthma and COPD outdated terms? *Thorax* 69:72–79. doi:[10.1136/thoraxjnl-2013-203602](https://doi.org/10.1136/thoraxjnl-2013-203602)
  35. Nakken N, Janssen DJA, van den Bogaart EHA et al (2014) An observational, longitudinal study on the home environment of people with chronic obstructive pulmonary disease: the research protocol of the Home Sweet Home study. *BMJ Open* 4:e006098. doi:[10.1136/bmjopen-2014-006098](https://doi.org/10.1136/bmjopen-2014-006098)
  36. Zhang H, Gustafsson M, Nestor C et al (2014) Targeted omics and systems medicine: personalising care. *Lancet Respir Med* 2:785–787. doi:[10.1016/S2213-2600\(14\)70188-2](https://doi.org/10.1016/S2213-2600(14)70188-2)
  37. Lacombe D, Tejpar S, Salgado R et al (2014) European perspective for effective cancer drug development. *Nat Rev Clin Oncol* 11:492–498. doi:[10.1038/nrclinonc.2014.98](https://doi.org/10.1038/nrclinonc.2014.98)
  38. European Commission, Health Directorate (2010) Workshop report: from systems biology to systems medicine. [http://ec.europa.eu/research/health/pdf/systems-medicine-workshop-report\\_en.pdf](http://ec.europa.eu/research/health/pdf/systems-medicine-workshop-report_en.pdf)
  39. The Federal Ministry of Education and Research (BMBF) (2012) Paving the way for systems medicine – the e:Med research and funding concept. [http://www.bmbf.de/pub/e\\_med\\_en.pdf](http://www.bmbf.de/pub/e_med_en.pdf)
  40. CASYM Europe: what is systems medicine? <https://www.casym.eu/what-is-systems-medicine>.
  41. Montecucco F, Carbone F, Dini FL et al (2014) Implementation strategies of systems medicine in clinical research and home care for cardiovascular disease patients. *Eur J Intern Med* 25:785–794. doi:[10.1016/j.ejim.2014.09.015](https://doi.org/10.1016/j.ejim.2014.09.015)
  42. Auffray C, Chen Z, Hood L (2009) Systems medicine: the future of medical genomics and healthcare. *Genome Med* 1:2. doi:[10.1186/gm2](https://doi.org/10.1186/gm2)

## Taking Bioinformatics to Systems Medicine

Antoine H.C. van Kampen and Perry D. Moerland

### Abstract

Systems medicine promotes a range of approaches and strategies to study human health and disease at a systems level with the aim of improving the overall well-being of (healthy) individuals, and preventing, diagnosing, or curing disease. In this chapter we discuss how bioinformatics critically contributes to systems medicine. First, we explain the role of bioinformatics in the management and analysis of data. In particular we show the importance of publicly available biological and clinical repositories to support systems medicine studies. Second, we discuss how the integration and analysis of multiple types of omics data through integrative bioinformatics may facilitate the determination of more predictive and robust disease signatures, lead to a better understanding of (patho)physiological molecular mechanisms, and facilitate personalized medicine. Third, we focus on network analysis and discuss how gene networks can be constructed from omics data and how these networks can be decomposed into smaller modules. We discuss how the resulting modules can be used to generate experimentally testable hypotheses, provide insight into disease mechanisms, and lead to predictive models. Throughout, we provide several examples demonstrating how bioinformatics contributes to systems medicine and discuss future challenges in bioinformatics that need to be addressed to enable the advancement of systems medicine.

**Key words** Bioinformatics, Information management, Biological networks, Multi-omics, Integrative bioinformatics, Top-down systems biology, Systems medicine

---

### 1 Introduction

Systems medicine finds its roots in systems biology, the scientific discipline that aims at a *systems-level* understanding of, for example, biological networks, cells, organs, organisms, and populations. It generally involves a combination of wet-lab experiments and computational (bioinformatics) approaches. Systems medicine extends systems biology by focusing on the application of systems-based approaches to clinically relevant applications in order to improve patient health or the overall well-being of (healthy) individuals [1]. Systems medicine is expected to change health care practice in the coming years. It will contribute to new therapeutics through the identification of novel disease genes that provide drug candidates

less likely to fail in clinical studies [2, 3]. It is also expected to contribute to fundamental insights into networks perturbed by disease, improved prediction of disease progression, stratification of disease subtypes, personalized treatment selection, and prevention of disease. To enable systems medicine it is necessary to characterize the patient at various levels and, consequently, to collect, integrate, and analyze various types of data including not only clinical (phenotype) and molecular data, but also information about cells (e.g., disease-related alterations in organelle morphology), organs (e.g., lung impedance when studying respiratory disorders such as asthma or chronic obstructive pulmonary disease), and even social networks. The full realization of systems medicine therefore requires the integration and analysis of environmental, genetic, physiological, and molecular factors at different temporal and spatial scales, which currently is very challenging. It will require large efforts from various research communities to overcome current experimental, computational, and information management related barriers. In this chapter we show how bioinformatics is an essential part of systems medicine and discuss some of the future challenges that need to be solved.

---

## 2 Bioinformatics and High-Throughput Experimental Technologies

### 2.1 *Bioinformatics in Biomedical Research*

To understand the contribution of bioinformatics to systems medicine, it is helpful to consider the traditional role of bioinformatics in biomedical research, which involves basic and applied (translational) research to augment our understanding of (molecular) processes in health and disease. The term “bioinformatics” was first coined by the Dutch theoretical biologist Paulien Hogeweg in 1970 to refer to the study of information processes in biotic systems [4]. Soon, the field of bioinformatics expanded and bioinformatics efforts accelerated and matured as the first (whole) genome and protein sequences became available. The significance of bioinformatics further increased with the development of high-throughput experimental technologies that allowed wet-lab researchers to perform large-scale measurements. These include determining whole-genome sequences (and gene variants) and genome-wide gene expression with next-generation sequencing technologies (NGS; *see* Table 1 for abbreviations and web links) [5], measuring gene expression with DNA microarrays [6], identifying and quantifying proteins and metabolites with NMR or (LC/GC-) MS [7], measuring epigenetic changes such as methylation and histone modifications [8], and so on. These, “omics” technologies, are capable of measuring the many molecular building blocks that determine our (patho)physiology. Genome-wide measurements have not only significantly advanced our fundamental understanding of the molecular biology of health and disease but

**Table 1**  
**Abbreviations and websites**

Abbreviation	Full	Website
ASD	Autism spectrum disorder	
BBMRI	Biobanking and BioMolecular resources Research Infrastructure	<a href="http://bbmri-eric.eu">http://bbmri-eric.eu</a>
CASyM	Coordinating Action Systems Medicine	<a href="https://www.casym.eu">https://www.casym.eu</a>
CGHub	The cancer genome hub	<a href="https://cghub.ucsc.edu">https://cghub.ucsc.edu</a>
DIGGIT	Driver-gene inference by genetical-genomics and information theory	
DREAM	Dialogue on reverse engineering assessment and methods	<a href="http://dreamchallenges.org">http://dreamchallenges.org</a>
EBI	European Bioinformatics Institute	<a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a>
ELIXIR	European life-sciences infrastructure for biological information	<a href="http://www.elixir-europe.org">http://www.elixir-europe.org</a>
ENCODE	Encyclopedia of DNA Elements	<a href="https://www.encodeproject.org">https://www.encodeproject.org</a>
eQTL	Expression quantitative trait loci	
GTEx	Genotype-Tissue Expression project	<a href="http://www.gtexportal.org/home">http://www.gtexportal.org/home</a>
GWAS	Genome wide association study	
ICGC	International Cancer Genome Consortium	<a href="https://icgc.org">https://icgc.org</a>
IMI	European Innovative Medicines Initiative	<a href="http://www.imi.europa.eu">http://www.imi.europa.eu</a>
IMPROVER	Industrial methodology for process verification	<a href="https://sbvimprover.com">https://sbvimprover.com</a>
ISCB	International Society of Computational Biology	<a href="http://www.iscb.org">http://www.iscb.org</a>
LC/GC MS	Liquid/gas chromatography - mass spectroscopy	
MGI	Mouse Genome Informatics	<a href="http://www.informatics.jax.org">http://www.informatics.jax.org</a>
NCBI	National Center for Biotechnology Information	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
NGS	Next generation sequencing	
NMR	Nuclear magnetic resonance	
PheWAS	Phenome-wide association study	
SIB	Swiss Institute of Bioinformatics	<a href="http://www.isb-sib.ch">http://www.isb-sib.ch</a>
SNP	Single nucleotide polymorphism	
TCGA	The Cancer Genome Atlas	<a href="http://cancergenome.nih.gov">http://cancergenome.nih.gov</a>
WGCNA	Weighted gene co-expression network analysis	<a href="http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork">http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork</a>

have also contributed to new (commercial) diagnostic and prognostic tests [9, 10] and the selection and development of (personalized) treatment [11]. Nowadays, bioinformatics is therefore defined as “Advancing the scientific understanding of living systems through computation” (ISCB), or more inclusively as “Conceptualizing biology in terms of molecules and applying ‘informatics techniques’ (derived from disciplines such as applied mathematics, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale” [12].

It is worth noting that solely measuring many molecular components of a biological system does not necessarily result in a deeper understanding of such a system. Understanding biological function does indeed require detailed insight into the precise function of these components but, more importantly, it requires a thorough understanding of their static, temporal, and spatial interactions. These interaction networks underlie all (patho)physiological processes, and elucidation of these networks is a major task for bioinformatics and systems medicine.

## **2.2 New Dimensions in Biomedical Research**

The developments in experimental technologies have led to challenges that require additional expertise and new skills for biomedical researchers:

- *Information management.* Modern biomedical research projects typically produce large and complex omics data sets, sometimes in the order of hundreds of gigabytes to terabytes of which a large part has become available through public databases [13, 14] sometimes even prior to publication (e.g., GTEX, ICGC, TCGA). This not only contributes to knowledge dissemination but also facilitates reanalysis and meta-analysis of data, evaluation of hypotheses that were not considered by the original research group, and development and evaluation of new bioinformatics methods. The use of existing data can in some cases even make new (expensive) experiments superfluous. Alternatively, one can integrate publicly available data with data generated in-house for more comprehensive analyses, or to validate results [15]. In addition, the obligation of making raw data available may prevent fraud and selective reporting. The management (transfer, storage, annotation, and integration) of data and associated meta-data is one of the main and increasing challenges in bioinformatics that needs attention to safeguard the progression of systems medicine.
- *Data analysis and interpretation.* Bioinformatics data analysis and interpretation of omics data have become increasingly complex, not only due to the vast volumes and complexity of the data but also as a result of more challenging research ques-

tions. Bioinformatics covers many types of analyses including nucleotide and protein sequence analysis, elucidation of tertiary protein structures, quality control, pre-processing and statistical analysis of omics data, determination of genotype-phenotype relationships, biomarker identification, evolutionary analysis, analysis of gene regulation, reconstruction of biological networks, text mining of literature and electronic patient records, and analysis of imaging data. In addition, bioinformatics has developed approaches to improve experimental design of omics experiments to ensure that the maximum amount of information can be extracted from the data. Many of the methods developed in these areas are of direct relevance for systems medicine as exemplified in this chapter.

Clearly, new experimental technologies have to a large extent turned biomedical research in a data- and compute-intensive endeavor. It has been argued that production of omics data has nowadays become the “easy” part of biomedical research, whereas the real challenges currently comprise information management and bioinformatics analysis. Consequently, next to the wet-lab, the computer has become one of the main tools of the biomedical researcher.

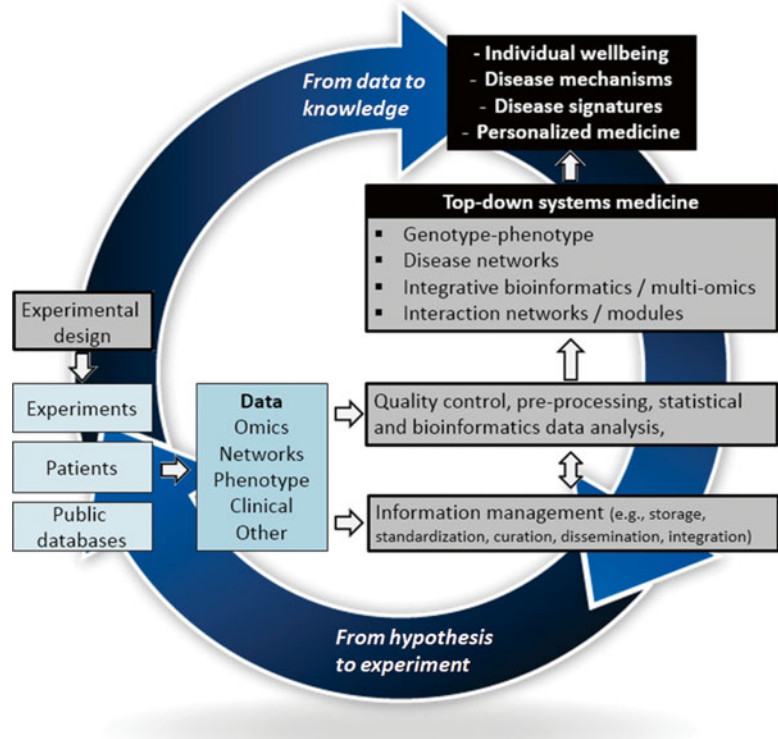
---

### 3 Bioinformatics and Systems Medicine

Bioinformatics enables and advances the management and analysis of large omics-based datasets, thereby directly and indirectly contributing to systems medicine in several ways (Fig. 1):

1. Design of new omics experiments [16–18].
2. Information management of omics and clinical data (Subheading 4).
3. Quality control and pre-processing of omics data. Pre-processing typically involves data cleaning (e.g., removal of failed assays) and other steps to obtain quantitative measurements that can be used in downstream data analysis.
4. (Statistical) data analysis methods of large and complex omics-based datasets. This includes methods for the integrative analysis of multiple omics data types (Subheading 5), and for the elucidation and analysis of biological networks (top-down systems medicine; Subheading 6).

Systems medicine comprises top-down and bottom-up approaches. The former represents a specific branch of bioinformatics, which distinguishes itself from bottom-up approaches in several ways [3, 19, 20]. Top-down approaches use omics data to obtain a holistic view of the components of a biological system and, in general, aim to construct system-wide static functional or physical interaction networks such as gene



**Fig. 1** The contribution of bioinformatics (*dark grey boxes*) to systems medicine (*black box*). (Omics) experiments, patients, and public repositories provide a wide range of data that is used in bioinformatics and systems medicine studies

co-expression networks and protein-protein interaction networks. In contrast, bottom-up approaches aim to develop detailed mechanistic and quantitative mathematical models for sub-systems. These models describe the dynamic and nonlinear behavior of interactions between known components to understand and predict their behavior upon perturbation. However, in contrast to omics-based top-down approaches, these mechanistic models require information about chemical/physical parameters and reaction stoichiometry, which may not be available and require further (experimental) efforts. Both the top-down and bottom-up approaches result in testable hypotheses and new wet-lab or in silico experiments that may lead to clinically relevant findings.

## 4 Information Management for Systems Medicine

### 4.1 Public Databases in Systems Medicine

Biomedical research and, consequently, systems medicine are increasingly confronted with the management of continuously growing volumes of molecular and clinical data, results of data analyses and in silico experiments, and mathematical models. Due



to policies of scientific journals and funding agencies, omics data is often made available to the research community via public databases. In addition, a wide range of databases have been developed, of which more than 1550 are currently listed in the Molecular Biology Database Collection [14] providing a rich source of biomedical information. Biological repositories do not merely archive data and models but also serve a range of purposes in systems medicine as illustrated below from a few selected examples. The main repositories are hosted and maintained by the major bioinformatics institutes including EBI, NCBI, and SIB that make a major part of the raw experimental omics data available through a number of primary databases including GenBank [21], GEO [22], PRIDE [23], and Metabolights [24] for sequence, gene expression, MS-based proteomics, and MS-based metabolomics data, respectively. In addition, many secondary databases provide information derived from the processing of primary data, for example pathway databases (e.g., Reactome [25], KEGG [26]), protein sequence databases (e.g., UniProtKB [27]), and many others. Pathway databases provide an important resource to construct mathematical models used to study and further refine biological systems [28, 29]. Other efforts focus on establishing repositories integrating information from multiple public databases. The integration of pathway databases [30–32], and genome browsers that integrate genetic, omics, and other data with whole-genome sequences [33, 34] are two examples of this. Joint initiatives of the bioinformatics and systems biology communities resulted in repositories such as BioModels, which contains mathematical models of biochemical and cellular systems [35], Recon 2 that provides a community-driven, consensus “metabolic reconstruction” of human metabolism suitable for computational modelling [36], and SEEK, which provides a platform designed for the management and exchange of systems biology data and models [37]. Another example of a database that may prove to be of value for systems medicine studies is MalaCards, an integrated and annotated compendium of about 17,000 human diseases [38]. MalaCards integrates 44 disease sources into disease cards and establishes gene-disease associations through integration with the well-known GeneCards databases [39, 40]. Integration with GeneCards and cross-references within MalaCards enables the construction of networks of related diseases revealing previously unknown interconnections among diseases, which may be used to identify drugs for off-label use. Another class of repositories are (expert-curated) knowledge bases containing domain knowledge and data, which aim to provide a single point of entry for a specific domain. Contents of these knowledge bases are often based on information extracted (either manually or by text mining) from literature or provided by domain experts



[41–43]. Finally, databases are used routinely in the analysis, interpretation, and validation of experimental data. For example, the Gene Ontology (GO) provides a controlled vocabulary of terms for describing gene products, and is often used in gene set analysis to evaluate expression patterns of groups of genes instead of those of individual genes [44] and has, for example, been applied to investigate HIV-related cognitive disorders [45] and polycystic kidney disease [46].

## **4.2 Phenotype Databases**

Several repositories such as miR2Disease [47], PeroxisomeDB [41], and Mouse Genome Informatics (MGI) [43] include associations between genes and disorders, but only provide very limited phenotypic information. Phenotype databases are of particular interest to systems medicine. One well-known phenotype repository is the OMIM database, which primarily describes single-gene (Mendelian) disorders [48]. ClinVar is another example and provides an archive of reports and evidence of the relationships among medically important human variations found in patient samples and phenotypes [49]. ClinVar complements dbSNP (for single-nucleotide polymorphisms) [50] and dbVar (for structural variations) [51], which both provide only minimal phenotypic information. The integration of these phenotype repositories with genetic and other molecular information will be a major aim for bioinformatics in the coming decade enabling, for example, the identification of comorbidities, determination of associations between gene (mutations) and disease, and improvement of disease classifications [52]. It will also advance the definition of the “human phenome,” i.e., the set of phenotypes resulting from genetic variation in the human genome. To increase the quality and (clinical) utility of the phenotype and variant databases as an essential step towards reducing the burden of human genetic disease, the Human Variome Project coordinates efforts in standardization, system development, and (training) infrastructure for the worldwide collection and sharing of genetic variations that affect human health [53, 54].

## **4.3 Clinical Data**

To implement and advance systems medicine to the benefit of patients’ health, it is crucial to integrate and analyze molecular data together with de-identified individual-level clinical data complementing general phenotype descriptions. Patient clinical data refers to a wide variety of data including basic patient information (e.g., age, sex, ethnicity), outcomes of physical examinations, patient history, medical diagnoses, treatments, laboratory tests, pathology reports, medical images, and other clinical outcomes. Inclusion of clinical data allows the stratification of patient groups into more homogeneous clinical subgroups. Availability of clinical data will increase the power of downstream data analysis and modeling to elucidate molecular mechanisms, and to identify molecular

biomarkers that predict disease onset or progression, or which guide treatment selection. In biomedical studies clinical information is generally used as part of patient and sample selection, but some omics studies also use clinical data as part of the bioinformatics analysis (e.g., [9, 55]). However, in general, clinical data is unavailable from public resources or only provided on an aggregated level. Although good reasons exist for making clinical data available (Subheading 2.2), ethical and legal issues comprising patient and commercial confidentiality, and technical issues are the most immediate challenges [56, 57]. This potentially hampers the development of systems medicine approaches in a clinical setting since sharing and integration of clinical and nonclinical data is considered a basic requirement [1]. Biobanks [58] such as BBMRI [59] provide a potential source of biological material and associated (clinical) data but these are, generally, not publicly accessible, although permission to access data may be requested from the biobank provider. Clinical trials provide another source of clinical data for systems medicine studies, but these are generally owned by a research group or sponsor and not freely available [60] although ongoing discussions may change this in the future ([61] and references therein).

Although clinical data is not yet available on a large scale, the bioinformatics and medical informatics communities have been very active in establishing repositories that provide clinical data. One example is the Database of Genotypes and Phenotypes (dbGaP) [62] developed by the NCBI. Study metadata, summary-level (phenotype) data, and documents related to studies are publicly available. Access to de-identified individual-level (clinical) data is only granted after approval by an NIH data access committee. Another example is The Cancer Genome Atlas (TCGA), which also provides individual-level molecular and clinical data through its own portal and the Cancer Genomics Hub (CGHub). Clinical data from TCGA is available without any restrictions but part of the lower level sequencing and microarray data can only be obtained through a formal request managed by dbGaP.

Medical patient records provide an even richer source of phenotypic information, and has already been used to stratify patient groups, discover disease relations and comorbidity, and integrate these records with molecular data to obtain a systems-level view of phenotypes (for a review see [63]). On the one hand, this integration facilitates refinement and analysis of the human phenome to, for example, identify diseases that are clinically uniform but have different underlying molecular mechanisms, or which share a pathogenetic mechanism but with different genetic cause [64]. On the other hand, using the same data, a phenome-wide association study (PheWAS) [65] would allow the identification of unrelated phenotypes associated with specific shared genetic variant(s), an effect referred to as pleiotropy. Moreover, it makes use of

information from medical records generated in routine clinical practice and, consequently, has the potential to strengthen the link between biomedical research and clinical practice [66]. The power of phenome analysis was demonstrated in a study involving 1.5 million patient records, not including genotype information, comprising 161 disorders. In this study it was shown that disease phenotypes form a highly connected network suggesting a shared genetic basis [67]. Indeed, later studies that incorporated genetic data resulted in similar findings and confirmed a shared genetic basis for a number of different phenotypes. For example, a recent study identified 63 potentially pleiotropic associations through the analysis of 3144 SNPs that had previously been implicated by genome-wide association studies (GWAS) as mediators of human traits, and 1358 phenotypes derived from patient records of 13,835 individuals [68]. This demonstrates that phenotypic information extracted manually or through text mining from patient records can help to more precisely define (relations between) diseases. Another example comprises the text mining of psychiatric patient records to discover disease correlations [52]. Here, mapping of disease genes from the OMIM database to information from medical records resulted in protein networks suspected to be involved in psychiatric diseases.

---

## 5 Integrative Bioinformatics

Integrative bioinformatics comprises the integrative (statistical) analysis of multiple omics data types. Many studies demonstrated that using a single omics technology to measure a specific molecular level (e.g., DNA variation, expression of genes and proteins, metabolite concentrations, epigenetic modifications) already provides a wealth of information that can be used for unraveling molecular mechanisms underlying disease. Moreover, single-omics disease signatures which combine multiple (e.g., gene expression) markers have been constructed to differentiate between disease subtypes to support diagnosis and prognosis. However, no single technology can reveal the full complexity and details of molecular networks observed in health and disease due to the many interactions across these levels. A systems medicine strategy should ideally aim to understand the functioning of the different levels as a whole by integrating different types of omics data. This is expected to lead to biomarkers with higher predictive value, and novel disease insights that may help to prevent disease and to develop new therapeutic approaches. Integrative bioinformatics can also facilitate the prioritization and characterization of genetic variants associated with complex human diseases and traits identified by GWAS in which hundreds of thousands to over a million SNPs are assayed in a large number of individuals. Although such studies lack the

statistical power to identify all disease-associated loci [69], they have been instrumental in identifying loci for many common diseases. However, it remains difficult to prioritize the identified variants and to elucidate their effect on downstream pathways ultimately leading to disease [70]. Consequently, methods have been developed to prioritize candidate SNPs based on integration with other (omics) data such as gene expression, DNase hypersensitive sites, histone modifications, and transcription factor-binding sites [71].

### **5.1 Data Integration**

The integration of multiple omics data types is far from trivial and various approaches have been proposed [72–74]. One approach is to link different types of omics measurements through common database identifiers. Although this may seem straightforward, in practice this is complicated as a result of technical and standardization issues as well as a lack of biological consensus [32, 75–77]. Moreover, the integration of data at the level of the central dogma of molecular biology and, for example, metabolite data is even more challenging due to the indirect relationships between genes, transcripts, and proteins on the one hand and metabolites on the other hand, precluding direct links between the database identifiers of these molecules.

Statistical data integration [72] is a second commonly applied strategy, and various approaches have been applied for the joint analysis of multiple data types (e.g., [78, 79]). One example of statistical data integration is provided by a TCGA study that measured various types of omics data to characterize breast cancer [80]. In this study 466 breast cancer samples were subjected to whole-genome and -exome sequencing, and SNP arrays to obtain information about somatic mutations, copy number variations, and chromosomal rearrangements. Microarrays and RNA-Seq were used to determine mRNA and microRNA expression levels, respectively. Reverse-phase protein arrays (RPPA) and DNA methylation arrays were used to obtain data on protein expression levels and DNA methylation, respectively. Simultaneous statistical analysis of different data types via a “cluster-of-clusters” approach using consensus clustering on a multi-omics data matrix revealed that four major breast cancer subtypes could be identified. This showed that the intrinsic subtypes (basal, luminal A and B, HER2) that had previously been determined using gene expression data only could be largely confirmed in an integrated analysis of a large number of breast tumors.

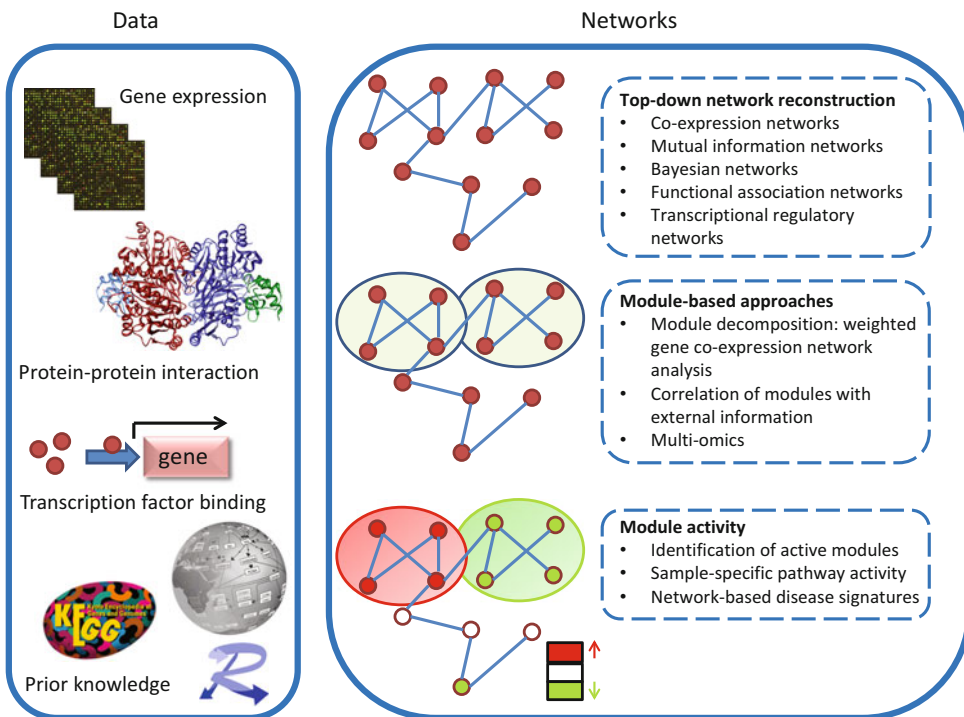
### **5.2 Multi-omics Disease Signatures**

Single-level omics data has extensively been used to identify disease-associated biomarkers such as genes, proteins, and metabolites. In fact, these studies led to more than 150,000 papers documenting thousands of claimed biomarkers. However, it is estimated that fewer than 100 of these are currently used for routine clinical

practice [81]. Integration of multiple omics data types is expected to result in more robust and predictive disease profiles since these better reflect disease biology [82]. Further improvement of these profiles may be obtained through the explicit incorporation of interrelationships between various types of measurements such as microRNA–mRNA target, or gene methylation–microRNA (based on a common target gene). This was demonstrated for the prediction of short-term and long-term survival from serous cystadenocarcinoma TCGA data [83].

## 6 Biological Networks

According to the recent CASyM roadmap: “Human disease can be perceived as perturbations of complex, integrated genetic, molecular and cellular networks and such complexity necessitates a new approach.” [84]. In this section we discuss how (approximations) to these networks can be constructed from omics data and how these networks can be decomposed in smaller modules. Then we discuss how the resulting modules can be used to generate experimentally testable hypotheses, provide insight into disease mechanisms, lead to predictive diagnostic and prognostic models, and help to further subclassify diseases [55, 85] (Fig. 2). Such top-down



**Fig. 2** Overview of network-based approaches for systems medicine (Subheading 6)

network-based approaches will provide medical doctors with molecular level support to make personalized treatment decisions.

### **6.1 Top-Down Network Reconstruction**

In a top-down approach the aim of network reconstruction is to infer the connections between the molecules that constitute a biological network. Network models can be created using a variety of mathematical and statistical techniques and data types. Early approaches for network inference (also called reverse engineering) used only gene expression data to reconstruct gene networks. Here, we discern three types of gene network inference algorithms using methods based on (1) correlation-based approaches, (2) information-theoretic approaches, and (3) Bayesian networks [86].

Co-expression networks are an extension of commonly used clustering techniques, in which genes are connected by edges in a network if the amount of correlation of their gene expression profiles exceeds a certain value. Co-expression networks have been shown to connect functionally related genes [87]. Note that connections in a co-expression network correspond to either direct (e.g., transcription factor-gene and protein-protein) or indirect (e.g., proteins participating in the same pathway) interactions. In one of the earliest examples of this approach, pair-wise correlations were calculated between gene expression profiles and the level of growth inhibition caused by thousands of tested anticancer agents, for 60 cancer cell lines [88]. Removal of associations weaker than a certain threshold value resulted in networks consisting of highly correlated genes and agents, called relevance networks, which led to targeted hypotheses for potential single-gene determinants of chemotherapeutic susceptibility.

Information-theoretic approaches have been proposed in order to capture nonlinear dependencies assumed to be present in most biological systems and that cannot be captured by correlation-based distance measures. These approaches often use the concept of mutual information, a generalization of the correlation coefficient which quantifies the degree of statistical (in)dependence. An example of a network inference method that is based on mutual information is ARACNe, which has been used to reconstruct the human B-cell gene network from a large compendium of human B-cell gene expression profiles [89]. In order to discover regulatory interactions, ARACNe removes the majority of putative indirect interactions from the initial mutual information-based gene network using a theorem from information theory, the data processing inequality. This led to the identification of *MYC* as a major hub in the B-cell gene network and a number of novel *MYC* target genes, which were experimentally validated. Whether information-theoretic approaches are more powerful in general than correlation-based approaches is still subject of debate [90].

Bayesian networks allow the description of statistical dependencies between variables in a generic way [91, 92]. Bayesian

networks are directed acyclic networks in which the edges of the network represent conditional dependencies; that is, nodes that are not connected represent variables that are conditionally independent of each other. A major bottleneck in the reconstruction of Bayesian networks is their computational complexity. Moreover, Bayesian networks are acyclic and cannot capture feedback loops that characterize many biological networks. When time-series rather than steady-state data is available, dynamic Bayesian networks provide a richer framework in which cyclic networks can be reconstructed [93].

Gene (co-)expression data only offers a partial view on the full complexity of cellular networks. Consequently, networks have also been constructed from other types of high-throughput data. For example, physical protein-protein interactions have been measured on a large scale in different organisms including human, using affinity capture-mass spectrometry or yeast two-hybrid screens, and have been made available in public databases such as BioGRID [94]. Regulatory interactions have been probed using chromatin immunoprecipitation sequencing (ChIP-Seq) experiments, for example by the ENCODE consortium [95].

Using probabilistic techniques, heterogeneous types of experimental evidence and prior knowledge have been integrated to construct functional association networks for human [96], mouse [97], and, most comprehensively, more than 1100 organisms in the STRING database [98]. Functional association networks can help predict novel pathway components, generate hypotheses for biological functions for a protein of interest, or identify disease-related genes [97]. Prior knowledge required for these approaches is, for example, available in curated biological pathway databases, and via protein associations predicted using text mining based on their co-occurrence in abstracts or even full-text articles. Many more integrative network inference methods have been proposed; for a review see [99]. The integration of gene expression data with ChIP data [100] or transcription factor-binding motif data [101] has shown to be particularly fruitful for inferring transcriptional regulatory networks. Recently, Li et al. [102] described the results from a regression-based model that predicts gene expression using ENCODE (ChIP-Seq) and TCGA data (mRNA expression data complemented with copy number variation, DNA methylation, and microRNA expression data). This model infers the regulatory activities of expression regulators and their target genes in acute myeloid leukemia samples. Eighteen key regulators were identified, whose activities clustered consistently with cytogenetic risk groups.

Bayesian networks have also been used to integrate multi-omics data. The combination of genotypic and gene expression data is particularly powerful, since DNA variations represent naturally occurring perturbations that affect gene expression detected as expression quantitative trait loci (eQTL). *Cis*-acting eQTLs



can then be used as constraints in the construction of directed Bayesian networks to infer causal relationships between nodes in the network [103].

## **6.2 Module-Based Approaches**

Large multi-omics datasets consisting of hundreds or sometimes even thousands of samples are available for many commonly occurring human diseases, such as most tumor types (TCGA), Alzheimer's disease [104], and obesity [105]. However, a major bottleneck for the construction of accurate gene networks is that the number of gene networks that are compatible with the experimental data is several orders of magnitude larger still. In other words, top-down network inference is an underdetermined problem with many possible solutions that explain the data equally well and individual gene-gene interactions are characterized by a high false-positive rate [99]. Most network inference methods therefore try to constrain the number of possible solutions by making certain assumptions about the structure of the network. Perhaps the most commonly used strategy to harness the complexity of the gene network inference problem is to analyze experimental data in terms of biological modules, that is, sets of genes that have strong interactions and a common function [106]. There is considerable evidence that many biological networks are modular [107]. Module-based approaches effectively constrain the number of parameters to estimate and are in general also more robust to the noise that characterizes high-throughput omics measurements. A detailed review of module-based techniques is outside the scope of this chapter (see, for example [108]), but we would like to mention a few examples of successful and commonly used modular approaches.

Weighted gene co-expression network analysis (WGCNA) decomposes a co-expression network into modules using clustering techniques [109]. Modules can be summarized by their module eigengene, a weighted average expression profile of all gene member of a given module. Eigengenes can then be correlated with external sample traits to identify modules that are related with these traits. Parikshak et al. [110] used WGCNA to extract modules from a co-expression network constructed using fetal and early postnatal brain development expression data. Next, they established that several of these modules were enriched for genes and rare de novo variants implicated in autism spectrum disorder (ASD). Moreover, the ASD-associated modules are also linked at the transcriptional level and 17 transcription factors were found acting as putative co-regulators of ASD-associated gene modules during neocortical development. WGCNA can also be used when multiple omics data types are available. One example of such an approach involved the integration of transcriptomic and proteomic data from a study investigating the response to SARS-CoV infection in mice [111]. In this study WGCNA-based gene and protein co-expression modules were constructed and



integrated to obtain module-based disease signatures. Interestingly, the authors found several cases of identifier-matched transcripts and proteins that correlated well with the phenotype, but which showed poor or anticorrelation across these two data types. Moreover, the highest correlating transcripts and peptides were not the most central ones in the co-expression modules. *Vice versa*, the transcripts and proteins that defined the modules were not those with the highest correlation to the phenotype. At the very least this shows that integration of omics data affects the nature of the disease signatures.

Identification of active modules is another important integrative modular technique. Here, experimental data in the form of molecular profiles is projected onto a biological network, for example a protein-protein interaction network. Active modules are those subnetworks that show the largest change in expression for a subset of conditions and are likely to contain key drivers or regulators of those processes perturbed in the experiment. Active modules have, for example, been used to find a subnetwork that is overexpressed in a particularly aggressive lymphoma subtype [112] and to detect significantly mutated pathways [113]. Some active module approaches integrate various types of omics data. One example of such an approach is PARADIGM [114], which translates pathways into factor graphs, a class of models that belongs to the same family of models as Bayesian networks, and determines sample-specific pathway activity from multiple functional genomic datasets. PARADIGM has been used in several TCGA projects, for example, in the integrated analysis of 131 urothelial bladder carcinomas [55]. PARADIGM-based analysis of copy number variations and RNA-Seq gene expression in combination with a propagation-based network analysis algorithm revealed novel associations between mutations and gene expression levels, which subsequently resulted in the identification of pathways altered in bladder cancer. The identification of activating or inhibiting gene mutations in these pathways suggested new targets for treatment. Moreover, this effort clearly showed the benefits of screening patients for the presence of specific mutations to enable personalized treatment strategies.

### **6.3 Network-Based Disease Signatures**

Often, published disease signatures cannot be replicated [81] or provide hardly additional biological insight. Also here (modular) network-based approaches have been proposed to alleviate these problems. A common characteristic of most methods is that the molecular activity of a set of genes is summarized on a per sample basis. Summarized gene set scores are then used as features in prognostic and predictive models. Relevant gene sets can be based on prior knowledge and correspond to canonical pathways, gene ontology categories, or sets of genes sharing common motifs in their promoter regions [115]. Gene set scores can also be

determined by projecting molecular data onto a biological network and summarizing scores at the level of subnetworks for each individual sample [116]. While promising in principle, it is still subject of debate whether gene set-based models outperform gene-based ones [117].

#### **6.4 Crossing the Species Boundary**

The comparative analysis of networks across different species is another commonly used approach to constrain the solution space. Patterns conserved across species have been shown to be more likely to be true functional interactions [107] and to harbor useful candidates for human disease genes [118]. Many network alignment methods have been developed in the past decade to identify commonalities between networks. These methods in general combine sequence-based and topological constraints to determine the optimal alignment of two (or more) biological networks. Network alignment has, for example, been applied to detect conserved patterns of protein interaction in multiple species [107, 119] and to analyze the evolution of co-expression networks between humans and mice [120, 121]. Network alignment can also be applied to detect diverged patterns [120] and may thus lead to a better understanding of similarities and differences between animal models and human in health and disease. Information from model organisms has also been fruitfully used to identify more robust disease signatures [122–125]. Sweet-Cordero and co-workers [122] used a gene signature identified in a mouse model of lung adenocarcinoma to uncover an orthologous signature in human lung adenocarcinoma that was not otherwise apparent. Bild et al. [123] defined gene expression signatures characterizing several oncogenic pathways of human mammary epithelial cells. They showed that these signatures predicted pathway activity in mouse and human tumors. Predictions of pathway activity correlated well with the sensitivity to drugs targeting those pathways and could thus serve as a guide to targeted therapies. A generic approach, Pathprint, for the integration of gene expression data across different platforms and species at the level of pathways, networks, and transcriptionally regulated targets was recently described [126]. The authors used their method to identify four stem cell-related pathways conserved between human and mouse in acute myeloid leukemia, with good prognostic value in four independent clinical studies.

#### **6.5 From Networks to Medicine**

We reviewed a wide array of different approaches showing how networks can be used to elucidate integrated genetic, molecular, and cellular networks. However, in general no single approach will be sufficient and combining different approaches in more complex analysis pipelines will be required. This is fittingly illustrated by the DIGGIT (Driver-gene Inference by Genetical-Genomics and Information Theory) algorithm [127]. In brief, DIGGIT identifies candidate master regulators from an ARACNe gene co-expression

network integrated with copy number variations that affect gene expression. This method combines several previously developed computational approaches and was used to identify causal genetic drivers of human disease in general and glioblastoma, breast cancer, and Alzheimer's disease in particular. This enabled identification of KLHL9 deletions as upstream activators of two previously established master regulators in a specific subtype of glioblastoma.

---

## 7 Discussion

Systems medicine is one of the steps necessary to make improvements in the prevention and treatment of disease through systems approaches that will (a) elucidate (patho)physiologic mechanisms in much greater detail than currently possible, (b) produce more robust and predictive disease signatures, and (c) enable personalized treatment. In this context, we have shown that bioinformatics has a major role to play.

Bioinformatics will continue its role in the development, curation, integration, and maintenance of (public) biological and clinical databases to support biomedical research and systems medicine. The bioinformatics community will strengthen its activities in various standardization and curation efforts that already resulted in minimum reporting guidelines [128], data capture approaches [75], data exchange formats [129], and terminology standards for annotation [130]. One challenge for the future is to remove errors and inconsistencies in data and annotation from databases and prevent new ones from being introduced [32, 76, 131–135]. An equally important challenge is to establish, improve, and integrate resources containing phenotype and clinical information. To achieve this objective it seems reasonable that bioinformatics and health informatics professionals team up [136–138]. Traditionally health informatics professionals have focused on hospital information systems (e.g., patient records, pathology reports, medical images) and data exchange standards (e.g., HL7), medical terminology standards (e.g., International Classification of Disease (ICD), SNOMED), medical image analysis, analysis of clinical data, clinical decision support systems, and so on. While, on the other hand, bioinformatics mainly focused on molecular data, it shares many approaches and methods with health informatics. Integration of these disciplines is therefore expected to benefit systems medicine in various ways [139].

Integrative bioinformatics approaches clearly have added value for systems medicine as they provide a better understanding of biological systems, result in more robust disease markers, and prevent (biological) bias that would possibly occur from using single-omics measurements. However, such studies, and the scientific community in general, would benefit from improved strategies to disseminate and share data which typically will be produced at multiple

research centers (e.g., <https://www.synapse.org>; [140]). Integrative studies are expected to increasingly facilitate personalized medicine approaches such as demonstrated by Chen and co-workers [141]. In their study they presented a 14-month “integrative personal omics profile” (iPOP) for a single individual comprising genomic, transcriptomic, proteomic, metabolomic, and autoantibody data. From the whole-genome sequence data an elevated risk for type 2 diabetes (T2D) was detected, and subsequent monitoring of HbA1c and glucose levels revealed the onset of T2D, despite the fact that the individual lacked many of the known non-genetic risk factors. Subsequent treatment resulted in a gradual return to the normal phenotype. This shows that the genome sequence can be used to determine disease risk in a healthy individual and allows selecting and monitoring specific markers that provide information about the actual disease status.

Network-based approaches will increasingly be used to determine the genetic causes of human diseases. Since the effect of a genetic variation is often tissue or cell-type specific, a large effort is needed in constructing cell-type-specific networks both in health and disease. This can be done using data already available, an approach taken by Guan et al. [142]. The authors proposed 107 tissue-specific networks in mouse via their generic approach for constructing functional association networks using low-throughput, highly reliable tissue-specific gene expression information as a constraint. One could also generate new datasets to facilitate the construction of tissue-specific networks. Examples of such approaches are TCGA and the genotype-tissue expression (GTEx) project. The aim of GTEx is to create a data resource for the systematic study of genetic variation and its effect on gene expression in more than 40 human tissues [143]. Regardless of the way how networks are constructed, it will become more and more important to offer a centralized repository where networks from different cell types and diseases can be stored and accessed. Nowadays, these networks are difficult to retrieve and are scattered in supplementary files with the original papers, links to accompanying web pages, or even not available at all. A resource similar to what the systems biology community has created with the BioModels database would be a great leap forward. There have been some initial attempts in building databases of network models, for example the CellCircuits database [123] (<http://www.cell-circuits.org>) and the causal biological networks (CBN) database of networks related to lung disease [144] (<http://causalbionet.com>). However, these are only small-scale initiatives and a much larger and coordinated effort is required.

Another main bottleneck for the successful application of network inference methods is their validation. Most network inference methods to date have been applied to one or a few isolated datasets and were validated using some limited follow-up experiments, for example via gene knockdowns, using prior knowledge from

databases and literature as a gold standard, or by generating simulated data from a mathematical model of the underlying network [145, 146]. However, strengths and weaknesses of network inference methods across cell types, diseases, and species have hardly been assessed. Notable exceptions are collaborative competitions such as the Dialogue on Reverse Engineering Assessment and Methods (DREAM) [147] and Industrial Methodology for Process Verification (IMPROVER) [146]. These centralized initiatives propose challenges in which individual research groups can participate and to which they can submit their predictions, which can then be independently validated by the challenge organizers. Several DREAM challenges in the area of network inference have been organized, leading to a better insight into the strengths and weaknesses of individual methods [148]. Another important contribution of DREAM is that a crowd-based approach integrating predictions from multiple network inference methods was shown to give good and robust performance across diverse data sets [149]. Also in the area of systems medicine challenge-based competitions may offer a framework for independent verification of model predictions.

Systems medicine promises a more personalized medicine that effectively exploits the growing amount of molecular and clinical data available for individual patients. Solid bioinformatics approaches are of crucial importance for the success of systems medicine. However, really delivering the promises of systems medicine will require an overall change of research approach that transcends the current reductionist approach and results in a tighter integration of clinical, wet-lab laboratory, and computational groups adopting a systems-based approach. Past, current, and future success of systems medicine will accelerate this change.

---

## Acknowledgements

We would like to thank Dr. Aldo Jongejan for his comments that improved the text.

## References

1. Wolkenhauer O, Auffray C, Jaster R et al (2013) The road from systems biology to systems medicine. *Pediatr Res* 73(4 Pt 2):502–507
2. Hood L, Auffray C (2013) Participatory medicine: a driving force for revolutionizing healthcare. *Genome Med* 5(12):110
3. Schneider HC, Klabunde T (2013) Understanding drugs and diseases by systems biology? *Bioorg Med Chem Lett* 23(5): 1168–1176
4. Hogeweg P (2011) The roots of bioinformatics in theoretical biology. *PLoS Comput Biol* 7(3):e1002021
5. Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11(1):31–46
6. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet* 21(1 Suppl):33–37
7. Lindon JC, Nicholson JK (2008) Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. *Annu Rev Anal Chem (Palo Alto Calif)* 1:45–69
8. Mensaert K, Denil S, Trooskens G et al (2014) Next-generation technologies and

- data analytical approaches for epigenomics. *Environ Mol Mutagen* 55(3):155–170
9. van't Veer LJ, Dai H, van de Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871):530–536
  10. Zanotti L, Bottini A, Rossi C et al (2014) Diagnostic tests based on gene expression profile in breast cancer: from background to clinical use. *Tumour Biol* 35(9):8461–8470
  11. Paik S, Shak S, Tang G et al (2004) A multi-gene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 351(27):2817–2826
  12. Luscombe NM, Greenbaum D, Gerstein M (2001) What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med* 40(4):346–358
  13. Baxevanis AD (2011) The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics*. Chapter 1:Unit 1 1
  14. Fernandez-Suarez XM, Rigden DJ, Galperin MY (2014) The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic Acids Res* 42(Database issue):D1–D6
  15. Rung J, Brazma A (2013) Reuse of public genome-wide gene expression data. *Nat Rev Genet* 14(2):89–99
  16. Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. *Biostatistics* 2(2):183–201
  17. Lambert CG, Black LJ (2012) Learning from our GWAS mistakes: from experimental design to scientific method. *Biostatistics* 13(2):195–203
  18. Robles JA, Qureshi SE, Stephen SJ et al (2012) Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics* 13:484
  19. Petranovic D, Vemuri GN (2009) Impact of yeast systems biology on industrial biotechnology. *J Biotechnol* 144(3):204–211
  20. Bruggeman FJ, Westerhoff HV (2007) The nature of systems biology. *Trends Microbiol* 15(1):45–50
  21. Benson DA, Clark K, Karsch-Mizrachi I et al (2014) GenBank. *Nucleic Acids Res* 42(Database issue):D32–D37
  22. Barrett T, Edgar R (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* 411:352–369
  23. Vizcaino JA, Cote RG, Csordas A et al (2013) The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* 41(Database issue):D1063–D1069
  24. Haug K, Salek RM, Conesa P et al (2013) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res* 41(Database issue):D781–D786
  25. Croft D, Mundo AF, Haw R et al (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res* 42(Database issue):D472–D477
  26. Kanehisa M, Goto S, Sato Y et al (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42(Database issue):D199–D205
  27. UniProt C (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 42(Database issue):D191–D198
  28. Buchel F, Rodriguez N, Swainston N et al (2013) Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC Syst Biol* 7:116
  29. Wrzodek C, Buchel F, Ruff M et al (2013) Precise generation of systems biology models from KEGG pathways. *BMC Syst Biol* 7:15
  30. Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. *Nucleic Acids Res* 34(Database issue):D504–D506
  31. Cerami EG, Gross BE, Demir E et al (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* 39(Database issue):D685–D690
  32. Stobbe MD, Swertz MA, Thiele I et al (2013) Consensus and conflict cards for metabolic pathway databases. *BMC Syst Biol* 7:50
  33. Flicek P, Amode MR, Barrell D et al (2014) Ensembl 2014. *Nucleic Acids Res* 42(Database issue):D749–D755
  34. Karolchik D, Barber GP, Casper J et al (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 42(Database issue):D764–D770
  35. Chelliah V, Laibe C, Le Novere N (2013) BioModels Database: a repository of mathematical models of biological processes. *Methods Mol Biol* 1021:189–199
  36. Thiele I, Swainston N, Fleming RM et al (2013) A community-driven global reconstruction of human metabolism. *Nat Biotechnol* 31(5):419–425
  37. Wolstencroft K, Owen S, du Preez F et al (2011) The SEEK: a platform for sharing data and models in systems biology. *Methods Enzymol* 500:629–655
  38. Rappaport N, Nativ N, Stelzer G et al (2013) MalaCards: an integrated compendium for diseases and their annotation. *Database (Oxford)* 2013:bat018
  39. Safran M, Dalah I, Alexander J et al (2010) GeneCards Version 3: the human gene integrator. *Database (Oxford)* 2010:baq020

40. Stelzer G, Dalah I, Stein TI et al (2011) In-silico human genomics with GeneCards. *Hum Genomics* 5(6):709–717
41. Schluter A, Real-Chicharro A, Gabaldon T et al (2010) PeroxisomeDB 2.0: an integrative view of the global peroxisomal metabolome. *Nucleic Acids Res* 38(Database issue):D800–D805
42. Geifman N, Rubin E (2013) The mouse age phenome knowledgebase and disease-specific inter-species age mapping. *PLoS One* 8(12):e81114
43. Shaw DR (2009) Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Curr Protoc Bioinformatics*. Chapter 1:Unit1 7
44. Nam D, Kim SY (2008) Gene-set approach for expression pattern analysis. *Brief Bioinform* 9(3):189–197
45. Levine AJ, Miller JA, Shapshak P et al (2013) Systems analysis of human brain gene expression: mechanisms for HIV-associated neurocognitive impairment and common pathways with Alzheimer's disease. *BMC Med Genomics* 6:4
46. Pandey P, Qin S, Ho J et al (2011) Systems biology approach to identify transcriptome reprogramming and candidate microRNA targets during the progression of polycystic kidney disease. *BMC Syst Biol* 5:56
47. Jiang Q, Wang Y, Hao Y et al (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 37(Database issue):D98–D104
48. Amberger J, Bocchini C, Hamosh A (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). *Hum Mutat* 32(5):564–567
49. Landrum MJ, Lee JM, Riley GR et al (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42(Database issue):D980–D985
50. Bhagwat M (2010) Searching NCBI's dbSNP database. *Curr Protoc Bioinformatics*. Chapter 1:Unit 1 19
51. Lappalainen I, Lopez J, Skipper L et al (2013) DbVar and DGVA: public archives for genomic structural variation. *Nucleic Acids Res* 41(Database issue):D936–D941
52. Roque FS, Jensen PB, Schmock H et al (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol* 7(8):e1002141
53. On not reinventing the wheel (2012) *Nat Genet* 44(3):233.
54. Kohonen-Corish MR, Smith TD, Robinson HM et al (2013) Beyond the genomics blueprint: the 4th Human Variome Project Meeting, UNESCO, Paris, 2012. *Genet Med* 15(7):507–512
55. Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 507(7492):315–322
56. Eichler HG, Abadie E, Breckenridge A et al (2012) Open clinical trial data for all? A view from regulators. *PLoS Med* 9(4):e1001202
57. Rodwin MA, Abramson JD (2012) Clinical trial data as a public good. *JAMA* 308(9):871–872
58. Artene SA, Ciurea ME, Purcaru SO et al (2013) Biobanking in a constantly developing medical world. *ScientificWorldJournal* 2013:343275
59. Yuille M, van Ommen GJ, Brechot C et al (2008) Biobanking for Europe. *Brief Bioinform* 9(1):14–24
60. Vickers AJ (2006) Whose data set is it anyway? Sharing raw data from randomized trials. *Trials* 7:15
61. Tudur SC, Dwan K, Altman DG et al (2014) Sharing individual participant data from clinical trials: an opinion survey regarding the establishment of a central repository. *PLoS One* 9(5):e97886
62. Tryka KA, Hao L, Sturcke A et al (2014) NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* 42(Database issue):D975–D979
63. Jensen PB, Jensen LJ, Brunak S (2012) Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 13(6):395–405
64. Oti M, Huynen MA, Brunner HG (2008) Phenome connections. *Trends Genet* 24(3):103–106
65. Denny JC, Ritchie MD, Basford MA et al (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover disease associations. *Bioinformatics* 26(9):1205–1210
66. Shah NH (2013) Mining the ultimate phenome repository. *Nat Biotechnol* 31(12):1095–1097
67. Rzhetsky A, Wajngurt D, Park N et al (2007) Probing genetic overlap among complex human phenotypes. *Proc Natl Acad Sci U S A* 104(28):11694–11699
68. Denny JC, Bastarache L, Ritchie MD et al (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 31(12):1102–1110
69. Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753
70. van der Sijde MR, Ng A, Fu J (2014) Systems genetics: from GWAS to disease pathways. *Biochim Biophys Acta* 1842(10):1903–1909

71. Hou L, Zhao H (2013) A review of post-GWAS prioritization approaches. *Front Genet* 4:280
72. Choi H, Pavelka N (2011) When one and one gives more than two: challenges and opportunities of integrative omics. *Front Genet* 2:105
73. Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* 7(3):198–210
74. Kristensen VN, Lingjaerde OC, Russnes HG et al (2014) Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer* 14(5):299–313
75. Sansone SA, Rocca-Serra P, Field D et al (2012) Toward interoperable bioscience data. *Nat Genet* 44(2):121–126
76. Stobbe MD, Houten SM, Jansen GA et al (2011) Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Syst Biol* 5:165
77. van Iersel MP, Pico AR, Kelder T et al (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11:5
78. Bjerrum JT, Rantalainen M, Wang Y et al (2014) Integration of transcriptomics and metabolomics: improving diagnostics, biomarker identification and phenotyping in ulcerative colitis. *Metabolomics* 10(2):280–290
79. Meng C, Kuster B, Culhane AC et al (2014) A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 15:162
80. Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418):61–70
81. Poste G (2011) Bring on the biomarkers. *Nature* 469(7329):156–157
82. Yuan Y, Van Allen EM, Omberg L et al (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* 32(7):644–652
83. Kim D, Shin H, Sohn KA et al (2014) Incorporating inter-relationships between different levels of genomic data into cancer clinical outcome prediction. *Methods* 67(3):344–353
84. The CASyM roadmap: Implementation of Systems Medicine across Europe, version 1.0 (2014)
85. Golub TR, Slonim DK, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439):531–537
86. Bansal M, Belcastro V, Ambesi-Impiombato A et al (2007) How to infer gene networks from expression profiles. *Mol Syst Biol* 3:78
87. Lee HK, Hsu AK, Sajdak J et al (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res* 14(6):1085–1094
88. Butte AJ, Tamayo P, Slonim D et al (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A* 97(22):12182–12186
89. Basso K, Margolin AA, Stolovitzky G et al (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet* 37(4):382–390
90. Song L, Langfelder P, Horvath S (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* 13:328
91. Friedman N, Linial M, Nachman I et al (2000) Using Bayesian networks to analyze expression data. *J Comput Biol* 7(3–4):601–620
92. Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. Adaptive computation and machine learning. MIT Press, Cambridge, MA
93. Kim SY, Imoto S, Miyano S (2003) Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief Bioinform* 4(3):228–235
94. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S et al (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res* 41(Database issue):D816–D823
95. Gerstein MB, Kundaje A, Hariharan M et al (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91–100
96. Franke L, van Bakel H, Fokkens L et al (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 78(6):1011–1025
97. Guan Y, Myers CL, Lu R et al (2008) A genomewide functional network for the laboratory mouse. *PLoS Comput Biol* 4(9):e1000165
98. Franceschini A, Szklarczyk D, Frankild S et al (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41(Database issue):D808–D815
99. De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. *Nat Rev Microbiol* 8(10):717–729
100. Bar-Joseph Z, Gerber GK, Lee TI et al (2003) Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 21(11):1337–1342
101. Ernst J, Beg QK, Kay KA et al (2008) A semi-supervised method for predicting transcription factor-gene interactions in *Escherichia coli*. *PLoS Comput Biol* 4(3):e1000044
102. Li Y, Liang M, Zhang Z (2014) Regression analysis of combined gene expression regula-



- tion in acute myeloid leukemia. *PLoS Comput Biol* 10(10):e1003908
103. Zhu J, Zhang B, Smith EN et al (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* 40(7):854–861
  104. Zhang B, Gaiteri C, Bodea LG et al (2013) Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell* 153(3):707–720
  105. Greenawalt DM, Dobrin R, Chudin E et al (2011) A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res* 21(7):1008–1016
  106. Alon U (2007) An introduction to systems biology: design principles of biological circuits, vol 10, Chapman & Hall/CRC mathematical and computational biology. Chapman & Hall/CRC, Boca Raton, FL
  107. Segal E, Friedman N, Kaminski N et al (2005) From signatures to models: understanding cancer using microarrays. *Nat Genet* 37(Suppl):S38–S45
  108. Mitra K, Carvunis AR, Ramesh SK et al (2013) Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet* 14(10):719–732
  109. Zhao W, Langfelder P, Fuller T et al (2010) Weighted gene coexpression network analysis: state of the art. *J Biopharm Stat* 20(2):281–300
  110. PARIKSHAK NN, LUO R, ZHANG A et al (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155(5):1008–1021
  111. Gibbs DL, Gralinski L, Baric RS et al (2014) Multi-omic network signatures of disease. *Front Genet* 4:309
  112. Dittrich MT, Klau GW, Rosenwald A et al (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 24(13):i223–i231
  113. Vandin F, Upfal E, Raphael BJ (2011) Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 18(3):507–522
  114. Vaske CJ, Benz SC, Sanborn JZ et al (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26(12):i237–i245
  115. Drier Y, Sheffer M, Domany E (2013) Pathway-based personalized analysis of cancer. *Proc Natl Acad Sci U S A* 110(16):6388–6393
  116. Chuang HY, Lee E, Liu YT et al (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3:140
  117. Staiger C, Cadot S, Györfy B et al (2013) Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front Genet* 4:289
  118. Ala U, Piro RM, Grassi E et al (2008) Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput Biol* 4(3):e1000043
  119. Clark C, Kalita J (2014) A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics* 30(16):2351–2359
  120. Berg J, Lassig M (2006) Cross-species analysis of biological networks by Bayesian alignment. *Proc Natl Acad Sci U S A* 103(29):10967–10972
  121. Kolar M, Meier J, Mustonen V et al (2012) GraphAlignment: Bayesian pairwise alignment of biological networks. *BMC Syst Biol* 6:144
  122. Sweet-Cordero A, Mukherjee S, Subramanian A et al (2005) An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat Genet* 37(1):48–55
  123. Bild AH, Yao G, Chang JT et al (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439(7074):353–357
  124. Anvar SY, Tucker A, Vinciotti V et al (2011) Interspecies translation of disease networks increases robustness and predictive accuracy. *PLoS Comput Biol* 7(11):e1002258
  125. Hu Y, Wu G, Rusch M et al (2012) Integrated cross-species transcriptional network analysis of metastatic susceptibility. *Proc Natl Acad Sci U S A* 109(8):3184–3189
  126. Altschuler GM, Hofmann O, Kalatskaya I et al (2013) Pathprinting: an integrative approach to understand the functional basis of disease. *Genome Med* 5(7):68
  127. Chen JC, Alvarez MJ, Talos F et al (2014) Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell* 159(2):402–414
  128. Taylor CF, Field D, Sansone SA et al (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* 26(8):889–896
  129. Chervitz SA, Deutsch EW, Field D et al (2011) Data standards for Omics data: the basis of data sharing and reuse. *Methods Mol Biol* 719:31–69
  130. Rubin DL, Shah NH, Noy NF (2008) Biomedical ontologies: a functional perspective. *Brief Bioinform* 9(1):75–90
  131. Joosten RP, Vriend G (2007) PDB improvement starts with data deposition. *Science* 317(5835):195–196

132. Karp PD (1998) What we do not know about sequence analysis and sequence databases. *Bioinformatics* 14(9):753–754
133. Schnoes AM, Brown SD, Dodevski I et al (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 5(12):e1000605
134. Stobbe MD, Houten SM, van Kampen AH et al (2012) Improving the description of metabolic networks: the TCA cycle as example. *FASEB J* 26(9):3625–3636
135. Wong WC, Maurer-Stroh S, Eisenhaber F (2010) More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput Biol* 6(7):e1000867
136. Kulikowski CA, Kulikowski CW (2009) Biomedical and health informatics in translational medicine. *Methods Inf Med* 48(1):4–10
137. Kulikowski CA, Shortliffe EH, Currie LM et al (2012) AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. *J Am Med Inform Assoc* 19(6):931–938
138. Martin-Sanchez F, Iakovidis I, Norager S et al (2004) Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *J Biomed Inform* 37(1):30–42
139. Crosswell LC, Thornton JM (2012) ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol* 30(5):241–242
140. Omberg L, Ellrott K, Yuan Y et al (2013) Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet* 45(10):1121–1126
141. Chen R, Mias GI, Li-Pook-Than J et al (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148:1293–1307
142. Guan Y, Gorenshteyn D, Burmeister M et al (2012) Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Comput Biol* 8(9):e1002694
143. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45(6):580–585
144. sbv IMPROVER project team, Ansari S, Binder J et al (2013) On crowd-verification of biological networks. *Bioinform Biol Insights* 7:307–325
145. Olsen C, Fleming K, Prendergast N et al (2014) Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics* 103(5–6):329–336
146. Meyer P, Alexopoulos LG, Bonk T et al (2011) Verification of systems biology research in the age of collaborative competition. *Nat Biotechnol* 29(9):811–815
147. Stolovitzky G, Monroe D, Califano A (2007) Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann NY Acad Sci* 1115:1–22
148. Marbach D, Prill RJ, Schaffter T et al (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci U S A* 107(14):6286–6291
149. Marbach D, Costello JC, Kuffner R et al (2012) Wisdom of crowds for robust gene network inference. *Nat Methods* 9(8):796–804

# Chapter 3

## Systems Medicine: The Future of Medical Genomics, Healthcare, and Wellness

Mansoor Saqi, Johann Pellet, Irina Roznovat, Alexander Mazein, Stéphane Ballereau, Bertrand De Meulder, and Charles Auffray

### Abstract

Recent advances in genomics have led to the rapid and relatively inexpensive collection of patient molecular data including multiple types of omics data. The integration of these data with clinical measurements has the potential to impact on our understanding of the molecular basis of disease and on disease management. Systems medicine is an approach to understanding disease through an integration of large patient datasets. It offers the possibility for personalized strategies for healthcare through the development of a new taxonomy of disease. Advanced computing will be an important component in effectively implementing systems medicine. In this chapter we describe three computational challenges associated with systems medicine: disease subtype discovery using integrated datasets, obtaining a mechanistic understanding of disease, and the development of an informatics platform for the mining, analysis, and visualization of data emerging from translational medicine studies.

**Key words** Systems medicine, P4 medicine, Disease subtyping, Patient stratification, Translational informatics

---

### 1 Introduction

The ability to collect large volumes of molecular data as well as detailed clinical measurements for patients will impact on disease classification and clinical management. Systems medicine is an approach that uses the concepts and methods of systems biology to understand disease conditions through an integration of data at multiple levels of biological organization [1]. Systems biology has progressed rapidly in recent years due to advances in technology that enable the rapid and increasingly inexpensive capture of multiple “omics” data (e.g., genomics, epigenomics, transcriptomics, proteomics, metabolomics), together with advances in computing that make it possible to store, query, and analyze the associated large datasets. An important feature of systems medicine is the interplay between biology, computation, and technology [2].

Many diseases are *heterogeneous*, that is, they are associated with a variety of phenotypes, which sometimes overlap. Such diseases represent a collection of subtypes, each characterized by, for example, different aberrant pathways and processes. Patient stratification involves identification of the particular subtype of disease from which a patient is suffering, and this can impact on drug discovery toward more personalized and effective treatments. If a disease condition is considered as a single *homogeneous* entity, potentially useful therapeutics could be discarded as they may show no overall beneficial effect in the cohort as a whole. However, these therapeutics might be of value to a selected group of patients with a particular disease subtype. In addition, an understanding of the mechanistic basis of disease subtypes could lead to the development of novel subtype-specific medicines.

When carried out on a large scale, the application of systems approaches to medicine offers the potential for the development of a new taxonomy of disease, namely, a taxonomy based on molecular mechanistic features rather than the presentation of clinical symptoms (see, e.g., [3, 4]). For some diseases, such a classification may reveal a number of subtypes, each involving different molecular pathways and processes. This new classification can lead to a more individualized approach to therapy, as the identification of disease subtypes can directly impact on clinical management, with therapeutic intervention tailored to the disease subtype of the patient. A new taxonomy might also suggest that several apparently different diseases, hitherto thought to be separate and distinct conditions, share common mechanisms at the molecular level, and such information could be useful for drug repurposing (see, e.g., [5]). Barabasi and coworkers [6] have used a network approach to connect human disease genes (the disease genome) with various human diseases (the disease phenome). The relationships are represented as a bipartite graph from which two networks are extracted, a human disease network (HDN) and a disease gene network (DGN). The modular structure of the HDN approach reveals connections between diseases that may appear different at the phenotypic level and that of the DGN shows groups of genes that share a disease phenotype.

Systems medicine involves the collection of large amounts of data, including clinical data, omics data, and, recently, data on patients' environment and activities collected through devices that make use of wearable sensor technologies. This has led to a new approach to personalized medicine, namely, P4 systems medicine, which is personalized, preventive, predictive, and participatory [7–10]. The aim is to develop personalized healthcare plans, to monitor the status of a patient's wellness so that early intervention can be made when appropriate, opening the possibility, through the identification of actionable molecular or lifestyle targets, to prevent the transition from wellness to disease, or to promote the

reversal from the early stage of disease onset to the normal condition. The ability to accurately predict disease progression would mean that unnecessary or inappropriate therapy could be avoided thereby saving costs and limiting exposure to potential side effects of the therapy. Additionally and importantly, P4 systems medicine is participatory, as it involves active engagement of researchers, clinicians, and patient groups empowered through social networks. Wearable sensor technologies will mean that patients can collect their own lifestyle and exposure measurements to score and comment on how they feel on a regular basis and share these data easily with the stakeholders they choose. The knowledge gained from community participation (e.g., response of individuals to particular therapies or other actionable interventions) can be fed back into computational systems biology models. P4 systems medicine will also need to address challenges associated with societal issues such as ethics, privacy, and data security or educational issues such as understanding how the patient can be seen as his/her own control in monitoring the transition from health to disease states [11].

Advances in genomics have triggered a revolution in medical genetics, reducing the cost of sequencing, accelerating health-improvement projects, and providing a comprehensive resource on human genetic variants establishing the link between the genotype and the phenotype [12–15]. The 1000 Genomes Project released a catalogue of validated loss-of-function (LoF) variants and naturally occurring “knockout” alleles for over 1000 human protein-coding genes; many of these genes have minimal functional annotation [16, 17]. Coding variants could affect human fitness with regard to responses against pathogens and heightened susceptibility to infection.

In oncology, there have been several recent examples of the clinical potential of a strategy involving diagnosis of subtype followed by a specific therapy. Most of these have involved the presence or absence of specific mutations or chromosomal rearrangements that can be indicative of disease prognosis or drug response. The relatively short time in which *crizotinib* [18] was demonstrated to be an effective therapy for a subset of patients with non-small cell lung cancer (NSCLC) indicates how patient mutational status can be one route to stratification. A subset of patients with NSCLC shows a chromosomal inversion that leads to the production of a fused protein encoded by a recombination of the echinoderm microtubule-associated protein-like 4 (ELM4) gene and the anaplastic lymphoma kinase (ALK) gene. This presence of the fusion protein can act as a diagnostic and also a target for the drug *crizotinib* [19]. Another example of a therapeutic strategy showing the potential of testing for given mutations combined with individualized therapies comes from colorectal cancer (CRC). In CRC, patients with particular mutations in the KRAS protein (which is involved in signaling pathways) show poor

response to the epidermal growth factor receptor (EGFR) inhibitor drugs, *panitumumab* and *cetuximab* [20].

Respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD) are examples of complex *heterogeneous* diseases. These diseases, characterized by airway inflammation and remodeling and airflow limitation, involve various types of interacting elements at the molecular level and provide illustrative examples for systems medicine approaches. For a number of patients suffering from severe asthma, the disease cannot be controlled by corticosteroid therapy. A recent systems medicine project U-BIOPRED (Unbiased BIOmarkers for the PREDiction of respiratory disease outcomes) aims at the identification of asthma subtypes and stratification of patients with respect to subtype that offers the prospect of more individualized approaches to treatment [21]. The Synergy-COPD project [22, 23] is another recent example of a systems medicine initiative aimed at understanding the heterogeneity of COPD and the associated patterns of comorbidity. As part of Synergy-COPD, Turan and coauthors [24] explored skeletal muscle wasting in COPD patients, by integrating physiological and gene expression data to build molecular networks which could then be tested for pathway and functional enrichment. The extent to which asthma and COPD share common mechanisms has been discussed in the literature [25] and illustrates some of the challenges of understanding complex *heterogeneous* disease. Specifically, the authors have curated common pathways and developed gene networks for four major respiratory diseases (asthma, COPD, tuberculosis, and essential hypertension) based on specialized studies from literature. The network overlap between these disease types has been analyzed, with the highest being identified between asthma and COPD. These results show stronger association between asthma and COPD than between the other analyzed phenotypes, suggesting the potential of developing therapeutic strategies to target both these diseases.

Finally, one of the most recent initiatives in P4 systems medicine is the pioneers of health and wellness pilot project [8]. The project targets deep characterization of wellness and proposes to (1) obtain a complete genome sequence for each individual; (2) follow with digital-monitoring devices that measure heart rate, activity, quality of sleep, weight, and blood pressure; and (3) follow every 3 months measurements of blood metabolites; blood organ-specific proteins for the brain, heart, and liver; the gut microbiome; salivary cortisol; white cell methylation; and telomeric lengths and clinical chemistries focused on nutrition, thus combining approaches that have proven to be effective by two pioneering individuals who monitored themselves for environmental factors and lifestyle [26] or multiple omics [27] which when combined with clinical assessments enabled the identification of early signs of

disease occurrence that they could counteract through appropriate individual adaptations.

Large-scale systems medicine studies will require robust computational pipelines. The application of computing methodologies to the domain of translational medicine research to enable the storage, mining, analysis, and visualization of large patient datasets is sometimes referred to as translational informatics. In the next sections, we describe three computational challenges associated with P4 systems medicine, namely:

1. Integrative approaches to subtype discovery
2. Obtaining a mechanistic understanding of disease subtypes
3. Developing a platform for translational informatics

We conclude with a brief discussion of some of the broader educational challenges associated with the development of systems medicine.

---

## 2 Integrative Approaches to Disease Subtype Discovery

Biomarkers are patterns that discriminate between disease and non-disease or between different disease subtypes. Identification of subtypes could suggest appropriate therapeutic strategies. Biomarkers can also be used prognostically, for example, to predict whether a patient may have an aggressive or benign form of a disease. The components of the patterns could be a set of genes or metabolites or other measured biological features.

Relatively few clinically useful biomarkers have been developed, although extensive work has been done on biomarker studies during the last decades [28]. Given that biological data is generally noisy (perhaps due to the *heterogeneous* nature of most complex diseases), it is difficult to obtain reproducible molecular signatures. Filtering and integrating, which exploit prior knowledge and the ability to group together data, can help to improve the signal to noise ratio [29]. In an analysis of breast cancer metastasis [30], gene expression patterns were integrated with protein-protein interaction networks and biomarkers were identified as subgraphs. These network-based signatures were more robust than markers based on individual genes, showing greater reproducibility across studies.

Biomarker identification is usually carried out in a supervised manner. Supervised approaches use the patient phenotype as a class label associated with each patient. This might be assigned from the clinical presentation of the patient and could, for example, be associated with severity or aggressiveness of the disease. Supervised approaches aim to find sets of features that distinguish between classes.

Unsupervised approaches to disease classification make no assumption as to patient phenotype, and, essentially, group data according to similarities among the molecular features (e.g., gene expression, metabolomics profiles), resulting in groups (*clusters*) that may represent disease subtypes. These may correspond to already known differences in the phenotypes or may represent some as-yet undiscovered subtypes reflecting some perturbation of underlying molecular processes and pathways that are not immediately apparent from the patient's clinical measurements.

Omics measurements from different platforms can provide complementary types of information, and the collection and integration of data from multiple omics platforms can suggest novel disease subtypes. Several cancer-related studies have recently been published with multiple omics data types collected for the same group of patients (e.g., refer to [31]). This strategy of data collection is likely to become increasingly common in translational medicine [32, 33]. In the context of asthma treatment, this approach has been taken, for example, with the U-BIOPRED consortium, which aims to develop molecular fingerprint and handprint signatures that will lead to a better classification of the different types of severe asthma [34]. Knowledge of these subtypes may help in the development of better types of treatment for asthma patients.

One might expect that the underlying biology associated with different disease subtypes would be reflected within different types of molecular data collected across the patient cohort. The identification of consistent patterns across different omics platforms may, therefore, reflect more reliable subgroupings. Alternatively, it may be the case that the signal from an individual platform is too weak to distinguish between disease subtypes, but taken together, the data might lead to the discovery of robust patterns that are useful diagnostically and prognostically. Disease subtype discovery using multiple genomics datasets of different types (e.g., gene expression, copy number variation, methylation) collected for the same set of patients can offer new insights into the taxonomy of disease. Each data type can be clustered separately, and the concordance and the conflict between clusters can be explored. An integrative approach should ideally identify (1) molecular patterns (signatures) that are common across the different omics datasets, (2) patterns that are specific to individual datasets, and also (3) patterns that only emerge after data integration [35]. We report below a few examples of integrative approaches to the subtype discovery problem. Patterns specific to a given data type can be termed *fingerprints*, and those associated with the integrated data can be termed *handprints*.

The multiple "omics" data types can include gene expression, proteomics, and metabolomics. The data matrices associated with each data type  $j$  are of dimensionality  $n$  by  $p_j$  where the number of patients is  $n$  and the index  $p_j$  refers to the number of features or



variables for data type  $j$ . Typically,  $p_j$  will be different across different platforms. Several exploratory data analysis methods have been proposed to compare two omics datasets ( $j=2$ ). These include *partial least squares (PLS)* and *canonical correlation analysis (CCA)*, and *sparse approaches* have been used to perform the integration and variable selection together [36].

Co-inertia analysis (CIA) [37] measures the degree to which two datasets are in concordance and is suitable for datasets where the number of features (variables) exceeds the number of samples (patients) which is usually the case with omics data. An approach using CIA has been applied to compare microarray data from two different platforms [38]. Recently, this approach has been extended to handle more than two omics datasets [39], and the multiple co-inertia analysis method is available as an R package (*omicade4*). For example, given the same set of ( $n$ ) patients for which multiple data types are available (such as gene expression, transcripts, methylation levels), the *omicade4* package performs co-inertia analysis by combining information from all these datasets provided for the  $n$  patients. Note that all datasets must have one common dimension (i.e., the patient number), while the second dimension can differ.

Shen and coworkers [35] proposed an integrative method called *iCluster* using a joint latent variable approach. *iCluster* performs a simultaneous clustering of omics datasets, which are represented by ( $p \times n$ ) matrices with  $p$  being the number of features (e.g., genes) and  $n$  the number of patients. The method simultaneously projects high-dimensional data matrices associated with various omics platforms and with different numbers of features onto a unified latent space of lower dimensionality. Simulations show that clustering in the latent space produces a better separation than using PCA [40]. A recent breast cancer study using *iCluster* demonstrated the value of integration through the identification of subtypes that were not suggested by the component data platforms of gene expression and copy number [41]. The program *iCluster+* (a further development to *iCluster*) permits integration of different data types, e.g., binary, categorical, and continuous [42]. The program is available as an R package.

Kirk and coauthors [43] use an unsupervised Bayesian correlated clustering approach for *multiple dataset integration (MDI)*. MDI allows the identification of subsets of samples that cluster across several different datasets. Lock and Dunson [44] described a flexible integrated approach that allows an overall clustering to identify shared structure and a clustering that is specific to each data modality.

A network-based approach to subtype discovery was developed by [45]. A similarity network is constructed for each data type, such as gene expression or DNA methylation. These individual networks are fused to form an integrated network. An advantage of this approach is that strong similarities supported by evidence from

several data modalities are retained, as well as some weak similarities that share a common tightly connected network neighborhood across the individual networks. The method has been demonstrated [45] to detect clinically relevant subtypes for a variety of cancer datasets from the Cancer Genome Atlas TCGA.

### 3 Toward a Mechanistic Understanding of Disease Subtypes

After identifying putative molecular signatures that are associated with disease subtypes, the next challenge for translational informatics is to use these signatures to get a mechanistic insight into disease *heterogeneity*. Pathways and networks describe a level of functional organization that is between molecular function and physiological function. As such, mapping genes, which are suspected to be involved in disease, to pathways and networks can give insight into potential mechanisms that may be involved in the disease process and could also suggest strategies for therapeutic intervention. Pathways and networks represent collections of molecular components that interact in some way and participate in a given process, and these collections can be represented in a variety of ways and at different levels of granularity. Traditional pathway and network maps captured metabolic reactions showing the associated reactants and products. These have been extended to include signaling pathways and pathways involved in disease.

The Systems Biology Graphical Notation (SBGN) project [46] aimed to provide a standard representation of molecular pathways and networks while recognizing that their complexity required different views depending on the visualization requirements. The SBGN standard [46] consists of three complementary languages: process description (PD), activity flow (AF), and entity relationship (ER). Each of these three complementary languages has certain purpose, advantages, and limitations (refer to Table 1). The process description is currently the most widely used language.

**Table 1**  
**Advantages and limitations of the three complimentary languages within SBGN**

Features	Process description	Activity flow	Entity relationship
Ambiguity	Unambiguous	Ambiguous	Unambiguous
Sequence of events	Sequential	Sequential	Nonsequential
Advantage	Clear sequence of events	Compactness	Can deal with combinatorial explosion
Limitation	Cannot deal with combinatorial explosion	Ambiguity	Sequence of events cannot be shown

It represents biological events such as metabolic reactions, protein phosphorylation, and complex formation in an unambiguous way and depicts causal sequences of events that are well-suited for mathematical modeling and simulation. This language is arguably the best environment for knowledge representation that can be used both by mathematicians and biologists to accurately express detailed information about a biological system and use it for model development, hypothesis generation, and predictions. One of the major limitations of this language, similarly to other pathway map visualizations, is that it cannot deal with potential combinatorial explosion [47]. The activity flow language can be seen as a simplified version of process description language with fewer details and the focus on activity transformation from one molecule to another in a pathway. This SBGN language is the closest to the commonly used signaling pathway diagrams, for example, in BioCarta. This level of representation is a good fit for omics data visualization and functional analysis. Similar less-detailed compact diagrams are used, for example, in Ingenuity Pathway Analysis (Qiagen) and MetaCore (Thomson Reuters). The entity relationship language loses the sequential expressiveness of the other two languages but instead can very well deal with combinatorial explosion. The entity relationship diagram cannot be read as a pathway or network but rather as a set of states one molecule can be in depending on the influences from other molecules. There are many software applications that support SBGN diagrams ([www.sbgn.org/SBGN\\_Software](http://www.sbgn.org/SBGN_Software)).

CellDesigner [48] is a software used for developing diagrams in a format compliant with SBGN process description language, with visual elements that in most cases correspond to visual elements of SBGN process description. These diagrams can be exported from CellDesigner in SBGN and SBML formats.

Two disease maps have recently been constructed in the area of neurodegenerative disease. Mizuno and coauthors [49] have developed a map of signaling pathways in Alzheimer's disease, and Fujita and coauthors [50] have constructed a Parkinson's disease (PD) map which captures known components involved in gene regulatory and metabolic processes associated with this disease. Both the Alzheimer's map and the PD map are developed in CellDesigner in SBGN compliant format. The PD map is coupled to bioinformatics tools that, for example, explore the overlay of differentially expressed genes from a gene expression study in order to allow identification of the main pathways that may be perturbed. The map can be extended by the addition of other data types such as protein interaction data and associations derived from text mining that can facilitate hypothesis discovery.

A major challenge in the construction of disease maps is the extraction of information from the scientific literature. While text-mining approaches have considerable potential [51], this needs to

be supplemented by manual curation by domain experts if a reliable high-quality map is to be developed. Another challenge is the level of detail to be included. A description of a disease condition at the molecular level ideally needs also to capture the effect of, for example, single-nucleotide polymorphism (SNP) or information relating to enrichment of gene expression in particular tissue types. Quantitative information about reaction kinetics could be included, which would allow simulations and modeling to be carried out. However, mostly only qualitative information about correlative relationships can be found. The biological expression language (BEL) [52] is a framework for capturing causal and correlative relationships and has the advantage that it is expressive and human readable and can be extended. The BEL framework does not have its own ontology but makes use of existing ontologies.

Disease maps represented at multiple levels of granularity are important to obtain mechanistic insight into disease subtypes and to put biological context around experimental results. Although details of sequential reaction steps and temporal and spatial information in relation to these processes are valuable, many relationships described in the scientific literature are at a much higher level of granularity suggesting that entity A has an effect on entity B. Malhotra and coauthors [53] describe an integrative approach to put functional context around putative biomarkers by capturing more speculative interactions and relationships using text mining and by including protein-protein interaction networks and gene expression data. This strategy is a synthesis of data driven and background knowledge-based approaches.

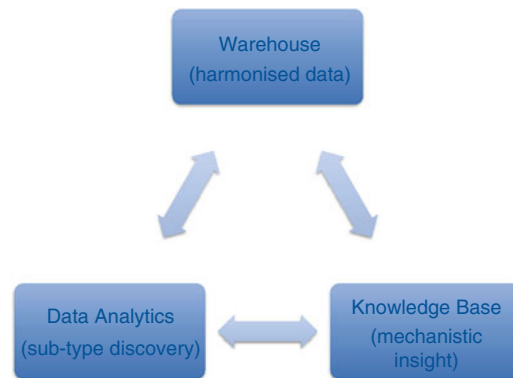
Multi-scale modeling, which aims to describe functionality of the whole system rather than of its components, is being increasingly used in systems medicine during recent years in order to facilitate exploration of key features at various scales, from molecular (e.g., metabolite network) to cellular (e.g., intercellular influences) and macroscopic levels (e.g., phenotype manifestation) [54]. Major advantages of developing multi-scale models in systems medicine are represented by the possibility of integrating experimental data at different scales, exploring features of phenomena across system layers, and investigating drug effects on the whole biological system. Although considered powerful, the multi-scale description approach is characterized by (1) a large set of parameters needed to represent system compounds and (2) various ranges of both spatial and temporal scales (e.g., from rapid dynamics and microscopic spaces, such as metabolic activity occurring at second order within the nucleus, to slow dynamics and extensive spaces, such as disease progression observed at the tissue and organ level over years). The multi-scale modeling approach has been used in cancer medicine to explore, e.g., abnormal phenomena during breast, colorectal, lung, and prostate cancers, reviewed in [54, 55] and references therein; in drug discovery and development research to investigate

the impact, e.g., of cytostatic agents on tumor progression [56] of danoprevir inhibitor on hepatitis C virus dynamics [57]; and in epigenetics research to analyze relationships between aging and aberrant modifications of the transgenerational epigenetic inheritance mechanisms [58]. A recent paper by [59] introduces a multi-scale computational model developed to predict lung functional deregulations during respiratory disease development by integrating information on CT images from COPD and asthma patients.

## 4 Developing a Platform for Translational Informatics

What systems need to be in place to enable the exploitation of the large datasets that are being collected as part of ongoing and planned systems medicine initiatives? Although the cost of genome sequencing has dropped dramatically in recent years to around 1000 US\$, the cost of analysis and interpretation is still considerably higher by two orders of magnitude [60]. Additionally, implementation of systems medicine approaches to health and wellness transitions, drug development, and treatment will require a computational infrastructure to allow for storage, retrieval, and mining of data in an integrated manner (refer to Fig. 1).

In addition to omics datasets being collected from different platforms (such as gene expression, copy number variation methylation, etc.), phenotypic data is also becoming richer. Instead of a single endpoint status representing the phenotype (e.g., disease or non-disease), a set of measurements may be collected (e.g., mild,



**Fig. 1** Three key components for a translational medicine platform: **(a)** A data warehouse for storage and querying of clinical and multi-omics data associated with patient cohorts; data harmonization is needed to ensure that the data conforms to standards, in order to facilitate comparison across different studies. **(b)** A data analytics component for visual exploration of the data and for subtype discovery. **(c)** A knowledge base to enable experimental results to be understood in the context of known disease pathways and processes and to suggest a possible mechanistic basis for subtypes

moderate, or severe disease). These can give a better description of the phenotype and may, if taken across time, describe disease progression or reversal [26, 27]. It is likely that such high-dimensional phenotype datasets (described by a large number of clinical measurements for each patient taken at given time points) will become increasingly important. As these datasets start to be routinely collected, each patient will be associated with a cloud of data consisting of millions to billions of data points, and the mining and analysis of this data is likely to offer insight into the onset of disease as well as transitions from health to disease and vice versa [8, 10].

Traditionally, various bioinformatics data repositories have tended to be centered around fixed data types. Until fairly recently, the identification of molecular signatures associated with disease conditions has been derived from analyses of gene expression data, with the main public repositories being the gene expression omnibus (GEO) [61] and ArrayExpress databases [62]. Other omics data types in addition to transcriptomics data can include proteomics, metabolomics, and data from genome structural variations, and these can provide additional molecular signatures (refer to, e.g., [63]). More recently, the database of genotypes and phenotypes (dbGAP) has been established as a repository for genotype-phenotype data and includes molecular data (e.g., expression, copy number variation, methylation) as well as phenotypic data and contextual information (e.g., research protocols).

A number of initiatives are underway aimed at making bioinformatics tools more accessible and at sharing analysis workflows and results: GenomeSpace ([www.genomespace.org](http://www.genomespace.org)) and Garuda ([www.garuda-alliance.org](http://www.garuda-alliance.org)) are frameworks for interoperability of bioinformatics tools; Galaxy [64] is a web-based platform for tools integration which also allows tracking of provenance and the sharing of workflows; Synapse [65] provides a framework for collaboration with particular emphasis on provenance and sharing, with respect to the data and also the results of analyses carried out on the data; cBioPortal [66] provides a web-based tool for analysis and visualization of cancer datasets of multiple data types, such as gene expression and copy number variation, and offers an R interface so that data in the repository can be queried from R scripts and a Matlab ([www.mathworks.com](http://www.mathworks.com)) toolbox to allow programmatic access from Matlab code ([www.mathworks.com](http://www.mathworks.com)); and OMICtools [67] is a manually curated repository for web-based tools related to “omics” data analysis.

Computational platforms to enable systems medicine will need to be accessible to researchers and clinicians and have advanced visualization functionalities to facilitate hypothesis generation. The data types that need to be stored, queried, and integrated include gene expression, proteomics, metabolomics data, and DNA structural variations such as chromosomal rearrangements, copy number variations, DNA methylation data,

microRNA data, as well as data associated with medical imaging and clinical data.

One of the main challenges in the development of a data repository for translational medicine studies is the semantic heterogeneity of the data. This means that the same concept (e.g., a clinical measurement) may be referred to by different names or different concepts may be referred to by the same name. The use of available standards for clinical data (e.g., CDISC) and for multi-omics data (ISA standards [68]) will help to address this challenge and will enable cross-study comparisons. However, there will remain a problem in the harmonization of legacy data, which do not conform to current standards, and this is likely to be resource intensive, involving semi-manual curation. A translational medicine platform also needs to be secure and to conform to legislation relating to data privacy.

The data analytics component of such a platform will need to handle common types of analyses, such as exploring attributes of the patient cohorts, required by clinicians and biomedical researchers, and include workflows for facilitating disease subtyping by the identification of molecular signatures from the omics datasets associated with each patient sample. Finally, the platform should put putative disease subtypes into biological context by using background knowledge in disease maps, to suggest a mechanistic basis for the subtypes.

An early example of the development of a translational informatics platform OncoPrint [69] aimed at the integration of microarray data from cancer studies. This platform attempted to address some of the challenges of semantic and syntactic heterogeneity of the data. The tranSMART platform [70, 71] was developed as a warehouse for both clinical and high-dimensional omics data such as gene expression and SNP data. The platform facilitates cohort selection and exploratory visual exploration of clinical data associated with the cohorts and has been integrated with more specialized analytics tools such as GenePattern ([www.genepattern.org](http://www.genepattern.org)). More recently, tranSMART has been integrated with Genedata Analyst for advanced analysis of a number of omics datasets [72]. The Innovative Medicines Initiative recently funded a project to build a platform for translational research eTRIKS (European Translational Information and Knowledge Management Services), which uses tranSMART at the core of its infrastructure.

To advance systems medicine clinical data, basic research data, mathematical modeling, and knowledge management need to be integrated and interlinked. Important prerequisites to achieve this are standards for data acquisition, harmonization of documentation, and a policy for data sharing [73]. BioXM, developed by Biomax Informatics, has been used for data management in the Synergy-COPD project [74], where information is represented as a network showing evidence that relates different biological



concepts [75]. This has been extended to allow the integration with applications for computational modeling and simulation and for clinical decision support systems [76]. Accessibility to a wide range of biomedical and clinical researchers can be an important feature in some environments for a data integration platform for translational medicine. T-MedFusion [77] is a system that integrates patient data, clinical measurements, and omics data that have been evaluated in use cases for psoriasis and rheumatoid arthritis. Recently, the STATegra project developed STATegra EMS to manage clinical with high-throughput omics data (RNA-seq, ChIP-seq, Methyl-seq, etc.) [78].

As large cohort longitudinal studies gain momentum, patients will be represented by millions to billions of data points based on the collection of increasingly complex and heterogeneous data types, and fast effective data analytics and visualization will be necessary if these platforms are to become clinical decision support tools. Cloud computing may provide a solution which is scalable and requires low start-up costs [79].

---

## 5 Conclusions

Systems medicine will involve the collection, integration, and analysis of large patient datasets. It offers the prospect of suggesting a new taxonomy of disease, from which patient-specific therapeutic strategies can be developed. It represents the best option to enable implementation of participatory, personalized, predictive, and preventive medicine, thus fostering the transition from the reactive practice of medicine and treating the symptoms when they have fully developed into a disease stage, to a proactive and anticipative medical practice based on a scientific understanding of wellness.

Large longitudinal studies will enable monitoring the transition from wellness to disease and identification of the associated perturbations in molecular networks. The application of genomic medicine will present several challenges. It will necessitate the education of both clinicians and the patient community. Modern medical curricula will need to reflect the importance of an integrative, holistic, patient-centered approach. It will also require developments in data integration, big data analytics, and visualization.

Personalized predictive approaches to medicine have the potential to impact significantly on patient wellness. The challenge to overcome, in order to make it endorsed and practiced across developed as well as very poor social environments and scientific and medical infrastructures, is to demonstrate that it could lead to reductions in healthcare costs through more effective strategies for disease management.



## Acknowledgments

This work was supported by the CNRS and in part by the EU grants to CA in the context of the U-BIOPRED consortium (Unbiased Biomarkers for the PREDiction of respiratory disease outcomes, Grant Agreement IMI n°115010), the MeDALL consortium (Mechanisms of the Development of Allergy, Grant Agreement FP7 n°264357), the eTRIKS consortium (European Translational Research Information & Knowledge Management Services, Grant Agreement n°115446), and the CASyM Coordinating Action Systems Medicine (Implementation of Systems Medicine across Europe, Grant Agreement n°305033).

## References

1. Auffray C, Chen Z, Hood L (2009) Systems medicine: the future of medical genomics and healthcare. *Genome Med* 1:2. doi:[10.1186/gm2](https://doi.org/10.1186/gm2)
2. Hood L (2013) Systems biology and p4 medicine: past, present, and future. *Rambam Maimonides Med J* 4:e0012. doi:[10.5041/RMMJ.10112](https://doi.org/10.5041/RMMJ.10112)
3. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease (2011) *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press, Washington, DC, USA
4. Loscalzo J, Barabasi A-L (2011) *Systems biology and the future of medicine*. Wiley Interdiscip Rev Syst Biol Med 3:619–627. doi:[10.1002/wsbm.144](https://doi.org/10.1002/wsbm.144)
5. Menche J, Sharma A, Kitsak M et al (2015) Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi:[10.1126/science.1257601](https://doi.org/10.1126/science.1257601)
6. Goh K-I, Cusick ME, Valle D et al (2007) The human disease network. *Proc Natl Acad Sci U S A* 104:8685–8690. doi:[10.1073/pnas.0701361104](https://doi.org/10.1073/pnas.0701361104)
7. Hood L, Tian Q (2012) Systems approaches to biology and disease enable translational systems medicine. *Genomics Proteomics Bioinformatics* 10:181–185. doi:[10.1016/j.gpb.2012.08.004](https://doi.org/10.1016/j.gpb.2012.08.004)
8. Hood L, Price ND (2014) Demystifying disease, democratizing health care. *Sci Transl Med* 6:225ed5. doi:[10.1126/scitranslmed.3008665](https://doi.org/10.1126/scitranslmed.3008665)
9. Flores M, Glusman G, Brogaard K et al (2013) P4 medicine: how systems medicine will transform the healthcare sector and society. *Pers Med* 10:565–576. doi:[10.2217/PME.13.57](https://doi.org/10.2217/PME.13.57)
10. Hood L, Auffray C (2013) Participatory medicine: a driving force for revolutionizing healthcare. *Genome Med* 5:110. doi:[10.1186/gm514](https://doi.org/10.1186/gm514)
11. Cesario A, Auffray C, Russo P, Hood L (2014) P4 medicine needs P4 education. *Curr Pharm Des.* 20(38):6071–2
12. Miller FA, Hayeems RZ, Bytautas JP et al (2014) Testing personalized medicine: patient and physician expectations of next-generation genomic sequencing in late-stage cancer care. *Eur J Hum Genet* 22:391–395. doi:[10.1038/ejhg.2013.158](https://doi.org/10.1038/ejhg.2013.158)
13. Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. *J Hum Genet* 59:5–15. doi:[10.1038/jhg.2013.114](https://doi.org/10.1038/jhg.2013.114)
14. Sachidanandam R, Weissman D, Schmidt SC et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933. doi:[10.1038/35057149](https://doi.org/10.1038/35057149)
15. Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. *Science* 291:1304–1351. doi:[10.1126/science.1058040](https://doi.org/10.1126/science.1058040)
16. 1000 Genomes Project Consortium, Abecasis GR, Auton A et al (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65. doi:[10.1038/nature11632](https://doi.org/10.1038/nature11632)
17. Clarke L, Zheng-Bradley X, Smith R et al (2012) The 1000 Genomes Project: data management and community access. *Nat Methods* 9:459–462. doi:[10.1038/nmeth.1974](https://doi.org/10.1038/nmeth.1974)
18. Gerber DE, Minna JD (2010) ALK inhibition for non-small cell lung cancer: from discovery

- to therapy in record time. *Cancer Cell* 18:548–551. doi:[10.1016/j.ccr.2010.11.033](https://doi.org/10.1016/j.ccr.2010.11.033)
19. Shaw AT, Yasothan U, Kirkpatrick P (2011) Crizotinib. *Nat Rev Drug Discov* 10:897–898. doi:[10.1038/nrd3600](https://doi.org/10.1038/nrd3600)
  20. Karapetis CS, Khambata-Ford S, Jonker DJ et al (2008) K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med* 359:1757–1765. doi:[10.1056/NEJMoa0804385](https://doi.org/10.1056/NEJMoa0804385)
  21. Auffray C, Adcock IM, Chung KF et al (2010) An integrative systems biology approach to understanding pulmonary diseases. *Chest* 137:1410–1416. doi:[10.1378/chest.09-1850](https://doi.org/10.1378/chest.09-1850)
  22. Roca J, Vargas C, Cano I et al (2014) Chronic obstructive pulmonary disease heterogeneity: challenges for health risk assessment, stratification and management. *J Transl Med* 12(Suppl 2):S3. doi:[10.1186/1479-5876-12-S2-S3](https://doi.org/10.1186/1479-5876-12-S2-S3)
  23. Gomez-Cabrero D, Lluch-Ariet M, Tegnér J et al (2014) Synergy-COPD: a systems approach for understanding and managing chronic diseases. *J Transl Med* 12(Suppl 2):S2. doi:[10.1186/1479-5876-12-S2-S2](https://doi.org/10.1186/1479-5876-12-S2-S2)
  24. Turan N, Kalko S, Stincone A et al (2011) A systems biology approach identifies molecular networks defining skeletal muscle abnormalities in chronic obstructive pulmonary disease. *PLoS Comput Biol* 7:e1002129. doi:[10.1371/journal.pcbi.1002129](https://doi.org/10.1371/journal.pcbi.1002129)
  25. Kaneko Y, Yatagai Y, Yamada H et al (2013) The search for common pathways underlying asthma and COPD. *Int J Chron Obstruct Pulmon Dis* 8:65–78. doi:[10.2147/COPD.S39617](https://doi.org/10.2147/COPD.S39617)
  26. Smarr L (2012) Quantifying your body: a how-to guide from a systems biology perspective. *Biotechnol J* 7:980–991. doi:[10.1002/biot.201100495](https://doi.org/10.1002/biot.201100495)
  27. Chen R, Mias GI, Li-Pook-Than J et al (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148:1293–1307. doi:[10.1016/j.cell.2012.02.009](https://doi.org/10.1016/j.cell.2012.02.009)
  28. McDermott JE, Wang J, Mitchell H et al (2013) Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert Opin Med Diagn* 7:37–51. doi:[10.1517/17530059.2012.718329](https://doi.org/10.1517/17530059.2012.718329)
  29. Ideker T, Dutkowski J, Hood L (2011) Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* 144:860–863. doi:[10.1016/j.cell.2011.03.007](https://doi.org/10.1016/j.cell.2011.03.007)
  30. Chuang H-Y, Lee E, Liu Y-T et al (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3:140. doi:[10.1038/msb4100180](https://doi.org/10.1038/msb4100180)
  31. Bass AJ, Thorsson V, Shmulevich I et al (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513:202–209. doi:[10.1038/nature13480](https://doi.org/10.1038/nature13480)
  32. Gomez-Cabrero D, Abugessaisa I, Maier D et al (2014) Data integration in the era of omics: current and future challenges. *BMC Syst Biol* 8:11. doi:[10.1186/1752-0509-8-S2-11](https://doi.org/10.1186/1752-0509-8-S2-11)
  33. Wheelock CE, Goss VM, Balgoma D et al (2013) Application of “omics” technologies to biomarker discovery in inflammatory lung diseases. *Eur Respir J* 42:802–825. doi:[10.1183/09031936.00078812](https://doi.org/10.1183/09031936.00078812)
  34. Bel EH, Sousa A, Fleming L et al (2011) Diagnosis and definition of severe refractory asthma: an international consensus statement from the Innovative Medicine Initiative (IMI). *Thorax* 66:910–917. doi:[10.1136/thx.2010.153643](https://doi.org/10.1136/thx.2010.153643)
  35. Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25:2906–2912. doi:[10.1093/bioinformatics/btp543](https://doi.org/10.1093/bioinformatics/btp543)
  36. Lê Cao K-A, Martin PGP, Robert-Granić C, Besse P (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* 10:34. doi:[10.1186/1471-2105-10-34](https://doi.org/10.1186/1471-2105-10-34)
  37. Dray S, Chessel D, Thioulouse J (2003) Co-inertia analysis and the linking of ecological data tables. *Ecology* 84:3078–3089
  38. Culhane AC, Perrière G, Higgins DG (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics* 4:59. doi:[10.1186/1471-2105-4-59](https://doi.org/10.1186/1471-2105-4-59)
  39. Meng C, Kuster B, Culhane AC, Gholami AM (2014) A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 15:162. doi:[10.1186/1471-2105-15-162](https://doi.org/10.1186/1471-2105-15-162)
  40. Shen R, Mo Q, Schultz N et al (2012) Integrative subtype discovery in glioblastoma using iCluster. *PLoS One* 7:e35236. doi:[10.1371/journal.pone.0035236](https://doi.org/10.1371/journal.pone.0035236)
  41. Curtis C, Shah SP, Chin S-F et al (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486:346–352. doi:[10.1038/nature10983](https://doi.org/10.1038/nature10983)
  42. Mo Q, Wang S, Seshan VE et al (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci U S A* 110:4245–4250. doi:[10.1073/pnas.1208949110](https://doi.org/10.1073/pnas.1208949110)

43. Kirk P, Griffin JE, Savage RS et al (2012) Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 28:3290–3297. doi:[10.1093/bioinformatics/bts595](https://doi.org/10.1093/bioinformatics/bts595)
44. Lock EF, Dunson DB (2013) Bayesian consensus clustering. *Bioinforma Oxf Engl* 29:2610–2616. doi:[10.1093/bioinformatics/btt425](https://doi.org/10.1093/bioinformatics/btt425)
45. Wang B, Mezlini AM, Demir F et al (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 11:333–337. doi:[10.1038/nmeth.2810](https://doi.org/10.1038/nmeth.2810)
46. Le Novère N, Hucka M, Mi H et al (2009) The systems biology graphical notation. *Nat Biotechnol* 27:735–741. doi:[10.1038/nbt.1558](https://doi.org/10.1038/nbt.1558)
47. Blinov ML, Faeder JR, Goldstein B, Hlavacek WS (2006) A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *Biosystems* 83:136–151. doi:[10.1016/j.biosystems.2005.06.014](https://doi.org/10.1016/j.biosystems.2005.06.014)
48. Kitano H, Funahashi A, Matsuoka Y, Oda K (2005) Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol* 23:961–966. doi:[10.1038/nbt1111](https://doi.org/10.1038/nbt1111)
49. Mizuno S, Iijima R, Ogishima S et al (2012) AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *BMC Syst Biol* 6:52. doi:[10.1186/1752-0509-6-52](https://doi.org/10.1186/1752-0509-6-52)
50. Fujita KA, Ostaszewski M, Matsuoka Y et al (2013) Integrating pathways of parkinson's disease in a molecular interaction map. *Mol Neurobiol*. doi:[10.1007/s12035-013-8489-4](https://doi.org/10.1007/s12035-013-8489-4)
51. Younesi E, Toldo L, Müller B et al (2012) Mining biomarker information in biomedical literature. *BMC Med Inform Decis Mak* 12:148. doi:[10.1186/1472-6947-12-148](https://doi.org/10.1186/1472-6947-12-148)
52. Slater T, Song D (2012) Saved by the BEL ringing in a common language for the life sciences. *Drug Discovery World*, Fall 2012
53. Malhotra A, Younesi E, Bagewadi S, Hofmann-Apitius M (2014) Linking hypothetical knowledge patterns to disease molecular signatures for biomarker discovery in Alzheimer's disease. *Genome Med* 6:97. doi:[10.1186/s13073-014-0097-z](https://doi.org/10.1186/s13073-014-0097-z)
54. Deisboeck TS, Wang Z, Macklin P, Cristini V (2011) Multiscale cancer modeling. *Annu Rev Biomed Eng.* doi:[10.1146/annurev-bioeng-071910-124729](https://doi.org/10.1146/annurev-bioeng-071910-124729)
55. Chakrabarti A, Verbridge S, Stroock AD et al (2012) Multiscale models of breast cancer progression. *Ann Biomed Eng.* doi:[10.1007/s10439-012-0655-8](https://doi.org/10.1007/s10439-012-0655-8)
56. Ribba B, Saut O, Colin T et al (2006) A multiscale mathematical model of avascular tumor growth to investigate the therapeutic benefit of anti-invasive agents. *J Theor Biol* 243:532–541. doi:[10.1016/j.jtbi.2006.07.013](https://doi.org/10.1016/j.jtbi.2006.07.013)
57. Dwivedi G, Fitz L, Hegen M et al (2014) A multiscale model of interleukin-6-mediated immune regulation in Crohn's disease and its application in drug discovery and development. *CPT Pharmacomet Syst Pharmacol* 3:1–9. doi:[10.1038/psp.2013.64](https://doi.org/10.1038/psp.2013.64)
58. Przybilla J, Rohlf T, Loeffler M, Galle J (2014) Understanding epigenetic changes in aging stem cells – a computational model approach. *Aging Cell* 13:320–328. doi:[10.1111/acel.12177](https://doi.org/10.1111/acel.12177)
59. Burrowes KS, Doel T, Brightling C (2014) Computational modeling of the obstructive lung diseases asthma and COPD. *J Transl Med* 12:S5. doi:[10.1186/1479-5876-12-S2-S5](https://doi.org/10.1186/1479-5876-12-S2-S5)
60. Mardis ER (2010) The \$1,000 genome, the \$100,000 analysis? *Genome Med* 2:84. doi:[10.1186/gm205](https://doi.org/10.1186/gm205)
61. Barrett T, Troup DB, Wilhite SE et al (2011) NCBI GEO: archive for functional genomics data sets – 10 years on. *Nucleic Acids Res* 39:D1005–D1010. doi:[10.1093/nar/gkq1184](https://doi.org/10.1093/nar/gkq1184)
62. Parkinson H, Kapushesky M, Shojatalab M et al (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35:D747–D750. doi:[10.1093/nar/gkl995](https://doi.org/10.1093/nar/gkl995)
63. Ballereau S, Glaab E, Kolodkin A et al (2013) Functional genomics, proteomics, metabolomics and bioinformatics for systems biology. In: Prokop A, Csukás B (eds) *Systems biology. Integrative biology and simulation tools*. Springer, New York, pp 3–41
64. Goecks J, Nekrutenko A, Taylor J, Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11:R86. doi:[10.1186/gb-2010-11-8-r86](https://doi.org/10.1186/gb-2010-11-8-r86)
65. Omberg L, Ellrott K, Yuan Y et al (2013) Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet* 45:1121–1126. doi:[10.1038/ng.2761](https://doi.org/10.1038/ng.2761)
66. Gao J, Aksoy BA, Dogrusoz U et al (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 6:pl1. doi:[10.1126/scisignal.2004088](https://doi.org/10.1126/scisignal.2004088)
67. Henry VJ, Bandrowski AE, Pepin A-S et al (2014) OMICtools: an informative directory for multi-omic data analysis. *Database* 2014:bau069. doi:[10.1093/database/bau069](https://doi.org/10.1093/database/bau069)
68. Rocca-Serra P, Brandizi M, Maguire E et al (2010) ISA software suite: supporting

- standards-compliant experimental annotation and enabling curation at the community level. *Bioinforma Oxf Engl* 26:2354–2356. doi:[10.1093/bioinformatics/btq415](https://doi.org/10.1093/bioinformatics/btq415)
69. Mathew JP, Taylor BS, Bader GD et al (2007) From bytes to bedside: data integration and computational biology for translational cancer research. *PLoS Comput Biol* 3:e12. doi:[10.1371/journal.pcbi.0030012](https://doi.org/10.1371/journal.pcbi.0030012)
70. Szalma S, Koka V, Khasanova T, Perakslis ED (2010) Effective knowledge management in translational medicine. *J Transl Med* 8:68. doi:[10.1186/1479-5876-8-68](https://doi.org/10.1186/1479-5876-8-68)
71. Athey BD, Braxenthaler M, Haas M, Guo Y (2013) tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research. *AMIA Summits Transl Sci Proc* 2013:6–8
72. Schumacher A, Rujan T, Hoefkens J (2014) A collaborative approach to develop a multi-omics data analytics platform for translational research. *Appl Transl Genomics* 3:105–108. doi:[10.1016/j.atg.2014.09.010](https://doi.org/10.1016/j.atg.2014.09.010)
73. Wolstencroft K, Owen S, du Preez F et al (2011) The SEEK: a platform for sharing data and models in systems biology. *Methods Enzymol* 500:629–655. doi:[10.1016/B978-0-12-385118-5.00029-3](https://doi.org/10.1016/B978-0-12-385118-5.00029-3)
74. Miralles F, Gomez-Cabrero D, Lluch-Ariet M et al (2014) Predictive medicine: outcomes, challenges and opportunities in the Synergy-COPD project. *J Transl Med* 12(Suppl 2):S12. doi:[10.1186/1479-5876-12-S2-S12](https://doi.org/10.1186/1479-5876-12-S2-S12)
75. Maier D, Kalus W, Wolff M et al (2011) Knowledge management for systems biology a general and visually driven framework applied to translational medicine. *BMC Syst Biol* 5:38. doi:[10.1186/1752-0509-5-38](https://doi.org/10.1186/1752-0509-5-38)
76. Cano I, Tényi Á, Schueller C et al (2014) The COPD knowledge base: enabling data analysis and computational simulation in translational COPD research. *J Transl Med* 12(Suppl 2):S6. doi:[10.1186/1479-5876-12-S2-S6](https://doi.org/10.1186/1479-5876-12-S2-S6)
77. Abugessaisa I, Saevarsdottir S, Tsipras G et al (2014) Accelerating translational research by clinically driven development of an informatics platform – a case study. *PLoS One* 9:e104382. doi:[10.1371/journal.pone.0104382](https://doi.org/10.1371/journal.pone.0104382)
78. Hernández-de-Diego R, Boix-Chova N, Gómez-Cabrero D et al (2014) STATegra EMS: an experiment management system for complex next-generation omics experiments. *BMC Syst Biol* 8(Suppl 2):S9. doi:[10.1186/1752-0509-8-S2-S9](https://doi.org/10.1186/1752-0509-8-S2-S9)
79. Dudley JT, Pouliot Y, Chen R et al (2010) Translational bioinformatics in the cloud: an affordable alternative. *Genome Med* 2:51. doi:[10.1186/gm172](https://doi.org/10.1186/gm172)

## Next-Generation Pathology

Peter D. Caie and David J. Harrison

### Abstract

The field of pathology is rapidly transforming from a semiquantitative and empirical science toward a big data discipline. Large data sets from across multiple omics fields may now be extracted from a patient's tissue sample. Tissue is, however, complex, heterogeneous, and prone to artifact. A reductionist view of tissue and disease progression, which does not take this complexity into account, may lead to single biomarkers failing in clinical trials. The integration of standardized multi-omics big data and the retention of valuable information on spatial heterogeneity are imperative to model complex disease mechanisms. Mathematical modeling through systems pathology approaches is the ideal medium to distill the significant information from these large, multi-parametric, and hierarchical data sets. Systems pathology may also predict the dynamical response of disease progression or response to therapy regimens from a static tissue sample. Next-generation pathology will incorporate big data with systems medicine in order to personalize clinical practice for both prognostic and predictive patient care.

**Key words** Histopathology, Integrative pathology, Systems pathology, Spatial heterogeneity, Predictive models, Cancer pathology, Multi-omics, Image analysis

---

### 1 Introduction

The manual, microscopic viewing of thinly cut and stained tissue sections by histopathologists has been the steadfast method of deciphering tissue architecture and concluding a prognosis for multiple diseases for over 100 years. The field of pathology recognizes the rich data source which lies within a tissue section. With the aid of specific histochemical stains, augmented by immunological and even mRNA- or DNA-based approaches, the pathologist takes into account the entire heterogeneous and heterotypic microenvironment and its interactions across the tissue section. Through experience, they are able to process this complex, sometimes subtle information and translate it in order to aid its diagnostic or prognostic conclusion. Research pathologists also apply this methodology to evaluate novel or significant prognostic features such as the tumor differentiation, tumor gland morphology at the invasive front, or immune infiltrate within the microenvironment.

The development of immunohistochemistry from the 1940s provided the pathologist with the ability to interrogate the tissue section with a further level of complexity where they could match biomarker expression with histopathological features and morphometry, although it took some time and the advent of monoclonal antibodies some 30 years or so later for a dramatic increase in routine use of the technology. The use of protein biomarkers, visualized through immunohistochemistry, allowed quantification at both spatial heterogeneity and subcellular resolution. Since the post-omics era, the field of modern pathology is experiencing an explosion of data across multiple but disparate omics strands. Most notably within the clinic is the genomic profiling of a patient's tissue sample through next-generation sequencing (NGS) where, in colorectal cancer, for example, EGFR and KRAS mutations now may be routinely tested for in order to predict the response to anti-EGFR antibody treatment. Single "magic bullet" biomarkers, however, have a limited use in clinical prognosis, drug prediction, and efficacy studies as they attempt to describe or modulate complex multi-pathway molecular and cellular interactions in an often too simplistic way.

Advances in the integration of genomics, proteomics, transcriptomics, epigenomics, and the emerging field of image analysis-based phenomics are now able to add valuable information to the hierarchical understanding of complex disease mechanisms. These molecular signatures correlated with morphological and clinical data have the ability to advance traditional diagnostic medicine from broad population-based prediction to a more personalized and precision-based science. Pathology has overcome the bottleneck of creating large, hierarchical, and complex "big data"; however, the challenge the field is now facing is how to handle this data in a meaningful manner which directly leads to translational impact. The overarching goal of modern big data pathology is to infer a dynamical prediction of disease from a static patient tissue sample. Systems pathology through mathematical modeling allows the integration, interrogation, and identification of significant parameters from large multi-omics data sets while having the ability to add a dynamic aspect to personalized medicine.

---

## 2 Tissue Is Heterogeneous

Tissue is extremely heterogeneous and cancer especially so; cancer heterogeneity can originate from multiple sources: cell of origin, clonal evolution, cancer stem cells (CSCs), response to microenvironment, and host factors as well as stromal or immune cell infiltrate. The clonal evolution theory states that the cancers build up heterogeneous subpopulations after concurrent mutations over multiple rounds of cell division due to the plasticity of

the cells through chromosomal and replicative instability or exogenous insults. These heterogeneous subpopulations are under the influence of natural selection where they may acquire mutations which ultimately lead to cell death, while others accumulate a specific set of driver mutations allowing the cancer cells to metastasize. CSCs may originate from healthy tissue stem cells or may have attained their stem-like phenotype through epigenetic alterations of the genome or through stromal cell interaction from their microenvironmental niche. The stem-like attributes associated with CSCs would confer a certain amount of plasticity upon it in order for it to evade aggressive treatment regimens or commit to the metastatic cascade. CSCs may have the ability to produce hierarchical heterogeneous cell subpopulation progenies of which only some are tumorigenic and others differentiated. CSCs are thought to initiate tumorigenesis and have the ability to propagate the cancer after chemotherapy, and a cure for the patient depends on the eradication of such self-renewing cells. CSCs also appear to be more resistant to radiation and chemotherapeutic treatment and may incur tumor recurrence even after a long period of remission and dormancy. More recently, the “Big Bang model” of intra-tumor heterogeneity has been described where tumors mainly grow as a single expansion and that intra-tumor heterogeneity within tumor subpopulations is high but occurs early on in the tumor’s evolution. In this model, aggressive subclones may not be predominant and can remain undetected although they would provide overall resistance to subsequent insult by treatment regimens [1].

The focus of cancer research for prognosis, prediction, and drug discovery has been on the tumor itself; however, this target is changing. It is becoming apparent that the tumor microenvironment as a whole, and more precisely the stromal and immune infiltrate, is increasingly important in tumor progression and evasion of chemotherapy. The host interaction on the tumor, its stem cell subpopulations, and its microenvironmental niche add a further level of heterogeneity to the tumor. Spatial heterogeneity within the stromal compartment of the tumor is a critical influence on the tumor, its subsequent progression, and potential resistance to therapy. The combination of the above creates a further level of complexity in the accurate understanding of disease and for its dynamic modeling.

---

### 3 Tissue Samples Are Imperfect

A wealth of prognostic and predictive information lies within the patient’s tissue sample. Classical histopathology strives to infer dynamical prediction of disease progression from the static artifact which is the tissue section. The pathologist directly observes



microscopically the complex diseased tissue and its interaction with the host microenvironment in order to mentally compute these multiple signals into a prognosis. This has long been the gold standard in clinical prognosis. Although multiple novel prognostic methodologies for colorectal cancer (CRC) have been developed to replace or augment classical pathology and while some show promise, for example, the gene expression signatures ColoPrint [2] and Oncotype DX [3], none has established itself within routine clinical prognosis. The classical Dukes and TNM morphological and histological staging of the disease remains steadfast in clinical pathology. One reason for this is standardization and the imperfection of tissue. The human eye can account for the variation and artifacts that occur from surgical removal of the tissue through to mounting sections onto microscope slides for analysis. Poor and small sample size, imperfection, and damage to tissue as well as poor tissue orientation can be easily disregarded by the pathologist, while they can glean the pertinent information from the final stained tissue section. Automated quantification of the tissue section, spanning the omics fields, is not able to be so selective and may therefore return variable results. The need for standardization across all aspects of automated tissue datafication is therefore essential.

Advances in extracting data in a meaningful and robust manner will add value to classical histopathology methodologies and provide greater impact and accuracy of patient stratification at a more personalized level than current population statistics, such as TNM staging. This is increasingly relevant when the quantification techniques take into account the heterogeneity of the disease and report on it. Datafication of tissue is the extraction of information in a fully quantifiable and standardized manner. This can take the form of quantifying a single biomarker to capturing a complex and hierarchical multimodal omics signature. Routinely, single readouts are extracted from a single tissue sample; however, advances in data-capturing technologies now allow multiple readouts captured across multiple omics fields which may be reported across distinct subpopulations identified through morphometric or biomarker expression. Big data pathology is now a reality, but creating standardized data sets amenable to complex modeling and which take into account the imperfection of tissue and its inherent heterogeneity is still in its infancy.

---

## 4 Quantifying Heterogeneity

Understanding tumor heterogeneity is important in striving toward an intelligent and individualized treatment strategy which translates into clinical impact. To truly fulfill a personalized medicine approach and select the correct combination therapy for a



patient, it is essential to know which mutational or epigenetic aberrations their cancer carries in both primary and distant disease and what the subsequent phenotypic and functional effects on the cells and their microenvironment are.

Multiple interactions at multiple levels occur in tissue architecture. Histopathology describes the end result but not the underlying molecular mechanisms. Since the post-“omics” era, scientists have been armed with a suite of new tools to identify biomarkers to subgroup a patient’s cancer at the molecular level. Using these tools, a raft of data and new biomarkers have been discovered over the last few decades and allowed genome-scale analysis and comparisons. The main disciplines to bear the wealth of the results are genomics, transcriptomics, proteomics, and epigenetics. Technologies such as NGS and array CGH allow the mutation and copy number status of the genome to be analyzed. RNA microarray chips and RNA sequencing technologies are employed to profile gene expression, whereas reverse phase protein array (RPPA) and mass spectrometry have brought proteomics into the field of big data pathology.

Inter-patient and intra-patient heterogeneity exists (Fig. 1), and the aim of all “omics” research is to identify biomarkers which can lead to targeted drug discovery programs or companion diagnostics which will allow the clinician and pathologist to make rapid informed decisions on the prognosis of the disease and to predict which treatment will display the greatest efficacy and best outcome as possible for the individual patient.

Although the above methodologies to quantify the molecular mode of action driving cancer subtypes have added significant value, they also hold disadvantages to assaying such complex material. To extract DNA, RNA, and protein molecules, these assays usually homogenize and destroy the tissue integrity. The tissue is literally “mashed and measured” mixing together any subpopulations of cancer and host cells expressing differential properties while losing spatial resolution. This results in one end point being reported for the whole tumor. Due to the nature of these applications, intra-tumoral heterogeneity of the tissue may be under-detected where the dominant or most abundant genotypes or phenotypes mask signal from smaller cell populations within the tumor. Healthy tissue and host cells from the tumor microenvironment are both also added to the molecular sample creating a further source of noise to the signal and could increase the reporting of false positive or negative results. Under-detection of tissue heterogeneity therefore leads to an urgent and difficult problem when treating a patient with combination therapy, as resistant subgroups could go unnoticed and untreated. There are, however, tools to overcome this problem which attempt to better quantify, and thus comprehend, the complexity of heterogeneous tumors. One such tool is the laser capture microdissection (LCM) which



Background signals from complex tissue can still create noise in these assays, and robust and sensitive data depends on the LCM technique as well as the specificity of probes, antibodies, and detection technology used. To avoid contamination of signals from heterogeneous subpopulations within tissue, *in situ* imaging of protein through immunohistochemistry (IHC) and genomics through fluorescence *in situ* hybridization (FISH) may be applied. This has advantages over destructive assays as the tissue structure, spatial orientation, and sub-localization of molecules are retained and heterogeneity can be visualized, compartmentalized, and quantified while providing insight into cellular interactions within the tumor and its microenvironment. IHC further allows the visualization of morphological status of the cells expressing the biomarker of interest and allows the observer to correlate morphometric and proteomic signatures at the cellular resolution. Spatial heterogeneity impacts the prognostic and predictive significance of biomarkers, and it is becoming increasingly apparent that this must be taken into consideration for the modeling of disease. The immunoscore in colorectal cancer, which quantifies the density and intratumoral location of CD3+ and CD8+ lymphocytes through image analysis, has been shown to hold a higher prognostic significance than the gold standard of TNM staging [4]. Similarly, the spatial heterogeneity of unbiased and automatically quantified lymphocytes in breast cancer tissue sections was statistically modeled and found to be associated to patient survival [5]. In the field of transcriptomics, it has recently been discovered that mesenchymal cell gene expression classifiers are linked to poor prognosis in colorectal cancer though it proves difficult to ascertain whether these classifiers are expressed by the tumor or the stromal cells; however, immunohistochemistry for mesenchymal proteins in tissue sections and laser capture microdissection have elucidated that the mesenchymal signatures originate from stromal cancer-associated fibroblasts and not from the tumor itself [6, 7].

Although the field of high-content analysis is not new, where multiple parameters and biomarkers are measured from fluorescently labeled cells [8], the discipline has been slow to translate to histopathology and the clinic. This has been in part due to the complexity of tissue and its imperfection compared to *in vitro* cell studies and the need for extensive validation and standardization for clinical use. This is now changing, and digital pathology as well as automated image analysis for tissue-based studies is rapidly emerging into the realm of clinical research. The integration of digital pathology with automated image analysis brings advantages to the field. These include the standardization of quantification where observer variability is excluded and the robust analysis of rare or complex features is captured. Traditionally, image analysis in histopathology concentrated on the quantification of protein expression through immunohistochemistry and

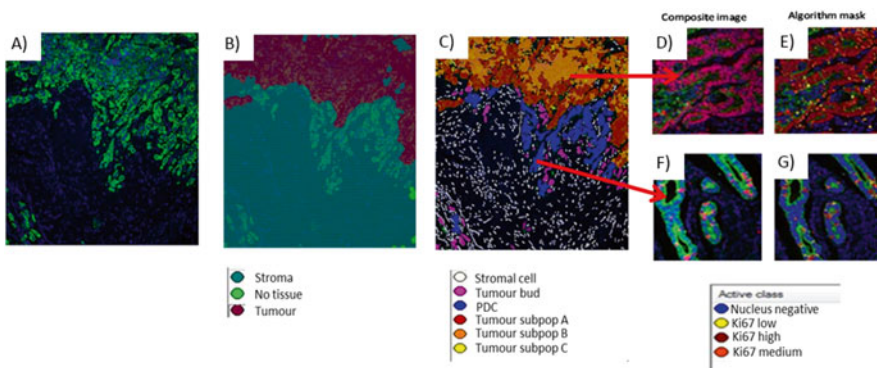
immunofluorescence (IF). This was to overcome the subjective manual and semiquantitative scoring of a 1+, 2+, 3+ system. Upon employing IF, image analysis software can perform fully quantified continuous data exports from which cutoffs can be calculated in order to stratify patient subgroups. Computer-based quantification of nuclear morphometry, however, has been practiced for over a decade. Continuous improvements to image analysis software now allow the simultaneous export of morphometric parameters of cells and histopathological features alongside biomarker quantification associated to this feature. In this co-registering methodology, it is possible to identify surrogate morphological features which correlate with molecular phenotype.

The market-leading tissue imaging platform manufacturers provide their own image analysis solutions for chromogenic and fluorescence assays which allow segmentation of cells and subcellular compartments and subsequent biomarker quantification within heterogeneous tissue. These software packages are designed to work in connection with the images captured from their own platforms and can sometimes be restrictive to the quantification of set assays, biomarkers, and parameters. Definiens (<http://www.definiens.com/>), Indica Labs (<http://indicalab.com/>), and Visiopharm (<http://www.visiopharm.com/>) offer image analysis packages which can import images from most microscopes and allow a more flexible image analysis environment to capture the complexity of the heterogeneous tissue section.

While the imaging of a single biomarker can yield predictive or prognostic information, the ability to multiplex two or more markers on a single tissue section becomes a much more powerful tool. An advantage of IF-based image analysis is the ability to multiplex, co-register, and quantify biomarkers at the cellular resolution. Multiplexing reports on protein interactions, pathway activation, and multiple cellular events. Accurate co-localization and spatial resolution of multiple biomarkers or histological features on the same section of tissue report a richer high-content and functional data than serial sections of one biomarker while saving the precious resource which is the tissue sample. Researchers can quantify multiple proteins on a per cell basis or accurately quantify multiple cell types within a heterogeneous population. Traditional multiplexing is limited by bleed through of fluorophores and chromogens as well as antibody cross-reactivity of secondary host species. Multispectral imaging and unmixing of chromogens and fluorophores allow an accurate spectral readout for each biomarker of interest, increase the multiplexing capacity, and negate any autofluorescence. Sophisticated image analysis software and multiplexed in situ labeling permit the big data capture from image analysis-based segmented tissue sections to quantify the data-rich histopathology and the interactions and spatial heterogeneity of the cancer microenvironment's phenotypic features. This involves

the extraction of complex and hierarchical data pertaining to a single segmented feature or set of features across the segmented tissue section. This data may be captured through co-registering of biomarkers as proteomic or genomic signals, multiple morphometric and texture parameters, or a combination of both, essentially extracting as much data as possible from each single segmented object within the image. A multi-parametric signature is therefore built up for each tissue sample which may be compiled of multi-omic image-based features. Tissue subpopulations may be identified in this manner and further mined through in situ labeling or microdissection to interrogate the patient's sample at the personalized level for predictive or prognostic pathology (Fig. 2).

Sophisticated data mining is required to identify the significant single or combination of parameters within the signature in order to stratify patients for prognostic or predictive purposes. Data mining techniques previously applied to identify significant parameters have been logistic regression analysis and ensemble decision tree models. Further advancements in in situ labeling and image analysis such as mass spectrometry imaging, next-generation immunohistochemistry [9], and multi-parametric data capture, where biomarkers are correlated to morphometry, are catapulting this field into the realm of true big data alongside the more traditional omics fields. Image analysis and in situ labeling of tissue sections coupled to spatial statistics will most probably factor highly when profiling a disease complex heterogeneous microenvironment in the future of systems pathology.



**Fig. 2** Subpopulation segmentation and biomarker quantification through image analysis. Tumor subpopulation segmentation and classification through whole slide image analysis of immunofluorescence-labeled colorectal cancer tissue utilizing Definiens image analysis software. (a) Raw image: DAPI (*blue*) and panCK (*green*). (b) Image analysis algorithm automatically segments tumor from stroma. (c) Tissue is further segmented into stromal cells, tumor buds, poorly differentiated clusters (PDC), and three tumor gland subpopulations. Ki67 (*red*) proliferation marker is quantified within separate subpopulations at the invasive front (f, g) and the tumor core (d, e)

---

## 5 Integrative Pathology

Traditional omics research attempts to identify single molecular or histopathological features which could be utilized for prognosis or prediction of response to drug therapy. Cancer is, however, a very complex disease with multiple molecular interactions within the cell and multiple cellular interactions within the microenvironment. Many single biomarkers never translate to the clinic, as they do not take into account the complexity and heterogeneity of the disease. Integrating large-scale data from multiple omics fields may help to address this problem as it will create a better understanding of the multiple molecular interactions occurring within the cell and how these translate to disease progression. This approach was exemplified in colorectal cancer where histopathological subtypes were integrated with methylation and mutation status to assess their correlation and impact on prognosis [10]. Integrative large-scale pathology has also been implemented in breast cancer where cellular resolution of in situ and co-registered genotype and phenotype was utilized to study intra-tumoral heterogeneity between primary and distant metastases for studies of prognosis and potential drug targets [11]. Finally, a further breast cancer study integrated a multi-omics signature and discovered JAK-STAT and TNF signaling pathways to be significant in triple-negative disease which could lead to novel and personalized drug treatments [12]. There is a wealth of data collected during classical histopathology which largely remains unused in clinical decision making. This clinical data is beginning to be integrated with the modern datafication modalities as a further hierarchical level of understanding of the disease from the tissue. In mucoepidermoid carcinoma, histopathologic, immunophenotypic, and cytogenetic parameters were integrated to identify a signature which was able to identify the pulmonary disease from other subtypes of lung cancer [13]. Clinical and molecular data is now also being integrated with the complex and data-rich image-based phenotypic signatures to investigate cancer heterogeneity and its interaction with the microenvironment. The morphometric signatures can also be correlated to the genomic profile and clinical outcome [5]. Computational IT solutions are also now available which allow the incorporation of multiscale omics data [14, 15] as well as integrate it with clinical information [16].

---

## 6 Systems Pathology

Pathology is now adept at creating large and complex data sources from across the omics fields and more recently including histopathology, morphometrics, and spatial heterogeneity. This data, however, must be integrated in a meaningful way which makes best use



of its complexity and is standardized, reproducible, and robust enough to be clinically relevant. The challenge ahead is how to incorporate this integrated data into models which can identify the optimal combinations of parameters to answer clinical questions in a robust and standardized manner. Systems medicine, and more recently systems pathology, takes a holistic view of tissue, the cell, and its multitude of interactions. Systems pathology requires a large amount of high-quality multiscale data to be extracted from tissue and which acts as input for predictive mathematical models. Although systems pathology has predominantly concentrated on molecular profiling of the genome, transcriptome, or proteome, image analysis-based multi-parametric biomarker and morphometry is perfectly matched to add to the hierarchical data within a systems model. This additional in situ information allows the retention of the valuable spatial heterogeneity within the disease microenvironment.

Essentially, a modern integrative pathology would adopt the principles of 4P medicine in a systems pathology approach. 4P medicine consists of prediction, personalization, prevention, and patient participation [17]. There are many definitions of systems medicine. Within Europe systems medicine is defined by the EU consortium CASyM ([www.casym.eu](http://www.casym.eu)), as stated within the first chapter of this book. In 2014 CASyM hosted a select meeting in the University of St Andrews where European pioneers in the field of Systems Pathology were invited. The consensus at the end of the meeting was that although there are multiple methods to create big datasets in pathology, for work to be classed as ‘Systems Pathology’ the methodology must involve dynamic mathematical modelling.

The principle of systems pathology is to predict a dynamic pathological response from static data sets. The more standardized and robust data which is used for input into the model directly relates to the quality of prediction within the model. Systems pathology is complex with the implementation of multiple differential equations into a multiscale dynamic model to predict a drug effect on a patient or inform how that patient will respond over time. Systems pathology, under this definition, was utilized to confirm the role of PTEN in trastuzumab drug resistance [18]. Systems pathology can also be implemented to track tumor evolution post-chemotherapy through intra-tumor heterogeneity and spatial distribution of phenotype and genotype at the cellular level [19]. In CRC, a systems pathology approach was employed to identify a disease recurrence signature in early-stage patients from a multi-omics data set where parameters associated with immune response were found to be the most significant predictors [20].

Systems pathology is therefore already making a valuable impact into the field of translatable clinical research. Systems pathology is the ideal tool to distill significant parameters with significant population cutoffs and which are therefore translatable to the clinic, from multiple integrated complex big data sets.

This is what we have termed “next-generation pathology.” The ultimate goal of next-generation pathology is to make use of this hierarchical data captured across multiple modalities from an imperfect and static tissue sample, in order to better understand both disease progression and a patient’s personalized response to treatment.

## References

1. Sottoriva A, Kang H, Ma Z, Graham TA (2015) A Big Bang model of human colorectal tumor growth. *Nat Genet* 47(3):209–216
2. Kopetz S, Taberero J, Rosenberg R, Jiang ZQ, Moreno V, Bachleitner-Hofmann T et al (2015) Genomic classifier ColoPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. *Oncologist* 20(2):127–133
3. Srivastava G, Renfro LA, Behrens RJ, Lopatin M, Chao C, Soori GS et al (2014) Prospective multicenter study of the impact of oncoPrint DX colon cancer assay results on treatment recommendations in stage II colon cancer patients. *Oncologist* 19(5):492–497
4. Galon J, Mlecnik B, Bindea G, Angell HK, Berger A, Lagorce C et al (2013) Towards the introduction of the “Immunoscore” in the classification of malignant tumors. *J Pathol* 232(2):199–209
5. Yuan Y (2015) Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. *J R Soc Interface*. doi:10.1098/rsif.2014.1153
6. Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A et al (2015) Stromal contribution to the colorectal cancer transcriptome. *Nat Genet* 47(4):312–319
7. Calon A, Lonardo E, Berenguer-Llgero A, Espinet E, Hernando-Mombona X, Iglesias M et al (2015) Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet* 47(4):320–329
8. Caie PD, Walls RE, Ingleston-Orme A, Daya S, Houslay T, Eagle R et al (2010) High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Mol Cancer Ther* 9(6):1913–1926
9. Rimm DL (2014) Next-gen immunohistochemistry. *Nat Methods* 11(4):381–383
10. Inamura K, Yamauchi M, Nishihara R, Kim SA, Mima K, Sukawa Y et al (2015) Prognostic significance and molecular features of signet-ring cell and mucinous components in colorectal carcinoma. *Ann Surg Oncol* 22(4):1226–1235
11. Almendro V, Kim HJ, Cheng YK, Gonen M, Itzkovitz S, Argani P et al (2014) Genetic and phenotypic diversity in breast tumor metastases. *Cancer Res* 74(5):1338–1348
12. Karagoz K, Sinha R, Arga KY (2015) Triple negative breast cancer: a multi-omics network discovery strategy for candidate targets and driving pathways. *Omics* 19(2):115–130
13. Roden AC, Garcia JJ, Wehrs RN, Colby TV, Khoor A, Leslie KO et al (2014) Histopathologic, immunophenotypic and cytogenetic features of pulmonary mucoepidermoid carcinoma. *Mod Pathol* 27(11):1479–1488
14. Le Cao KA, Gonzalez I, Dejean S (2009) integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* 25(21):2855–2856
15. Day RS, McDade KK, Chandran UR, Lisovich A, Conrads TP, Hood BL et al (2011) Identifier mapping performance for integrating transcriptomics and proteomics experimental results. *BMC Bioinformatics* 12:213
16. Miyoshi NS, Pinheiro DG, Silva WA Jr, Felipe JC (2013) Computational framework to support integration of biomolecular and clinical data within a translational approach. *BMC Bioinformatics* 14:180
17. Hood L, Friend SH (2011) Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol* 8(3):184–187
18. Faratian D, Goltsov A, Lebedeva G, Sorokin A, Moodie S, Mullen P et al (2009) Systems biology reveals new strategies for personalizing cancer medicine and confirms the role of PTEN in resistance to trastuzumab. *Cancer Res* 69(16):6713–6720
19. Almendro V, Cheng YK, Randles A, Itzkovitz S, Marusyk A, Ametller E et al (2014) Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Rep* 6(3):514–527
20. Madhavan S, Gusev Y, Natarajan TG, Song L, Bhuvaneshwar K, Gauba R et al (2013) Genome-wide multi-omics profiling of colorectal cancer identifies immune determinants strongly associated with relapse. *Front Genet* 4:236



## Training in Systems Approaches for the Next Generation of Life Scientists and Medical Doctors

Damjana Rozman, Jure Acimovic, and Bernd Schmeck

### Abstract

The last decades have challenged us with novel technologies and abilities to look deep into the human genome of each individual, to screen for numerous metabolites in the blood, to search organs by imaging techniques, and much more. We can collect data from all these measurements from humans with different (usually multifactorial) diseases and, in accordance with ethical rules, store the data safely, locally, or remotely. The human data is growing exponentially, from giga- ( $10^9$ ), tera- ( $10^{12}$ ) to peta- ( $10^{15}$ ) and zettabytes ( $10^{21}$ ), both in public and restricted access settings. However, everyone agrees that the technological and information booms have not yet sufficiently reached medicine and have so far only barely influenced the clinical settings. In this chapter, we discuss the opinion that without topping up the education system, it will be difficult to catch up. We propose that in addition to the classical medical education, which is traditionally good and highly respected in European countries, we must find ways on how to introduce into the medical curricula mathematical and big data aspects and insights. Only a global (systems) view on physiology and pathophysiology can break the Gordian knot of many multifactorial diseases where we still don't understand the complexity of disease causes nor we can predict or cure the disease. We believe that the breakthrough is in the systems and interdisciplinary education and training, as early as possible in professional careers. If medical and related students and professionals would be formally educated in such interdisciplinary manner, they could take this knowledge further towards applications in their daily medical practice. We describe the current challenges and scattered best practices of introducing the wider systems medicine topics into the medical education as well as possibilities for systems medicine training at the doctoral and lifelong levels.

**Key words** Multifactorial disease, Interdisciplinary, Medicine, Education, Systems biology

---

### 1 Introduction

In the twenty-first century, the view on diseases is rapidly changing. Novel high-throughput technologies and advanced computation with possibilities to sequence and understand the context of each individual human genome have been introduced only a few years ago. Their impact on medicine is so quick and so influential that it is difficult to speak about *evolution* of biomedical sciences,

but rather about *revolution*. Despite the technological boom and technical abilities to sequence, clone, and map virtually everything from human individuals, we face a number of multifactorial and complex diseases which we do not understand completely nor we can predict or treat them. In multifactorial conditions, the combinations of genetic and environmental factors interplay with each other leading to distinct disease phenotypes. Due to the individuality of humans and combinatorial effects, it is virtually impossible to predict all combinations that can lead to a multifactorial disease phenotype. This limits also the ability to predict the individual's disease progression and to discover/apply efficient treatments.

Among multifactorial diseases are also major noncommunicable diseases, such as cardiovascular disease, cancer, chronic respiratory disease, diabetes, rheumatologic disorders, mental disorders, etc., which represent currently the predominant health burden. The World Health Organization mentioned this in the 2008 Action Plan [1] and in the United Nations 2010 Resolution, followed by the 2014 Council conclusions on nutrition and physical activity [2].

The novel trends for managing noncommunicable diseases should involve integrative, holistic approaches, and such diseases can be considered as a single expression (phenotype) of a disease with different risk factors and pathology dimensions, including the molecular ones.

Currently, an innovative integrated health system is being built around systems that can combat such multifactorial diseases. Training in systems approaches plays a special role since it targets the current and future generations of medical doctors, other biomedical scientists, and health-care professionals. It is important to note that international perspectives of such efforts are very important since they can apply to different countries and communities [3].

A major challenge of today's medicine is thus to incorporate the technological revolution accompanied with expansion of various data into the everyday clinical practice. One to two generations might be needed for a major change of mind to enter the clinics. The duty of the society is to introduce into education the novelties (novel technological, big data, and computation principles) and link it with classical medicine towards a win-win situation. While the medical community is becoming increasingly aware of the new educational needs, the paths of how (and when) to introduce the novel subjects are not so obvious. One view that will be presented also in this chapter is to apply systems biology approaches and tools to biomedical problems and to start educating biomedical students in an interdisciplinary manner as early as possible. In addition, these educational efforts have to take into account ethical concerns as well as economic circumstances and specific aspects of the different health-care systems. Therefore, stakeholders of (bio) medical education and health-care delivery have to be included in these processes.

---

## 2 From Systems Biology to Systems Medicine: The Ceiling of Data

Systems biology is an interdisciplinary field bridging biological sciences, applied mathematics, and computational and engineering sciences and is focused at systems' understanding of biological processes. Leroy Hood and colleagues described this in 2001 as follows: "Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations" [4]. If we focus on the systems' understanding in humans, we come from the broader *systems biology* to a more specialized *systems medicine*. It is clear that we cannot apply directly all the systems biology principles to humans. For example, it is for humans not possible to "perturb them (biologically, genetically, or chemically)" to collect the desirable data for modeling. The major difference between *systems biology* and *systems medicine* seems to be in the scale of the available data where for systems biology the data can be obtained from model organisms and cell cultures. There are limitations also for model organisms, but we are still very flexible in measurements *in vivo*; we can design time series experiments and work with multiple organisms respecting ethical standards, such as the Amsterdam protocol on animal protection and welfare and other European Commission policies. The situation with collecting human biological data is very different. Due to ethical reasons, only selected measurements/trials are approved for humans. It is, for example, impossible to count on kinetic data from human organs *in vivo*. Even *ex vivo* studies relying on data from human organs, such as the human liver, are frequently small and difficult to compare with one another [5]. A way forward would be to search further for common roots in pathogenesis pathways, such as the lately identified overlaps between pathways related to oncological, metabolic, and inflammatory diseases [6].

Despite this, it appears that in individuals a different combination of genetic and environmental factors might define the pathology progress which accumulates with age. We are faced with a challenging situation where on one hand there is a large progress in understanding the molecular players of disease stages and overlapping with other diseases, while the inconsistencies from different studies and different populations leave the impression that we are indeed at the start. For example, in liver diseases the genome-wide association studies, transcriptome analyses, meta-analyses, and other clinical studies in different populations and ethnic backgrounds were until 2014 concordant in polymorphisms of a single gene *PNPLA3* whose function was not clear at that time [7].

Except for cases with noninvasive data collection such as in circadian studies including the human chronotherapy efforts [8, 9], it is unlikely to have access to the human time series measurements. Consequently, the computation approaches applied in systems biology are not necessarily identical to approaches required to deal with the limiting human data: systems medicine requires novel or adapted algorithms to fit the human data. The existing computational tools could be developed further and applied for evaluating the disease risk, managing the disease, individualizing the diagnosis, prognosis, etc. We should also not forget the electrophysiological roots of systems medicine where already in 1960 Denis Noble developed the computer model of a heart pacemaker [10].

---

### 3 How to Teach Systems Medicine?

Being able to apply, analyze, integrate, model, and understand enormous quantities and variety of data would require a new type of physician—one with a grasp of modern computational sciences and omic technologies (genomics, proteomics, metabolomics, transcriptomics, etc.). These all represent parts of a systems approach to medicine where the new tools arise from the intersection of research across a variety of disciplines and are difficult to capture in traditional education curricula, especially not in traditional medicine curricula [11]. If it is aimed that clinicians would take advantage of systems approaches, we have to educate them properly. For the complex, multidisciplinary nature of systems medicine, the ideal training setting might be if both teachers and students would form multidisciplinary teams, especially since there is no established routine of systems medicine education and training.

Debates about what the best approach to teaching systems medicine should be are vivid and continuing. Some parts of the training, such as modeling, may require intensive efforts from the tutors. A solution could be a pre-training session prepared with the aid of web-based materials available to the trainees in advance. Trainees would work on problems beforehand and, at the on-site part of the training, get guidance and feedback. This is nowadays discussed as the trendy “flip training.” The struggle of tutors here is how to get students to do the preparatory work in advance. Based on my personal experience being a professor at Faculty of Medicine, University of Ljubljana, and experiences of many of my colleagues, much depends on the “maturity” of students irrespective of the efforts of professors. However, a majority of medical (and likely other) students will do the preparatory work in advance if they are directly motivated by grading their work, meaning that they can be scored for the work they have done off-site.

Until 2015, there are only a few best practices of training in systems medicine, so there are still multiple opportunities for paving the path. The European Consortium gathered within Coordinating Action Systems Medicine (CASyM) (<https://www.casym.eu/>) decided to establish a European plan for education in systems medicine. This should base on existing programs and experiences gained from relevant training concepts as described later in this chapter and should result in sustainable education programs for systems medicine training of medical doctors and related scientists. Since it is not agreed what should be the best age/level to start with systems approaches in biomedical education, the European CASyM consortium proposed several possibilities. One is to design tailored interdisciplinary programs and implement training at the master's, doctoral, and postdoctoral levels, as well as specifically for clinicians at different stages of their careers. To achieve this, one needs to specify course modules relevant for systems medicine which could be incorporated into curricula for medical and other biomedical fields. Whether this could result in a widely accepted independent "systems medicine" curriculum is still not clear. It is thus important that the systems medicine courses are accredited by the European Credit Transfer and Accumulation System (ECTS) [12] in the path towards the formalization of this interdisciplinary education branch.

To reach the population of already established scientists and medical doctors, lifelong education possibilities can be offered, such as systems medicine meetings, expert-guided workshops and summer schools, targeted lecture series, etc. Medical doctors can be attracted if the courses (online and face-to-face) are accredited to offer the continuous professional development (CPD) credits which for medical doctors mean the continuing medical education (CME) credits that are in several European countries required for maintaining the practitioner license. Table 1 illustrates the requirements for CME credits in CASyM countries and selected associates [13].

---

## 4 The Current Practices in Systems Medicine Education and Challenges to Meet

Systems medicine has the potential to make medical care and practice more patient centered, more (cost-)effective, and more holistic (a more efficient integration of a variety of components), also achieving a better control of potential side effects. However, to develop and implement systems strategies in medicine, we do not only need medical literacy in basic scientists performing systems biology, but also trained and experienced clinical practitioners. What challenges do we have to meet on this way?

By definition, systems medicine involves the implementation of systems biology approaches in medical concepts, research, and

**Table 1**  
**The requirements for CME credits in European countries [13]**

Country	CME requirement	Credit/year	CME scheme delivered by	CME activities accredited by	Sanctions
Austria	Voluntary	50	Regulator	Regulator	No sanctions
Belgium	Voluntary	20	Accredited providers	Regulator	No sanctions
France	None	–	–	–	–
Germany	Compulsory	50	Accredited providers	Regulator body (regional)	License loss/fees reduced
Greece	Compulsory	20	Accredited providers	Medical association	No sanctions
Hungary	Compulsory	50	Accredited providers	Medical societies	Retake examinations
Italy	Compulsory	50	Accredited providers	Regulator body (regional)	–
Netherlands	Compulsory	40	Professional providers + providers	Professional societies	Annual registration
Norway	Compulsory	40	Universities + societies	Medical association	Loss of status + fees
Poland	Compulsory	40	Regulator (regional)	Regulator body (regional)	–
Slovakia		50	Accredited providers + council	Regulator	–
Slovenia	Compulsory	15	Accredited providers	Regulator	–
Spain	Voluntary	–	–	–	–
Sweden	Voluntary	50 (10 days)	Accredited providers	Professional association	–
USA	Compulsory	12–50		Medical association/committee	Varies—fine, reprimand

Different European countries apply very different practices, some requiring CME for prolonging licenses and some not. Also, the number of yearly credits differs, and sanctions of not achieving this can be as severe as loss of license in, i.e., Germany

practice. This could be achieved by iterative and reciprocal feedback between data-driven computational and mathematical models and model-driven translational and clinical investigations [14]. Final outcomes could be examples of personalized medicine or the so-called P4 medicine (predictive, preventive, personalized, and participatory) [15].

Therefore, key components are, e.g., the development of multidisciplinary training and professional dissemination of concepts, creating and shaping a sustainable European community of systems

medicine [16]. At the CASyM ICSB2013 training workshop and the Ljubljana CASyM course 2013, important issues have been discussed and goals have been defined [17]: (1) Systems medicine should span all aspects of medical education as a framework for integration of all (pre)clinical disciplines. (2) Systems medicine-facilitated courses of “traditional” topics should aim at understanding complex topics with the help of dynamic systems approaches and (visualization-based) gadgets. (3) Research physicians and clinical practitioners should be educated more thoroughly in statistics, bioinformatics, and -omics technologies and should be open minded for the use of systems biology modeling for medical purposes. (4) Software should be adapted for practical usage by clinicians.

Tasks we have to meet on the way to physicians with literacy in and probably affinity to systems medicine start during medical school. One possible choice is that medical education could be based on a 4-year graduation (bachelor) focused on basic and natural sciences (in different proportions according to personal predispositions) followed by a 4-year medical degree (MD) in clinical medicine. This US type of curriculum could leave more space to hard sciences, seen as a good introduction to systems medicine, for a minority of specifically interested MDs. On the other hand, new concepts, e.g., in Germany and Slovenia, aim at a 6-year master’s program starting with clinical examples on the first day, focusing on “problem-oriented learning,” i.e., signs and symptoms, rather than pathophysiological systems (cancer, proliferation, inflammation, etc.). Whether this is compatible with or in favor of systems medicine-oriented training might be doubted. In addition, many European countries face a shortage of physicians, especially in the field of general practitioners. This has motivated representatives of health insurance companies, political parties, and patient organizations to demand a faster and less “science-oriented” medical education. The desired result would be an increased output of practitioners that can recognize the most frequent signs and symptoms and prescribe a standardized, established, and cost-effective therapy—if available. Moreover, the aforementioned shortage in physicians leads in many countries to (1) a high and still increasing workload for physicians and (2) a significant disparity between basic researchers and clinical practitioners in terms of income as well as the possibility to get tenured positions. Consequently, in many areas, it becomes increasingly difficult to find medical students or practitioners, which are willing and/or able to take an interest in or participate in new (scientific) developments. Very often, young clinicians that would be ideal candidates to become “systems physicians,” both as researchers and practitioners, are nowadays forced to and rewarded for fast, standardized, and unquestioned application of (observational) “evidence-based medicine.”

In conclusion, besides the development of systems medicine-facilitated courses of “traditional” and integrative topics, we should focus on two additional goals: (1) to increase the awareness for the mid- and long-term benefits of this way among students, academic teachers, and clinicians, but also representatives of health insurance companies, political parties, and patient organizations, and (2) to give both medical students and young physicians a framework and protected area to train and participate both in the development and application of systems-based strategies. In the end, we absolutely need their positive input and enthusiasm.

Below, we describe some of the education and training programs in systems medicine which are in 2015 available in Europe and beyond. They are divided into master’s, doctoral, and other types of systems medicine education suitable also for lifelong training.

#### **4.1 Systems Medicine at Master’s Studies of Medicine**

*Linköping University, Sweden.* The medical curriculum of Linköping University currently includes a 1-hour introductory lecture in systems medicine during the fourth semester.

*University of Ljubljana, Faculty of Medicine.* The medical master’s curriculum is composed of obligatory and elective courses that represent up to 10 % per each study year (1–6). Several systems medicine quantitative and data-oriented topics are already offered within the elective courses, each of 3 ECTS credits, such as “application of physics and biophysics in diagnostics and treatment,” “mathematical principles in biochemistry,” “basics of computer-based imaging methods in medicine,” “e-learning and e-materials in medicine,” “health information practicum,” “molecular modeling in biochemistry,” “computer simulations of dynamical processes in biochemistry,” “application of bioinformatics tools in medicine,” “contemporary informatics in biomedicine,” “functional genomics in medicine,” etc. Biophysics is offered in year 1 as an obligatory course, while students are not offered mathematics. Students can also choose elective research projects which include systems medicine projects that can be performed also at another university, clinic, or accredited research site. The current goal is to follow how many students are interested in education on the quantitative and systems approaches in medicine and upon that decide whether to offer a “systems medicine elective course module” running from year 1–6.

At the *Philipps-University Marburg, Germany*, the Systems Biology Platform ([www.i-lung.de](http://www.i-lung.de)) of the German Center for Lung Diseases is implementing a facultative curricular course in systems medicine for medical students and students in the master’s program “human biology.” It will start in 2015 and grant six ECTS credit points. The course covers modules on clinical medicine and pathophysiology, molecular regulatory circuits and technology, and statistics, bioinformatics, and modeling. In addition, the



platform offers a clinician scientist program for young physicians in cooperation with the University Medical Center, Marburg. It provides training in systems medicine and free time for own research.

The *Georgetown University, Washington, DC, USA*, offers a dual master's degree program (MD/MS) in systems medicine which seems to be the most comprehensive formal program of systems medicine up to date. Medical students may choose to learn genomics, proteomics, translational bioinformatics, metabolomics, systems biology, pharmacogenomics, epigenomics, and biomedical informatics, all in the context of clinical decision-making. In addition to the course curriculum, the students also experience a year-long practicum wherein they apply informatics methodologies to clinical data. While the experience is still nascent, it appears that graduates are selecting careers in which these new systems medicine skills will be relevant. Further information is available on <https://gumc.georgetown.edu/spi/systemsmedicine>.

#### **4.2 Doctoral Training towards MD/PhD Title**

The *University of Ljubljana* offers doctoral training to MDs within doctoral studies of biomedicine, where MDs can choose courses for PhD in basic medicine or clinical medicine. This 3.5-year doctoral study since 2014 offers also ten ECTS modules on systems medicine. Doctoral students could choose this module that includes lectures, hands-on computation tutorials, and systems medicine project works that are graded ([www.uni-lj.si/elektronskeknjige/02%20Biomedicina%20angl/Biomedicine.html#p=2](http://www.uni-lj.si/elektronskeknjige/02%20Biomedicina%20angl/Biomedicine.html#p=2)).

The *University College London* took the lead and is since 2013 already offering a systems biology course (SysMIC) to UK students of different universities. They run a web-based training which offers multimedia content with guides for self-study and self-assessment. The courses had so far over 700 bioresearchers inscribed. This shows the growing awareness in life scientists about the need of systems skills. It is interesting to note that students funded by BBRSC (Biotechnology and Biological Sciences Research Council) in the field of life sciences are obliged to take this e-course. Several topics of the SysMIC course are relevant for systems medicine, especially all introduced mathematical concepts. What is missing is the disease-oriented problem work that could be introduced with the aid of CASyM partners.

The *Imperial College London* runs the Stratified Medicine Graduate Training Programme in Systems Medicine and Spectroscopic Profiling (STRATiGRAD) PhD program (<http://www1.imperial.ac.uk/stratigrad/about/>) addressing the work at the interface of disciplines that collectively drive new discoveries at the systems level. Through the collaborative network of research organizations with common interests in the areas of stratified medicine, clinical diagnostics, prognostics and theranostics biomarker discovery, novel therapeutic development, etc., they train PhD students in the fields of molecular phenotyping, systems modeling,

and stratified medicine. The program applies e-training resources and computational/analytical infrastructures in systems biology at the Imperial College ([www1.imperial.ac.uk/computationalsystemsmedicine/phenomicsandmodelling/](http://www1.imperial.ac.uk/computationalsystemsmedicine/phenomicsandmodelling/)).

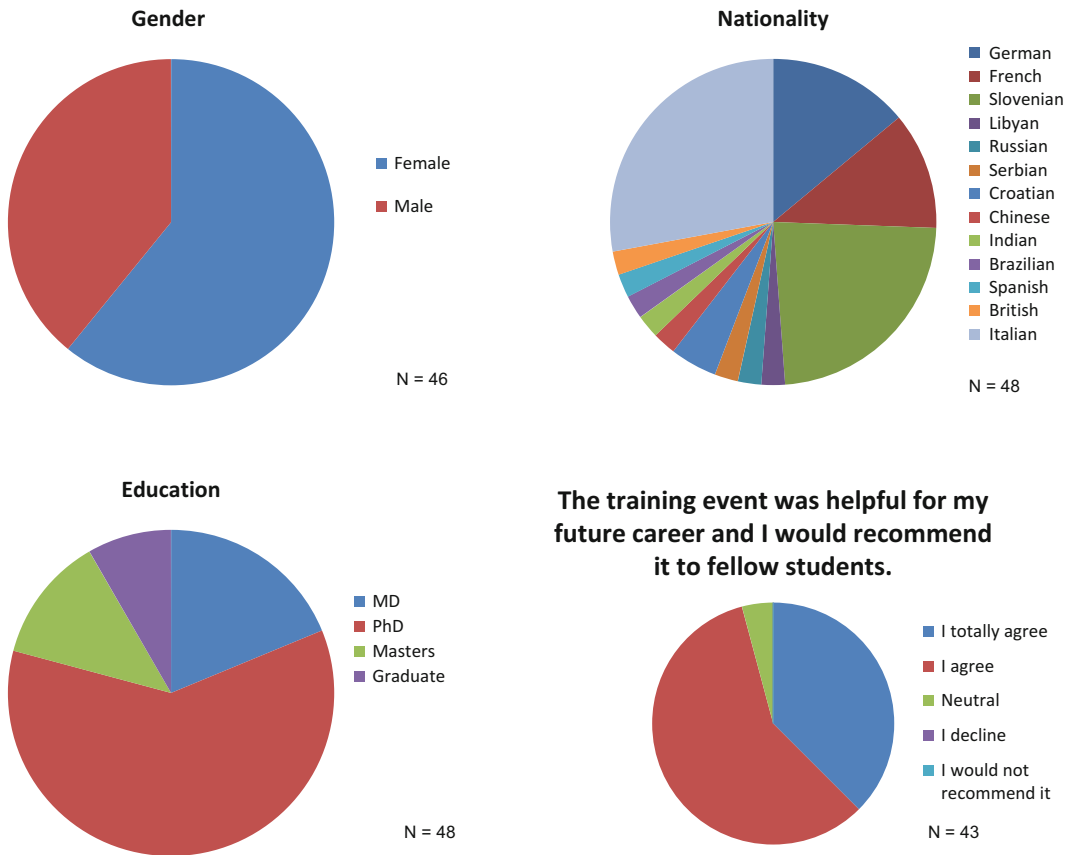
The *Helmholtz Graduate School* “Molecular Cell Biology” is a collaboration between the Max Delbrück Center for Molecular Medicine (MDC) and the Humboldt-Universität zu Berlin (HU), Freie Universität Berlin, and Charité-Universitätsmedizin Berlin Medical Faculty. The graduate school offers an interdisciplinary structured PhD research training and currently supports 350 PhD students. Research training is supplemented with lectures and workshops on methods and technologies, combined with soft skills courses in career development. The graduate school applies the credit point system which helps the students to structure their training according to needs and interests. Students may choose to apply to international exchange programs, such as the MDC-NYU PhD Exchange Program in Medical Systems Biology ([www.mdc-berlin.de/en/bimsb/phd\\_program/index.html](http://www.mdc-berlin.de/en/bimsb/phd_program/index.html)). The research offered at the MDC covers several multifactorial disease areas, such as Cardiovascular and Metabolic Research, Cancer Biology and Immunology, Neurobiology, Medical Systems Biology and Bioinformatics, etc.

The *University College Dublin* has been running a Bioinformatics and Systems Biology structured PhD program since 2009, with an emphasis on equipping its students with interdisciplinary “wet” and “dry” lab training. Research focuses on cancer and infection biology, and the program is run in conjunction with collaborators at *Trinity College Dublin* and the *Royal College of Surgeons in Ireland*.

### **4.3 Other Types of Systems Medicine Education and Training for Preclinical and Clinical MDs**

Other types of systems medicine education include workshops for different levels of students and professionals. A successful example of this was the CASyM workshop in Ljubljana in 2013 (<http://cfgbc.mf.uni-lj.si/2013anniv8casym>) that was accredited with 5 ECTS credits by the University of Ljubljana and with 20 CME credits by the Medical Chamber of Slovenia. Other hands-on tutorial examples are:

- The CASyM training tutorial “Modeling Tools for Pharmacokinetics and Systems Medicine” which was adjoined to the “20th International Symposium on Microsomes and Drug Oxidations,” Leinfelden-Echterdingen, Germany (2014). The event was accredited with 6 European CME credits (ECMEC) by the European Accreditation Council for Continuing Medical Education (EACCME).
- The 1st SyBSyM Como School 2014 “Systems Biology and Systems Medicine: Precision Biotechnology and Therapies,” Lake Como, Italy.
- The CASyM training event “Systems approaches to biological clocks and diseases” adjoined to the 44th congress of the



**Fig. 1** Participants of the systems medicine tutorials presented in this chapter according to their gender, education, and nationality

French-speaking society of chronobiology (Société Francophone de Chronobiologie, SFC), Paris, France (2014).

- The “Advanced Summer School in Systems Medicine: Implementation of Systems Medicine across Europe. A FEBS Advanced Lecture Course.” Djurhamn, Sweden (2015).

The distribution of the attendees of the systems medicine tutorials according to their gender, education, nationality, and overall satisfaction is presented in Fig. 1. Similar workshops are planned in various European locations for the coming years.

## 5 Which Benefits Are Expected from the Interdisciplinary Systems Medicine Training?

Within the systems medicine training procedures, the added value of systems biology will be highlighted and focused towards the pending clinical issues on every complex disease.

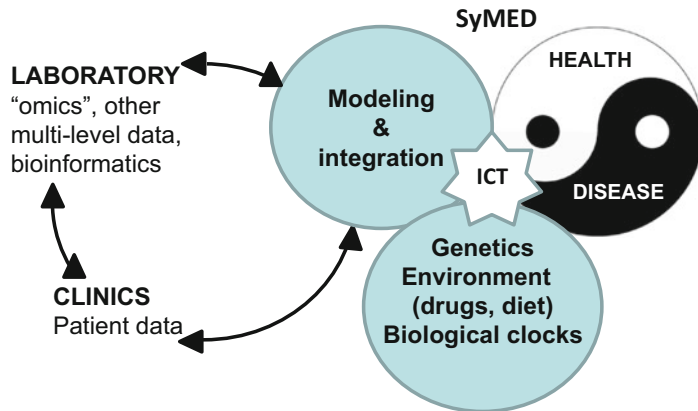
New educational concepts will be proposed by stakeholders including systems medicine centers, major clinical centers in Europe, and life sciences industries. Their involvement is necessary for systems medicine to reach the practice. Since the training should address primarily clinicians and their clinical needs with a translational focus on personalized (P4) medicine, this should accelerate paths towards the discovery of novel diagnostics and biomarkers and also towards a rational design of combinatorial and chronotherapies encompassing biological clocks [8]. The systems approaches facilitate early intervention, anticipation, and/or prevention of disease risk and/or onset and aid in the development of safer and more efficient personalized treatments. Such training will broaden the pool of biomedical researchers that combine quantitative techniques and systems approaches in translational and clinical medicine settings and integrate them with health applications of information and communication technologies. Consequently, this should contribute to the reduction of chronic disease-related health-care costs. Last but not least, the students with competences in computational modeling on top of their basic medical background are more flexible and more attractive for the job market. Pharmaceutical companies are increasing their modeling groups and are searching for medical doctors to add the systems medical viewpoint [18].

---

## 6 Where Should We Go (Where Is the Sky)?

Despite numerous efforts to promote topics of the systems medicine education, we were so far unable to make a massive move towards the formalization of the systems medicine education. For the master's and graduate levels, we can likely progress further in collaboration with some professional associations, such as AMSE, the Association of Medical Schools in Europe ([www.amse-med.eu/](http://www.amse-med.eu/)), and AMEE, the Association for Medical Education in Europe ([www.amee.org](http://www.amee.org)). A unifying curriculum for systems medicine is not expected in the near future since also the basic medical curricula differ between different medical schools. However, it is likely achievable that the interdisciplinary systems approaches presented in Fig. 2 would be presented within the medicine educational pillars, to accelerate the move of systems understanding into the clinical practice.

For the continuous professional education level, a real breakthrough would be if systems medicine training tutorials would become satellites of some larger European or world medical congresses, such as the International Liver Congress that yearly attract over 10,000 hepatologists, doctors, and medical professionals. To achieve this, we would need precise input from medical



**Fig. 2** Outline of the envisioned systems medicine education and training. Interactions and iterative cycles between modeling and experimentation in the laboratory and in clinical settings are envisioned to drive the education. *ICT* information and communication technologies

professionals for every medical discipline, to identify missing systems expertise that is relevant for their field. It might seem a long way to go. However, each path is made of many small steps.

## Acknowledgment

The authors would like to acknowledge all partners of the CASyM consortium and colleges from Georgetown University, Washington, DC, for their contributions to CASyM reports where parts are reviewed in this book chapter. Special thanks to Dr. Silvio Parodi from University of Genoa, for reading this chapter and giving critical comments. Part of this work has been funded by the BMBF (e:bio miRSys—FKZ 0316175B and e:Med CAPSYS—FKZ 01X1304E) and HMWK (LOEWE Medical RNomics—FKZ 519/03/00.001-(0003)) to B.S., from the FP7 CASyM grant (J.A.) and the Slovenian Research Agency grant P1-0390 (D.R.).

## References

1. WHO (2009) 2008-2013 Action plan for the global strategy for the prevention and control of noncommunicable diseases. WHO Document Production Services, Geneva, Switzerland. [http://apps.who.int/iris/bitstream/10665/44009/1/9789241597418\\_eng.pdf](http://apps.who.int/iris/bitstream/10665/44009/1/9789241597418_eng.pdf)
2. UNION (2014) T.C.O.E. Council conclusions on nutrition and physical activity. [http://www.consilium.europa.eu/uedocs/cms\\_data/docs/pressdata/en/lsa/143285.pdf](http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/lsa/143285.pdf)
3. Bousquet J et al (2014) Systems medicine approaches for the definition of complex phenotypes in chronic diseases and ageing. From concept to implementation and policies. *Curr Pharm Des* 20(38):5928–5944
4. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2:343–372
5. Rozman D (2014) From nonalcoholic fatty liver disease to hepatocellular carcinoma: a systems understanding. *Dig Dis Sci* 59(2):238–241

6. Barrenas F et al (2012) Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. *Genome Biol* 13(6):R46
7. Naik A, Kosir R, Rozman D (2013) Genomic aspects of NAFLD pathogenesis. *Genomics* 102(2):84–95
8. Levi F et al (2014) Wrist actimetry circadian rhythm as a robust predictor of colorectal cancer patients survival. *Chronobiol Int* 31(8):891–900
9. Davies SK et al (2014) Effect of sleep deprivation on the human metabolome. *Proc Natl Acad Sci U S A* 111(29):10761–10766
10. Noble D (1960) Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations. *Nature* 188:495–497
11. MacArthur L et al (2013) Systems medicine, a new model of health care. In: Sturmborg JP, Martin CM (eds) *Handbook of systems and complexity in health*. Springer, New York, Imprint, Springer, New York, NY. p. XXII, 954 p. 269 illus., 165 illus. in color
12. [http://ec.europa.eu/education/tools/ects\\_en.htm](http://ec.europa.eu/education/tools/ects_en.htm)
13. Murgatroyd GB (2011) Continuing professional development, the international perspective. [http://www.gmc-uk.org/CPD\\_\\_\\_The\\_International\\_Perspective\\_Jul\\_11.pdf\\_44810902.pdf](http://www.gmc-uk.org/CPD___The_International_Perspective_Jul_11.pdf_44810902.pdf)
14. Auffray C et al (2010) From systems biology to systems medicine, European Commission, DG Research, Directorate of Health. Brussels 14–15 June 2010. Workshop report. [ftp://ftp.cordis.europa.eu/pub/fp7/health/docs/final-report-systems-medicine-workshop\\_en.pdf](ftp://ftp.cordis.europa.eu/pub/fp7/health/docs/final-report-systems-medicine-workshop_en.pdf)
15. Sobradillo P, Pozo F, Agusti A (2011) P4 medicine: the future around the corner. *Arch Bronconeumol* 47(1):35–40
16. CASyM (2012) European Systems Medicine road-map discussion. <https://www.casym.eu/index.php?index=90>
17. CASyM (2013) CASyM ICSB2013 training workshop report – should systems medical training be integrated for basic and clinical researchers? <https://www.casym.eu/index.php?index=90>
18. Damjana Rozman UPZ, Acimovic J, Kirschner M, Tegnér J, Auffray C, Kolch W, Lévi F, Benson M, Byrne H (2014) CASyM report; D2.2 Plan for tailored interdisciplinary exchange programmes

## Systems Medicine in Pharmaceutical Research and Development

Lars Kuepfer and Andreas Schuppert

### Abstract

The development of new drug therapies requires substantial and ever increasing investments from the pharmaceutical company. Ten years ago, the average time from early target identification and optimization until initial market authorization of a new drug compound took more than 10 years and involved costs in the order of one billion US dollars. Recent studies indicate even a significant growth of costs in the meanwhile, mainly driven by the increasing complexity of diseases addressed by pharmaceutical research.

Modeling and simulation are proven approaches to handle highly complex systems; hence, systems medicine is expected to control the spiral of complexity of diseases and increasing costs. Today, the main focus of systems medicine applications in industry is on mechanistic modeling. Biological mechanisms are represented by explicit equations enabling insight into the cooperation of all relevant mechanisms. Mechanistic modeling is widely accepted in pharmacokinetics, but prediction from cell behavior to patients is rarely possible due to lacks in our understanding of the controlling mechanisms. Data-driven modeling aims to compensate these lacks by the use of advanced statistical and machine learning methods. Future progress in pharmaceutical research and development will require integrated hybrid modeling technologies allowing realization of the benefits of both mechanistic and data-driven modeling. In this chapter, we sketch typical industrial application areas for both modeling techniques and derive the requirements for future technology development.

**Key words** Pharmaceutical R&D, Mechanistic modeling, Data-driven modeling, Physiologically based pharmacokinetics (PBPK), Pharmacodynamics, Hybrid modeling, Data mining

---

### 1 Introduction

Lack of efficacy or concerns about patient safety are major threats in the development of new drugs. This is even more since the development of novel compounds requires substantial and ever increasing investments from the pharmaceutical company. The average time from early target identification and optimization until initial market authorization of a new drug compound takes currently more than 10 years and involves costs in the order of one billion US dollars [1]. The rising costs are driven by a multitude of factors, e.g., attrition risk, regulatory requirements, and time to market.

Notably, most of the budget is spent in late clinical phases with successful transition rates from one stage to the next only reaching between 34 % (phase 2) and 70 % (phase 3) [2], which has a high impact to the overall costs due to the late failures. Given this high attrition rates in pharmaceutical development, especially in the late phases, novel, knowledge-driven concepts for preclinical and clinical research are hence clearly needed.

Systems medicine aims for the translation of mechanistic insights gained in fundamental research toward clinical applications. This involves in particular a sufficient understanding of human physiology and its pathologies to support development of novel treatment strategies by predictive computational models. For mechanistic modeling providing most insight into the biological system, the respective predictive computational models must represent all relevant physiological processes in an appropriate quantitative way. This requires the modeling of physiological processes at the cellular level and the organ scale as well as their integration into a whole-body context. Due to the tremendous complexity of the underlying mechanisms and their mutual cooperation, mechanistic modeling on the full scale is rarely appropriate alone to represent the complexity of therapies from molecular to the macroscopic scale. Hence, mechanistic modeling can focus on subsystems of drug action which are of high relevance for specific applications throughout the drug R&D workflow, such as physiologically based pharmacokinetics or metabolic processes. Moreover, specific biomarkers can also be identified for diagnostic purposes from computational simulations of the underlying mechanisms.

However, a full mechanistic understanding is almost never available in biology and medicine. Data-driven approaches establishing correlations between the genetic, multi-omics, and physiology data must hence be used as surrogates to fill the gaps of mechanistic knowledge. Such combinations of mechanistic and data-driven models are developed for specific classes of applications, such as modeling predisposition of patients with a corresponding pathogenesis or development of disease models to design targeted treatment regimes.

An important benefit of modeling, most prominently mechanistic modeling, is the translation of knowledge throughout a project. Due to the mechanistic information included, computational models can be used for the translation of existing knowledge to new indications or patient subgroups as well as through the more than 10-year average time of a drug R&D project. The latter case is of very high relevance for the application of computational models in the pharmaceutical industry, and it provides a powerful memory of the very heterogeneous knowledge gained along a project as well as its integration.



### **1.1 Computational Modeling: From Systems Biology to Systems Medicine**

Biology has seen the advent of a multitude of novel experimental techniques in recent years allowing unprecedented insights in the various levels of biological organization. It has long been understood that both the sheer amount of experimental information and the inherent complexity of biological processes require adequate tools for data integration, data analysis, and knowledge extraction. In this regard, systems biology has fundamentally contributed to transfer biology from a rather qualitative and descriptive discipline to a quantitative and explanatory science.

Following this holistic concept, systems medicine integrates different levels of biological organization to enable mechanistic analyses of diseases and pathogenesis which will significantly support the development of targeted therapeutic strategies in the future. Along with advances in biology, the mechanistic understanding of processes underlying the multifaceted roles of human physiology has increased considerably in recent years. This scientific progress has been mainly driven by new experimental approaches ranging from omics technologies at the cellular scale to novel imaging techniques at the organ and the organism level. Each of these approaches, however, focuses inevitably on a specific physiological scale as such resembling previous reductionistic approaches in biology. To foster a systemic understanding of the interplay of processes in the human body, existing knowledge from different layers of physiological organization needs to be integrated into a mechanistic framework. It is only by such comprehensive approaches that pathogenesis of complex diseases may be understood and the influence of patient-specific cofactors such as genetic predisposition on clinical end points may be described quantitatively.

---

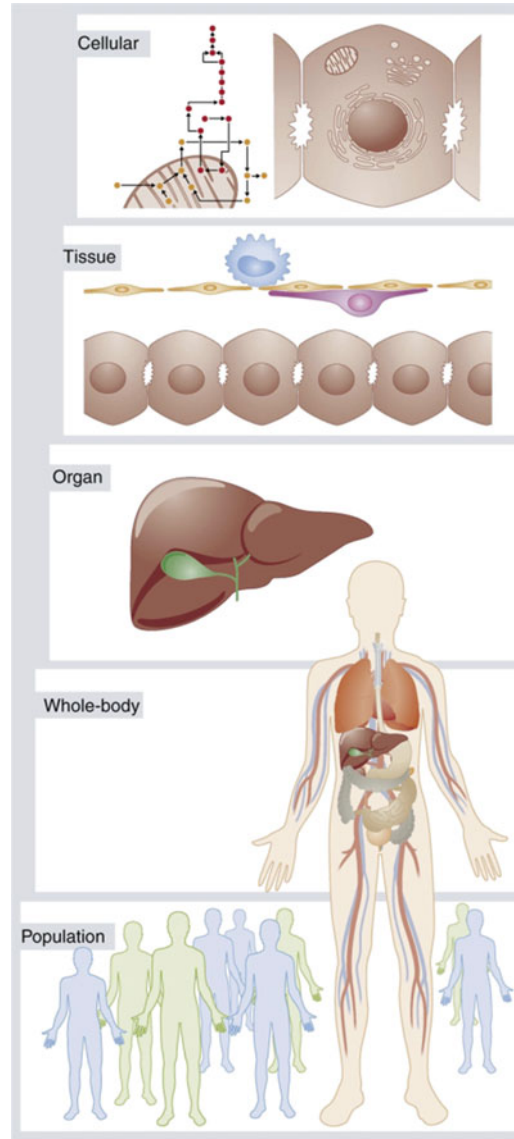
## **2 Mechanistic Modeling**

### **2.1 Covering Multiple Scales of Biological Organization in Systems Medicine**

Systems medicine aims for a multiscale representation of human physiology [3]. This requires the development of computational models from different scales of biological organization ranging from the cellular level up to the whole-body level (Fig. 1). In the following, we will briefly discuss modeling approaches for cellular and organ models before introducing physiologically based pharmacokinetic modeling as an approach to consider the organism level. We will also give application examples before outlining potential approaches for the vertical integration of model approaches across different scales [4].

### **2.2 Cellular Scale: Interaction-Based Graphs, Stoichiometric Networks, and Dynamic Models**

Computational models at the cellular scale are well established in systems biology. In brief, there are three main approaches: (1) interaction-based graphs which qualitatively describe the connection between two components of the graph (proteins, genes, etc.) [5], (2) stoichiometric models which represent cellular carbon biochemistry [6], and (3) dynamic models describing either cellular metabolism or



**Fig. 1** A multiscale representation of human physiology across different levels of biological organization [4]

intracellular signaling [7]. Each of the three approaches has in common that they were initially developed in microbial systems biology, but that there has been a consistent shift to applications in human biology and medicine in the last years [8–10]. Since there are many excellent reviews available, computational models at the cellular scale will not be elaborated on here in more detail.

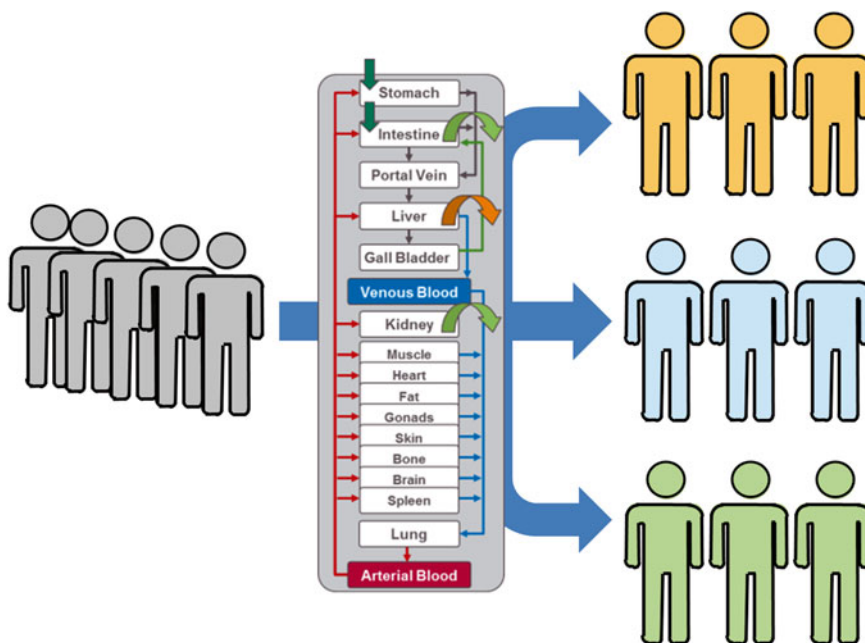
### **2.3 Tissue and Organ Level: Spatial-Temporal Modeling**

Computational models at the tissue and organ scale largely focus on physical conservation laws such as mass, momentum, and charge in order to describe specific physiological functions [11]. Since spatial architecture and organ morphology are important

boundary conditions in such computational models, they are frequently based on partial differential equations. This allows, for example, quantitative simulations of blood flow within spatially resolved representations of the surrounding tissue. To date, several organ systems have been developed such as the heart model [12], models of liver regeneration [13] and perfusion [14], or models of kidney [15] or of the lung [16]. These models are mainly driven by mechanistic representations of specific organ functions which ultimately drive the structure of the underlying modeling approach. The models are valuable tools for the investigation of the physiological functionality in healthy or even diseased individuals.

#### **2.4 Organism Level: Physiologically Based Pharmacokinetic (PBPK) Modeling**

Physiologically based pharmacokinetic (PBPK) modeling aims for a detailed mechanistic representation of human physiology at the whole-body level. This is because the various organs in the human body are explicitly represented in PBPK models (Fig. 2) allowing for the simulation of time-concentration profiles in specific tissues. On the one hand, PBPK models are based on large-scale collections of physiological parameters such as organ volumes or surface areas which are provided to the user by the modeling software itself. On the other hand, physicochemical properties of a compound such as lipophilicity or molecular weight are used to



**Fig. 2** Physiologically based pharmacokinetic (PBPK) modeling allows integration of experimental data from different levels of biological organization to quantify drug concentration profiles in the plasma and different tissues. PBPK model can also be used for building individualized models and to allow for patient subgroup stratification (adapted from Kuepfer et al. 2014)

parameterize the distribution model describing the underlying mass balance in PBPK models. Hence, even though PBPK models may comprise several hundreds of ordinary differential equations, the number of independent model parameters is usually small, i.e., around 5–10, due to the large degree of prior physiological information contained in the models.

For PBPK model building, prior information about the governing physiological processes underlying adsorption, distribution, metabolization, and excretion of a specific compound is required. The physicochemical properties of a compound are used to parameterize the basic distribution model quantifying in particular tissue permeation by passive distribution. Active processes such as enzyme-catalyzed metabolization or transporter-mediated drug uptake or secretion may be described by using quantifying tissue-specific protein abundance [17]. The degree of biological knowledge included in PBPK models hence ranges from tissue-specific gene expression profiles [17], organ blood flow, and perfusion rates to exhaustive collections of anthropometric parameters. For model establishment and validation, time-concentration profiles are required. Ideally, multiple doses should be considered as well as different routes of drug administration such as intravenous or oral applications allow identification of the independent model parameters with sufficient accuracy. PBPK is nowadays routinely used in pharmaceutical development specifically supporting the various stages in the R&D process. In the following, we will briefly review previous successful applications in preclinical and clinical applications.

#### 2.4.1 *Animal Models and Preclinical Research*

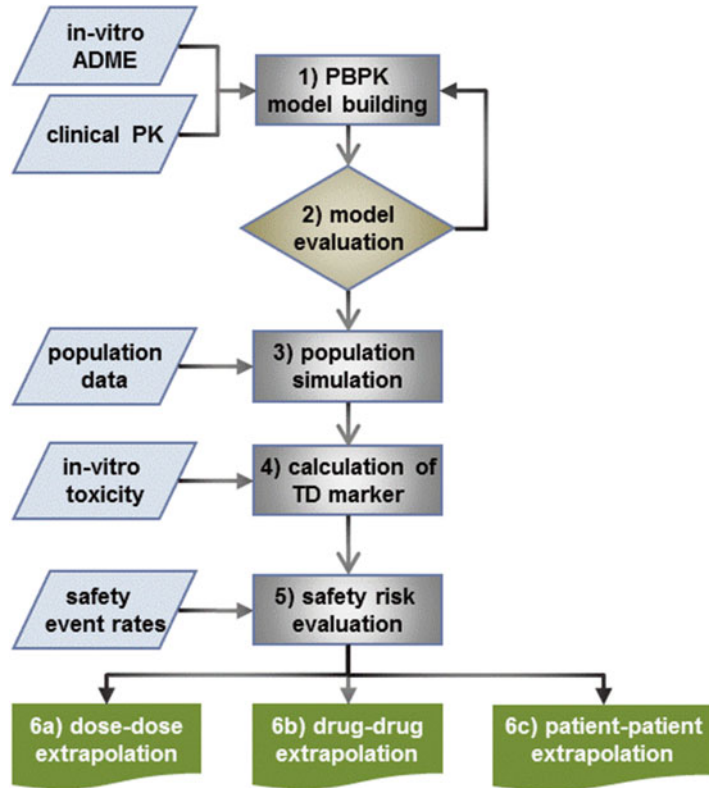
A prominent application of PBPK modeling is the analysis of animal models in preclinical research following in vitro lead identification and optimization. Still, novel compounds are routinely tested in different laboratory species such as mice, rats, dogs, or monkeys imposing an enormous ethical burden due to the considerable animal sacrifice. While no computational model can nowadays completely avoid testing of toxicity in living organisms, targeted simulations can nevertheless help to significantly decrease the number of animals needed. With regard to PBPK modeling, this is in particular the case since the basic distribution model reflecting the physiology of the organism can be easily re-parameterized to consider another species. Notably, most PBPK software packages contain such information. Keeping track of the preclinical R&D process, animal PBPK models hence can be implemented one by one, representing on the one hand specific physiological information for each organism and creating on the other hand a mechanistic record of the relevant physiological processes in different species. This cumulative integration ensures a rational experimental planning thereby decreasing significantly animal sacrifice.

A recent study in mice and humans considering ten arbitrary chosen drugs showed that PBPK-based cross-species extrapolation may lead to about 83.5 % of model agreement when species-specific physiology is considered together with kinetic rate constants in active processes, gene expression profiles, and species-specific plasma protein binding [18]. The analysis hence suggested that PBPK modeling can be used for the integration of preclinical data, the translation of mechanistic knowledge from one species to another, and ultimately the prediction and planning of first-in-man studies.

#### 2.4.2 PBPK Modeling of Special Populations

The initial step in PBPK-based analyses is usually the establishment of a reference model for an average patient using parameters derived from means of available data. Notably, the anthropometric information quantifying the basic physiology of this standardized individual is usually taken from database information provided in the PBPK software package. Following the establishment of a mean reference model, special subgroups of patients such as diseased individuals or specific genotypes can subsequently be considered. Since PBPK models describe physiology at a large level of detail, they allow a straightforward modification of parameters representing specific physiological properties such as hepatic or renal impairment [19] or a specific genotype [20]. Based on a reference PBPK model, the impact of physiological changes on the pharmacokinetic behavior of a specific compound can directly be simulated as such allowing either a model-based analysis of clinical trial data or, in turn, a model-based design of future trials. An important application of PBPK modeling of special populations is risk assessment for special patient cohorts. Generally speaking, PBPK models can be used to quantify the therapeutic window characterized by the lower level of efficacy, below which the drug shows no therapeutic effect, and the upper level of safety, above which toxic events may occur. In the past, PBPK has among others been used to describe physiological change in cirrhotic patients with different Child-Pugh scores [19], fatal levels of morphine in the mother's milk of breast-feeding women with a particular phenotype [21], or the analysis of statin-induced myopathy in rare genotype subgroups of patients [22].

In the latter study, a generic, PBPK-based workflow was proposed (Fig. 3) which allows to estimate the occurrence rate of adverse events in high-risk patient cohorts. Initially, carefully validated mean reference models are established for the dominant genotype subgroups. Subsequently, simulations of virtual populations are performed to capture the effect of interindividual variability in various physiological parameters at the PK levels. In the next step of the workflow, a toxicodynamic marker is derived based on the simulated tissue concentrations. Ultimately, the distribution of the toxicodynamic marker is normalized to clinical incidence rates in the dominant genotype subgroups thereby



**Fig. 3** A workflow for PBPK-based risk assessment [22]

allowing the prediction of the occurrence rates of adverse events in different patient subgroups, for different doses and even different drugs [22].

#### 2.4.3 Individualized Therapeutic Designs

Despite an increasing understanding of the physiological processes underlying drug pharmacokinetics, many pharmaceutical therapies are still based on the “one-size-fits-all” paradigm. This largely neglects significant variability between different human individuals due to physiology, gender, genotype, health state, or lifestyle. This negligence of available information leads to a rather cautious design of therapeutic doses, which rather underestimates the therapeutic window in order not to interfere with patient safety at any rate. If available, physiological information of individual patients can however well be included in PBPK models since the different anthropometric properties are explicitly represented. Bayesian PBPK modeling has been used, for example, to identify individualized parameterizations from clinical PK data explaining among others the functional genotype of the study participants in a liver uptake transporter [23].

#### 2.4.4 Pediatric Scaling

The usual trial design in clinical research foresees assessment of general safety in healthy volunteers in phase 1, evaluation of efficacy in patients in phase 2, and randomized trials in large groups of patients in phase 3. Obviously, such well-designed research programs are not possible when it comes to dose identification in children due to severe ethical issues. While phase 1 trials are forbidden in children, a limited number of patients in phase 2 or 3 would also significantly complicate dose identification for novel drugs if they were performed in a similar way as for adults. With respect to pediatrics, two important things need to be considered: (1) the body mass composition is different in children, particularly in newborns since, for example, the lipid/water content is significantly changed, and (2) the enzymes catalyzing active processes undergo a maturation process. Children are hence not just small adults, also from a pharmacokinetic point of view. Just like for the special populations described above, PBPK modeling offers the unique opportunity to integrate physiological information from specific data collections describing both physiology in different age groups and the corresponding gene expression [24, 25]. Notably, PBPK-based concepts for pediatric scaling are nowadays recommended by the Food and Drug Administration (FDA) to support an efficient study design in children.

#### 2.5 Multiscale Modeling in Systems Medicine

PBPK models provide the unique opportunity to describe the organism level into which cellular or tissue levels can be integrated. Using multiscale models as a structural template for the systematic integration of experimental measurements at the cellular level, organ level or at the whole-body scale provides an analytical framework for mechanistic analyses from a systems point of view. Vertical model integration has been applied successfully in the past in various cases: (1) A model of the MAPK signaling cascade has been integrated in a PBPK model describing PK of a prodrug and its corresponding metabolite [26]. The overall model was used in a case study to describe different therapeutic outcomes in the treatment of pancreatic tumors in specific genotype subgroups. (2) PBPK models have also been used to simulate stoichiometric network models at the cellular scale [27] within a whole-body context [28]. The resulting model, which was stepwise integrated to allow coupling of the different mathematical modeling formalisms, was exemplarily used for the analysis of allopurinol treatment of hyperuricemia or occurrence of paracetamol-induced intoxication [28]. (3) PBPK modeling was also used to describe mass flow in a spatially resolved model of a mouse liver [14]. The vascular tree of the liver was built based on micro-CT imaging such that hepatic mass transfer in the resulting spatially resolved models is governed by the physiological architecture and the composition of the connecting hepatic tissue. The model was used to describe first pass perfusion in the liver and to quantify the impact of steatosis and CCl<sub>4</sub>-induced necrosis on hepatic capacity.



---

### 3 Data-Driven Modeling

#### 3.1 *Data-Driven Modeling to Link Cell and Physiology Data*

Prediction of the efficacy of new drug-based therapies on complex diseases in humans, either using a single drug or a combination of drugs, requires handling and the integration of the multiscale complexity of response of living systems on abiotic stress in the physiological environment.

Human diseases result from abnormalities in an extremely complex system of molecular processes. In these processes, virtually no molecular entity acts in isolation, and complexity is caused by the vast amount of dependencies between molecular and phenotypological features. In a large-scale meta-analysis [29], the mutual involvement of genes and diseases has been analyzed on a genome-wide level. Apparently one-to-one relations between genes and diseases showed to be exceptions, restricted to the Mendelian diseases. Complex diseases, however, like cancers, metabolic disorders, autoimmune diseases, or psychotic disorders, cannot be associated to single genes only. Although associations between the genotype and physiology may neglect relevant biological mechanisms such as epigenetic control, transcriptomics, protein phosphorylation, etc., the studies showed that complex diseases are in any case associated with large sets of genes. Moreover, the relations between diseases or groups of diseases and the respective gene sets are not one-to-one. Most of the disease-related gene sets showed significant overlaps indicating relationships of heterogeneous complex diseases on the genome level. Using these overlaps, both on the level of the diseases and on the genome level, a relationship network can be established, where either two diseases are connected if the respective gene sets show a significant overlap or two genes are connected if they are related to the same disease. However, the resulting networks showed a very high degree of connectivity such that it proved to be hard to extract clinically relevant conclusions from networks connecting only genotype and clinical phenotype. Moreover, recent results show that genome-wide association studies (GWAS) linking complex diseases and genotype could not lead to results of therapeutic relevance [30]. Hence, multilevel network approaches linking genomics, proteomics, and metabolomics with clinical phenotypes are required to realize the promises of systems biology in drug research and development.

One gap of knowledge is due to the high amount of functional components of cells and their imbedding into tissues. Almost never all variables describing the systems can be assessed in an affordable experimental setting, even if the experimental measurement technologies were available. Hence, any model will neglect variables, the latent variables, leading to the so-called closure problem. This means that a model, which describes the behavior of only a subset



of all variables, will be predictive if the measured subset contains the full information with respect to the cell state and the neglected variables cannot force the cell into a qualitatively new phenotype. Hence, the proper identification of a set of “relevant” variables representing the full information with respect to the cell state is critical for any type of predictive modeling. This set of variables may be heterogeneous; it may contain genotype, expression, and physiology data. For example, if we are interested in the propensity of a specific cancer cell type to respond to a targeted drug, this might be effected by the genetic background of the cell, the mix of growth factors in the culture, and the relative strength of proliferation and resistance-related modules in the cell. Hence, we have to identify and integrate heterogeneous variables representing all these impact factors on cell fate into the model in an appropriate manner. It is not necessary that these variables play a causal role in the mechanism to be modeled, it is sufficient that they allow the observation of cell fate. Such systematic integration of heterogeneous data, however, is not established today except for very specific use cases.

An application which is tackled today in support of target search and combinatorial therapy prediction is the reconstruction of the activity of signaling networks, which provide a mechanism for regulating cellular cross talk and gene transcription, quantified by, e.g., protein phosphorylation. Efficient modeling and simulation of the response of signaling protein phosphorylation to multiple, complex combinations of stimuli and inhibitors are crucial for improved research for targeted drugs.

Methods for network reconstruction for signaling from time series data are established. However, in many applications dynamic data are not available, e.g., if the impact of mutations on the network of drug action has to be modeled or for the prediction of cross-effects between drugs. Here, the reconstruction of functional networks based on combinatorial but stationary data is the method of choice. The established network reconstruction algorithms solve a combinatorial, mixed-integer optimization problem in order to minimize the error of a network-based signaling model with given experimental data [31]. For example, nodes represent signaling proteins, and edges (connections between nodes) represent the cascade direction of stimulated protein phosphorylation. If the number  $n$  of network nodes increases, then the number of potential networks to be analyzed will increase at least exponentially with  $n$ . Thus, any algorithm using an exhaustive search analyzing all possible networks with  $n$  nodes will become impractical even at modest  $n$ . Since most mechanisms which are relevant for applications involve multiple pathways and their cross talk, there is a need for algorithms which avoid the pitfalls of detailed network reengineering in only one step.

Therefore, a fine-grained model may provide very detailed insight; however, it requires to identify networks with very high complexity including a large number of nodes and edges as well as feedback loops. Moreover, as proteins may be taken into account whose phosphorylation levels have not been measured, the direct network reengineering algorithms may become ill posed hampering the stability and numerical efficiency of the network reconstruction. Additionally, incorrect signal transfer models along edges can result in ill-posed network identification leading to ambiguous network models as well.

### **3.2 Structured Hybrid Modeling**

In order to reduce the complexity, structured hybrid modeling (SHM) can be applied [32]. An iterative refinement workflow starts with a pure black box modeling of the phosphorylations as function of the stimuli and inhibitions using black boxes. Then, an iterative refinement of the black box model by iterative introduction of intermediate black boxes as well as their connections can be performed. In each refinement step, the model is trained, and the structure with the minimum residuals can be selected for further refinement. This combinatorial approach follows a bottom-up strategy starting with a network with minimal complexity, so the combinatorial efforts can be controlled in each step.

In order to establish pharmacodynamics models allowing the prediction of drug efficacy in man, however, modeling signal transduction and its inhibition by the drugs is relevant but not sufficient. To translate cell culture-based data toward clinics, models integrating heterogeneous levels of biological functionalities, such as integration of genotype, epigenetics, expression, and physiology, are inevitable. Due to a significant lack of mechanistic understanding of these multiscale systems for most applications, “big data” approaches appear to be the method of choice. New experimental high-throughput experimentation technologies, such as next-generation sequencing, enable the generation of extremely large data sets from each sample of biomaterial containing thousands or millions of variables providing hope to extract the required information out of the data.

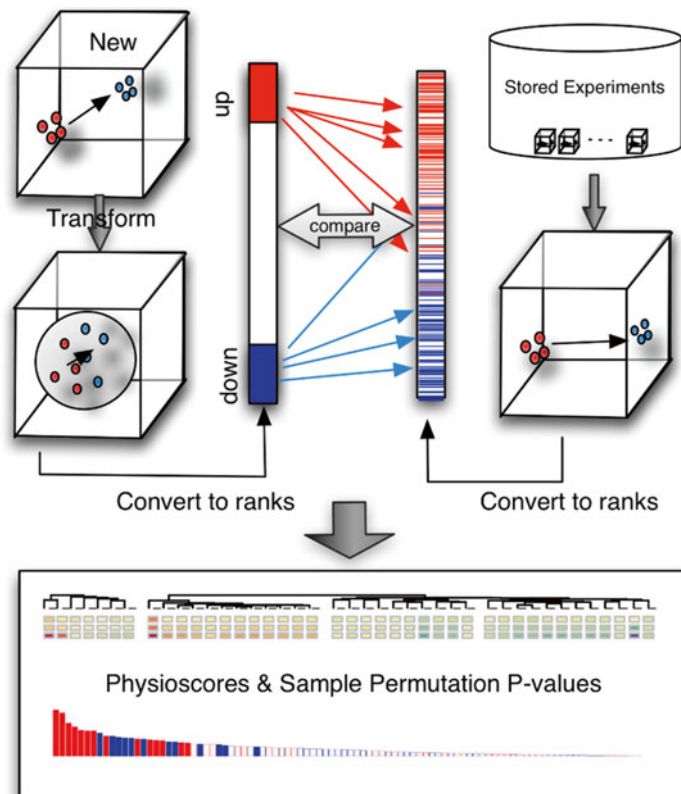
However, in almost all applications the number of biomaterial samples is much smaller than the number of variables assessed, such that a statistical  $N \ll p$ -problem has to be tackled. So “big data analysis” in biomedical applications is characterized by high dimensionality of data and comparably low size of samples, in contrast to most other applications of “big data” today, where large sample sizes characterized by low-dimensional parameters have to be analyzed. In this sense, it is questionable whether biomedical applications of this type should be referred to as “big data” applications. Due to the high dimensionality of the set of variables, all black box machine learning algorithms suffer from the “curse of dimensionality” hampering the reliability of the outcome patterns

of data analysis. To overcome the curse of dimensionality, hybrid approaches combining black box and white box models in a structured functional network may be the method of choice. For systems with a priori known internal structure, it has been shown that the effective dimensionality of the model identification problem can be reduced to the dimension of the largest functional node in the system when the system structure is explicitly implemented into the model identification algorithm. As the effective dimension of the input parameter space strongly controls the number of data required for the identification of black box models (curse of dimensionality), this structured hybrid modeling (SHM) approach can reduce the curse of dimensionality [32]. In SHM, hierarchical networks with a fixed interaction structure can be represented by a directed graph. Each node represents an unknown, nonlinear function depending only on the variables which are represented by the directed edges into the node. Hence, if the structure of a hierarchical system can be properly represented by such a network, then (under some restrictions of the structure of the system graph) the network model can be properly identified, and the number of data required for the identification can be reduced dramatically compared to unstructured model types. Therefore, the hierarchical network-based SHM models allow an interpolation in the modeling technologies between unstructured models (like neural networks) and fully mechanistic models. In addition, it allows a systematic integration of variables from heterogeneous biological functionalities in a model without the associated combinatorial explosion of complexity of pure machine learning approaches. Moreover, recently it has been shown that intrinsic properties of SHM models can be used for efficient, direct reengineering of functional network structures from data avoiding fitting of models with high complexity [33]. This can lead to improved stability of the resulting functional network structures compared to the standard network reengineering approach.

### **3.3 *PhysioSpace* Concepts**

An alternative approach to tackle the high dimensionality of omics data compared to sample sizes in order to translate information from cell line experiments to clinics is to reduce the original data sets by projection onto genome-/proteome-wide patterns representing the relevant physiological features. Recently, it has been shown that the first five components from principal component analyses (PCA) provide relevant information for classification of most tissues and disease states [34]. Further results showed that clinical omics data sets can be properly characterized using a few principal components only [35] and a few variables are sufficient to control biological phenotypes [36]. These results indicate that appropriate projections of lab data onto patterns generated from PCA on clinical data sets might be sufficient to translate lab data to the world of clinical research.

So, relating expression signatures from different sources such as cell lines, in vitro cultures from primary cells, and biopsy material is an important task in drug development and translational medicine as well as for tracking of cell fate and disease progression. The biological rationale underlying these projection methods is the existence of shared generic processes driving the physiologically relevant mechanisms. The respective signature relation approaches require robust statistical methods to account for the high biological heterogeneity in clinical data and must cope with small sample sizes in lab experiments and common patterns of co-expression in ubiquitous cellular processes. Recently, a novel method was described [37] allowing to position dynamics of time series data derived from cellular dynamics experiments, such as differentiation of stem cells, and disease progression in a genome-wide expression space. As depicted in Fig. 4, the PhysioSpace is defined by a compendium of publicly available gene expression signatures representing a large set of biological phenotypes. The mapping of gene expression changes onto the PhysioSpace leads to a robust ranking of physiologically relevant signatures, as



**Fig. 4** Algorithm of PhysioSpace enabling to map expression data from heterogeneous public data sources and specific lab experiments [37]

rigorously evaluated via sample-label permutations. A spherical transformation of the data improves the performance, leading to stable results even in case of small sample sizes. Using PhysioSpace with clinical cancer data sets reveals that such data exhibits large heterogeneity in the number of significant signature associations. This behavior was closely associated with the classification end point and cancer type under consideration, indicating shared biological functionalities in disease-associated processes. Even though the time series data of cell line differentiation exhibited responses in larger clusters covering several biologically related patterns, top-scoring patterns were highly consistent with a priori known biological information and separated from the rest of response patterns.

---

## 4 Future Trends and Challenges

The relevance of computational models in drug development is constantly increasing, and computational models are already considered as a valuable complement to experimental data by both pharmaceutical companies and regulatory bodies like the US Food and Drug Administration (FDA). PBPK-based predictions for drug-drug interactions have recently been accepted by the FDA without clinical validation as an element in the drug label for ibrutinib. This example nicely outlines how targeted simulations can be used as a valuable tool to complement experimental data. In the future, computational models will routinely be used in drug development for data integration, analysis, and processes thereby markedly providing benefits to industry, patient, and careers.

Recently, a Challenge Workshop “Mathematics for Health Care” was organized by the German Committee for Mathematical Modeling, Simulation and Optimization (KoMSO) in Heidelberg (<http://www.komso.org/>). Scientists from biological research, systems biology, pharmaceutical and food industry, clinics, and mathematics discussed and evaluated in detail the needs for mathematics arising from future trends in biology and clinics. In summary, a more intensive use of mathematics, combined with the development of problem-specific novel mathematical algorithms and modeling methods, will be a key driver for future development in biology and medicine.

In detail, open challenges for novel mathematical development cover a broad range of applications as well as mathematical methodologies:

- Clinicians see a tremendously increasing need for novel, efficient methods for the analysis of highly multivariate patient data in order to predict the time course of the disease under therapy. Their data show an extremely heterogeneous structure. Genotype data must be combined with physiological

parameters and overall diagnostic health parameters, thus linking together rather heterogeneous biological functionalities. Moreover, longitudinal studies must be linked with very large data sets which are available from patients who are monitored only at single time points. Despite significant investments in the past into clinical data analysis using machine learning approaches, the success stories are limited indicating the existence of intrinsic, structural challenges in the unraveling of biological systems.

- There is a strong need for efficient model identification and validation approaches. Dynamic models, based on ordinary differential equations, are the common approach for systems biology. However, even if the system to be modeled is sufficiently closed, systems identification remains an open challenge. Mostly, systems are identified by minimization of residues of simulation compared to experimental data. However, rarely the autocorrelation of the residues is seldom properly analyzed. Often the residues are reasonably low in  $L_2$  norm, but the distribution of the residues along the trajectories shows autocorrelations which hardly fit to white noise. This suggests either that the parameter fit is not sufficient or, even worse, that the assumed system structure does not sufficiently fit to biology, both probably leading to erroneous conclusions from the model predictions. Hence, tools for structural validation are indispensable for further success in more complex applications.
- Drug research will significantly benefit from efficient analysis of mechanisms on the cellular level, either for mechanisms of drugs or toxic action as well as reprogramming of cells. The cells respond to the disturbance by the drug or toxin; hence, the signals arise from both the primary mode of action and the secondary cellular stress response. Therefore, identification of the mechanistic functional networks which are relevant for drug action requires the deconvolution of primary and secondary signals. Basically, two approaches which might allow the identification of functional networks are known today: assessment of the time evolution of the cellular parameters under drug action or, alternatively, the analysis of cellular response on combinatorial stress factors. Both approaches require a high degree of experimental investment and suffer from limitations in the size of the networks which can be identified. Moreover, the deconvolution of primary and secondary response remains unsolved yet.

Although the challenges of complex diseases, namely, the extreme complexity of the respective systems, may require novel mathematical approaches and frameworks, mathematics provide highly beneficial solutions for specific applications even today.

An application-oriented, well-established modeling and simulation workflow must overcome the limitations of traditional statistical processes. The methods to be developed may integrate expert knowledge, public and published information with data from the research, and clinical research and development programs. Consistent data preparation and mathematical formalization ensure an impartial basis for decision-making.

## References

1. DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. *J Health Econ* 22:151–185. doi:[10.1016/S0167-6296\(02\)00126-1](https://doi.org/10.1016/S0167-6296(02)00126-1)
2. Paul SM, Mytelka DS, Dunwiddie CT et al (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov*. doi:[10.1038/nrd3078](https://doi.org/10.1038/nrd3078)
3. Wolkenhauer O, Auffray C, Brass O et al (2014) Enabling multiscale modeling in systems medicine. *Genome Med* 6:21. doi:[10.1186/gm538](https://doi.org/10.1186/gm538)
4. Kuepfer L (2010) Towards whole-body systems physiology. *Mol Syst Biol*. doi:[10.1038/msb.2010.70](https://doi.org/10.1038/msb.2010.70)
5. Lee D-S, Park J, Kay KA et al (2008) The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci* 105:9880–9885. doi:[10.1073/pnas.0802208105](https://doi.org/10.1073/pnas.0802208105)
6. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol*. doi:[10.1038/msb4100162](https://doi.org/10.1038/msb4100162)
7. Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK (2006) Physicochemical modelling of cell signalling pathways. *Nat Cell Biol* 8:1195–1203. doi:[10.1038/ncb1497](https://doi.org/10.1038/ncb1497)
8. Shlomi T, Cabili MN, Ruppin E (2009) Predicting metabolic biomarkers of human inborn errors of metabolism. *Mol Syst Biol*. doi:[10.1038/msb.2009.22](https://doi.org/10.1038/msb.2009.22)
9. Thiele I, Swainston N, Fleming RMT et al (2013) A community-driven global reconstruction of human metabolism. *Nat Biotechnol* 31:419–425. doi:[10.1038/nbt.2488](https://doi.org/10.1038/nbt.2488)
10. Agren R, Mardinoglu A, Asplund A et al (2014) Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol Syst Biol* 10:721. doi:[10.1002/msb.145122](https://doi.org/10.1002/msb.145122)
11. Hunter PJ, Borg TK (2003) Integration from proteins to organs: the Physiome Project. *Nat Rev Mol Cell Biol* 4:237–243. doi:[10.1038/nrml054](https://doi.org/10.1038/nrml054)
12. Noble D (2002) Modeling the heart – from genes to cells to the whole organ. *Science* 295:1678–1682
13. Hoehme S, Brulport M, Bauer A et al (2010) Prediction and validation of cell alignment along microvessels as order principle to restore tissue architecture in liver regeneration. *Proc Natl Acad Sci* 107:10371–10376. doi:[10.1073/pnas.0909374107](https://doi.org/10.1073/pnas.0909374107)
14. Schwen LO, Krauss M, Niederalt C et al (2014) Spatio-temporal simulation of first pass drug perfusion in the liver. *PLoS Comput Biol* 10:e1003499. doi:[10.1371/journal.pcbi.1003499](https://doi.org/10.1371/journal.pcbi.1003499)
15. Randall Thomas S (2009) Kidney modeling and systems physiology. *Wiley Interdiscip Rev Syst Biol Med* 1:172–190. doi:[10.1002/wsbm.14](https://doi.org/10.1002/wsbm.14)
16. Tawhai MH, Bates JHT (2011) Multi-scale lung modeling. *J Appl Physiol* 110:1466–1472. doi:[10.1152/jappphysiol.01289.2010](https://doi.org/10.1152/jappphysiol.01289.2010)
17. Meyer M, Schneckener S, Ludewig B et al (2012) Using expression data for quantification of active processes in physiologically based pharmacokinetic modeling. *Drug Metab Dispos* 40:892–901. doi:[10.1124/dmd.111.043174](https://doi.org/10.1124/dmd.111.043174)
18. Thiel C, Schneckener S, Krauss M et al (2015) A systematic evaluation of the use of physiologically based pharmacokinetic modeling for cross-species extrapolation. *J Pharm Sci* 104:191–206. doi:[10.1002/jps.24214](https://doi.org/10.1002/jps.24214)
19. Edginton AN, Willmann S (2008) Physiology-based simulations of a pathological condition: prediction of pharmacokinetics in patients with liver cirrhosis. *Clin Pharmacokinet* 47:743–752. doi:[10.2165/00003088-200847110-00005](https://doi.org/10.2165/00003088-200847110-00005)
20. Eissing T, Lippert J, Willmann S (2012) Pharmacogenomics of codeine, morphine, and morphine-6-glucuronide: model-based analysis of the influence of CYP2D6 activity, UGT2B7 activity, renal impairment, and CYP3A4 inhibition. *Mol Diagn Ther* 16:43–53. doi:[10.1007/BF03256429](https://doi.org/10.1007/BF03256429)



21. Willmann S, Edginton AN, Coboeken K et al (2009) Risk to the breast-fed neonate from codeine treatment to the mother: a quantitative mechanistic modeling study. *Clin Pharmacol Ther* 86:634–643. doi:[10.1038/clpt.2009.151](https://doi.org/10.1038/clpt.2009.151)
22. Lippert J, Brosch M, von Kampen O et al (2012) A mechanistic, model-based approach to safety assessment in clinical development. *CPT Pharmacomet Syst Pharmacol* 1:e13. doi:[10.1038/psp.2012.14](https://doi.org/10.1038/psp.2012.14)
23. Krauss M, Burghaus R, Lippert J et al (2013) Using Bayesian-PBPK modeling for assessment of inter-individual variability and subgroup stratification. *Silico Pharmacol* 1:6. doi:[10.1186/2193-9616-1-6](https://doi.org/10.1186/2193-9616-1-6)
24. Maharaj AR, Barrett JS, Edginton AN (2013) A workflow example of PBPK modeling to support pediatric research and development: case study with lorazepam. *AAPS J* 15:455–464. doi:[10.1208/s12248-013-9451-0](https://doi.org/10.1208/s12248-013-9451-0)
25. Maharaj AR, Edginton AN (2014) Physiologically based pharmacokinetic modeling and simulation in pediatric drug development. *CPT Pharmacomet Syst Pharmacol* 3:e148. doi:[10.1038/psp.2014.45](https://doi.org/10.1038/psp.2014.45)
26. Eissing T, Kuepfer L, Becker C et al (2011) A computational systems biology software platform for multiscale modeling and simulation: integrating whole-body physiology, disease biology, and molecular reaction networks. *Front Physiol* 2:4. doi:[10.3389/fphys.2011.00004](https://doi.org/10.3389/fphys.2011.00004)
27. Gille C, Bölling C, Hoppe A et al (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol Syst Biol*. doi:[10.1038/msb.2010.62](https://doi.org/10.1038/msb.2010.62)
28. Krauss M, Schaller S, Borchers S et al (2012) Integrating cellular metabolism into a multi-scale whole-body model. *PLoS Comput Biol* 8:e1002750. doi:[10.1371/journal.pcbi.1002750](https://doi.org/10.1371/journal.pcbi.1002750)
29. Schadt EE, Lamb J, Yang X et al (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 37:710–717. doi:[10.1038/ng1589](https://doi.org/10.1038/ng1589)
30. Couzin-Frankel J (2010) Major heart disease genes prove elusive. *Science* 328:1220–1221. doi:[10.1126/science.328.5983.1220](https://doi.org/10.1126/science.328.5983.1220)
31. Mitsos A, Melas IN, Siminelakis P et al (2009) Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS Comput Biol* 5:e1000591. doi:[10.1371/journal.pcbi.1000591](https://doi.org/10.1371/journal.pcbi.1000591)
32. Schuppert AA (2011) Efficient reengineering of meso-scale topologies for functional networks in biomedical applications. *J Math Ind* 1:6. doi:[10.1186/2190-5983-1-6](https://doi.org/10.1186/2190-5983-1-6)
33. Balabanov S, Wilhelm T, Venz S et al (2013) Combination of a proteomics approach and reengineering of meso scale network models for prediction of mode-of-action for tyrosine kinase inhibitors. *PLoS One* 8:e53668. doi:[10.1371/journal.pone.0053668](https://doi.org/10.1371/journal.pone.0053668)
34. Lukk M, Kapushesky M, Nikkilä J et al (2010) A global map of human gene expression. *Nat Biotechnol* 28:322–324. doi:[10.1038/nbt0410-322](https://doi.org/10.1038/nbt0410-322)
35. Schneckener S, Arden NS, Schuppert A (2011) Quantifying stability in gene list ranking across microarray derived clinical biomarkers. *BMC Med Genomics* 4:73. doi:[10.1186/1755-8794-4-73](https://doi.org/10.1186/1755-8794-4-73)
36. Müller F-J, Schuppert A (2011) Few inputs can reprogram biological networks. *Nature* 478:E4. doi:[10.1038/nature10543](https://doi.org/10.1038/nature10543)
37. Lenz M, Schuldt BM, Müller F-J, Schuppert A (2013) PhysioSpace: relating gene expression experiments from heterogeneous sources using shared physiological processes. *PLoS One* 8:e77627. doi:[10.1371/journal.pone.0077627](https://doi.org/10.1371/journal.pone.0077627)



# **Part II**

## **Opinions and Persepectives**

# Chapter 7

## Systems Medicine and Infection

Ruth Bowness

### Abstract

By using a systems-based approach, mathematical and computational techniques can be used to develop models that describe the important mechanisms involved in infectious diseases. An iterative approach to model development allows new discoveries to continually improve the model and ultimately increase the accuracy of predictions.

SIR models are used to describe epidemics, predicting the extent and spread of disease. Genome-wide genotyping and sequencing technologies can be used to identify the biological mechanisms behind diseases. These tools help to build strategies for disease prevention and treatment, an example being the recent outbreak of Ebola in West Africa where these techniques were deployed.

HIV is a complex disease where much is still to be learned about the virus and the best effective treatment. With basic mathematical modeling techniques, significant discoveries have been made over the last 20 years. With recent technological advances, the computational resources now available, and interdisciplinary cooperation, further breakthroughs are inevitable.

In TB, modeling has traditionally been empirical in nature, with clinical data providing the fuel for this top-down approach. Recently, projects have begun to use data derived from laboratory experiments and clinical trials to create mathematical models that describe the mechanisms responsible for the disease.

A systems medicine approach to infection modeling helps identify important biological questions that then direct future experiments, the results of which improve the model in an iterative cycle. This means that data from several model systems can be integrated and synthesized to explore complex biological systems.

**Key words** Infection, Mathematical, Modeling, Epidemic, Tuberculosis, HIV

---

## 1 Introduction

Infectious diseases continue to be major worldwide health concerns: hepatitis C, malaria, the human immunodeficiency virus (HIV), and tuberculosis (TB) are ongoing pandemics. A third of the world population is currently infected with the TB bacillus, and even though therapeutic drugs have slowed the threat of HIV, there is no definitive cure or viable vaccine in sight. A new way of approaching the problem is needed; this can be achieved with a

systems approach. By utilizing expertise across disciplines, fresh perspective and new insights can be gained.

Current modeling efforts span multiple levels in the disease system, population dynamics where the focus is on disease transmission, the progression from population level to individual level where heterogeneity is included, and moving to the pathogen-host interactions ranging from molecular to cellular to whole organism levels. Both biological experiments and mathematical modeling have been successful at elucidating the properties of a disease at a particular level, but a full understanding requires the integration of all scales. This remains a major challenge for systems medicine. Infectious diseases reflect an equilibrium between the host and the pathogen that is established and maintained by a broad network of interactions that occur on such scales. Maintenance and evolution of these interactions over a prolonged time frame adds further complexity to persistent infections. The understanding of a biological system requires the integration of data that are used to construct predictive models of the dynamic interactions between biological components of the complex pathogen-host system.

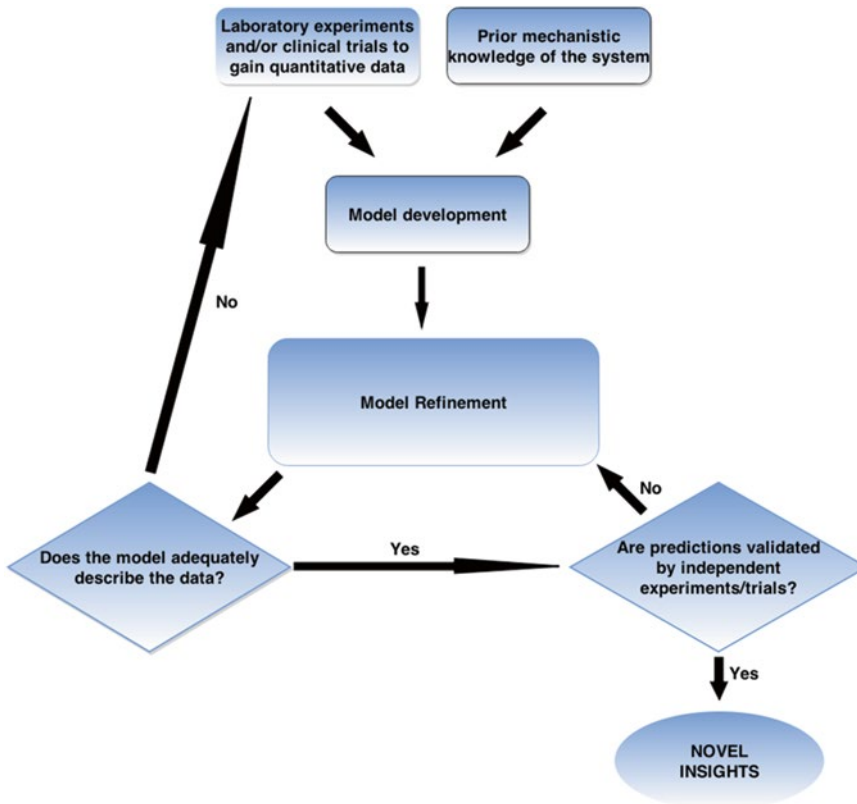
Mathematical and computational techniques, together with available *in vitro* and *in vivo* experimental results, can be used to generate realistic pathogen-host interaction models. Often these models involve iterative rounds of development, with testing and refinement using, for example, Bayesian methods. Models created in such a manner can be used to create testable hypotheses, with much learned about the biological system in the process (*see* Fig. 1). Ultimately, predictions can be made that can guide clinical practice. This iterative cycle is key to the systems approach and requires efficient integration of data with firm collaboration across disciplines. The technologies and computational resources now exist to realize this systems approach; we must take advantage of the potential these provide.

In this chapter, we review some basic epidemic modeling (Subheading 2) before using two infectious disease case studies, HIV (Subheading 3) and tuberculosis (Subheading 4), to demonstrate how a systems approach, using simple mathematical techniques, can be used to further biological understanding and ultimately lead to changes in clinical practice. The focus in these sections is at the cellular level, describing viral/bacterial load dynamics.

---

## 2 Epidemic Modeling

One motivating reason for modeling the spread of infectious diseases is to understand how future outbreaks can be prevented. This can be achieved in several different ways, such as isolation or imposed travel restrictions. These measures aim at reducing contact rates, *i.e.*, to reduce the reproduction rate of the pathogen.



**Fig. 1** This schematic describes dynamic model development. Prior knowledge and data from laboratory results help to develop an initial model. Once constructed, this model is refined. After this, an iterative cycle begins where the model is tested against the available data until the model is deemed to adequately describe this data. Once this cycle is complete, the last iterative cycle is embarked upon where predictions are made and ultimately validated against an independent data set

The effect of these depends on the particular disease and the community under consideration. Vaccination is an alternative preventive measure, which reduces the pool of susceptible individuals by imparting immunity.

In classical mathematical epidemic models [1], the total population number is assumed to be constant. A small group of infected people is introduced into a large population, and a model is used to describe the spread of infection within that population as a function of time.

The model consists of three subpopulations of individuals:

$S(t)$  which denotes the number of people susceptible to the disease

$I(t)$  which denotes the number of people who are infected and can transmit the disease

$R(t)$  which denotes the number of removed people, i.e., those who were infected but are now recovered, are immune, or have been isolated until they are recovered

Such models are known as SIR models, where transfer between populations is restricted to  $S \rightarrow I$  and  $I \rightarrow R$ .

Some diseases include a class in which the disease is latent, E. Such models are known as SEIR models.

Assumptions of the SIR model:

- The gain in the infective class is at a rate that is proportional to the number of infectious and susceptible people,  $rSI$ , where  $r > 0$  is a constant parameter.
- The susceptible population is lost at the same rate. The transfer of infected individuals to the removed class is proportional to the number of people who are infected,  $aI$ , where  $a > 0$  is a constant.  $1/a$  is a measure of the time spent in the infectious state.
- The incubation period is short enough to be negligible; hence, a susceptible person who catches the disease is infected immediately.
- Every person has equal probability of coming into contact with one another.

The model mechanism [2] is therefore

$$\begin{aligned} \frac{ds}{dt} &= -rSI, \\ \frac{dI}{dt} &= rSI - aI, \\ \frac{dR}{dt} &= aI, \end{aligned}$$

where  $r > 0$  is the infection rate and  $a > 0$  is the removal rate of infectious individuals.

Initial conditions are defined as  $S(0) = S_0 > 0$ ,  $I(0) = I_0 > 0$ , and  $R(0) = 0$ , and the constant total population is built into the system via  $\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$  which implies that  $S + I + R = N$ , where  $N$  is the total population.

With this model, questions can be answered such as whether an infection will spread or not. If it does, how does it develop with time and, critically, when will it start to decline? Given  $r$ ,  $a$ ,  $S_0$ , and  $I_0$ , these questions can be addressed.

This type of stochastic modeling looks at the properties of an epidemic by studying a given model and its parameters. It can be shown that the most important parameter in these SIR models is the basic reproduction rate of the disease,  $R_0 = \frac{rS_0}{a}$ , where it can be shown that when  $R_0 > 1$ , an epidemic ensues. Another important quantity when trying to avoid an outbreak is the critical vaccination coverage, which is defined as

$$v_c = 1 - \frac{1}{R_0} \text{ if } R_0 > 1; \text{ otherwise, } v_c = 0.$$

Even when trying to include as many realistic features in a model as possible, there is a limit to how close a model can get to reality, and models can never completely predict what will happen in a given situation. It is, for example, nearly impossible to predict how people will adapt and change behavior as a disease starts spreading. Having said this, models can still be very useful as guidance for health professionals when deciding about preventive measures aiming at reducing the spread of a disease. Much has been written describing mathematical models for infectious disease spread [3, 4], including stochastic epidemic models [5–7].

From traditional mathematical modeling, it is clear that understanding and incorporating information from multiple scales can dramatically increase the power of such approaches. Thus, for example, emerging genome-wide genotyping and sequencing technologies are used to identify the biological mechanisms underlying the development of complex diseases and traits among populations. This allows models to build in host and pathogen features but also points to the inclusion of larger, societal, or population data. Such approaches have started to be deployed, for example, in the recent epidemic outbreaks of Ebola in West Africa. Collectively, these approaches can help inform strategies for disease prevention and treatment.

---

### 3 HIV Modeling

In 2013, ~1.5 million people died from HIV-related causes globally, having claimed an overall 39 million lives so far (World Health Organization 2014). There is no cure for HIV infection, but antiretroviral treatment (ART) can be used to control the disease. There is much still to be learned about the virus and how best to administer effective treatment. Mathematical models developed over the last 20 years have helped to further understand and provide new insights into the disease.

One of the simplest mathematical models to describe HIV was proposed in 1995, where a simple linear first-order equation is used to describe viral load over time [8]:

$$\frac{dV}{dt} = P - cV,$$

where  $P$  represents the viral production rate and  $c$  is the viral clearance rate. Immune cells, fluid flow, and absorption into other cells are combined to give the overall clearance of viral peptides,  $c$ .

After introduction of the protease inhibitor, it is assumed that the drug is completely effective, so the drug will block all viral production after being introduced. Under this assumption,  $P = 0$  which leaves the simple equation

$$\frac{dV}{dt} = -cV \Rightarrow V(t) = V_0 e^{-ct},$$

where  $V_0$  is the mean viral concentration in the plasma before treatment.

Using linear regression to examine the relationship between  $\ln V$  against  $t$  gives an estimate for  $c$  and hence for the half-life of the virus in the plasma,  $t_{\frac{1}{2}} = \ln \frac{2}{c}$ .

If an assumption is then made that the levels of viral load measured in the plasma remain fairly constant before treatment begins, i.e., the patient is in a quasi-steady state,  $\frac{dV}{dt} = 0$ , then by

knowing  $c$  and the initial virus concentration  $V_0$ , the viral production rate before therapy can be computed, i.e.,  $P = cV_0$ . It should be noted, however, that as this calculation is based on the assumption that the drug completely blocks virus production, this assumption is unlikely to hold, and so, experiments would measure the rate of virus clearance in the face of some residual production, and the gradient of viral decline would therefore be a lower bound of the true clearance rate. Further models with added complexities have been developed from this simple framework (*see* [9] for a comprehensive review).

By using relatively simple modeling techniques and trivial mathematics, fundamental information has been gained about the underlying biological mechanisms of the disease. These developments have made a substantial impact on our thinking and understanding of HIV infection. For example, because the disease can take around 10 years to develop, many thought that the disease process would be slow and treatment could be delayed until symptoms appeared. Patients were therefore not monitored very aggressively. Modeling, coupled with appropriate experiments, has revealed that HIV is a dynamic disease encompassing various different time scales. An extremely simple model involving a single linear ordinary differential equation (described above), when applied to the interpretation of clinical data, gave the first quantitative estimate of how rapidly HIV was being produced and cleared [8, 10].

Rigorous analysis of previously conducted in vivo experiments also prompted the practice of using prolonged therapy with effective drug combinations. Calculations revealed that in an average HIV-infected person, around  $10^{10}$  viral particles are produced and released into bodily fluids per day [11]. It was therefore calculated that an infected person could go through about 200 replication cycles per year, with the possibility of mutating at each replication. With this new information, the rapid evolution of HIV could easily be understood. Therapy with a single drug, in which a few mutations

were all that were required for resistance to arise, was concluded to be a poor strategy. This demonstrated the need for combination therapy.

Adding complexity into the modeling by considering multiple cell populations, it was found that the virus concentration in plasma has a two-phase decline. By using data obtained from patients responding well to combination therapy, estimates of how long therapy would need to be given to clear the cells responsible for producing the observed levels of virus were calculated. This led to the practice of antiretroviral drugs being taken for at least 2 or 3 years after the virus is no longer detectable in the blood [9]. This modeling work also began the process of quantifying both the level and the role of latently and long-lived infected cell populations in HIV infection.

These important breakthroughs in the understanding of the HIV and its treatment have arisen from relatively trivial mathematical modeling exercises. This effectively demonstrates how systems-based interdisciplinary approaches can make huge advances in medical treatment.

---

## 4 Tuberculosis Modeling

Tuberculosis remains one of the leading causes of death by infectious disease, second only to HIV/AIDS. It is estimated that one third of the world population is latently infected by *Mycobacterium tuberculosis*, with 2012 seeing 8.6 million people falling ill and 1.3 million dying from the disease (World Health Organization 2014).

Although well-administered short-course chemotherapy is clinically effective [12], there are several concerns surrounding current TB treatment. The emergence of multidrug and extensive drug resistance is a major burden since it could lead to an increase of tuberculosis cases that are hard or impossible to treat [13]. Another major issue in tuberculosis treatment is its duration, which is currently a minimum of 6 months. Shortening the duration of effective TB therapy would mean better patient compliance and lower rates of relapse and drug resistance.

Traditional approaches in tuberculosis research are based on preclinical experiments, in vitro and in vivo. These systems each have limitations, and, often, desired experiments are not feasible due to laboratory or ethical issues. Mathematical modeling alongside these assays allows hypotheses to be developed and tested and hence further understanding of the disease by suggesting innovative approaches.

In the last 20 years, mathematical models have provided major insights in the knowledge of tuberculosis pathogenesis [14–16]. During this time, progress has also been made in the quantitative



description of both pharmacokinetics (PK) and in vitro pharmacodynamics (PD) of antituberculosis drugs [17–19]. Mathematical models that arise from a systems approach offer a unique potential to establish quantitative links across multiple biological scales. Different mathematical systems capture biological complexity best at individual scales. Once mathematical models adequately describe these biological complexities on an individual scale, integrating the scales will then be vital in order to understand the overall dynamics of this infectious disease. In this chapter, we focus on modeling at the cellular level, modeling bacterial load detected in clinical sputum samples.

It has long been noted in tuberculosis patients that decline in bacterial numbers appears to have two phases of decline. Many studies have employed nonlinear mixed effects modeling techniques to fit bi-exponential functions to clinical trial data [20, 21]. For example, in [21], a bi-exponential model of the form

$$\log_{10} \text{CFU} = \log_{10}(e^{0_1} e^{-\text{day}e^{0_2}} + e^{0_3} e^{-\text{day}e^{0_4}})$$

was used, where a  $\log_{10}$  transformation of the response and a variance function were used to account for heteroscedasticity and an exponentiated parameterization was used to enforce positivity of the parameters. These authors showed that the bi-exponential models provide a significantly better fit than a mono-exponential model. As it is known from in vitro studies that *Mycobacterium tuberculosis* bacteria exist in more than one cell state, the interpretation of the two-phase decline is that two subpopulations of *Mycobacterium tuberculosis* are present in the sputum, each declining at different rates. This has led to hypotheses about “dormant” cells being responsible for latent disease and relapse. Consequently huge effort has gone into researching this less metabolically active subpopulation of cells.

Traditional statistical modeling techniques have also been useful for identifying trends in clinical data sets. As tuberculosis is a slow-growing organism (with a generation time between 17.5 and 56 h [22, 23]), it can take time to obtain microbiological results to assess the progress of treatment. For this reason, modeling techniques have been used to identify biomarkers of success. Thus, the 8-week biomarker of culture conversion was created as the most used indicator of treatment outcome [24]. More recent modeling efforts have been employed to investigate the reliability of using baseline bacterial load as an indicator for later relapse [25].

Although statistical modeling techniques are very powerful, fitting to the clinical data empirically in order to provide future predictions, the differential equations describing the system are employed without reference to the mechanisms underlying the

biological system, and so little can be learned about the basic biology using these methods.

In contrast to statistical modeling, mathematical or mechanistic models can summarize current knowledge, and in their development, a greater understanding of the biological system can be gained. They can be used to highlight gaps in current knowledge and identify tractable biological questions. The mechanistic approach also allows us to predict how a system will shift when underlying processes change. If mechanistic models are correctly specified, they should provide better simulation and prediction properties than many current empirical models.

In TB, mechanistic models use available clinical and preclinical data to predefine parameter values before solving the set of differential equations describing the biological system. Although there is often uncertainty when assigning parameter values, common parameter estimation techniques such as profile likelihood can be used to analyze parameter sensitivity in the system. A huge advantage of mechanistic modeling is that we are able to analyze the effect that parameters have on the bacterial load. For example, by altering killing parameters, the effect of new regimens on time to culture conversion can be analyzed. Drugs targeting bacteria in different cell states can also be investigated. These simulations provide a surrogate for experiments that would not be feasible in vivo or in vitro.

This systems medicine approach means that during development of mechanistic models, important biological questions can be identified and therefore direct future experiments, the results of which will feed back into the model (*see* Fig. 1). Hence, data from several model systems (animal, human, bacterial, and computational) can be integrated and synthesized to explore the complex biological system and address relevant questions.

---

## 5 TB-HIV Coinfection

For tuberculosis patients coinfecting with the HIV, the disease becomes more complex to treat. Most TB-HIV models address the epidemiology of coinfection. In 1992, the first model was published to quantify the consequences of the emerging TB-HIV epidemic [26]. In 2003, it was found that antiretroviral therapy (ART) must be started early if expansion to the access of the therapy was to result in the increased control of TB [27]. Isoniazid preventative therapy (IPT) has also been modeled [28, 29], resulting in predictions of 34–100 % of reduction of risk of TB. Models have also been used to explore enhancements to DOTS-based programs [30] and also to inform policy on implementation of

new interventions and tools [31]. In a recent article, [32] suggest a modeling TB-HIV research agenda, based on expert discussions at a meeting convened by the TB Modelling and Analysis Consortium. Here they identified high-priority areas for future modeling efforts, the difficult diagnosis and high mortality of TB-HIV, the high risk of disease progression, TB health systems in high HIV prevalence settings, uncertainty in the natural progression of TB-HIV, and combined interventions for TB-HIV. The realization of these aims relies on a systems-based approach, where modelers' and key stakeholders' collaborative efforts result in tangible progress.

Models that address the pathology of TB-HIV-infected patients are also presented in the literature. Some studies attempt to understand how *Mycobacterium tuberculosis* affects the dynamic interaction of HIV-1 and the immune system [33, 34]. More modeling in this area is needed if we are to understand the complex interactions between these two pathogens and hence start to provide recommendations for improvement in treatment plans.

---

## 6 Conclusions

Infectious diseases such as HIV and tuberculosis are major global health concerns and much is still unknown about the diseases. In order to aid improvements in treatment, mathematical and computational models are needed to provide a new insight.

Interdisciplinary research is often criticized as researchers from other specialties are ignorant of many fundamental concepts in the discipline they are new to. However, this can also bring advantages. In model development, there is a danger that existing beliefs can inappropriately influence judgments about model assumptions and results. When mathematicians and computer scientists are used to develop models, however, they are not influenced by current dogmas in the field and can provide a fresh perspective on the problem.

Mathematics provides a precise quantitative language to describe the relation between variables and changes in states, and in medicine we can represent mathematically the clinical course of disease, the distribution of disease across populations and over time, and the mechanisms that generate disease. We have seen in this chapter that relatively simple mathematics has led to huge leaps forward in the understanding of both HIV and tuberculosis infection. With more sophisticated analysis and ever increasing computational power, the possibility of significant breakthroughs using this systems-based approach to medical research presents an exciting prospect.

## References

1. Murray JD (2002) *Mathematical biology*. Springer, New York
2. Kermack WO, McKendrick AG (1932) Contributions to the mathematical theory of epidemics. II. The Problem of endemicity. *Proc R Soc London A* 138:55–83. doi:10.1098/rspa.1932.0171
3. Anderson RM, May RM (1991) *Infectious diseases of humans: dynamics and control*. Oxford University Press, Oxford, p 757
4. Diekmann O, Heesterbeek J (2000) *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Wiley, Chichester, p 303
5. Britton T (2010) Stochastic epidemic models: a survey. *Math Biosci* 225:24–35. doi:10.1016/j.mbs.2010.01.006
6. Bailey NT (1987) *The mathematical theory of infectious diseases*. Macmillan Publishing Company, New York
7. Becker NG (1989) *Analysis of infectious disease data, monographs on statistics and applied probability*. Chapman & Hall, London, UK
8. Ho DD, Neumann AU, Perelson AS, Chen W (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* 373(6510):123–126
9. Perelson AS, Nelson PW (1999) Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Rev* 41:3–44. doi:10.1137/S0036144598335107
10. Wei X, Ghosh SK, Taylor ME et al (1995) Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* 373(6510):117–122
11. Perelson AS, Essunger P, Cao Y et al (1997) Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* 387:188–191. doi:10.1038/387188a0
12. Mitchison DA (2005) The diagnosis and therapy of tuberculosis during the past 100 years. *Am J Respir Crit Care Med* 171:699–706. doi:10.1164/rccm.200411-1603OE
13. Gandhi NR, Moll A, Sturm AW et al (2006) Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *Lancet* 368:1575–1580. doi:10.1016/S0140-6736(06)69573-1
14. Fang X, Wallqvist A, Reifman J (2009) A systems biology framework for modeling metabolic enzyme inhibition of *Mycobacterium tuberculosis*. *BMC Syst Biol* 3:92. doi:10.1186/1752-0509-3-92
15. Segovia-Juarez JL, Ganguli S, Kirschner D (2004) Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent-based model. *J Theor Biol* 231:357–376. doi:10.1016/j.jtbi.2004.06.031
16. Wigginton JE, Kirschner D (2001) A model to predict cell-mediated immune regulatory mechanisms during human infection with *Mycobacterium tuberculosis*. *J Immunol* 166:1951–1967. doi:10.4049/jimmunol.166.3.1951
17. Gumbo T (2010) New susceptibility breakpoints for first-line antituberculosis drugs based on antimicrobial pharmacokinetic/pharmacodynamic science and population pharmacokinetic variability. *Antimicrob Agents Chemother* 54:1484–1491. doi:10.1128/AAC.01474-09
18. Jayaram R, Gaonkar S, Kaur P et al (2003) Pharmacokinetics-pharmacodynamics of rifampin in an aerosol infection model of tuberculosis. *Antimicrob Agents Chemother* 47:2118–2124. doi:10.1128/AAC.47.7.2118-2124.2003
19. Jayaram R, Shandil RK, Gaonkar S et al (2004) Isoniazid pharmacokinetics-pharmacodynamics in an aerosol infection model of tuberculosis. *Antimicrob Agents Chemother* 48:2951–2957. doi:10.1128/AAC.48.8.2951-2957.2004
20. Rustomjee R, Lienhardt C, Kanyok T et al (2008) A Phase II study of the sterilising activities of ofloxacin, gatifloxacin and moxifloxacin in pulmonary tuberculosis. *Int J Tuberc Lung Dis* 12:128–138
21. Davies GR, Brindle R, Khoo SH, Aarons LJ (2006) Use of nonlinear mixed-effects analysis for improved precision of early pharmacodynamic measures in tuberculosis treatment. *Antimicrob Agents Chemother* 50:3154–3156. doi:10.1128/AAC.00774-05
22. O’Sullivan DM, McHugh TD, Gillespie SH (2010) Mapping the fitness of *Mycobacterium tuberculosis* strains: a complex picture. *J Med Microbiol* 59:1533–1535. doi:10.1099/jmm.0.019091-0
23. Shorten RJ, McGregor AC, Platt S et al (2013) When is an outbreak not an outbreak? Fit, divergent strains of *Mycobacterium tuberculosis* display independent evolution of drug resistance in a large London outbreak. *J Antimicrob Chemother* 68:543–549. doi:10.1093/jac/dks430
24. Aber VR, Nunn AJ (1978) Short term chemotherapy of tuberculosis. Factors affecting relapse following short term chemotherapy. *Bull Int Union Tuberc* 53(4):276–280
25. Perrin FM, Woodward N, Phillips PP, McHugh TD, Nunn AJ, Lipman MC, Gillespie SH (2010) Radiological cavitation, sputum mycobacterial load and treatment response in pulmonary tuberculosis. *Int J Tuberc Lung Dis* 14(12):1596–1602

26. Schulzer M, Radhamani MP, Grzybowski S et al (1994) A mathematical model for the prediction of the impact of HIV infection on tuberculosis. *Int J Epidemiol* 23:400–407
27. Williams BG, Dye C (2003) Antiretroviral drugs for tuberculosis control in the era of HIV/AIDS. *Science* 301:1535–1537. doi:[10.1126/science.1086845](https://doi.org/10.1126/science.1086845)
28. Wilton P, Smith RD, Coast J, Millar M, Karcher A (2001) Directly observed treatment for multi-drug-resistant tuberculosis: an economic evaluation in the United States of America and South Africa. *Int J Tuberc Lung Dis* 5(12):1137–1142
29. Murray M (2002) Determinants of cluster distribution in the molecular epidemiology of tuberculosis. *Proc Natl Acad Sci U S A* 99:1538–1543. doi:[10.1073/pnas.022618299](https://doi.org/10.1073/pnas.022618299)
30. Corbett EL, Watt CJ, Walker N et al (2003) The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* 163:1009–1021. doi:[10.1001/archinte.163.9.1009](https://doi.org/10.1001/archinte.163.9.1009)
31. Baltussen R, Floyd K, Dye C (2005) Cost effectiveness analysis of strategies for tuberculosis control in developing countries. *BMJ* 331:1364. doi:[10.1136/bmj.38645.660093.68](https://doi.org/10.1136/bmj.38645.660093.68)
32. Houben RMGJ, Dowdy DW, Vassall A et al (2014) How can mathematical models advance tuberculosis control in high HIV prevalence settings? *Int J Tuberc Lung Dis* 18:509–514. doi:[10.5588/ijtld.13.0773](https://doi.org/10.5588/ijtld.13.0773)
33. Kirschner D (1999) Dynamics of co-infection with M. Tuberculosis and HIV-1. *Theor Popul Biol* 55:94–109. doi:[10.1006/tpbi.1998.1382](https://doi.org/10.1006/tpbi.1998.1382)
34. Magomedze G, Garira W, Mwenje E (2006) Modelling the human immune response mechanisms to Mycobacterium tuberculosis infection in the lungs. *Math Biosci Eng* 3:661–682

## Systems Medicine for Lung Diseases: Phenotypes and Precision Medicine in Cancer, Infection, and Allergy

Bernd Schmeck, Wilhelm Bertrams, Xin Lai, and Julio Vera

### Abstract

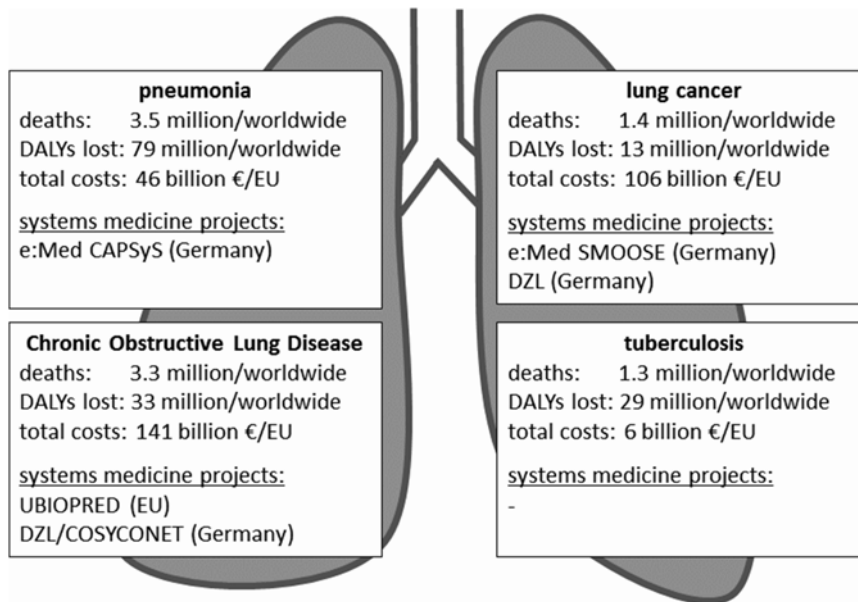
Lung diseases cause an enormous socioeconomic burden. Four of them are among the ten most important causes of deaths worldwide: Pneumonia has the highest death toll of all infectious diseases, lung cancer kills the most people of all malignant proliferative disorders, chronic obstructive pulmonary disease (COPD) ranks third in mortality among the chronic noncommunicable diseases, and tuberculosis is still one of the most important chronic infectious diseases. Despite all efforts, for example, by the World Health Organization and clinical and experimental researchers, these diseases are still highly prevalent and harmful. This is in part due to the specific organization of tissue homeostasis, architecture, and immunity of the lung. Recently, several consortia have formed and aim to bring together clinical and molecular data from big cohorts of patients with lung diseases with novel experimental setups, biostatistics, bioinformatics, and mathematical modeling. This “systems medicine” concept will help to match the different disease modalities with adequate therapeutic and possibly preventive strategies for individual patients in the sense of precision medicine.

**Key words** Asthma, Allergy, Infection, Pneumonia, Mycobacteria, Immunology, Transplantation, Model organisms, Gas exchange, Personalized medicine

---

### 1 Introduction

Among the ten most common causes of death worldwide, there are four pulmonary diseases, killing 9.5 million people per year (Fig. 1) [1]. In addition, pneumonia, COPD, tuberculosis, and lung cancer are also among the ten most common causes of disability-adjusted life years (DALYs) lost worldwide. Pneumonia alone is the single most important cause of DALYs lost, more important than HIV/AIDS, ischemic heart disease, cerebrovascular disease, or diarrhea [2]. The total cost of lung diseases in the European Union amounts to more than €380 billion annually. In addition to the abovementioned diseases, asthma alone causes costs of over €72 billion, consisting of €19.5 billion for direct medical costs, €14.4 billion for lost productivity (work absence, early



**Fig. 1** Frequent lung diseases and related systems medicine consortia. Depicted are the four lung diseases with the highest mortality worldwide, including death worldwide 2011, disability-adjusted life-years (DALYs) lost 2008, aggregated annual total (direct and indirect) costs and the value of DALYs lost for EU countries 2011, and related systems medicine projects according to internet and database research [1, 2]

retirement, etc.), and €38.3 billion for monetized value of DALYs. Lung cancer causes the greatest socioeconomical loss from disability and premature mortality. Because of late diagnosis and limited treatment options, most treatment costs are concentrated within the year of the diagnosis. Therefore, the total annual costs per case are calculated as €364,000 for lung cancer [2].

The lung and airways provide many unique features in terms of anatomy, physiology, and immunology. This paves the way for manifold pathologies and presents challenges for pulmonary clinicians and researchers. The lung is the body's largest organ, but contains about four liters of air and only about half a liter of tissue and the same amount of blood [3]. Therefore, the organ tissue has to be organized in a sophisticated and delicate way. The main function of the lung is the exchange of two gases, O<sub>2</sub> and CO<sub>2</sub>, between the air that we breathe and the bloodstream. Three main processes are involved in this: ventilation of air along the bronchial tree, passive diffusion of gases, and perfusion of blood through the alveolar capillaries [4]. These processes are tightly regulated, and not every part of the lung is equally ventilated and perfused. Every day, about 10,000 l of air—containing pollutants, allergens, pathogens, etc.—is ventilated through the about 23 generations of dichotomically dividing airways and over a lung surface of about 130 m<sup>2</sup> that is condensed mainly in over 300 million alveoli. The capillary surface has about the same size. Diffusion of oxygen critically depends on



a short diffusion distance over the air–blood barrier [3]. In healthy lungs, the mean thickness of this tissue barrier is about 0.6  $\mu\text{m}$ , consisting of very thin parts of alveolar epithelial type I cells and capillary endothelial cells sharing one single layer of basement membrane [5]. This delicate architecture is only possible due to the surfactant layer that lines the alveolar surface, reduces its surface tension, and is released by alveolar epithelial type II cells.

---

## 2 Clinical Challenges

### 2.1 *Acute Infection: Pneumonia*

Community-acquired pneumonia is a high incidence disease which results in more hospital admissions per year in some industrialized countries than myocardial infarction [6, 7]. It has the highest mortality rate worldwide of all infectious diseases. Its lethality ranges from 1 % in outpatient settings to 35 % in high-risk patients [8]. UNICEF (the United Nations Children’s Fund) data suggest that pneumonia kills more children under the age of 5 than malaria, AIDS, and measles together [9]. Notorious problems are emerging new pathogens that may combine a high mortality with an easy transmission, e.g., the SARS and MERS coronaviruses or certain pandemic influenza strains [10]. Severe pneumonia can lead to sepsis and septic shock requiring intensive care treatment with artificial organ support, causing extremely high costs [11]. Recently, it has been observed that pneumonia patients face an increased death risk for several months after the acute infection [8].

### 2.2 *Chronic Inflammation: Asthma*

Asthma is classically defined as an inflammatory chronic airway disease characterized by reversible airway obstruction and airway hyperresponsiveness [12]. This disease affects 200–300 million people worldwide, and its prevalence has been increasing over the last decades [13]. Up to 10 % of asthmatics are considered as severe cases. Typically, the inflammation in asthma is described to be allergic, eosinophilic, IgE dependent, and Th2 driven [14]. Therefore, therapy besides bronchodilation targets mainly this eosinophilic inflammation, either unspecific with topical or systemic glucocorticoids, with antileukotriene drugs, or with “biological therapies” specifically addressing IgE, or probably in the future IL-13, IL-5, and others [15]. However, sufficient control of asthma symptoms is impossible in many patients, in part due to more than 50 % of asthma patients that do not show a persistent eosinophilic inflammation [16], and this seriously challenges the classical pathophysiological concepts. Systems-based approaches can help to stratify patients to their phenotypes and respective available therapies and to identify new targets for the treatment of patients that are unresponsive to existing drugs [17].



### **2.3 Chronic Inflammation: COPD**

Chronic obstructive pulmonary disease (COPD) is the third most common noncommunicable disease with a very high prevalence worldwide. It causes significant disability, mortality, and health-care costs, e.g., for lifelong medication, lung transplantation, or mechanical ventilation. The most common causes for the development of COPD are long-term exposure to primary or secondary tobacco smoke, combustion of biomass, e.g., by open-fire cooking in developing countries, or genetic predisposition. Increasing evidence indicates that COPD is rather a syndrome than a solitary disease. The pathophysiological hallmarks are (1) chronic, cortisone-insensitive inflammation which causes mucus hypersecretion and fixed bronchoconstriction and remodeling as well as (2) irreversible tissue destruction in terms of bronchiectasis and emphysema [18]. They result in respiratory and ventilatory failure. An aggravating aspect is the vicious circle of impaired innate immunity, chronic bacterial colonization, and recurrent viral or bacterial infections called exacerbations and often resulting in hospitalization [19]. Besides substance avoidance and exercise training, COPD symptoms are typically treated by long-acting airway dilators on a lifelong basis. The underlying pathophysiology of chronic airway inflammation can be targeted by either topic or systemic glucocorticoids or phosphodiesterase inhibitors. However, these treatments are hampered by side effects and low effectivity. Increasing numbers of patients with severe ventilatory insufficiency receive lifelong mechanical ventilation, which is expensive and requires special infrastructure. On the other hand, the chances of COPD patients to receive lung transplantation are decreasing due to high risk of, e.g., cardiovascular or metabolic comorbidities.

### **2.4 Pulmonary Hypertension**

Pulmonary hypertension is a disease of the lung vasculature. Hallmarks are a deregulated proliferation of different vascular cell types and a progressive obliteration of vessels [20]. This often results in an increased pulmonary vascular resistance, increased right heart afterload, and *cor pulmonale*. Pulmonary hypertension is caused by a combination of genetic and environmental factors. It occurs in a variety of clinical situations and heterogeneous phenotypes. Several histological patterns of abnormalities have been described. Despite significant progress that has been made in understanding the pathogenesis and the development of new methods for delaying the progression of the disease, there is still no cure for it [21].

### **2.5 Cancer**

Lung cancer displays the highest neoplasia-related mortality in man. Over the last years, its mortality is also increasing in women. Cure rates and prognosis are generally poor due to late diagnosis. Therefore, a major focus lies on screening and early diagnosis [22]. Another important challenge is the correct molecular diagnosis for a targeted therapy. Up to now, lung cancer is subdivided in

so-called small cell lung cancer (SCLC, of neuroendocrine origin) and non-small cell lung cancer (NSCLC, e.g., squamous cell cancer, adenocarcinoma). Recently, certain mutations have been found to be predictive for the sensitivity to new targeted therapies [23]. Therefore, cancer might be the first entity of lung diseases matching the concept of precision medicine [24]. Lung cancer is mainly caused by exposure to primary or secondary tobacco smoke, biomass combustion, or naturally occurring radon [25].

---

### 3 Methodological Challenges: “The Mouse Trap”

The basic aim of systems medicine is to model human pathophysiology and disease to advance our understanding and to improve clinical diagnosis and treatment. Therefore, it is a logical and straightforward strategy to establish big and well-characterized patient cohorts to collect as many clinical and molecular data as possible—and suitable—as a solid base for this modeling process. However, certain processes or features cannot be observed or tested neither in healthy volunteers nor in patients. This includes the early origins of disease, e.g., environmental influence on asthma predisposition that may be epigenetically transferred from mother to child, the initial pathophysiological events in the alveolus during development of an influenza virus-induced pneumonia, and many more. In part, these problems can be resolved by the use of cell culture or tissue culture models [26]. However, some aspects can only be addressed in a living organism. In the past, for academic experimental research as well as for drug candidate studies for subsequent human trials, mice have been used to model human diseases because of the practical convenience (easy handling and low costs), the possibility to generate transgenic or gene knockout animals, and the availability of molecular and immunological tools [27]. However, certain caveats have to be kept in mind: Firstly, some of the lung diseases, e.g., asthma, are unique to humans and do not occur naturally in mice. Secondly, the anatomy and (patho-) physiology may differ significantly between mice and humans. Besides the obvious size difference—the human lung has a volume of 5 l, and the murine lung has a volume of 1 ml—there are also significant anatomical and physiological differences. For instance, the human airways have 23 generations, and the murine has only 13–16, the human airway division is dichotomic, and the murine airway division is monopod; mice do not have respiratory bronchioli, and the cellular composition and vascularization of the lung are different for both species [5, 27, 28]. Moreover, mice strains differ in their likelihood to react to stressors with certain pathophysiological events. In addition, there is increasing evidence that the immune systems of mice and men differ significantly. For trauma, burns, and endotoxemia, Seok and co-workers compared

gene response patterns between human subjects and corresponding mouse models: They found that these stressors resulted in highly similar patterns in humans, whereas the responses in corresponding mouse models showed only poor correlations with human conditions and also among each other [29]. Accordingly, there are several examples where the reliance on animal models misguided the pathophysiological understanding [28], led to unsuccessful clinical trials [29], or even resulted in disastrous outcomes of clinical studies [30]. This may be in part due to differences in the molecular repertoire of immune cells in mice and humans: For example, the inherent versatility of macrophages harbors the potential for a plethora of different activation subtypes [31]. Inappropriate polarization can be detrimental to the host, as macrophages can potentiate an inapt immune response and thus aggravate a pathological condition: In rodents, alveolar macrophages have been found to play an important role for the development of airway hyperresponsiveness in allergic animals [32, 33]. In mice, solid markers for macrophage polarization are established, as is exemplified by the well-described induction of NOS2 in M1 (IFN $\gamma$ ) and of Arg1 in M2 (IL-4) macrophages. Corresponding functional markers in the human system are yet to be found, and there are notable differences between human and murine macrophage activation patterns on the transcriptional level [34]. Accordingly, a comparative study of M2 (IL-4) polarization in mouse and man shows very limited interspecies consistency, as only transglutaminase 2 (TGM2) was found to be a functional marker shared by both [35].

---

## 4 Case Studies

### 4.1 *Modeling Lung Infection*

The most frequent cause of community-acquired pneumonia is pneumococcal infection. Smith and co-workers established a mathematical model to predict the outcome of pneumococcal pneumonia with the two possible states (1) bacterial clearance, or (2) sustained bacterial growth [36]. The model is based on data of pulmonary bacterial replication from a mouse model of pneumococcal pneumonia. Using ordinary differential equations (ODEs), it describes three lines of barrier defense: First, the initial alveolar macrophage response mounts a fast but weak defense and is described by only one equation for the bacterial population. Second, the early recruitment of neutrophils consist of cytokine release by different populations of alveolar macrophages and epithelial cells, the influx of neutrophils, neutrophil apoptosis, and debris removal by alveolar macrophages. Third, a subsequent recruitment of monocyte-derived macrophages has been included that contributes to bacterial killing. This model provides some interesting insights and sufficient accuracy for certain questions,

although it is of moderate complexity and based on limited experimental data. Some more data from previously published mouse studies have been included in another effort to model pneumococcal pneumonia: Bacterial numbers in the lung and the blood and also neutrophil levels from infection experiments with four differently susceptible mouse strains [37] have been used to calibrate this ODE model. It consists of four equations describing the time evolution of the number of pathogens in the lung and in the blood, damaged lung epithelial cells, and total activated phagocytes. The model has been validated on other published data sets, and its predictions are consistent with most experimental observations. However, no study so far has modeled pneumonia dynamics in patients. Therefore, a new consortium (“Medical Systems Biology of Pulmonary Barrier Failure in Community Acquired Pneumonia; e:Med CAPSYS” [38]) aims at multiscale modeling of pulmonary barrier failure in bacterial pneumonia based on comprehensive physiological, proteomic, and transcriptomic data sets from clinical cohorts, complex mouse models, and human cell culture models (Fig. 1). It will include three cohorts: 10,000 well-phenotyped pneumonia patients of the CAPNETZ study (clinical, biochemical, and genetic data) [39], more than 1000 patients with uncomplicated or severe pneumonia or pneumogenic sepsis (several visits, data from genotyping, expression profiling, and proteomics), and a newly recruited deep-phenotyping cohort of about 100 patients that will undergo, e.g., bronchoscopy for microbiome and exosome analysis.

Several mathematical models have been developed to study and understand host immune response mechanisms in pulmonary *Mycobacterium tuberculosis* (Mtb) infection. Marino and Kirschner [40] used a two compartment model to investigate the human immune response to Mtb in the lung. By performing bifurcation analysis of the model, the authors identified key processes of cellular activation and priming that occur between the lung and the nearest draining lymph node that have the potential to determine different outcomes of the Mtb infection. To identify control mechanisms of granuloma formation during Mtb infection in the lung, Segovia-Juarez et al. [41] built a complex agent-based model which accounts for interactions between Mtb, immune effectors such as chemokines and cytokines, and immune cells like macrophages, CD4<sup>+</sup>, and CD8<sup>+</sup> T cells. With the help of the model, the authors identified several issues that are crucial for granuloma formation during the course of Mtb infection, including efficiency of chemokine diffusion, prevention of macrophage overcrowding within the granuloma, arrival time, location, and number of T cells within the granuloma, as well as overall host ability to activate macrophages. To investigate the contribution of CD8<sup>+</sup> T cells to control Mtb infection, Sud et al. [42] built an ODE model of the immune response to Mtb in the lung. Using the model, the authors

examined the importance of CD8<sup>+</sup> T cells in the control of the infection and determined putative minimum T cell levels providing effective protection following vaccination. A model of differential equations was also developed to investigate the different roles played by alternatively activated macrophages (AAM) versus classically activated macrophages (CAM) in the early stages of Mtb infection in the lung. The model described the interactions among cells, bacteria, and cytokines involved in the activation of AAM and CAM and was a useful tool to analyze strategies for reducing the switching time (i.e., when CAM become more dominant than AAM), which ensures an adequate immune response to the pathogen [43]. Similarly, Kirschner's group built two ODE models to investigate the function of macrophage (CAM)-activating cytokines (i.e., TNF $\alpha$  and IFN $\gamma$ ) in Mtb infection. One model was used to test the ability of macrophages to kill Mtb under different scenarios, in which the macrophage activation is characterized by the timing of IFN $\gamma$  and TNF $\alpha$  signaling relative to the infection [44]. The model simulations unraveled a preferred host strategy for mycobacterial control that is implemented via the direct entry of macrophages into a granuloma site from lung vascular sources. The other model was used to predict the contribution of multiple TNF $\alpha$  activities to the control of Mtb infection within the granuloma, with the assumption that macrophage activation is a key effector mechanism for controlling bacterial growth in the lung. The simulation results suggested that bacterial numbers are a strong contributing factor to granuloma structure with TNF, and TNF-dependent apoptosis can reduce inflammation at the cost of impairing mycobacterial clearance [45].

One of the fundamental challenges in the control of pulmonary Mtb infection is to understand molecular mechanisms involved in the onset of latency and/or reactivation of Mtb after the initial infection. Magombedze and Mulder [46] built a mathematical model to simulate all possible Mtb latency occurrence scenarios in the lung based on the profile of differentially expressed genes. Their ODE model was used to simulate observed gene expression changes in *in vitro* latency models which allow for illustrating all possible latency/dormancy occurrence scenarios and latency reactivation. In a subsequent study, the same author used a systems biology approach combining both bioinformatics and mathematical modeling to identify potential drug target genes in the Mtb latency program. Boolean modeling of the data-driven regulatory network related to mycobacterial latency in the lung revealed a bistable switch between latent and actively replicating phases of Mtb [46].

## **4.2 Modeling Asthma**

Multiscale models of the lung have been developed and applied to gain a better understanding of asthma in several aspects. These models incorporate and couple multiple spatial scales (molecules, cells, tissues, and the lung) underlying airway hyperresponsiveness

to simulate the complex physiological response to, e.g., allergens in asthma. Venegas et al. carried out a study that probed the scale of ventilation heterogeneity in asthmatic subjects using positron emission tomography imaging and that modeled complex interdependent behavior in the lung [47]. The authors found that ventilation is not uniform within bronchoconstricted regions, and within the ventilation defects themselves, there is considerable ventilation heterogeneity. Brook et al. developed an axisymmetric two-layer model of an airway wall to represent both lung slices and an intact airway in vivo, which resolves connective tissue and muscle cell properties within a composite muscle layer [48]. The model predicted that different types of airway remodeling in asthma lead to significantly different contractile responses and stress environments. For better understanding of airway hyperresponsiveness in asthmatic airways, a multiscale model of partial and ordinary differential equations was developed, which linked regulatory processes occurring at molecular and cellular level ( $\text{Ca}^{2+}$  and crossbridge dynamics) with physiological phenomena occurring at the organ level (lung deformation) [49, 50]. Chernyavsky et al. developed a mathematical model that qualitatively describes the growth dynamics of airway smooth muscle cells (ASM) over short and long terms in the normal and inflammatory environments typically observed in asthma [51]. This model allowed possible ASM accumulation scenarios to be explored and suggested possible new targets for diagnosis and prevention of ASM remodeling in asthma.

Recently, new consortia have started to apply systems biology strategies to asthma in a clinical context: Within the Innovative Medicines Initiative, a European project aims at a personalized management approach for patients with severe asthma (Unbiased Biomarkers for the Prediction of Respiratory Disease Outcome Consortium; U-BIOPRED) [52]. It involves scientists from universities, research institutes, the pharmaceutical industry, and small companies and plans to define phenotypes with respect to therapeutic efficacy by integrating -omics data from invasively and noninvasively obtained patient material and modeling of the underlying pathologies [53]. Another project aims to develop validated models that predict disease progression and response to treatment in asthma and COPD by integrating expertise in physiology, radiology, image analysis, bioengineering, data harmonization, security and ethics, computational modeling, and systems biology (Airway Disease Predicting Outcomes through Patient Specific Computational Modeling Consortium; AirPROM) [54]. This project is funded by the European Union 7th Framework Programme. Recently, the German Center for Lung Research (DZL) has established a systems biology platform [55] to integrate patient data from its cohorts for childhood wheezing and severe asthma in adults with experimental models by means of multiscale modeling.

### **4.3 Modeling Other Lung Diseases**

To investigate airway disease resulting from inflammation and fibrosis following particulate exposure, Brown et al. used an agent-based model, which focuses on a limited number of relevant interactions, specifically those among macrophages, fibroblasts, pro-inflammatory ( $\text{TNF}\alpha$ ) and anti-inflammatory cytokines ( $\text{TGF}\beta 1$ ), collagen deposition, and tissue damage [56]. The model predicted three distinct states of inflammation whose developments depend primarily on the degree and duration of particulate exposure. The predictions were consistent with *in vivo* experimental observations obtained after exposing mice lung tissue to particulate matter.

A mathematical model composed of partial differential equations that describe the interactions among immune cells and cytokines related to sarcoidosis in the lung was built by Hao et al. [57]. The model was calibrated and validated using clinical data on cytokine levels in healthy and diseased lung tissues and further used to explore the effect of potential treatments (such as anti- $\text{TNF}\alpha$ , anti-IL-12, anti- $\text{IFN}\gamma$ , and  $\text{TGF}\beta$  enhancement) that may reduce the disease activity through decreasing the size of sarcoid granulomas. Taken together, the constructed model is a step toward a more comprehensive study of sarcoidosis and its treatment.

Many end-stage respiratory diseases require lung transplantation as a last resort. However, with 27 % overall 10-year survival, this procedure shows the poorest long-term survival of all solid organ transplantations [58]. The main reason for this is the development of chronic lung allograft dysfunction (CLAD) by over 50 % of all lung transplant recipients within 5 years. Therefore, 14 lung transplantation centers teamed up to build a computational model to estimate the personal recipient risk to develop CLAD within 3 years after the transplantation (systems biology of CLAD, SysCLAD). They will analyze clinical, environmental, and immunological data, the microbiome and different -omics data both from donors and recipients [59].

One question that is still unsolved in lung cancer is how circulating tumor cells can develop at the primary site and traverse the circulatory systems. Having in mind the difficulties to generate suitable *in vivo* data to elucidate this question, mathematical modeling under the systems biology paradigm seems to be a good methodological option. In line with this, Kuhn's group used a Markov chain Monte Carlo model that describes cancer progression to identify and quantify the multidirectional pathways and timescales associated with metastatic spread from primary lung cancer [60, 61]. In contrast to the traditional view of cancer metastasis as a unidirectional process starting at the primary site and spreading to distant sites as time progresses, the authors quantified three types of multidirectional mechanisms of cancer progression based on large autopsy data sets: (1) self-seeding of the primary tumor, (2) reseeding of the primary tumor from a metastatic site,



and (3) reseeding of metastatic tumors [61]. By simulating the model, the authors showed that for lung cancer, the main spreaders (i.e., the distant site that has higher probability of transmitting than keeping circulating tumor cells from the primary site) are the adrenal gland and kidney, whereas the main sponges (i.e., the distant site that has lower probability of transmitting than keeping circulating tumor cells from the primary site) are regional lymph nodes, the liver, and bone.

#### **4.4 Modeling Gas Exchange within the Lung and the Dynamics of Inhaled Pharmaceuticals for Lung Diseases**

Mathematical models have been utilized not only for a system-level understanding of the whole respiratory system but also for a detailed understanding of several functions that contribute to gas exchange within the lung (reviewed by Ben-Tal and Tawhai [62]). Particularly, two studies have been carried out to investigate the effect of ventilation/perfusion mismatch on hepatopulmonary syndrome and lung inflammation, respectively. Chakraborty et al. developed a differential equation model of pulmonary oxygen uptake by considering three disparate scales, namely, micro (red blood cell), meso (capillary and alveolus), and macro (lung) [63]. The authors used the model to quantify the oxygen uptake abnormalities in patients with hepatopulmonary syndrome as a result of functional intrapulmonary right to left shunting of pulmonary blood flow, as well as spatial heterogeneity of ventilation/perfusion mismatch in the lung. Furthermore, the quantified pulmonary gas exchange abnormalities in the patients were used for stratifying them into two categories—those who are oxygen responsive and those who are oxygen nonresponsive with intractable hypoxemia. Reynolds et al. [64] developed a multi-compartment model of ODEs for gas exchange with focus on inflammation in acute lung injury. Using the model, the authors explored effects of inflammation on ventilation/perfusion distribution and the resulting pulmonary venous partial pressure oxygen level during systemic inflammatory stresses.

In the last years, a number of compartment-based pharmacokinetic (PK) models accounting for the kinetics of inhaled pharmaceuticals have been published. For instance, Sturm [65] developed a stochastic model describing mucociliary clearance in cystic fibrosis patients and its development with progressing course of the disease. The model showed that patients with cystic fibrosis have a higher risk of inhaled particle accumulation and related particle overload in specific lung compartments than healthy subjects. Markovetz et al. built a more complex model to describe the mucociliary clearance and absorption of aerosolized radiolabeled particles and small molecular probes from human subjects with and without cystic fibrosis [66]. This model captured the mucociliary clearance and liquid dynamics of the hyperabsorptive state in cystic fibrosis airways and the mitigation of that state by hypertonic saline treatment.



---

## 5 Perspectives

Despite all clinical and scientific efforts so far, lung diseases cause an enormous suffering and death toll from patients and socioeconomical costs for our societies and health-care systems, especially in the fields of infection, malignancies, chronic noncommunicable diseases, and allergy. This may be in part due to complicated, even prenatal, timelines, and heterogeneous clinical phenotypes. On the other hand, new scientific insights, e.g., in the role of the microbiome and noncoding RNA, and new technological developments, e.g., the new sequencing technologies, may help us to improve our clinical performance in future. But these tremendous amounts of clinical and molecular data require a new way of organizational, technological, and intellectual cooperation between many clinical, experimental, and theoretical disciplines, called “systems medicine.” Many questions remain to be answered: What type and amount of clinical data do we need and can we afford to collect and analyze? What will be the role of animal models, human tissue models, complex cell culture models, or even artificial organs on a chip? How can we bring all these complex and multilevel data together by means of mathematics and computer science? Improving clinical practice in respiratory medicine will require enthusiasm and hard work from all participating physicians and scientists, as well as sustained support by our governments, funding agencies, and all stakeholders of our health-care systems.

---

## Acknowledgments

We thank many collaborators for fruitful discussion, especially Annalisa Marsico, Brian Caffrey, and Martin Vingron. Part of this work has been funded by BMBF (e:bio miRSys, FKZ 0316175B, and e:Med CAPSYS, FKZ 01X1304E) to B.S. and J.V., and DFG (SFB/TR-84) and HMWK (LOEWE Medical RNomics—FKZ 519/03/00.001-(0003)) to B.S. We would like to apologize to all colleagues whose excellent contributions to the field could not be included in this text due to space constraints.

## References

1. WHO (2011) World Health Statistics 2011. [http://www.who.int/gho/publications/world\\_health\\_statistics/EN\\_WHS2011\\_Full.pdf?ua=1](http://www.who.int/gho/publications/world_health_statistics/EN_WHS2011_Full.pdf?ua=1)
2. Gibson GJ, Loddenkemper R, Sibille Y, Lundbäck B (2013) The European lung white book. European Respiratory Society, Sheffield
3. Murray JF (2010) The structure and function of the lung. *Int J Tubercul Lung Dis* 14:391–396
4. Wagner PD (2015) The physiological basis of pulmonary gas exchange: implications for clinical interpretation of arterial blood gases. *Eur Respir J* 45:227–243

5. Weibel ER (2013) It takes more than cells to make a good lung. *Am J Respir Crit Care Med* 187:342–346
6. Angus DC, Marrie TJ, Obrosky DS et al (2002) Severe community-acquired pneumonia: use of intensive care services and evaluation of American and British Thoracic Society Diagnostic criteria. *Am J Respir Crit Care Med* 166:717–723
7. Ewig S, Birkner N, Strauss R et al (2009) New perspectives on community-acquired pneumonia in 388 406 patients. Results from a nationwide mandatory performance measurement programme in healthcare quality. *Thorax* 64:1062–1069
8. Restrepo MI, Faverio P, Anzueto A (2013) Long-term prognosis in community-acquired pneumonia. *Curr Opin Infect Dis* 26:151–158
9. UNICEF/WHO (2006) Pneumonia, the forgotten killer of children. [http://whqlibdoc.who.int/publications/2006/9280640489\\_eng.pdf](http://whqlibdoc.who.int/publications/2006/9280640489_eng.pdf)
10. Horby PW, Pfeiffer D, Oshitani H (2013) Prospects for emerging infections in East and Southeast Asia 10 years after severe acute respiratory syndrome. *Emerg Infect Dis* 19:853–860
11. Angus DC, Van Der Poll T (2013) Severe sepsis and septic shock. *N Engl J Med* 369:840–851
12. Busse WW, Lemanske RF Jr (2001) Asthma. *N Engl J Med* 344:350–362
13. Eder W, Ege MJ, Von Mutius E (2006) The asthma epidemic. *N Engl J Med* 355:2226–2235
14. Accordini S, Corsico AG, Braggion M et al (2013) The cost of persistent asthma in Europe: an international population-based study in adults. *Int Arch Allergy Immunol* 160:93–101
15. Haldar P, Pavord ID, Shaw DE et al (2008) Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 178:218–224
16. Barnes PJ (2008) Immunology of asthma and chronic obstructive pulmonary disease. *Nat Rev Immunol* 8:183–192
17. Gonem S, Desai D, Siddiqui S et al (2011) Evidence for phenotype-driven treatment in asthmatic patients. *Curr Opin Allergy Clin Immunol* 11:381–385
18. McDonough JE, Yuan R, Suzuki M et al (2011) Small-airway obstruction and emphysema in chronic obstructive pulmonary disease. *N Engl J Med* 365:1567–1575
19. Sethi S, Murphy TF (2008) Infection in the pathogenesis and course of chronic obstructive pulmonary disease. *N Engl J Med* 359:2355–2365
20. Huang JB, Liang J, Zhao XF et al (2013) Epigenetics: novel mechanism of pulmonary hypertension. *Lung* 191:601–610
21. Colvin KL, Yeager ME (2015) Proteomics of pulmonary hypertension: could personalized profiles lead to personalized medicine? *Proteomics Clin Appl* 9:111–120
22. Prosch H, Schaefer-Prokop C (2014) Screening for lung cancer. *Curr Opin Oncol* 26:131–137
23. Reck M, Heigener DF, Mok T et al (2013) Management of non-small-cell lung cancer: recent developments. *Lancet* 382:709–719
24. Collins FS, Varmus H (2015) A new initiative on precision medicine. *N Engl J Med* 372:793–795
25. Kurmi OP, Arya PH, Lam KB et al (2012) Lung cancer risk and solid fuel smoke exposure: a systematic review and meta-analysis. *Eur Respir J* 40:1228–1237
26. Hocke AC, Becher A, Knepper J et al (2013) Emerging human Middle East respiratory syndrome coronavirus causes widespread infection and alveolar damage in human lungs. *Am J Respir Crit Care Med* 188:882–886
27. Holmes AM, Solari R, Holgate ST (2011) Animal models of asthma: value, limitations and opportunities for alternative approaches. *Drug Discov Today* 16:659–670
28. Mullane K, Williams M (2013) Alzheimer's therapeutics: continued clinical failures question the validity of the amyloid hypothesis-but what lies beyond? *Biochem Pharmacol* 85:289–305
29. Seok J, Warren HS, Cuenca AG et al (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A* 110:3507–3512
30. Horvath C, Andrews L, Baumann A et al (2012) Storm forecasting: additional lessons from the CD28 superagonist TGN1412 trial. *Nat Rev Immunol* 12:740
31. Peters-Golden M (2004) The alveolar macrophage: the forgotten cell in asthma. *Am J Respir Cell Mol Biol* 31:3–7
32. Zaslona Z, Przybranowski S, Wilke C et al (2014) Resident alveolar macrophages suppress, whereas recruited monocytes promote, allergic lung inflammation in murine models of asthma. *J Immunol* 193:4245–4253
33. Pouliot P, Spahr A, Careau E et al (2008) Alveolar macrophages from allergic lungs are

- not committed to a pro-allergic response and can reduce airway hyperresponsiveness following *ex vivo* culture. *Clin Exp Allergy* 38:529–538
34. Martinez FO, Gordon S (2014) The M1 and M2 paradigm of macrophage activation: time for reassessment. *F1000prime reports* 6:13
  35. Martinez FO, Helming L, Milde R et al (2013) Genetic programs expressed in resting and IL-4 alternatively activated mouse and human macrophages: similarities and differences. *Blood* 121:e57–69
  36. Smith AM, McCullers JA, Adler FR (2011) Mathematical model of a three-stage innate immune response to a pneumococcal lung infection. *J Theor Biol* 276:106–116
  37. Mochan E, Swigon D, Ermentrout GB et al (2014) A mathematical model of intrahost pneumococcal pneumonia infection dynamics in murine strains. *J Theor Biol* 353:44–54
  38. CAPSyS. Systems Medicine of Community Acquired Pneumonia. <http://www.capsys.imise.uni-leipzig.de/>
  39. Klapdor B, Ewig S, Pletz MW et al (2012) Community-acquired pneumonia in younger patients is an entity on its own. *Eur Respir J* 39:1156–1161
  40. Marino S, Kirschner DE (2004) The human immune response to *Mycobacterium tuberculosis* in lung and lymph node. *J Theor Biol* 227:463–486
  41. Segovia-Juarez JL, Ganguli S, Kirschner D (2004) Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent-based model. *J Theor Biol* 231:357–376
  42. Sud D, Bigbee C, Flynn JL et al (2006) Contribution of CD8+ T cells to control of *Mycobacterium tuberculosis* infection. *J Immunol* 176:4296–4314
  43. Day J, Friedman A, Schlesinger LS (2009) Modeling the immune rheostat of macrophages in the lung in response to infection. *Proc Natl Acad Sci U S A* 106:11246–11251
  44. Ray JC, Wang J, Chan J et al (2008) The timing of TNF and IFN- $\gamma$  signaling affects macrophage activation strategies during *Mycobacterium tuberculosis* infection. *J Theor Biol* 252:24–38
  45. Ray JC, Flynn JL, Kirschner DE (2009) Synergy between individual TNF-dependent functions determines granuloma performance for controlling *Mycobacterium tuberculosis* infection. *J Immunol* 182:3706–3717
  46. Magombedze G, Mulder N (2013) Understanding TB latency using computational and dynamic modelling procedures. *Infect Genet Evol* 13:267–283
  47. Venegas JG, Winkler T, Musch G et al (2005) Self-organized patchiness in asthma as a prelude to catastrophic shifts. *Nature* 434:777–782
  48. Brook BS, Peel SE, Hall IP et al (2010) A biomechanical model of agonist-initiated contraction in the asthmatic airway. *Respir Physiol Neurobiol* 170:44–58
  49. Politi AZ, Donovan GM, Tawhai MH et al (2010) A multiscale, spatially distributed model of asthmatic airway hyperresponsiveness. *J Theor Biol* 266:614–624
  50. Lauzon AM, Bates JH, Donovan G et al (2012) A multi-scale approach to airway hyperresponsiveness: from molecule to organ. *Front Physiol* 3:191
  51. Chernyavsky IL, Croisier H, Chapman LC et al (2014) The role of inflammation resolution speed in airway smooth muscle mass accumulation in asthma: insight from a theoretical model. *PLoS One* 9:e90162
  52. Bel EH, Sousa A, Fleming L et al (2011) Diagnosis and definition of severe refractory asthma: an international consensus statement from the Innovative Medicine Initiative (IMI). *Thorax* 66:910–917
  53. Auffray C, Adcock IM, Chung KF et al (2010) An integrative systems biology approach to understanding pulmonary diseases. *Chest* 137:1410–1416
  54. AirPROM. <http://www.europeanlung.org/en/projects-and-research/projects/airprom/home>
  55. iLung—Institute for Lung Research. [www.i-lung.de](http://www.i-lung.de)
  56. Brown BN, Price IM, Toapanta FR et al (2011) An agent-based model of inflammation and fibrosis following particulate exposure in the lung. *Math Biosci* 231:186–196
  57. Hao WR, Crouser ED, Friedman A (2014) Mathematical model of sarcoidosis. *Proc Natl Acad Sci U S A* 111:16065–16070
  58. Wolfe RA, Roys EC, Merion RM (2010) Trends in organ donation and transplantation in the United States, 1999–2008. *Am J Transplant* 10:961–972
  59. Pison C, Magnan A, Botturi K et al (2014) Prediction of chronic lung allograft dysfunction: a systems medicine challenge. *Eur Respir J* 43:689–693
  60. Newton PK, Mason J, Bethel K et al (2012) A stochastic Markov chain model to describe lung cancer growth and metastasis. *PLoS One* 7:e34637

61. Newton PK, Mason J, Bethel K et al (2013) Spreaders and sponges define metastasis in lung cancer: a Markov chain Monte Carlo mathematical model. *Cancer Res* 73:2760–2769
62. Ben-Tal A, Tawhai MH (2013) Integrative approaches for modeling regulation and function of the respiratory system. *Wiley Interdiscip Rev Syst Biol Med* 5:687–699
63. Chakraborty S, Balakotaiah V, Bidani A (2007) Multiscale model for pulmonary oxygen uptake and its application to quantify hypoxemia in hepatopulmonary syndrome. *J Theor Biol* 244:190–207
64. Reynolds A, Bard Ermentrout G, Clermont G (2010) A mathematical model of pulmonary gas exchange under inflammatory stress. *J Theor Biol* 264:161–173
65. Sturm R (2012) An advanced stochastic model for mucociliary particle clearance in cystic fibrosis lungs. *J Thorac Dis* 4:48–57
66. Markovetz MR, Corcoran TE, Locke LW et al (2014) A physiologically-motivated compartment-based model of the effect of inhaled hypertonic saline on mucociliary clearance and liquid transport in cystic fibrosis. *PLoS One* 9:e111972

## **Third-Kind Encounters in Biomedicine: Immunology Meets Mathematics and Informatics to Become Quantitative and Predictive**

**Martin Eberhardt\*, Xin Lai\*, Namrata Tomar\*, Shailendra Gupta, Bernd Schmeck, Alexander Steinkasserer, Gerold Schuler, and Julio Vera**

### **Abstract**

The understanding of the immune response is right now at the center of biomedical research. There are growing expectations that immune-based interventions will in the midterm provide new, personalized, and targeted therapeutic options for many severe and highly prevalent diseases, from aggressive cancers to infectious and autoimmune diseases. To this end, immunology should surpass its current descriptive and phenomenological nature, and become quantitative, and thereby predictive.

Immunology is an ideal field for deploying the tools, methodologies, and philosophy of systems biology, an approach that combines quantitative experimental data, computational biology, and mathematical modeling. This is because, from an organism-wide perspective, the immunity is a biological system of systems, a paradigmatic instance of a multi-scale system. At the molecular scale, the critical phenotypic responses of immune cells are governed by large biochemical networks, enriched in nested regulatory motifs such as feedback and feedforward loops. This network complexity confers them the ability of highly nonlinear behavior, including remarkable examples of homeostasis, ultra-sensitivity, hysteresis, and bistability. Moving from the cellular level, different immune cell populations communicate with each other by direct physical contact or receiving and secreting signaling molecules such as cytokines. Moreover, the interaction of the immune system with its potential targets (e.g., pathogens or tumor cells) is far from simple, as it involves a number of attack and counterattack mechanisms that ultimately constitute a tightly regulated multi-feedback loop system. From a more practical perspective, this leads to the consequence that today's immunologists are facing an ever-increasing challenge of integrating massive quantities from multi-platforms.

In this chapter, we support the idea that the analysis of the immune system demands the use of systems-level approaches to ensure the success in the search for more effective and personalized immune-based therapies.

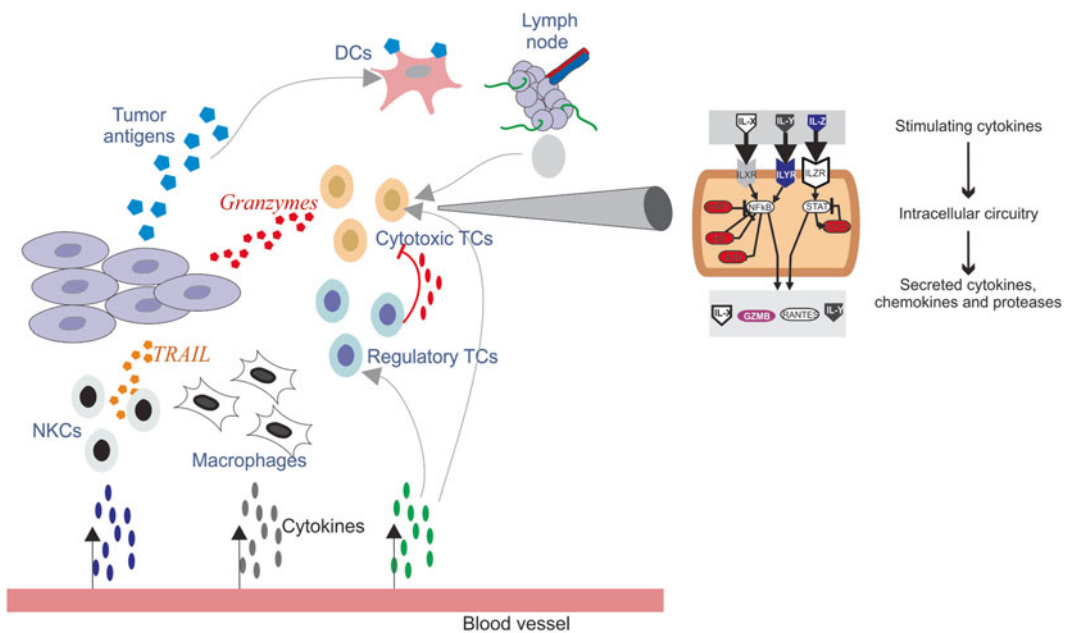
**Key words** Immunoinformatics, Systems immunology, Network reconstruction, Immunogenicity, Kinetic modeling, Immune intervention

---

\*These authors contributed equally.

## 1 Introduction: Immunity—A System of Systems

When observed in its entirety, the immune system is a marvelous biological “system of systems” (SoS), a multi-level ensemble of specialized biological entities that interact and **cross-talk** to create an adaptive biological system, with the ability to combat the diverse internal and external threats encountered by the organism. From a structural perspective, the immunity is a multi-scale SoS, composed of many populations of interacting and highly specialized immune cells (Fig. 1). Interestingly, the fate of these cell populations, their ability to proliferate, migrate, or differentiate to attack pathogens, is tightly regulated by a multiplicity of control structures often called **regulatory motifs**, the most prominent of which is the **feedback loop**. Within each immune cell, there is a large and highly interconnected **biochemical network** in which transcriptional, signaling, and metabolic circuits enriched in regulatory motifs cross-talk to evoke fine-tuned phenotypic responses. Furthermore, each individual immune cell relies on the interaction with other cells to determine its fate, either by physical cell–cell



**Fig. 1** Immunity: a system of systems. The immune system is a multi-scale system of systems, composed of many populations of interacting immune cells. Within each one of these cells, biochemical networks cross-talk to control, modulate, and evolve fine-tuned phenotypic responses. Each individual immune cell relies on the interaction with other cells to determine its fate, either by direct contact or through the local/global secretion of signaling molecules, especially cytokines. The system becomes more complicated when considering the interaction with pathogens or tumor cells. In this figure, the tumor and the immune cells communicate through chemical signals to affect each other’s fate in multiple ways. *DCs* dendritic cells, *NKCs* natural killer cells, *TCs* T cells

contacts or through the local/global secretion of signaling molecules, especially cytokines. In this way, regulatory motifs are also fundamental to the exquisite ability of different immune cell populations to coordinate, amplify, and compensate the effect of other types of immune cells, in a sort of complex intercellular and interpopulation communication system. When one links all these and views them from a global perspective, the system of systems of immunity becomes evident.

In the last decade, exhaustive analyses of immunity have made necessary the use of high-throughput data (HTD) techniques. These are either identical to the -omics constellation of experimental techniques or have been customized for the purpose of immunological research. The use of HTD has increased the magnitude of the challenge that modern immunologists are facing: unraveling a complex, multi-scale, tangled system through massive amounts of multi-scale experimental data generated from different platforms. Tackling this sort of structural and operative complexity is a dilemma faced in other fields of biomedical research a decade ago and needs the introduction of the revolutionary concept and tools of systems biology.

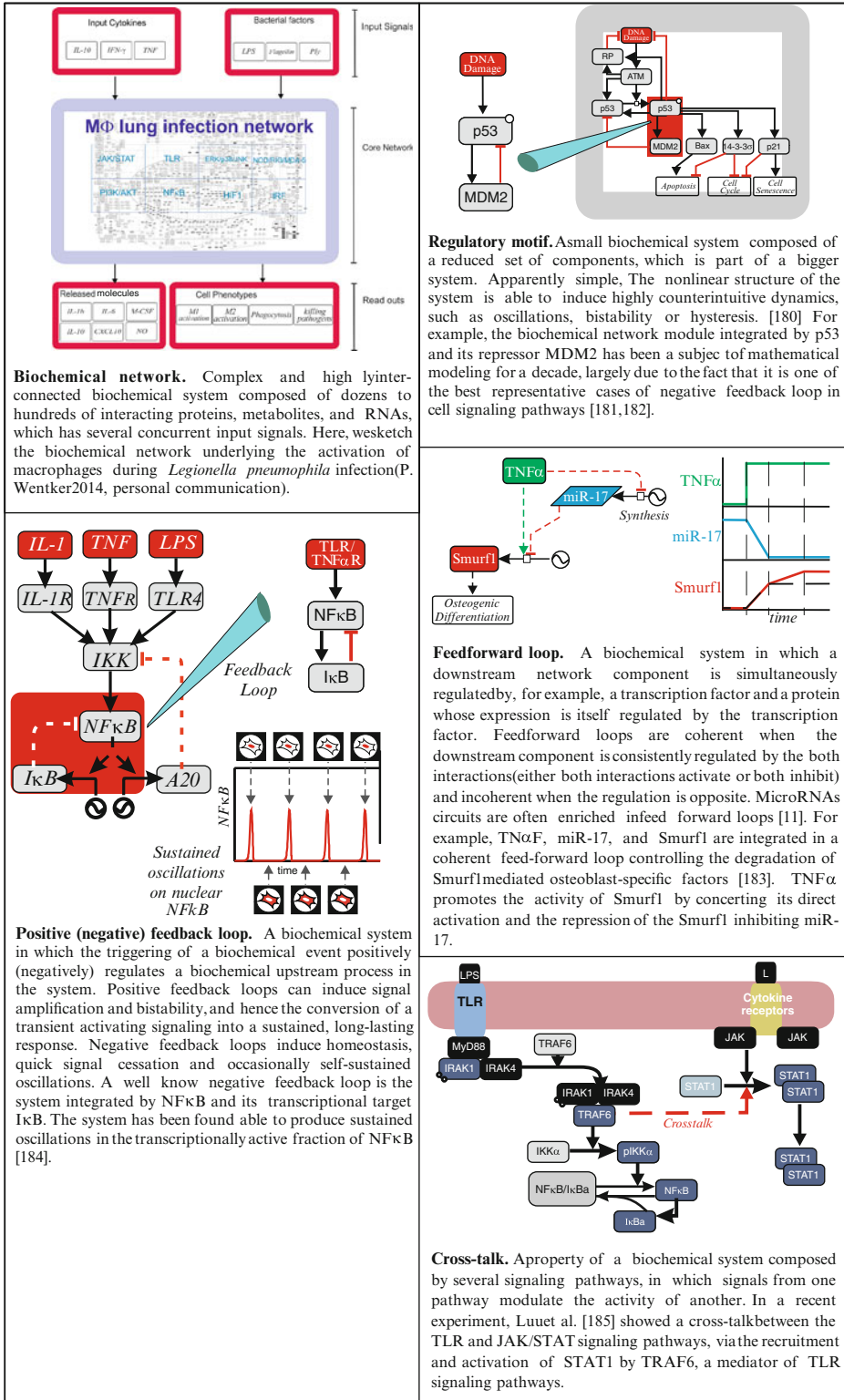
---

## 2 Systems Biology in Nine Sentences

To get a deeper understanding of the philosophy and the use of systems biology methods in biomedicine, the reader is referred to more detailed publications [1–5]. In the following, the systems biology approach is defined in a nutshell and a text box (*see* Textbox 1):

Systems biology is a methodology that employs mathematical modeling and computational biology tools to integrate and analyze quantitative biological data. It is *the optimal* strategy when one plans to (a) elucidate the function and regulation of biochemical networks enriched in regulatory motifs; (b) analyze high-throughput quantitative biological data; or (c) integrate quantitative biological data accounting for different temporal and spatial scales of the same biological phenomenon.

A key element of the method is the use of **mathematical modeling**, employed to analyze or integrate data, design experiments, formulate hypotheses, or perform quantitative simulations for a biotechnological or biomedical purpose. The term systems biology was initially coined for the application of kinetic models to the analysis of the dynamics and regulation of biochemical pathways [6, 7], but it has extended its meaning to refer to the use of statistical models, bioinformatics, or advanced computational biology tools in the analysis of large sets of quantitative biological, biotechnological, or biomedical data. In the context of immunology, one can find at least three different branches which conflate into systems biology:



**Textbox 1** Short visual vocabulary for systems biology newcomers



1. An advanced version of immunoinformatics [8], in which computational biology tools are used in a high-throughput fashion to systematically elucidate structural properties and immunogenicity of antigens, as well as to cross-validate these predictions with quantitative data.
2. The generation and analysis, by means of customized biostatistical methods, of single or multiple types of omics data, used to quantify a plethora of parameters accounting for the activity of the immune system at a genome-wide scale [9].
3. *Stricto sensu*, systems biology deploys mathematical models (preferentially kinetic ones) to dissect the nature of biochemical networks, to detect and elucidate complex non-intuitive relations between their components and to provide support in making hypotheses and designing experiments. It is a concept primarily adopted from a strategy used in the last century in physics, chemistry, and engineering, which relies on the modeling of natural or artificial systems containing regulatory motifs to study their dynamics, regulation, and controllability [1, 4].

In the following sections of the chapter, we review some of the results published over the last years in immunology using (1) immunoinformatics, (2) omics data, or (3) mathematical modeling. Furthermore, we here and in our work maintain the notion that in the ultimate version of the systems biology workflow, tools from all these disciplines will be combined to obtain a comprehensive perspective on the structure, regulation, and phenotypic response of the immune system [10, 11].

---

### 3 Determining the Structure and Properties of Immunogenic Epitopes via Immunoinformatics

Although wider definitions have been proposed for the term “immunoinformatics” [12–14], in what it follows immunoinformatics is a branch of computational biology, devoted to the development and use of computational methods to analyze those properties of antigens which may affect their immunogenicity: their ability to trigger an adaptive immune response via activation of B or T cells. A primary aim of this field is to predict the structure of epitopes, parts of given antigens identified by the immune system via antibodies, B cells, or T cells. The features and efficiency of the interaction between the epitopes and the cited elements of the immune system rely strongly on the 3D structure of the antigen to which they belong. Importantly, epitopes can display a linear structure (a.k.a. continuous or sequential epitopes, a linear chain of amino acid residues that maintain proximity in the 3D conformation of the antigen tertiary structure), or discontinuous structure (a.k.a. conformational epitopes, composed of distant sections of the pri-

many antigen's amino acid residues that are brought into spatial proximity by protein folding within the folded 3D protein structure).

Epitopes are not only important for understanding disease mechanisms or host–pathogen interactions, but also for antimicrobial target discovery and vaccine design. They also have a strategic biotechnological importance: epitopes can be synthesized or, in case of a protein, its gene can be cloned into an expression vector, and thereby they can replace an antigen in the process of either antibody production or antibody detection. The experimental techniques for elucidating epitope structure and properties are expensive and time consuming and do not scale well to large sets of antigens. In the last decades a plethora of computational methods have been developed to determine the structure of epitopes and their binding affinity to the targeted immune system compounds based on established computational biology techniques. These include matrix-driven methods, finding structural binding motifs, quantitative structure-activity relationship (QSAR) analysis, homology modeling, protein threading, docking, and several machine learning algorithms and tools. The overall computational approach for designing epitope-based vaccines is highlighted in Fig. 2. The following text of this section is an updated version of previous articles contributed by one of the authors [15, 16]. We have classified the information into B cell epitopes, T cell epitopes, and allergy-related epitopes.

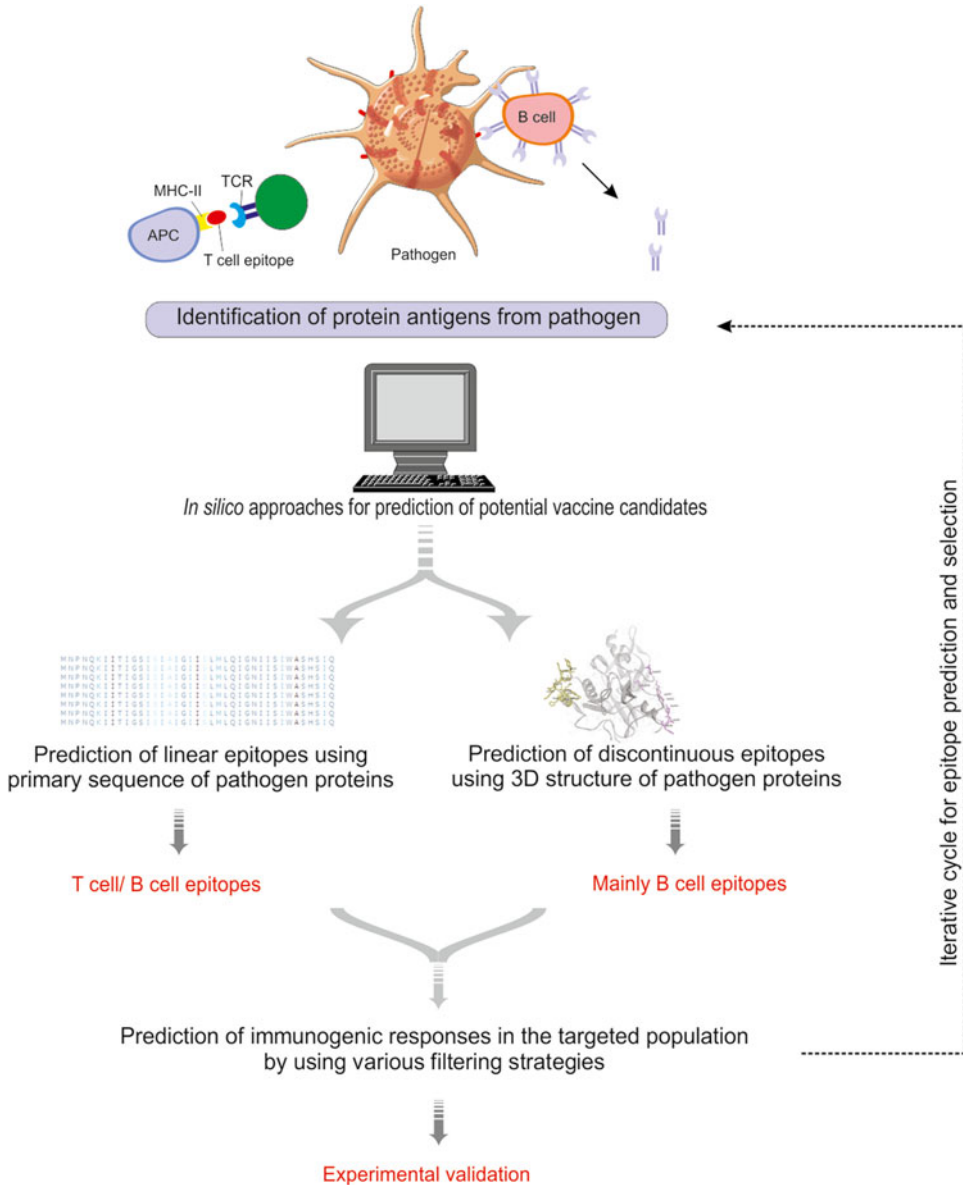
### 3.1 B Cell Epitopes

B cell epitopes play an important role in vaccine design, immunodiagnostic tests, and antibody production. As shown in Fig. 3, B cell epitopes can be linear or discontinuous [17] based on the spatial arrangement of amino acid residues on the surface of an antigen protein. B cell epitopes are antigenic determinants on the surface of pathogens that interact with B cell receptors (BCRs). The BCR-binding site is hydrophobic, having six hypervariable loops of variable length and amino acid composition. There are both sequence-based and structure-based prediction tools for predicting B cell epitopes [18, 19].

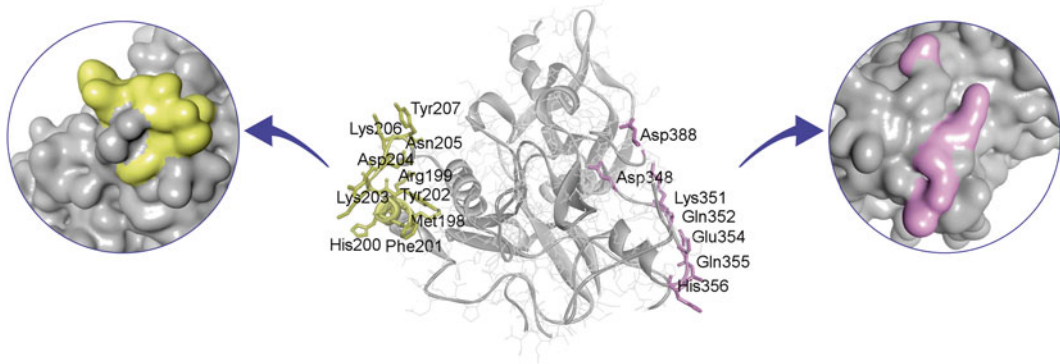
#### 3.1.1 Prediction of Linear B Cell Epitopes

Methodologies for the prediction of continuous B cell epitopes involve sequence-based methods, amino acid propensity scale-based methods, and machine learning-based methods.

*Sequence-based methods* find the epitope surface that is accessible for antibody binding and thereby the usage of these methods is limited to continuous epitope prediction. For instance, sequence-based methods have been tested for predicting two protective linear epitopes known in influenza A virus hemagglutinin HA1 [20]: the first one is the 91–108 epitope, protective in rabbit and able to elicit antibodies neutralizing infectivity of influenza viruses [21],



**Fig. 2** Overview of computational prediction of epitope-based vaccine. The workflow starts with the identification of potential antigen proteins from pathogen or proteins that are only expressed in case of tumor phenotype. Various computational tools and methods are available for the prediction of T and B cell epitopes reviewed in detail in this section of the chapter. These epitopes require various filtering strategies in order to select a minimal set of potential epitopes for synthetic vaccine development. Some of the filtering criteria for T cell epitopes include MHC allele frequencies in the targeted population, T cell receptors (TCR) expressed, mutations in the TCR, and so on, while for B cell epitopes, important filtering strategies include subcellular localization of the protein, solvent accessibility, population subsets based on B cell receptors (BCR) sequences, mutations in BCRs, etc. Selection of T and B cell epitopes is an iterative process



**Fig. 3** Continuous versus discontinuous epitopes. Here, sequential and conformational epitopes are highlighted on *P. falciparum* apical membrane antigen-1 (AMA1) protein (PDB ID: 1Z40). The sequential/continuous epitope is shown in *yellow* color (amino acid residues from 198 to 207) on Domain-I of the protein [178] along with the surface representation in the *circle* on *left*. Discontinuous/conformational epitope is highlighted in *pink* color (amino acid residues 348, 351–52, 354–356, 385, and 388–389 [179]) with surface structure on *right*

and the second one is the 127–133 epitope, protective against the influenza strain A/Achi/2/68 (H3N2) in mouse [22]. Similarly, Gupta and coworkers also predicted six unique B cell epitopes using sequence-based methods from a pool of 15 immunogenic consensus peptide fragments in globally distributed influenza-A H1N1 neuraminidase [10].

*Amino acid propensity scale-based methods and other structure prediction methods* – The classical methods of identifying potential linear B cell epitopes from antigenic sequences typically rely on the use of amino acid propensity scales for parameters such as hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity, and antigenic propensity of polypeptides chains. These methods apply amino acid scales to compute the scores of a residue  $i$  in a given protein sequence. In these methods, the  $i-(n-1)/2$  neighboring residues on each side of residue  $i$  are used to compute the score for residue  $i$  in a window of size  $n$ . The final score for residue  $i$  is the average of the scale values for  $n$  amino acids in the window. In line with this, El-Manzalawy et al. [23] compared propensity scale-based methods using a Naive Bayes classifier and used two datasets, protectivity [24] and BciPep [19]. The protectivity dataset is comprised of 57 non-redundant pathogen proteins parsed from the IEDB database [25]. The BciPep dataset composed of 125 non-redundant antigens at 30 % sequence similarity cutoff was derived from the BciPep database [19]. Interestingly, Bepitope predicts continuous epitopes using more than 30 propensity scale values based on the prediction of protein turns [26].

Ponomarenko et al. implemented a Web-based tool, Ellipro, that combines several methods for structure prediction and visu-

alization to predict and display the antibody–epitope binding complex in protein sequence and structure [27]. It predicts both linear and discontinuous antibody epitopes, where it accepts as an input either a protein structure (in a PDB format) or a protein sequence. It associates each predicted epitope with a score, defined as a PI (Protrusion Index) value averaged over epitope residues. To be very precise, this tool implements three algorithms for approximation of the protein shape as an ellipsoid, calculation of the residue PI, and clustering of neighboring residue based on their PI values.

*Machine learning-based methods.* Machine learning is a family of computational and statistical algorithms programmed to perform a sort of automatic ‘learning process’ from input data sets, used later on to make predictions on additional data sets. In case of continuous B cell epitope prediction, machine learning methods are being successfully used by several researchers. For instance, Saha and Raghava [28] used feed forward and recurrent neural networks to predict continuous B cell epitopes. Sweredoski and Baldi [29] implemented a two-step system for prediction of continuous B cell epitopes. In the first step of this method, a fragment epitopic propensity score is assigned to protein sequence fragments using a support vector machine (SVM, a type of machine learning algorithm), while in the second step, an epitopic propensity score is calculated for each residue based on the SVM scores.

Newer algorithms for determining linear epitopes combine the results obtained from two or more of the mentioned classes of methods. For example, Larsen and coauthors [30] computed results using several propensity scale methods and applied a hidden Markov model (HMM, another machine learning method) to find optimal parameters. The results from the HMM were combined with the ones obtained from preselected propensity scale methods to get more accurate predictions. Currently, approximately 65 % of accuracy has been achieved in the prediction of linear epitopes applying combinations of amino acid scales or machine learning techniques. Higher accuracy could be reached by improving the quality of existing B cell epitope data sets [17].

### 3.1.2 Prediction of Discontinuous B Cell Epitopes

The most accurate way to identify B cell epitopes is through X-ray crystallography, but this is a time- and resource-consuming, not always successful, and not scalable method. Hence, several groups put their efforts to successfully develop methods for the elucidation and screening of discontinuous B cell epitopes.

*3D structure-based methodologies* – B cell epitope prediction using computational methods based on the 3D structure of an antigen is a research area under development. However, there are some available prediction methods for conformational B cell epit-

opes that deserve attention. Anderson et al. presented a method, DiscoTope, that combines amino acid statistics, spatial information, and surface exposure [30]. The method detects 15.5 % of residues located in discontinuous epitopes with a specificity of 95 %. It is said to be the first method developed for the prediction of discontinuous B cell epitopes with better performance than methods based only on sequence data. An updated version of the method has incorporated a novel spatial neighborhood definition and half-sphere exposure as surface measure [31]. Sweredoski and Baldi use a weighted linear combination of amino acid propensity scores and half-sphere exposure values [32], which encode side chain orientation and solvent accessibility of amino acid residues for the prediction of conformational epitopes [33]. Authors have also reported its improvement in performance over the DiscoTope method. Bublil et al. developed a methodology, based on the hypothesis that the simplest meaningful fragment of an epitope is an ‘amino acid pair’ (AAP) of residues lying within the epitope as a result of folding [34]. In their work, they obtained a set of affinity-isolated peptides by screening the phage display peptide libraries with the antibody of interest. This set was used as input data and to obtain 1–3 epitope candidates on the surface of the atomic structure of the antigens. A recent approach [35] has focused on the impact of interior residues, different contributions of adjacent residues, and the imbalanced data which contain much more non-epitope residues than epitope and applied random forest (RF) algorithm for the prediction of conformational B cell epitope [36].

*Mimotope-based methodology* – Mimotopes are bio-macromolecules, often peptides, that mimic basic features of the structure of a natural epitope, and therefore are able to trigger an antibody response similar to the one promoted by the natural epitope. Interestingly, mimotopes and real epitopes can combine the same paratope (paratope is the part of an antibody which recognizes an antigen, the antigen-binding site of an antibody) of monoclonal antibody and thereby cause immune response [37]. Besides, the selected mimotopes commonly share high sequential similarity, which implies that certain key binding motifs and physicochemical preferences exist during the interaction. Therefore, one may find the real epitopes more accurately after mapping the mimotopes back to the source antigen. The mimotope-based prediction is a methodology that combines both antibody affinity-selected peptides and the 3D structure of the targeted antigen as input. Pizzi et al. [38] derived an approach that relies on the generation of a phage display library of random peptides, which is scanned against a desired antibody aiming at obtaining mimotopes of high-affinity binding to the antibody. The authors assumed that the set of high affinity-selected mimotopes maintain critical physicochemical properties and spatial organization of the genuine epitopes [37, 39, 40], and hence the

libraries of mimotopes obtained in this way can be mined to predict real epitope properties.

Other software tools applying a similar strategy are MIMOP [37] (based on degenerated alignment analyses and consensus identification); MIMOX [41] (able to map a single mimotope or a consensus sequence of a set of mimotopes onto the corresponding antigen structure and look for the clusters of residues that could be part of the native epitope); Pepitope [42] (which works by mapping mimotopes onto the solved structure of antigen surface); or Pep-3D-Search [43] (conceived to localize the surface region mimicked by the mimotope).

Sometimes linear peptides mimic conformational epitopes and methodologies have been developed to make profit of this property. For example, Schreiber et al. developed a tool able to detect linear peptide sequences within 3D structures of proteins, which was tested in the localization of mimotopes from HIV-positive patient plasma within the 3D structure of gp120 [44]. In a similar manner, Huang et al. developed a method that aligns mimotope motifs to the antigen sequence directly and rates the best matching sequences as epitope candidates [45].

*Hybrid prediction methods* – One increasingly interesting option is to combine or integrate the results obtained from several prediction methods [41, 46]. In line with this idea, several integration strategies are possible to generate a consensus set of epitope predictions. Options are, for example, to use (a) majority voting (it selects epitopes predicted by a majority of the methods employed), (b) weighted linear combinations (the consensus prediction is obtained via a weighted sum of the predictions taken from the set of predictors) or (c) meta-learning (a meta-classifier is trained on a training dataset using the outputs of the predictors on each input sample as input to the classifier and the corresponding class label as the desired output of the classifier) [41], (d) nearest neighbor, and (e) decision tree [47].

For example, an ensemble of linear B cell epitope optimal performing classifiers was developed by Sollner [46]. It was based on a proposed majority voting strategy known as positive unanimity voting.

### 3.1.3 Databases for B Cell Epitope Search

In direct relation with the methodologies discussed above, a large number of databases have been published for B cell epitopes. These databases contain (a) experimentally verified B cell epitopes and their mapping on antigen sequences; (b) known antigenic residues and their interacting antibodies, obtained from Protein Data Bank structures or via curation of literature; (c) predicted linear or discontinuous B cell epitopes; and (d) predicted mimotopes. In Table 1, the reader can find a selection of these databases.

## 3.2 T Cell Epitopes

The basis of the mechanism behind the recognition of pathogens (and tumor cells) by cytotoxic cells is the efficient binding of anti-



**Table 1**  
**Databases on B cell epitopes, T cell epitopes, allergen, and molecular evolution of immune system components**

Databases	Names	URLs
B cell epitopes	CED	<a href="http://www.immunet.cn/ced/log.html">http://www.immunet.cn/ced/log.html</a>
	Bcipep	<a href="http://www.imtech.res.in/raghava/bcipep">http://www.imtech.res.in/raghava/bcipep</a>
	Epiotme	<a href="http://www.rostlab.org/services/epitome/">http://www.rostlab.org/services/epitome/</a>
	IEDB	<a href="http://www.immuneepitope.org/">http://www.immuneepitope.org/</a>
	IMGT®	<a href="http://www.imgt.org">http://www.imgt.org</a>
	MimoPro	<a href="http://www.informatics.nenu.edu.cn/MimoPro">http://www.informatics.nenu.edu.cn/MimoPro</a>
	Mimodb	<a href="http://www.immunet.cn/mimodb">http://www.immunet.cn/mimodb</a>
T cell epitopes	Syfpethi	<a href="http://www.syfpethi.de">http://www.syfpethi.de</a>
	IEDB	<a href="http://www.immuneepitope.org/">http://www.immuneepitope.org/</a>
	IMGT®	<a href="http://www.imgt.org">http://www.imgt.org</a>
	TEPITOPEpan	<a href="http://www.biokdd.fudan.edu.cn/Service/TEPITOPEpan/">http://www.biokdd.fudan.edu.cn/Service/TEPITOPEpan/</a>
Allergen	Database of IUIS	<a href="http://www.allergen.org">http://www.allergen.org</a>
	SDAP	<a href="http://www.fermi.utmb.edu/SDAP/">http://www.fermi.utmb.edu/SDAP/</a>
	Allergome	<a href="http://www.allergome.org">http://www.allergome.org</a>
Information related to molecular evolution of immune system components	ImmTree	<a href="http://www.bioinf.uta.fi/ImmTree">http://www.bioinf.uta.fi/ImmTree</a>
	Immunome database	<a href="http://www.bioinf.uta.fi/Immunome/">http://www.bioinf.uta.fi/Immunome/</a>
	ImmunomeBase	<a href="http://www.bioinf.uta.fi/ImmunomeBase">http://www.bioinf.uta.fi/ImmunomeBase</a>
	Immunome KnowledgeBase	<a href="http://www.bioinf.uta.fi/IKB/">http://www.bioinf.uta.fi/IKB/</a>

genic peptides with the major histocompatibility complex molecules (MHC) allocated at the surface of T cell plasma membrane. Having this in mind, a number of methods have been developed in the last decade to predict the binding affinity of epitopes to the MHC molecules. These methods rely on either machine learning-based methods like matrix-driven algorithms, HMMs, artificial neural networks, or structural-biology methods like those based on the prediction of the epitope structure or the molecular dynamics behind epitope-MHC binding process.

*Machine learning-based methods* – Huang and Dai [48] first developed a matrix-driven method for elucidating the binding between epitopes and MHC molecules, based on a BLOSUM matrix with the amino acid indicator vectors for direct prediction of T cell epitopes. For information, it should here be mentioned that in quantitative matrix-driven methods, the contribution of binding from



each amino acid at each peptide position within the binding groove is quantified [49]. In line with this and using quantitative matrices accounting for 47 MHC alleles, Bhasin et al. [47] implemented a server-based method able to predict the mutated promiscuous and high-affinity MHC binding peptides.

Transfer-associated protein (TAP) is an ATP-binding-cassette transporter able to translocate cytosolic peptides of 8–40 amino acids in size to the endoplasmic reticulum, where these peptides are loaded into nascent MHC I molecules for further presentation. Zhang et al. [50] implemented and tested an artificial neural network (ANN) and HMM-based method to compute peptide binding to human TAP in the form of PRED<sup>TAP</sup>. Technically speaking, their method includes a three-layer back propagation network for the development of the PRED<sup>TAP</sup> server, with the sigmoid activation function. It uses binary strings representing nonamer peptides as input data to the ANN. It has a Web user interface also, which uses graphical user interface forms.

In ANN-based methods, a neural network is trained to associate given input antigenic peptide sequences to a measurement of the binding affinity of these sequences to the considered MHC molecule [51]. To mention a paradigmatic example, Nielsen et al. described an improved algorithm to predict T cell class I epitopes, based on the combination of neural networks derived using different sequence-encoding schemes [52]. In their method, a sparse encoding, BLOSUM encoding for sequences, and input obtained from a HMM are integrated. The input dataset for their method is a 528-amino acid nanomer set with known binding affinities to the HLA A\*0204. They illustrate the use of their method in the prediction of T cell epitopes for the hepatitis C virus.

Other machine learning methods have been used in the elucidation of T cell epitopes. For example, Ant colony search (ACS) has been found useful for the identification of a multiple alignment of a set of peptides [53]. SVMs have also been used, for example, to find the correlation between nine physiochemical properties of antigenic peptide sequences and the TAP binding affinity [54].

*Structure-based prediction methods* – Current structure-based prediction methods for T cell epitopes are based on the computation of the peptide binding affinity to MHC molecules. Quantitative structure-activity relationship (QSAR) is a well-established approach to correlate the changes in the structure of series of compounds with changes in their biological activities. Several predictive computational models have been developed for epitopes and MHC binding affinity based on QSAR approaches [55–58]. 2D-QSAR methods are based on the calculation of physicochemical properties affecting the epitope-MHC interaction. The features obtained are used to predict the values of the epitope-MHC binding affinity. In line with this, it is generally accepted that only peptides that bind to MHC with IC<sub>50</sub> value below a threshold value

(typically 500nM) function as T cell epitopes. 2D-QSAR methods largely rely on the quality and amount of data that comprise the training set of epitopes bound to MHC molecules. In comparison to 2D-QSAR models, 3D-QSAR methods are more reliable in predicting the binding affinity of epitopes by mapping 3D interaction potentials on the structure of molecules being investigated. Kangueane and Sakharkar developed the web service T-EPITOPE Designer for the *in silico* prediction of promiscuous and allele-specific HLA II-restricted T cell epitopes [59]. T-EPITOPE Designer's user interface displays and compares pocket profiles, and finds similar HLA II differing in their binding capacity for a given peptide sequence. The main drawback is that the method can be applied to only 51 out of over 700 known HLA-DR molecules. Zhang et al. used a 3D QSAR-based method to extend these results to 700 HLA-DR molecules [60]. Their method makes use of the interaction potential around aligned sets of 3D peptide structures to describe the peptide-MHC binding. Guan et al. [61] made a Perl implementation of 2D QSAR application to peptide-MHC prediction and covers both class I and class II MHC allele peptide specificity models. Jojic et al. [62] developed an improved structure-based model which makes use of known 3D structures of MHC-peptide complexes, MHC class I sequences and binding energies for MHC-peptide complexes, and a training binary dataset containing information from genuine epitopes (considered as strong binder peptides and non-binders (peptides that have a low affinity for given MHC molecules)). Their method can make epitope-MHC binding predictions for alleles, for which little data is available beyond just their sequence, including prediction for alleles for which 3D structures are not available [63]. Other computational methods, such as molecular docking and molecular dynamics simulation, which are well established for protein-ligand interactions, were successfully implemented to screen for potential epitopes that bind to MHC molecules [64]. The impact of epigenetic changes on the binding affinity of MHC-molecules too can be investigated using docking and simulation methods. The only limitation of these methods is the availability of a reliable 3D structure of the targeted molecules.

Tumor immunology is an active field in the search for MHC epitopes because tumor peptides on the tumor cell surface that can bind to MHC have the potential to initiate a host immune response against the tumor by activating cytotoxic T cells, or can be used in the design of therapeutic anticancer vaccines. In line with this idea, Schueler-Furman et al. [65] developed a structure-based algorithm for the prediction of MHC binding epitopes, derived from tumor-specific antigenic proteins. They illustrated the use of their method with the identification of putative nine amino acid epitopes with potential to bind to the MHC molecule HLA-A2, which were derived from the sequence of the cancer-testis antigen KU-CT-1.

*Hybrid methods* – In case of T cell epitopes also some hybrid approaches combining results from several prediction methods have been developed. Brusic et al. developed a hybrid method for the prediction of MHC class II binding peptides. The method integrates experimental data and expert knowledge of binding motifs, together with results coming from evolutionary algorithms and ANNs [66]. Doytchinova et al. implemented a multi-step algorithm for T cell epitope prediction, based on quantitative matrices, which belongs to the next generation of in silico T cell epitope identification methods [67]. Lundegaard et al. [68] derived an algorithm for high-accuracy estimation of peptide-MHC binding that combines ANNs trained on data from 55 MHC alleles and position-specific scoring matrices for additional 67 HLA alleles.

### 3.2.1 Databases for T Cell Epitope Search

One can find in the literature a number of databases for T cell epitope search, the most prominent of which are listed in Table 1. To cite some of them: (a) Syfpeithi has information on MHC class I and II anchor motifs and binding specificity for some important model species like apes and mouse [69]; (b) IEDB has more than 123,978 peptidic epitopes [70]; (c) IMGT<sup>®</sup> contains an extensive collection of IG, TR, MHC, and other proteins related to the human immune system, available through online tools for sequence, genome, and 3D structure analysis [71]; (d) IMGT/HLA provides a specialist database with 9232 HLA Class I alleles and 3010 HLA Class II alleles, as well as 164 non-HLA alleles [72].

### 3.3 Allergy Epitopes

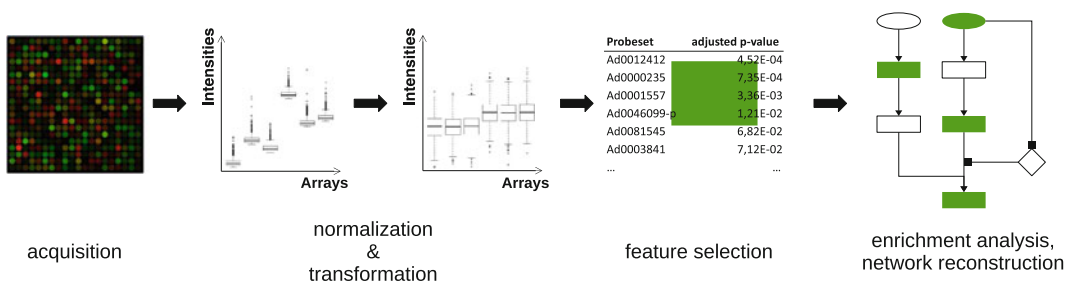
A large number of online databases and prediction tools are available for predicting allergens and cross-reactivity between known allergens. For example, the Allergen Nomenclature database of the International Union of Immunological Societies contains an allergen database [73]. The AllergenPro database contains information concerning 2434 known allergens in rice microbes, animals, and plants [74]. The web server Allergome provides an exhaustive repository of IgE-binding compounds data that contains up to 2265 allergen sources [75]. The Structural database of Allergenic Proteins (SDAP) [76] is offered in a web server that integrates a database containing allergenic proteins with various computational tools assisting in the elucidation of allergen structures. In the last version available, it contains 92 allergens whose structure is accessible in the PDB and another 1312 protein sequences for allergens and isoallergens. It needs to be mentioned that several computational approaches for predicting allergenicity of proteins have been developed; however, their performance and shortcomings are required to be compared.

## 4 Analysis of Immunological High-Throughput Data

In the last decade it became technically and economically feasible to perform quantitative, high-throughput experiments to produce HTD accounting for the genomic, transcriptomic, and post-transcriptomic (mature RNAs and proteins) levels. These techniques can be applied to generate HTD for the detailed analysis of the dynamics of activation or differentiation of immune cells traced in *in vitro* experiments, but also to generate HTD from cohorts of patients suffering from infectious or immune-related diseases, or cancer patients treated with immunotherapies. Having in mind that we are here handling massive volumes of quantitative data, any attempt to interpret them requires the use of sophisticated statistical and computational methods.

Note that in the following discussion and in accordance with the statistical literature, “feature” is a generic word used for a gene, transcript, or protein. Large biomedical HTD sets usually are within the category of “low sample size–high feature number,” for instance a sample of 100 patients with 50,000 measured features each. This kind of problem, not common in many other fields of application of statistics, required and facilitated the development of statistical approaches that can derive meaningful results from this “inverted” situation. A common approach used in the analysis of biomedical HTD is given as a flowchart in Fig. 4.

Some of the general guidelines when dealing with data sets of this size have been established for a long time. Analyses generally rely on background correction and inter-array **normalization** [77, 78], **log-transformation** of the data, **moderation** of the selected statistical model by borrowing information across features and samples, and correction for multiple testing [79] or **control of the false discovery rate** [80]. There is a plethora of algorithms for determining which features to call differentially expressed between groups, among them SAM [81], linear models combined with Bayesian estimators [82], shrinkage [83], and others (for a review, *see* for example Ref. 84). Sometimes, a fold-change cutoff is used



**Fig. 4** Flowchart for a general statistical analysis of high-throughput data

as a secondary condition to prune the results, although this approach has been criticized [81]. The methods developed originally for microarray analysis were later adapted for next-generation sequencing (NGS) data [85].

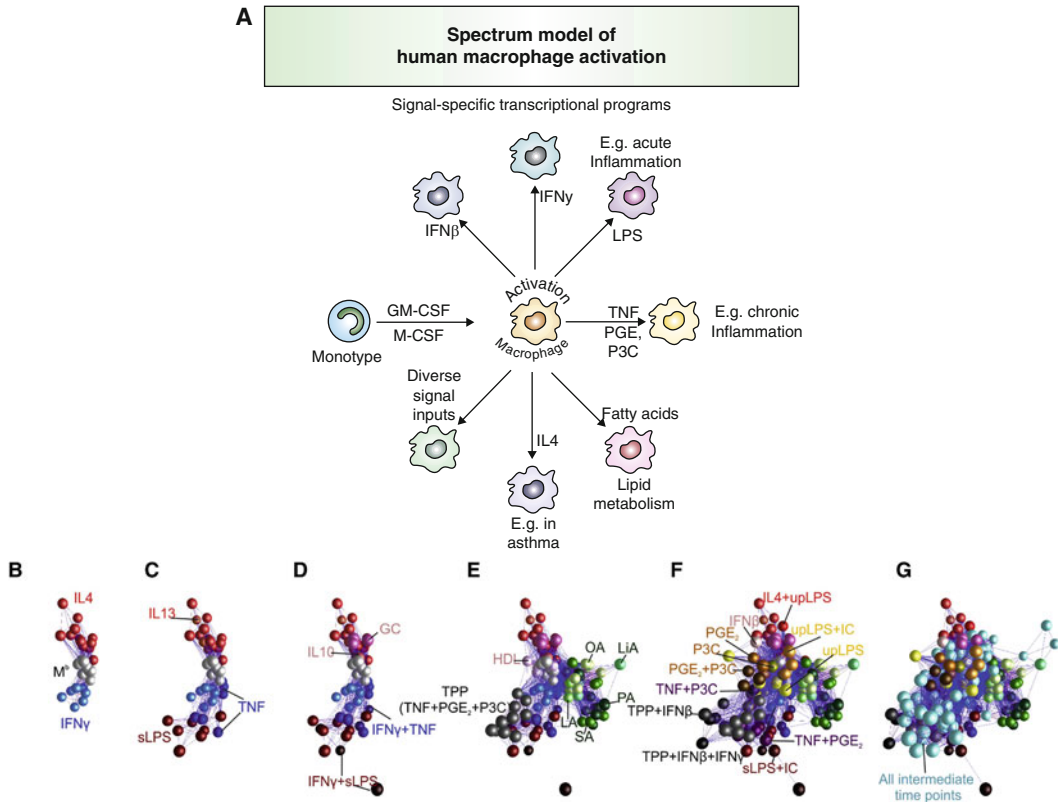
After extracting the interesting features, a number of well-established enrichment analysis methods can be performed to account for functional connections between features [86, 87]. These analyses focus on pathways [88], gene ontologies [89], or gene sets on the genomic level [90]. Complementary methods for interactions on the transcriptomic (e.g., inhibition through microRNAs) and proteomic (e.g., correlations of posttranslational modifications, isoform realization) levels are expected to become more widely adapted once the availability of proteomic data rises.

In a further step, networks between the extracted features can be constructed. Based on a priori known interactions [91] or a similarity score that is calculated from experimental measurements [92], these networks can reveal higher order information about the data set, and can be used in topological analyses and for the generation of mathematical models. In the following we review some recent results, in which these techniques have been used to analyze immunological HTD.

#### **4.1 HTD in Cell Differentiation Analysis: Macrophages and Dendritic Cells**

*Regulation of macrophage function and differentiation.* Great effort has been poured into the elucidation of macrophage function. Macrophages develop from monocytes and survey their surroundings as agents of the immune system. They are, together with neutrophils, among the first cells to encounter molecular patterns associated with threats, and are equipped to mount an inflammatory response. While one of their main tasks is the clearance of microbes, they also serve to coordinate the immune response together with other cell types. Macrophages themselves are subject to differentiation, also called polarization in this case, according to the type of threat they encounter. The differentiation options have been extended over the last decades, and the analysis of HTD has given more insights into available paths.

In 2012, Beyer and coworkers published a refined inspection of the canonical M1–M2 polarization in humans and identified potential surface marker proteins [93]. At the same time, they compared data recorded by microarray with data produced by RNA-seq from a different sample set. They came to the conclusion that while RNA-seq and microarray fold changes show good correlation, RNA-seq offers a higher dynamic range of measurement attributed to lower noise [93]. Also, the RNA-seq analysis revealed a larger set of differentially expressed genes between M1 and M2 compared to the microarray data analysis, which was not, however, simply a superset. Both results shared 595 genes although the total number was 900 from microarray and 1500 from RNA-seq. From the list of differentially expressed genes, a-priori-knowl-



**Fig. 5** Transcriptomic data can contribute to the identification and discrimination of macrophage polarization types. (a) Overview over macrophage stimulation conditions. The tips of the vertical axis correspond to M1 (top) and M2 (bottom) polarization, respectively. (b–g) A correlation network of stimulated macrophage transcriptomes was reconstructed (conditions are indicated), with (g) or without (b–f) measurements at intermediate time points. From left to right, the network is expanded with additional stimulatory conditions. The M1–M2 axis is visible in (b). For details, see original work. Modified from original work [94] under a CC BY 3.0 license

edge-based networks were created, on which the extent and direction of differential expression for each gene were mapped.

Other studies have strived to deepen the knowledge of macrophage differentiation and look for a finer-grained classification. In a recent example [94], a broadened spectrum of differentiation modes were proposed after analysis of human macrophages stimulated with different cues (Fig. 5). The authors used co-regulation analysis to project the macrophage profiles into three-dimensional space, both reproducing and expanding the previously known two-dimensional M1–M2 axis. After sorting genes into modules according to a gene set enrichment analysis, they employed a wide variety of methods from the bioinformatics field to examine exemplary non-classic stimulation conditions and to discern gene expression differences compared to M1–M2 macrophages.

Moreover, the investigation of literature data sets for human alveolar macrophages from COPD patients and healthy individuals

revealed results that did not fully confirm the expected patterns. Rather, a loss of M1-specific activation together with a diminished profile in antigen processing, inflammatory response, and regulation of immune response was observed [94]. A similar result was obtained for a smoker vs. nonsmoker analysis. Finally, the authors proposed a strategy for comparing murine and human macrophage gene expression. From this, they unearthed a set of macrophage-specific surface markers that is supposed to be applicable in both species, namely *CD14*, *FCGR2A*, *MERTK*, *FCGR1A*, and *CD13* [94].

The activation of macrophages has also been studied from the perspective of the cross talk among its core pathways [95]. In this work, signaling through toll-like receptor (TLR), interferon, NF- $\kappa$ B, and apoptosis pathways was integrated to build a comprehensive map of the core module of macrophage activation. The authors generated HTD from mouse bone marrow macrophages activated with INF $\gamma$  and used the map for interpretation of the transcriptomic changes at several time points, highlighting how different segments of the pathways are regulated in a time-dependent manner.

While transcriptomic data has helped in the elucidation of many aspects of cellular regulation, it cannot account for several levels of posttranscriptional regulation that work on the protein population. Among those levels are miRNA-mediated translational inhibition, co- and posttranslational modifications, and controlled protein decay. Proteomic techniques offer the chance to remedy this shortcoming but are less widely employed because they require a higher commitment in funds and training. One protein group of particular interest is phosphoproteins whose participation in signaling pathways and metabolism has been well established. Phosphorylation is one of the most common chemical modifications that can switch a protein between different states, such as active/inactive or affine/non-affine. Two specialized enzyme classes, kinases and phosphatases, catalyze and coordinate the cellular phosphoproteomic content.

The phosphoproteome of murine macrophages activated by microbial lipopolysaccharide (LPS) has been surveyed in a study by Weintz and coworkers [96]. Using the established SILAC (stable isotope labeling with amino acids in cell culture) approach, they identified a set of approximately 7000 phosphorylation sites of which 60 % were not registered at that time. In comparison of unstimulated and stimulated macrophages, the time series data showed more phosphorylation than dephosphorylation events and a dynamic regulation of sites (only about one-third of phosphorylations were sustained over a 4-h time window). Functionally, the observed changes were linked to proteins associated with LPS-mediated TLR4 signaling including the MAPK/ERK and AKT/mTOR pathways, cytoskeleton restructuration, and cell proliferation; however, only the cytoskeleton module reached statistical significance in the accompanying GO terms. It should



be pointed out that the statistical model chosen here relied on GO term overrepresentation relative to the identified non-phosphorylated sites [96] rather than relative to the genomic background. In an effort to examine the link between phosphoproteome changes and transcriptional activity, the authors then combined transcriptomic profiles from microarray measurements with proteomic data. They identified additional potential regulators in the families of OCT, HOXC, and SORY transcription factors and calcium-dependent modules.

*Regulation of dendritic-cell function.* Dendritic cells are another cell type that can be derived from blood monocytes. Their classical function as professional antigen-presenting cells has made them an attractive target for immune-based therapies, e.g., in cancer. In a recent paper [97], the maturation of human blood monocytes into dendritic cells was investigated in a time-dependent manner. During differentiation with a standard protocol, cells were harvested after 0, 4, 8, and 24 h and subjected to microarray expression analysis. Of note, in addition to messenger RNAs (mRNA), microRNAs (miRNAs) and cytokines were also considered in this study. The authors identified five gene clusters in the differentially expressed mRNAs, while only two were found in their miRNA counterpart. The functional stratification of the mRNA clusters, as revealed by pathway enrichment analysis, reflected the time-course stratification to some extent, with immune-related genes coming up early and metabolism-related genes following. This argues for the mixed model of activation, where the initial phase is followed by a general revision of the cellular biochemistry to prepare for a change in function.

Additionally, the examination of monocytes and their leukocyte relatives in peripheral blood has revealed general principles of regulation. Chaussabel and coworkers [98] partitioned the transcriptome of human PBMCs into functional units by clustering microarray expression data associated with ten disorders along the gene axis and assembling all genes with a near-identical cluster profile into the same module. Subsequently, they exploited these expression modules for the generation of disease- and disease-state-specific supra-gene-level fingerprints [98] in systemic lupus erythematosus. The authors proposed a twofold application of the identified modules: as composite biomarkers, and as basis for the generation of a multivariate score of disease progression. They employed spider plots for the visualization of comparative expression on a relevant subset of modules.

Finally, in an approach to unravel the role stochastic noise in gene expression plays in cell differentiation, time-course microarray measurements of clonally expanded cell populations showed that initially distinct clones from one cell population over time tend to assume the same variability that was observed in their



population of origin [99]. This result was achieved thanks to the application of SAM and correlation distance calculation.

#### **4.2 Medical Immune Interventions Analyzed through HTD**

The spectrum of therapies that are linked to the immune system stretches from vaccination with pathogen-derived particles to the contemporary application of ex vivo-primed patient-derived immune cells. Here, the interplay between the immunogenic agent and the immune system is at the focus of attention.

*Vaccination against infectious diseases.* In an example of omics technique application, the transcriptional response to vaccination has been studied in connection with an influenza H1N1 vaccine in 63 healthy individuals [100]. The contributions of inter- and intrasubject variation (the latter referring to changes over time) were elucidated by an ANOVA model, covering PBMC subset composition, gene expression, and antibody response. Besides the expected time-coordinated activation of the innate and adaptive immune system, a negative correlation between initial antibody titer and antibody titer increase was observed. Another ANOVA model revealed that the observed differences in vaccination efficacy are due to initial antibody titer rather than macroscopic phenotypical features (e.g., age, sex, and ethnicity). After removal of macroscopic features, predictive modeling with robust correlation analysis between pre- and post-vaccinated samples identified genes involved in ER stress, N-glycan biosynthesis, and cell cycle as good predictors, and pointed towards plasmablasts as major source of the observed changes. In a similar manner and more strikingly, the authors used only pre-vaccination PBMC subsets and scored pathway activities in their model and identified 12 cell populations with good predictive power. Furthermore, partitioning PBMCs into temporally unstable and temporally stable subpopulations uncovered their distinct contributions to immune status. It has to be emphasized that for the latter analyses, antigen specificity did not play a role.

In a similar work, Querec and coworkers identified genetic expression signatures that predict the efficacy of vaccination in yellow fever [101]. They used human PBMC microarray data from two sets of vaccinated subjects as a starting point, and then carried out differential expression analyses on multiple input sets to compare the overlap of the called genes. After identifying a small set of 65 relevant genes, they performed pathway enrichment analysis, transcription factor-binding site analysis, and network reconstruction, resulting in a network of 50 genes related to interferon and viral response signaling. Interestingly, these genes were not related with the strength of the specific CD8 T cell responses, which led to a new round of analysis. From a new set of 839 genes identified by correlation between CD8 T cell count and expression change, classifiers were built according to the classification to nearest centroid (ClANC)

and discriminant analysis via mixed integer programming (DAMIP). Seven of the nine transcribed loci in the selected set were genes with known function. After cross-validation, the classifiers reached accuracies of 80 % and more. A similar approach was employed to integrate the neutralizing antibody response into the results.

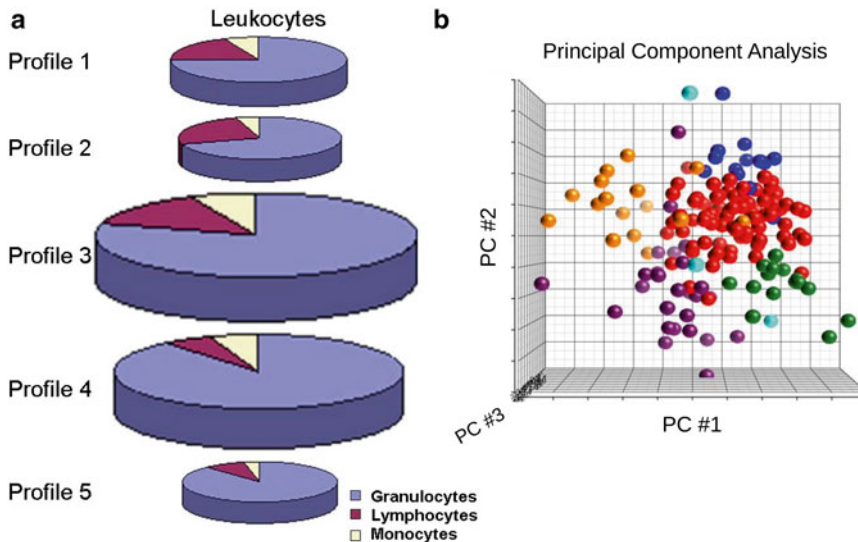
In a related study on malaria vaccination the authors used the SAM method [81] to select differentially expressed genes, which were used as input for principal component analysis (PCA) to cluster subjects [102]. The result showed very good separation of the cohort's classes. As the selected genes were associated with major histocompatibility complex type I peptide processing, the authors suggested to use upregulation of those genes as a surrogate marker for malaria protection.

A detailed look at the investigation of vaccination with systems biology approaches can be found in recent reviews like [103] and [104].

*The tumor-immunity interaction.* In the field of cancer studies, the resistance to therapy has acquired a spot in the limelight over the last decades. One publication by Zhang and coworkers [105] discusses how the tumor microenvironment protects chronic myeloid leukemia (CML) stem cells from tyrosine kinase inhibitors. After finding experimental evidence for the role bone marrow stromal cells (BMSC) play in the protection of their malignant neighbors, the effect of BMSC presence on transcription in CML co-cultured cells was examined. Microarray analysis unearthed a small set of genes of restricted origins: they either participate in signaling through cadherins and the Wnt pathway or regulate self-renewal capacity and cell adhesion. The identification of Wnt signaling components prompted experiments that confirmed the involvement of  $\beta$ -catenin in malignant perseverance.

Gustafson and coworkers in a study on cancer [106] classified patients into five categories according to their profile of absolute concentrations of leukocyte subspecies in whole blood (Fig. 6). A survival analysis revealed that the profiles may hold predictive power for long-term clinical outcome. Also, the authors suggested that the immune system reacts to challenges by switching among a finite set of stable states that are reflected in the absolute peripheral blood leukocyte subspecies profiles.

In a recent paper, Montoya and coauthors [107] analyzed biopsy samples from metastatic-melanoma patients. They strived to detect a pretreatment gene expression signature that is able to predict the response to an anticancer immunotherapy from a Phase II study (recombinant MAGE-A3 antigen combined with an immunostimulant, AS15 or AS02B). Biopsies from participating patients were obtained, processed, quantified using microarray analysis, and validated with quantitative PCR. In the next step, the authors combined the pretreatment expression profiles with data on survival after therapy. Data clustering techniques identified a



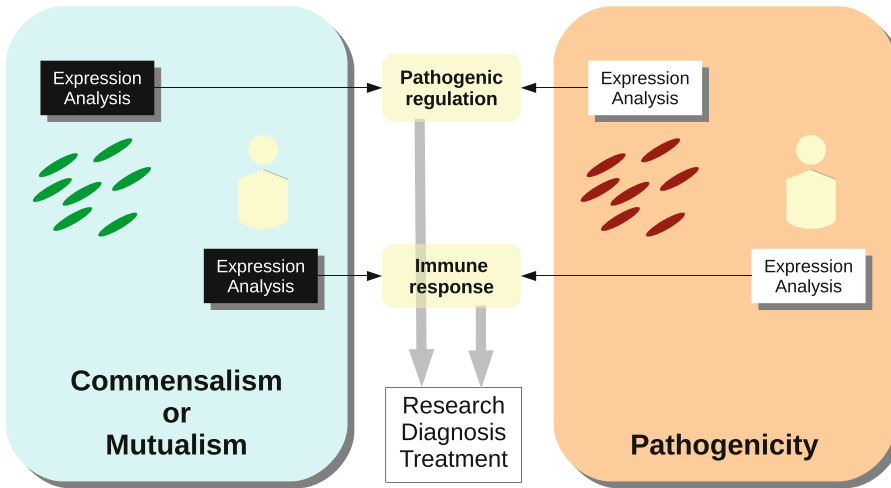
**Fig. 6** The absolute concentrations of leukocytes in human blood might be an indicator for cancer survival chances. **(a)** Visualization of the identified five whole-blood profiles as pie charts. Relative size of the pie chart scales with total leukocyte concentration. **(b)** Three-dimensional projection plot of 40 controls and 120 patients based on PCA of ten selected leukocyte subspecies blood concentrations. The data points are *colored* according to the individual's classification into one of the identified profiles (cf. panel (a)). Modified from original work [106] under a CC BY 2.0 license

signature of 84 genes whose expression was potentially associated with clinical benefit of the immunotherapy [107]. Interestingly, many of these genes were clearly associated with the tumor-immunity interaction through their involvement in mechanisms that provide tumor cells with immune evasion capabilities.

#### 4.3 Microbe–Host Interactions

There are usually two players across an immunologic interface: self and non-self. Much work has been done to explain the regulation on the side of the self, but the other side is just as important. The diversity in microbes and other potential dangers that the immune system encounters routinely does not only constitute a risk, but it also serves to prime and fine-tune immune responses in an organism-wide context. The targets of the immune system thus play a dual role as regulators of its activity parameter (Fig. 7). In healthy human individuals, microbes colonize different habitats according to their adaptation [108]. In the above study, analysis of a sample's ecological composition, isolation, and sequencing of 16S rRNAs (a ubiquitous constituent of prokaryotic ribosomes) are routinely used to differentiate phyla. PCA along the sample axis revealed clustering according to the habitat of origin (skin, mouth, stool, and others). In addition, the authors reconstructed networks for habitat similarity and correlation of presence per habitat.

A habitat that has traditionally been of particular interest for research is the human gut microbiome. It has been linked to



**Fig. 7** Both host and pathogen factors must be analyzed to understand their mutual interaction. Not only do pathogenic and nonpathogenic colonizing organisms differ in their gene expression, but they may also induce distinct changes in the expression pattern of their human host. By comparing high-throughput data from both sources, it is possible to identify the crucial effectors and interactions that mediate the struggle. Further experiments can then elucidate ways to recognize the differences for diagnosis or manipulate them via therapeutic intervention

multiple immune-related observations, like blood group classification and autoimmune tolerance. In a study of colorectal carcinogenesis in mice, Arthur and coworkers [109] used multidimensional scaling and analysis of similarity to compare the prokaryote composition in the intestinal lumen between wild-type and IL-10 knockout mice. They observed a difference in cluster structure and a reduction in diversity in the knockout group, both apparently the result of the chronic inflammation evoked by the absence of IL-10. As the majority of knockout strains developed facultative colitis that progresses, under pharmacological challenge, to colorectal tumors, the authors compared the impact of different *Escherichia coli* strains on disease progression. They found that a strain lacking the genetic island responsible for producing the genotoxic substance colibactin induced fewer neoplastic lesions during pharmacological challenge. The association between colibactin and carcinogenesis was also observable in human patient cohorts.

It has been reported that anti-apoptotic signals are upregulated after macrophages phagocyte *Staphylococcus spp.* in a germ-induced process that interferes with the supposed chain of events and helps the parasite proliferate in a protected compartment [110]. SAM of the accompanying microarray data revealed *MCL1* as the most promising candidate for regulation, which was confirmed by quantitative PCR [110]. Similarly, a highly pathogenic

*Leishmania* strain did not induce IL-12 secretion in the host macrophage, thereby avoiding the mounting of an immune response [111]. The elucidation of the mechanism behind this finding was facilitated by FDR-corrected ANOVA of microarray recordings that implied a lack of activated interferon signaling. Successful defense against *M. tuberculosis* is linked to vitamin D levels [112]. In a work that focused on the macrophage transcriptional response to 1,25-dihydroxyvitamin D, the authors also examined the cross talk between macrophages and epithelial cells during the course of infection, pointing to a broader role for vitamin D in human immunity [113]. In a study on the eukaryotic parasite *Toxoplasma gondii*, microarray analysis corroborated the finding that the host cell transcription factor c-Myc is upregulated in fibroblasts shortly after infection [114].

Other studies have focused on the identification of biomarkers in microbe-associated diseases. A set of eight miRNAs came up as possible candidates in a study of septic conditions which used pathway enrichment analysis and a protein–protein interaction network of the relevant miRNA targets to visualize the impact of its findings [115].

---

## 5 Mathematical Modeling of Immune-Related Pathways and Systems

Mathematical modeling has been employed for decades as a tool for investigating complex and tightly regulated physicochemical systems containing regulatory motifs. Regulatory motifs are small subsystems displaying structural complexity that provides them with nonlinear, often non-intuitive dynamical behavior, the simplest but most pervasive of them being the feedback loop. Biological systems are in one sense a more complex instance of these systems. The regulation of basic cell processes like differentiation or proliferation is controlled by complex biochemical networks, composed of large numbers of molecular species, such as transcription factors, interacting proteins, as well as protein-coding and noncoding RNAs. These networks are enriched in different kinds of regulatory motifs, such as positive and negative feedback loops, and coherent and incoherent feedforward loops. In addition, clusters and hubs are also common and central in the architecture of these biological networks.

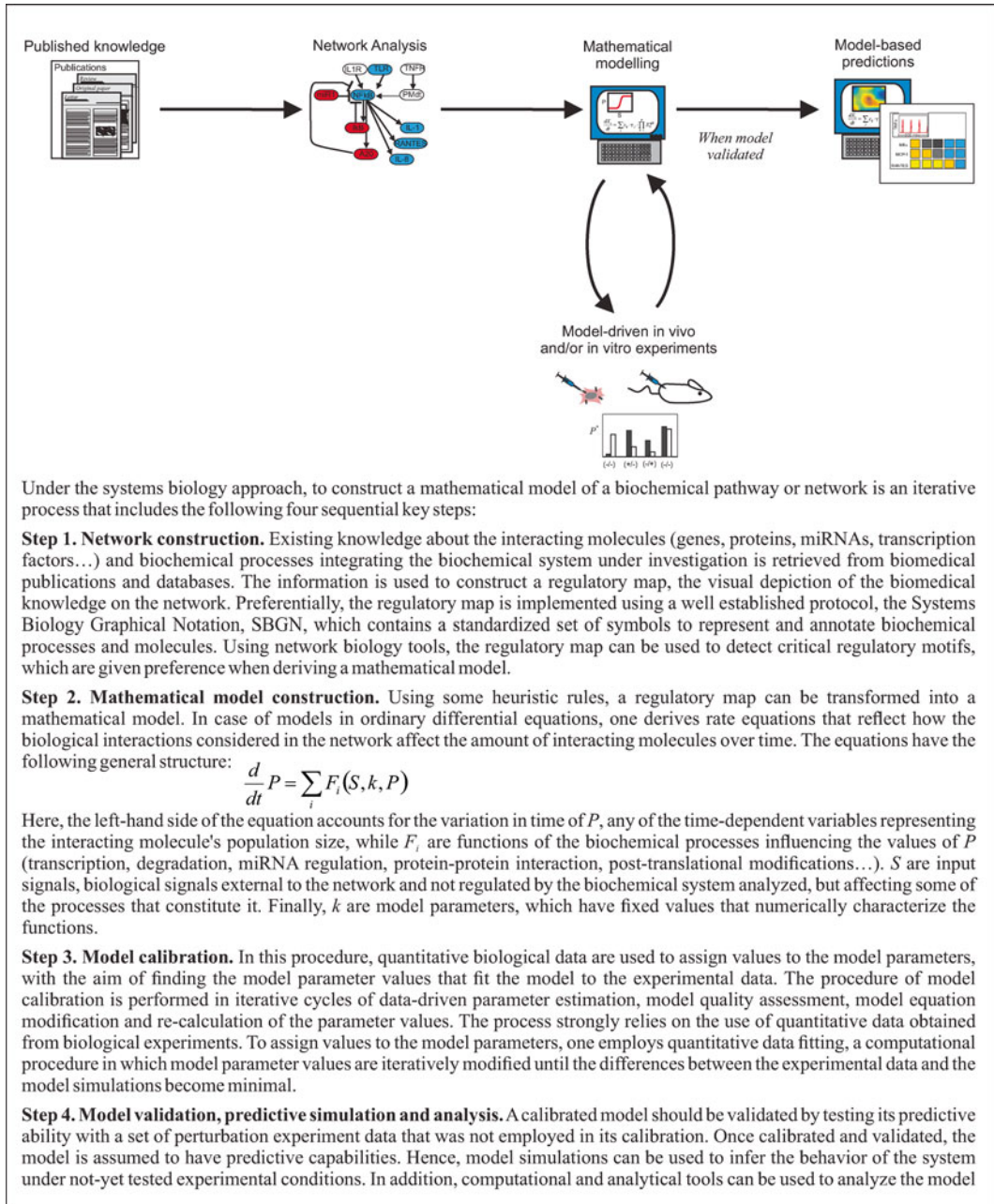
Immune cells are not an exception to this rule. Over the last decade, accumulated evidence has shown that the activation of key immune cell phenotypes is controlled by biochemical networks enriched in different types of regulatory motifs, which very often cross-talk or synergize to regulate immune-related gene expression programs. When trying to find a mechanistic interpretation on the structure of these large networks or integrating large amounts of quantitative data accounting for their regulation, human intuition

and conventional data analysis fail because there are too many interacting variables [1, 116]. In immunology, this problem scales up one step because immune cells are often engaged in complex networks of cell-to-cell interactions becoming interesting examples of multi-level regulatory loops. Under these circumstances, mathematical modeling is proving to be a powerful means to analyze biochemical or cell-to-cell networks, to integrate multiscale data, and to support the generation of hypotheses and the design of experiments (*see* Textbox 2). In the following sections, we are going to illustrate this idea with several examples from literature.

### 5.1 Modeling the Networks Regulating Immune Cells

*T cells.* Several computational models have been developed to investigate positive and negative feedback loops that contribute to the response variability of T cells. Chan et al. [117] used a differential equations model to elucidate the role of phosphatase/kinase feedbacks in T cell receptor (TCR) discrimination. The model showed that positive feedback of protein phosphorylation events activated by the TCR results in **hysteresis**, a nonlinear phenomenon enabling the TCR signaling to act as a **bistable switch**. This result explained the previous counterintuitive observation that low-affinity TCR engagement can significantly sustain T cell signaling. Altan-Bonnet and Germain [118] developed a mathematical model and used it to simulate the competition between two feedback loops: a digital positive feedback loop based on ERK activity and an analog negative feedback involving SHP1 in TCR ligand discrimination. According to their simulations, the authors found that the response of ERK phosphorylation of T cells displays a digital behavior (i.e., on or off), which is critical for defining a sharp TCR ligand-discrimination threshold. Informed by the model predictions, subsequent experimental work identified key parameters (i.e., concentrations of key effectors, such as SHP1 and CD8) that can quantitatively tune the immune response in either a digital or analog way [119]. By combining modeling and experimentation, the authors found that Ras activation in lymphoid cells displays the features of digital-like signaling and hysteresis as a result of an SOS-mediated positive feedback loop. They claim that these characteristics, together with analog signaling of Ras activation via Ras-GRP, enable the efficient and varied cellular responses of T lymphocytes to invading pathogens [120, 121]. Hong et al. [122] modeled differentiation of naive CD4<sup>+</sup>T cells into different effector T cell subsets via upregulation of phenotype-specific TFs whose activation is governed by two positive feedback loops. The study showed how the loops between the TFs allow for the differentiation of a particular subset of T effectors. More recently, mathematical modeling of SHP1-mediated feedback loop for early T cell activation provided a quantitative explanation for the





**Textbox 2** Road map to modeling in systems biology

discrimination between self and foreign peptides in the early immune response [123].

Mathematical modeling has also been used to decode the control of phosphorylation events during TCR signaling. Mukhopadhyay et al. [124] developed a model that explicitly accounts for the regulation of TCR signaling by the CD3 phos-

phorylation cascade, LCK, ZAP70, and CD45. The model showed that the switch-like response of TCR could arise from the multiple phosphate acceptor sites on CD3 $\zeta$ , sequential phosphorylation of these sites by LCK, and protection from dephosphorylation by ZAP70. In another recent contribution, proteomics and mathematical modeling were integrated to unravel the positive role of the phosphatase SHP1 and the shortcut recruitment of the actin regulator WAS in TCR signaling [125].

*B cells.* A series of studies have proven that mathematical modeling of population kinetics using in vivo bromodeoxyuridine (BrdU) labeling data is a powerful tool to study B cell population dynamics during an immune response [126–128]. In the B cell lineage, a mathematical model helped to elucidate how gene regulatory networks allow the intensity of B cell receptor (BCR) signaling to modulate cell fate. The results showed that the strength of BCR stimulation determined the levels and dynamics of IRF4 expression, and that modulation of these expression levels was sufficient to predict distinct cell fate transitions of B cells [129]. Stochastic modeling of B cell differentiation after stimulation was used to explain how the allocation of B cells to distinct cell fates is achieved to ensure proper immune regulation [130]. A computational model of B cell migration between light and dark zones in germinal centers showed that compared with random migration the observed directed migration has a strong impact on arrival time in the light zone [131]. Meyer-Hermann and coauthors proved that mathematical modeling is the right tool to elucidate the B cell selection mechanisms happening in the well-characterized germinal center [132, 133]. According to their model, a theory of germinal center B cell selection, division, and exit was concluded. Particularly, B cells selected on the basis of successful antigen processing always return to the dark zone for asymmetric division, and the antigen-retaining B cells differentiate into plasma cells and leave the germinal center through the dark zone [133].

*NK cells.* Mathematical methods have been used to analyze the regulation of NK cells by various processes including education/licensing, priming, integration of positive and negative signals through an array of activating and inhibitory receptors, and the development of memory-like functionality (reviewed in Ref. 134). Kaplan et al. [135] used an agent-based model to simulate the NK cell immunological synapse, a transient structure with which NK cells select and kill susceptible target cells, regulated by signals from activating and inhibitory receptors. The simulation confirmed the experimental observation that every bound inhibitory receptor acts on activating receptors within a certain radius around it to regulate the NK cell immune synapse. By



using a mathematical model, Almeida et al. [136] substantiated several experimental findings on NK regulation, such as the observed delay between the formation of NK-target cell conjugates and target cell lysis is required for the activation or priming of NK cells before initiating their cytotoxic response. Mathematical modeling together with experiments was employed to show the crucial role of switch-like regulation of VAV1 phosphorylation by SRC-family kinases in determining the cytotoxic activity of NK cells [137].

## **5.2 Cytokine Dynamics and the Regulation of the Immune Response**

Understanding how the immune system decides between tolerance and activation by antigens requires addressing cytokine regulation as a highly dynamic and modulatory process. Mathematical modeling of CD4<sup>+</sup> T cell differentiation has been used in the last decade to link the dynamics of cytokines (IL-4 and IFN $\gamma$ ) and TFs (TBET or GATA3) to the phenotypic composition of Th1 and Th2 cells during differentiation and reprogramming [138–140]. Using *in silico* analysis and modeling, a positive feedback loop via IL-2 signaling was proposed to mediate a digital switch for the proliferation of Th cells and also to function as an analog amplifier for the IL-2 uptake capacity of Treg cells [141]. Single-cell quantification of IL-2 response in a population of T cells unraveled a mechanism by which heterogeneous IL-2 receptor expression allows plasticity of T cells in immune response [142]. Logical modeling of the merged TCR and IL-2 receptor signaling pathways elucidated the potential cross talk between the two pathways in T cell reprogramming [143]. By using *in silico* modeling of the dynamic interplay between T cells, Treg cells, and IL-2, Khailaie delineated three regimes of adaptive immune activation, which depend on the levels of antigen stimulation: (1) a subcritical stimulation, insufficient for pathogen clearance; (2) a threshold stimulation inducing a proper immune response; and (3) an overcritical stimulation leading to chronic coexistence of antigen and immune activity [144]. Through a combination of experimental and computational analyses, Tkach and coauthors found a tight correlation between antigen load and IL-2 accumulation, which was regulated by two IL-2 feedback loops; this system ensures that large pathogenic challenges are communicated through large IL-2 availability in a way a robust immune response is generated [145]. A deterministic model was established to elucidate the role of IL-7 in maintaining naive T cell homeostasis in healthy adults and children [146, 147]. A mathematical model-driven analysis of clinical data revealed the potential for IL-7 to achieve sustained CD4<sup>+</sup> T cell restoration with limited IL-7 exposure in HIV1-infected patients which experienced immune failure despite anti-retroviral therapy [148].

### **5.3 Modeling the Tumor-Immunity Interaction and Anticancer Immunotherapies**

Mathematical models are also useful for discovering and clarifying interactions between the immune system and cancer cells. Many types of modeling approaches, such as **ODE**, **PDE**, and **agent-based models**, have been employed.

Spatial-oriented approaches like PDE and agent-based models have been used to investigate the spatial features of the interaction between the immune system and a tumor. A mathematical model of the spatial interactions of macrophages, tumor cells, and normal tissue cells showed that normal tissue is susceptible to the introduction of mutant cells despite the ability of macrophages to kill these cells (i.e., an immune response), and also that the composition of the resulting tumor can be significantly altered by the mutant cells [149]. A mathematical model describing the spatiotemporal dynamics of a solid tumor in vivo under the control of cytotoxic T lymphocytes was used to identify critical processes that determine the complex behavior of the immune–tumor interacting system [150]. A two-compartment model describing the tumor-immunity interaction at the tumor site and at the draining lymph node was developed to simulate the capability of cytotoxic T lymphocytes to eliminate incipient micro-tumors before clinical detection [151].

On the other hand, ODE-based models have been preferentially used to elucidate the details of the temporal dynamics of the tumor-immunity interaction, but also to assess, explore, and combine immunotherapies for cancer. A concept mathematical model for guiding the development of combination therapies such as immunotherapy, vaccine, and chemotherapy treatments was established to illustrate situations for which neither chemotherapy nor immunotherapy alone are sufficient to control tumor growth. Interestingly, the results indicate that a well-designed therapy combination is able to deplete tumors [152]. Sensitivity analysis of an ODE model of dendritic cell vaccination in melanoma revealed patient-specific parameters that have the greatest impact on treatment efficacy [153]. Joshi and coauthors developed a mathematical model to describe the interaction of cancer cells with CD8<sup>+</sup> cytotoxic T lymphocytes and professional antigen-presenting cells (APCs) in relatively small and multicellular tumors [154]. The model simulations indicated that active vaccination with tumor-antigen-pulsed APCs is more effective than adoptive immunotherapy protocols in inhibiting tumor growth and recurrence. A personalized mathematical model was developed to predict tumor response based on data collected during early phases of the treatment for prostate cancer patients and applied to real-time treatment personalization [155]. In order to understand how a combined treatment contributes to tumor clearance, Wilson and Levy [156] modeled the effects of anti-TGF $\beta$  treatment when used in conjunction with vaccine-induced tumor cytotoxicity for sup-

pressing tumor growth. Glauche and coworkers developed a mathematical model to study treatments against human chronic myeloid leukemia, CML, based on the combination of tyrosine kinase inhibitors and IFN $\alpha$  [157]. With the help of model simulations, the authors suggested a combination treatment in CML patients that requires the simultaneous administration of both molecular species in overlapping time intervals for further clinical implementation. Parra-Guillen and coworkers implemented a model to characterize the mechanisms implied in tumor growth dynamics after the administration of a CyaA-E7 vaccine, which is able to trigger a potent immune response through targeting antigens to DCs [158]. The model analysis indicated that the combination with IL-12 increases the antitumor effect of the vaccine.

#### **5.4 The Modeling of the Host–Pathogen Interaction Dynamics**

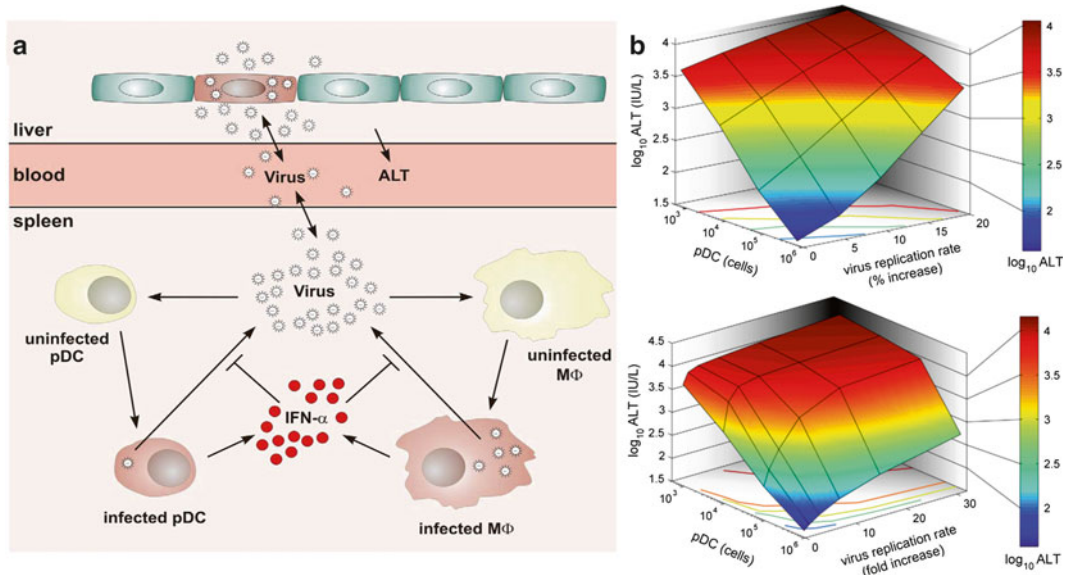
Mathematical modeling has proven to be a useful tool to understand the intricate interaction between the host and the pathogen in the course of an infection. By modeling hepatitis B virus dynamics during the acute stages of the infection, Ciupe and coauthors found that a cell-mediated immune response plays an important role in controlling the virus after the peak in viral load [159]. Binder and coworkers developed a detailed mathematical model of the initial dynamic phase of the intracellular RNA replication of the hepatitis C virus (HCV) [160], an infection that becomes chronic in a high percentage of patients. They found that several host factors involved in the formation of a protective replication compartment determine cellular permissiveness to HCV replication. In silico modeling of human immunodeficiency virus infection elaborated dynamics of immune escape and helped the authors to interpret longitudinal data of the infection [161]. A two-compartment model that quantifies the interplay between influenza A virus infection and adaptive immunity was developed to predict, assess, and design prophylactic and therapeutic strategies against the infection [162]. A mechanistic model was used to describe the expansion, trafficking, and disappearance of activated virus-specific CD8<sup>+</sup> T cells in lymph nodes, spleens, and lungs of mice during primary influenza A infection [163].

A series of studies were carried out to enable an understanding of fundamental characteristics of host immune response during *Chlamydia trachomatis* infection [164–166]. Integrative modeling of the response of pulmonary macrophages to the porcine respiratory and reproductive syndrome virus infection was used to identify the immune mechanism determining the infection duration, and to explore the variability in pathogen virulence and host susceptibility [167]. To better understand the dynamics of *Mycobacterium tuberculosis* (Mtb) infection and immunity, Marino and coworkers developed an ODE model which quantitatively

characterizes the cellular and cytokine control network during the infection [168]. The model was able to reproduce typical disease progression scenarios including primary infection, latency, and clearance; furthermore, it was able to predict key processes determining different disease trajectories. A Boolean model was built to study the host–pathogen interactome during Mtb infection. The model simulations indicated that the clearance of the bacteria is impaired by suppressing some processes such as phagocytosis and phagolysosome fusion or cytokines such as  $\text{TNF}\alpha$  and  $\text{IFN}\gamma$ , while removing cytokines such as IL-10 alongside bacterial defense proteins such as SapM can greatly favor the clearance [169]. In Carbo et al. [170], the authors developed a computational model to investigate the role of IL-21 in the maintenance of effector  $\text{CD4}^+$  T cell responses during chronic *Helicobacter pylori* infection. Their results indicated that IL-21 regulates Th1 and Th17 effector responses during chronic infection in a STAT1- and STAT3-signaling-dependent manner, therefore playing a major role controlling *Helicobacter pylori* infection and gastritis.

### **5.5 Multi-Scale Modeling in Immunology**

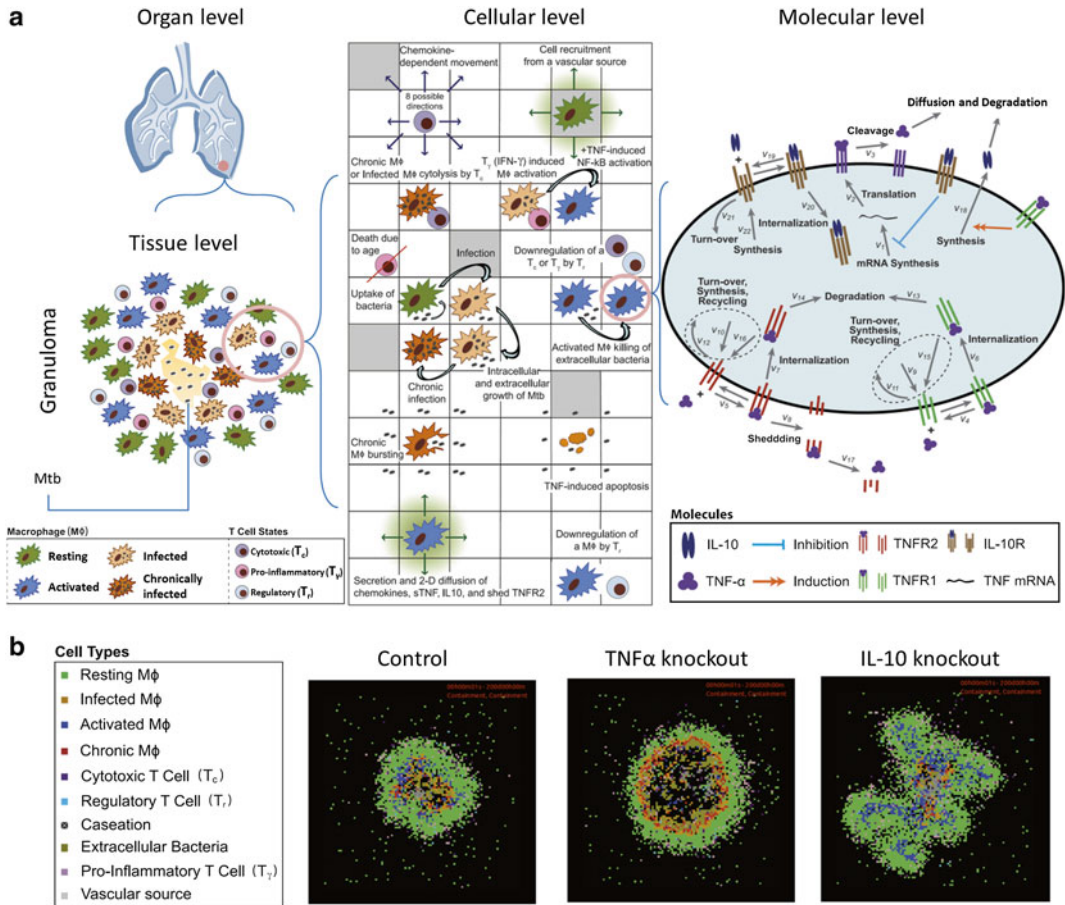
Simulating the immune system across multiple biological scales in space and time is increasingly being recognized as a powerful tool for data integration, refinement of hypotheses, and experimental design. One can find several instances of this strategy in immunology. Multi-scale modeling of plasmacytoid-DC-mediated protection against mouse hepatitis virus infection was used to show that the plasmacytoid DC population in spleen ensures a robust protection against virus variants, which substantially down-regulates  $\text{IFN}\alpha$  secretion [171] (Fig. 8). A multi-scale model comprised of the target organ, where the immune response takes place, circulating blood, lymphoid T, and lymphoid B tissue, was built to describe qualitative predictions of global immune system responses arising from tuberculosis infection, tumor rejection, or a blood-borne pathogen [172]. Multi-scale modeling has proven as well to be a good strategy to simulate the immune responses under Mtb infection (Fig. 9a). Fallahi-Sichani et al. [173, 174] developed a multi-scale model, which describes the dynamics of granuloma formation and the role of TNF receptor signaling and trafficking. Granulomas are clusters of immune cells and mycobacteria formed in the course of the immune response to Mtb infection, and they are essential to control the infection. Model analysis showed that the spatial organization of immune cells as well as certain molecular level mechanisms (i.e., TNF–TNF receptor binding) are significant factors for regulating bioavailability in granulomatous tissue. Subsequently, the model was extended by considering key anti- and pro-inflammatory mediators (IL-10 and  $\text{TNF}\alpha$ , respectively) that are elicited during the host immune response to Mtb. The



**Fig. 8** (a) Scheme of plasmacytoid-DC-mediated protection against mouse hepatitis virus infection. The multi-scale model considered processes such as virus replication, target cell turnover, and IFN $\alpha$  decay, as well as the production of virus and IFN $\alpha$  by infected macrophages (M $\Phi$ ) and plasmacytoid dendritic cells (pDC). (b) The simulation results show that splenic pDCs protect (denoted by the number of pDCs in the spleen) against severe disease (denoted by the level of alanine transaminase, abbreviated as ALT) for an up to 30-fold increase in the viral replication rate in splenic macrophages (*bottom*). However, such a protection is less efficient for a global increase of viral replication rate in the liver (*top*). The figure is adapted from Bocharov et al. [171]

model showed the importance of the balance between TNF $\alpha$  and IL-10 for the formation and maintenance of granulomas [175] (Fig. 9b). Paiva and coworkers [176] developed a multi-scale mathematical model to study how the immune response interferes with oncolytic virotherapy, which utilizes viruses that specifically kill tumor cells to treat cancer. The simulations suggested that reprogramming the immune microenvironment in tumors (such as in situ virus-mediated impairing of CD8<sup>+</sup> T cell motility, or blockade of B and T lymphocyte recruitment) can substantially enhance the effects of the therapy [176]. Dwivedi and coauthors [177] developed a multi-scale system of IL-6-mediated immune regulation in Crohn's disease. The model integrated intracellular signaling with organ-level dynamics of pharmacological markers underlying the disease. Model-based analysis suggested a dual-targeting strategy for suppressing pharmacological markers of Crohn's disease.





**Fig. 9 (a)** Overview of the multi-scale modeling of tuberculosis granulomas in the lung deployed in Cilfone et al. [175]. After Mtb reaches the lung, immune cells, such as macrophages, take up the bacteria and travel to the draining lymph node, leading to T cell priming. The T cells travel through the blood to the infected site, resulting in the formation of granulomas. During this process, molecular events, such as IL-10 and TNF receptor binding, determine immune cell behavior (e.g., cell death, survival, or activation) and cell behavior influences tissue-level events (e.g., T cell priming and granuloma formation). **(b)** Simulations of granuloma development in different biological scenarios. After constructing the model using the known interactions and behaviors of immune cells and the binding and trafficking reactions of TNF $\alpha$  and IL-10, the authors simulated the development of granuloma in three scenarios: control, TNF $\alpha$  knockout, and IL-10 knockout. In comparison to the control scenario, knockout of TNF $\alpha$  or IL-10 resulted in increased upload of bacteria and an irregular shaped granuloma, respectively. For detailed descriptions of the rules defined in the model and the model equations, the reader is referred to [175], from where this figure was taken and modified

## 6 Outlook: Great Expectations in the Integration of Systems Approaches in the Future of Immunology

In this chapter, we support the idea that the immune system is a paradigmatic case of a biological system of systems, a multi-level ensemble of specialized biological entities that interact to create the complex and adaptive system we know as immunity. However, the deep under-

standing of the structure, function, and organization of this special kind of multi-scale system demands the use of systems biology, a methodological approach that combines quantitative experimental data, computational biology, and mathematical modeling.

We here discussed three avenues in which systems biology has evolved over the last decade: (a) that of advanced immunoinformatics; (b) the one involving the generation and analysis of immune-related omics data; and finally (c) the one that employs mathematical modeling to investigate the structure and function of biochemical networks, and their abilities to modulate phenotypes. However, none of them alone is a definitive approach to tackle the complexity behind the regulation of the immune system.

Advanced immunoinformatics is providing us with a plethora of excellent methods and databases for the fine-grained analysis of antigen-mediated immunogenicity. However, without considering how the antigen–antibody interaction participates in the regulation of the large biochemical networks underlying the immune response, their results are nothing but a kind of computationally sophisticated reductionism. The finding of omics-derived gene signatures accounting for the regulation of critical disease-related phenotypes from *in vitro*, *in vivo*, or clinical data has been considered the Holy Grail of modern biomedicine for decades. Yet without experimental elucidation of the biological mechanisms behind them, most of those signatures are not at all the true endpoint of the story in many diseases. Full-detail, multi-scale mathematical models describing the communication among immune cells in connection with the regulation of their inner biochemical networks promises to be a better approach to elucidate the mechanism governing immunity. However, we have to assume that it could take decades to establish the computational methods and experimental technologies necessary to feed those large and detailed mathematical models with enough quantitative data. A realistic approach should integrate tools from these three avenues to obtain a more global perspective on the structure, regulation, and phenotypic response constituting immunity. The details of such an integrative approach are yet to be developed.

---

## Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) as part of the projects eBio:miRSys [0316175A to J.V. and B.S.], and eMed:CAPSyS [01ZX1304F to J.V. and B.S.]. Julio Vera is funded by the Erlangen University Hospital (ELAN funds, 14-07-22-1-Vera-González) and the German Research Foundation (DFG) through the project SPP 1757/1 (VE 642/1-1 to J.V.).

---

## 7 Glossary

**Normalization in differential gene expression analysis** Analyses almost always combine measurements from multiple microarray chips or NGS runs which will differ in parameters that dictate background strength and signal ranges for each particular feature. To make those measurements comparable with each other, inter-measurement normalization techniques have been developed. The most widely employed is probably quantile normalization, an algorithm that orders features by their expression value separately for each sample, averages the expression for each rank across all samples, and redistributes the averages to the original feature position in each sample. In many software tools, the “normalization” step offers additional computations, often including the transformation into the logarithmic scale (see below).

**(Log-)transformation** The distribution of signal values in a sample covers several orders of magnitude. Usually, very low or undetectable (i.e., zero) signals are highly overrepresented while a few signals are found as upper outliers, and both skew the gross signal distribution. To remedy this and make the distribution more malleable to established methods of statistical testing, a simple option is transforming the data into log scale. Although the base of the logarithm is an arbitrary choice, base 2 is used in many software tools to make fold-change calculations more intuitive.

**Moderation** In nontechnical terms, moderation is a statistical method that manipulates a distribution in a way that is conceived to yield better results. In the case of microarrays, moderation can be used to decrease the relative difference in feature variances so that the conditions for further statistical tests are (better) met.

**False discovery rate (FDR)** The FDR is the fraction of falsely called features in a set of called features. For example, if differential expression analysis yields a list of 100 called features but only 90 of them can be validated experimentally, the empiric FDR is 10 %.

**Correction for multiple testing/control of the FDR** For a differential expression analysis on a set of 50,000 features, 50,000 tests must be performed. At a significance level of 0.05, we expect 2500 of those test results to be true just by chance. Such a large number of false positives will drown out the true positive features with no direct way to distinguish between both. There are ways for dealing with this unfortunate consequence by adjusting the test scenario in such a way that the proportion of falsely called features in the results is statistically limited (“controlled”).

**Correlation network** A network that is based on—and perhaps spatially arranged according to—the correlation values between its nodes.

**ODE model** A system of mechanistic ordinary differential equations that determine the temporal state of the corresponding system of biochemical reactions.

**PDE model** A system of partial differential equations is a quantitative description of how a biochemical system changes in space and time.

**Agent-based model** Agent-based modeling is a rule-based, discrete-event, and discrete-time computational modeling methodology that employs abstract objects and focuses on the interactions among the individual objects (i.e., agents) of a system.



**Boolean model** A Boolean network consists of a set of nodes whose state (0 or 1) is determined by linking other nodes in the network through Boolean functions, such as AND and OR.

**Deterministic model** A dynamic system is deterministic if its trajectory is uniquely determined by the initial state and a given parameter set.

**Stochastic model** In contrast to deterministic models, for which the output will be identical using the same initial state and parameter set as the input, a stochastic system, at a given initial state in the phase space, can end with different states with different probabilities. In other words, the same input given to a stochastic system several times can result in different outputs.

**Sensitivity analysis** Determining the change in model variables (such as concentrations of molecular species) influenced by changes in parameter values (such as velocities of biochemical reactions).

**Bifurcation analysis** An analysis that shows how the qualitative behavior (e.g., the loss of stability and appearance of sustained oscillations) of a model changes as a function of critical model parameters.

**Bistability** For a certain set of parameters, a system has two stable steady states that are separated by an unstable steady state.

**Hysteresis** A hallmark of bistability: as a critical parameter (i.e., bifurcation parameter) increases beyond a particular value, the system jumps to an alternative steady state from the original steady state; then, if the parameter decreases below the value, the system jumps back to the original steady state.

**In silico** An expression used to mean “performed” on a computer or via computer simulation.

## References

- Vera J, Wolkenhauer O (2008) Chapter 17: a system biology approach to understand functional activity of cell communication systems. *Methods Nano Cell Biol* 90:399–415. doi:10.1016/s0091-679x(08)00817-0
- Wolkenhauer O, Auffray C, Baltrusch S et al (2009) Systems biologists seek fuller integration of systems biology approaches in new cancer research programs. *Cancer Res* 70:12–13. doi:10.1158/0008-5472.can-09-2676
- Wolkenhauer O, Auffray C, Jaster R et al (2013) The road from systems biology to systems medicine. *Pediatr Res* 73:502–507. doi:10.1038/pr.2013.4
- Vera J, Gupta SK, Wolkenhauer O, Schuler G (2014) Envisioning the application of systems biology in cancer immunology. *Cancer Immunol* 429–449. doi:10.1007/978-3-662-44006-3\_23
- Voit E (2012) *A first course in systems biology*, 1st edn. Garland Science, New York
- Mesarović MD (1968) Systems theory and biology—view of a theoretician. In: Mesarović MD (ed) *Syst theory biol*. Springer, Berlin, pp 59–87
- Savageau MA, Rosen R (1976) *Biochemical systems analysis: a study of function and design in molecular biology*. Addison-Wesley, Reading, MA
- Yan Q (2010) Immunoinformatics and systems biology methods for personalized medicine. *Syst Biol Drug Discov Dev* 662:203–220. doi:10.1007/978-1-60761-800-3\_10
- Kidd BA, Peters LA, Schadt EE, Dudley JT (2014) Unifying immunology with informatics and multiscale biology. *Nat Immunol* 15:118–127. doi:10.1038/ni.2787
- Gupta SK, Gupta SK, Smita S et al (2011) Computational analysis and modeling the effectiveness of “Zanamivir” targeting neuraminidase protein in pandemic H1N1 strains. *Infect Genet Evol* 11:1072–1082. doi:10.1016/j.meegid.2011.03.018
- Vera J, Schmitz U, Lai X et al (2013) Kinetic modeling-based detection of genetic signatures that provide chemoresistance via the E2F1-p73/DNp73/mbox-miR-205 network. *Cancer Res* 73:3511–3524. doi:10.1158/0008-5472.can-12-4095

12. Blythe MJ, Doytchinova IA, Flower DR (2002) JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* 18:434–439
13. Rammensee H-G (2003) Immunoinformatics: bioinformatic strategies for better understanding of immune function. Introduction. *Novartis Found Symp* 254:1–2
14. Brusica V, Petrovsky N (2003) Immunoinformatics—the new kid in town. *Novartis Found Symp* 254:3–13, discussion 13–22, 98–101, 250–2
15. Tomar N, De RK (2010) Immunoinformatics: an integrated scenario. *Immunology* 131:153–168. doi:10.1111/j.1365-2567.2010.03330.x
16. Tomar N, De RK (2014) Immunoinformatics: a brief review. *Methods Mol Biol* 1184:23–55. doi:10.1007/978-1-4939-1115-8\_3
17. Greenbaum JA, Andersen PH, Blythe M et al (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J Mol Recognit* 20:75–82. doi:10.1002/jmr.815
18. Tong JC, Ren EC (2009) Immunoinformatics: current trends and future directions. *Drug Discov Today* 14:684–689. doi:10.1016/j.drudis.2009.04.001
19. Saha S, Bhasin M, Raghava GPS (2005) Bcipep: a database of B-cell epitopes. *BMC Genomics* 6:79. doi:10.1186/1471-2164-6-79
20. Bui H-H, Peters B, Assarsson E et al (2007) Ab and T cell epitopes of influenza A virus, knowledge and opportunities. *Proc Natl Acad Sci U S A* 104:246–251. doi:10.1073/pnas.0609330104
21. Müller GM, Shapira M, Arnon R (1982) Anti-influenza response achieved by immunization with a synthetic conjugate. *Proc Natl Acad Sci U S A* 79:569–573
22. Naruse H, Ogasawara K, Kaneda R et al (1994) A potential peptide vaccine against two different strains of influenza virus isolated at intervals of about 10 years. *Proc Natl Acad Sci U S A* 91:9588–9592
23. El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting protective linear B-cell epitopes using evolutionary information. 2008 IEEE Int Conf Bioinforma Biomed. doi:10.1109/bibm.2008.80
24. Sollner J, Grohmann R, Rapberger R et al (2008) Analysis and prediction of protective continuous B-cell epitopes on pathogen proteins. *Immunome Res* 4:1. doi:10.1186/1745-7580-4-1
25. Vita R, Overton JA, Greenbaum JA et al (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 43:D405–D412. doi:10.1093/nar/gku938
26. Odorico M, Pellequer J-L (2003) BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J Mol Recognit* 16:20–22. doi:10.1002/jmr.602
27. Ponomarenko J, Bui H-H, Li W et al (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* 9:514. doi:10.1186/1471-2105-9-514
28. Saha S, Raghava GPS (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* 65:40–48. doi:10.1002/prot.21078
29. Sweredoski MJ, Baldi P (2009) COBEpro: a novel system for predicting continuous B-cell epitopes. *Protein Eng Des Sel* 22:113–120. doi:10.1093/protein/gzn075
30. Larsen J, Lund O, Nielsen M (2006) Improved method for predicting linear B cell epitopes. *Immunome Res* 2:2. doi:10.1186/1745-7580-2-2
31. Kringelum JV, Lundegaard C, Lund O, Nielsen M (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol* 8, e1002829. doi:10.1371/journal.pcbi.1002829
32. Hamelryck T (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins* 59:38–48. doi:10.1002/prot.20379
33. Sweredoski MJ, Baldi P (2008) PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics* 24:1459–1460. doi:10.1093/bioinformatics/btn199
34. Bublil EM, Freund NT, Mayrose I et al (2007) Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm. *Proteins* 68:294–304. doi:10.1002/prot.21387
35. Zhang W (2012) Bpredictor (<https://code.google.com/p/my-project-bpredictor/>). Accessed 1 Apr 2015
36. Zhang W, Xiong Y, Zhao M et al (2011) Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. *BMC Bioinformatics* 12:341. doi:10.1186/1471-2105-12-341

37. Moreau V, Granier C, Villard S et al (2006) Discontinuous epitope prediction based on mimotope analysis. *Bioinformatics* 22:1088–1095. doi:[10.1093/bioinformatics/btl012](https://doi.org/10.1093/bioinformatics/btl012)
38. Pizzi E, Cortese R, Tramontano A (1995) Mapping epitopes on protein surfaces. *Biopolymers* 36:675–680. doi:[10.1002/bip.360360513](https://doi.org/10.1002/bip.360360513)
39. Evans MC (2008) Recent advances in immunoinformatics: application of in silico tools to drug development. *Curr Opin Drug Discov Dev* 11:233–241
40. Mayrose I, Shlomi T, Rubinstein ND et al (2007) Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. *Nucleic Acids Res* 35:69–78. doi:[10.1093/nar/gkl975](https://doi.org/10.1093/nar/gkl975)
41. Huang J, Gutteridge A, Honda W, Kanehisa M (2006) MIMOX: a web tool for phage display based epitope mapping. *BMC Bioinformatics* 7:451. doi:[10.1186/1471-2105-7-451](https://doi.org/10.1186/1471-2105-7-451)
42. Mayrose I, Penn O, Erez E et al (2007) Pepitope: epitope mapping from affinity-selected peptides. *Bioinformatics* 23:3244–3246. doi:[10.1093/bioinformatics/btm493](https://doi.org/10.1093/bioinformatics/btm493)
43. Huang YX, Bao YL, Guo SY et al (2008) Pep-3D-search: a method for B-cell epitope prediction based on mimotope analysis. *BMC Bioinformatics* 9:538. doi:[10.1186/1471-2105-9-538](https://doi.org/10.1186/1471-2105-9-538)
44. Schreiber A, Humbert M, Benz A, Dietrich U (2005) 3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins. *J Comput Chem* 26:879–887. doi:[10.1002/jcc.20229](https://doi.org/10.1002/jcc.20229)
45. Huang J, Ru B, Zhu P et al (2012) MimoDB 2.0: a mimotope database and beyond. *Nucleic Acids Res* 40:D271–D277. doi:[10.1093/nar/gkr922](https://doi.org/10.1093/nar/gkr922)
46. Söllner J (2006) Selection and combination of machine learning classifiers for prediction of linear B-cell epitopes on proteins. *J Mol Recognit* 19:209–214. doi:[10.1002/jmr.770](https://doi.org/10.1002/jmr.770)
47. Bhasin M, Raghava GPS (2003) Prediction of promiscuous and high-affinity mutated MHC binders. *Hybrid Hybridomics* 22:229–234. doi:[10.1089/153685903322328956](https://doi.org/10.1089/153685903322328956)
48. Huang L, Dai Y (2006) Direct prediction of T-cell epitopes using support vector machines with novel sequence encoding schemes. *J Bioinform Comput Biol* 4:93–107
49. Parker JM, Guo D, Hodges RS (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry (Mosc)* 25:5425–5432
50. Zhang GL, Petrovsky N, Kwoh CK et al (2006) PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res* 2:3. doi:[10.1186/1745-7580-2-3](https://doi.org/10.1186/1745-7580-2-3)
51. Buus S, Lauemøller SL, Worning P et al (2003) Sensitive quantitative predictions of peptide-MHC binding by a “Query by Committee” artificial neural network approach. *Tissue Antigens* 62:378–384
52. Nielsen M, Lundegaard C, Worning P et al (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* 12:1007–1017. doi:[10.1110/ps.0239403](https://doi.org/10.1110/ps.0239403)
53. Nanni L (2006) Machine learning algorithms for T-cell epitopes prediction. *Neurocomput* 69:866–868. doi:[10.1016/j.neucom.2005.08.005](https://doi.org/10.1016/j.neucom.2005.08.005)
54. Bhasin M, Raghava GPS (2005) Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Res* 33:W202–W207
55. Lapinsh M, Prusis P, Gutcaits A et al (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim Biophys Acta* 1525:180–190
56. Doytchinova I, Flower D (2008) QSAR and the prediction of T-cell epitopes. *CP* 5:73–95. doi:[10.2174/157016408784911945](https://doi.org/10.2174/157016408784911945)
57. Tian F, Lv F, Zhou P et al (2008) Toward prediction of binding affinities between the MHC protein and its peptide ligands using quantitative structure-affinity relationship approach. *Protein Pept Lett* 15:1033–1043
58. Zhao C, Zhang H, Luan F et al (2007) QSAR method for prediction of protein-peptide binding affinity: application to MHC class I molecule HLA-A\*0201. *J Mol Graph Model* 26:246–254. doi:[10.1016/j.jmgm.2006.12.002](https://doi.org/10.1016/j.jmgm.2006.12.002)
59. Kanguane P, Sakharkar MK (2005) T-Epitope designer: a HLA-peptide binding prediction server. *Bioinformation* 1:21–24
60. Zhang W, Niu Y, Xiong Y et al (2012) Computational prediction of conformational B-cell epitopes from antigen primary structures

- by ensemble learning. *PLoS One* 7:e43575. doi:[10.1371/journal.pone.0043575](https://doi.org/10.1371/journal.pone.0043575)
61. Guan P, Doytchinova IA, Zygouri C, Flower DR (2003) MHCpred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res* 31:3621–3624. doi:[10.1093/nar/gkg510](https://doi.org/10.1093/nar/gkg510)
  62. Jojic N, Reyes-Gomez M, Heckerman D et al (2006) Learning MHC I-peptide binding. *Bioinformatics* 22:e227–e235. doi:[10.1093/bioinformatics/btl255](https://doi.org/10.1093/bioinformatics/btl255)
  63. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256:623–644. doi:[10.1006/jmbi.1996.0114](https://doi.org/10.1006/jmbi.1996.0114)
  64. Gupta SK, Singh A, Srivastava M et al (2009) In silico DNA vaccine designing against human papillomavirus (HPV) causing cervical cancer. *Vaccine* 28:120–131. doi:[10.1016/j.vaccine.2009.09.095](https://doi.org/10.1016/j.vaccine.2009.09.095)
  65. Schueler-Furman O, Altuvia Y, Sette A, Margalit H (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci* 9:1838–1846. doi:[10.1110/ps.9.9.1838](https://doi.org/10.1110/ps.9.9.1838)
  66. Brusic V, Rudy G, Honeyman G et al (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* 14:121–130. doi:[10.1093/bioinformatics/14.2.121](https://doi.org/10.1093/bioinformatics/14.2.121)
  67. Doytchinova IA, Guan P, Flower DR (2006) EpiJen: a server for multistep T cell epitope prediction. *BMC Bioinformatics* 7:131. doi:[10.1186/1471-2105-7-131](https://doi.org/10.1186/1471-2105-7-131)
  68. Lundegaard C, Lamberth K, Harndahl M et al (2008) NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res* 36:W509–W512. doi:[10.1093/nar/gkn202](https://doi.org/10.1093/nar/gkn202)
  69. Rammensee H, Bachmann J, Emmerich NP et al (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50:213–219
  70. Sathiamurthy M, Peters B, Bui H-H et al (2005) An ontology for immune epitopes: application to the design of a broad scope database of immune reactivities. *Immunome Res* 1:2. doi:[10.1186/1745-7580-1-2](https://doi.org/10.1186/1745-7580-1-2)
  71. Lefranc M-P, Giudicelli V, Ginestoux C et al (2009) IMGT, the international ImmunoGeneTics information system. *Nucleic Acids Res* 37:D1006–D1012. doi:[10.1093/nar/gkn838](https://doi.org/10.1093/nar/gkn838)
  72. Robinson J, Mistry K, McWilliam H et al (2011) The IMGT/HLA database. *Nucleic Acids Res* 39:D1171–D1176. doi:[10.1093/nar/gkq998](https://doi.org/10.1093/nar/gkq998)
  73. King TP, Hoffman D, Lowenstein H et al (1994) Allergen nomenclature. WHO/IUIS Allergen Nomenclature Subcommittee. *Int Arch Allergy Immunol* 105:224–233
  74. Kim C, Kwon S, Lee G et al (2009) A database for allergenic proteins and tools for allergenicity prediction. *Bioinformation* 3:344–345
  75. Mari A, Scala E, Palazzo P et al (2006) Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. The Allergome platform as a model. *Cell Immunol* 244:97–100. doi:[10.1016/j.cellimm.2007.02.012](https://doi.org/10.1016/j.cellimm.2007.02.012)
  76. Ivanciuc O, Schein CH, Braun W (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res* 31:359–362
  77. Irizarry RA, Hobbs B, Collin F et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–264. doi:[10.1093/biostatistics/4.2.249](https://doi.org/10.1093/biostatistics/4.2.249)
  78. Lim WK, Wang K, Lefebvre C, Califano A (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics* 23:i282–i288. doi:[10.1093/bioinformatics/btm201](https://doi.org/10.1093/bioinformatics/btm201)
  79. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6:65–70
  80. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 57:289–300
  81. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98:5116–5121. doi:[10.1073/pnas.091062498](https://doi.org/10.1073/pnas.091062498)
  82. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:1–25. doi:[10.2202/1544-6115.1027](https://doi.org/10.2202/1544-6115.1027)
  83. Opgen-Rhein R, Strimmer K (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat Appl Genet Mol Biol* 6:1544–6115. doi:[10.2202/1544-6115.1252](https://doi.org/10.2202/1544-6115.1252)
  84. Boulesteix A-L, Slawski M (2009) Stability and aggregation of ranked gene lists. *Brief*

- Bioinform 10:556–568. doi:[10.1093/bib/bbp034](https://doi.org/10.1093/bib/bbp034)
85. Sonesson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14:91. doi:[10.1186/1471-2105-14-91](https://doi.org/10.1186/1471-2105-14-91)
  86. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37:1–13. doi:[10.1093/nar/gkn923](https://doi.org/10.1093/nar/gkn923)
  87. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57. doi:[10.1038/nprot.2008.211](https://doi.org/10.1038/nprot.2008.211)
  88. Croft D, Mundo AF, Haw R et al (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res* 42:D472–D477. doi:[10.1093/nar/gkt1102](https://doi.org/10.1093/nar/gkt1102)
  89. Drághici S, Khatri P, Martins RP et al (2003) Global functional profiling of gene expression. *Genomics* 81:98–104. doi:[10.1016/S0888-7543\(02\)00021-6](https://doi.org/10.1016/S0888-7543(02)00021-6)
  90. Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 102:15545–15550. doi:[10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102)
  91. Franceschini A, Szklarczyk D, Frankild S et al (2013) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41:D808–D815. doi:[10.1093/nar/gks1094](https://doi.org/10.1093/nar/gks1094)
  92. Saris CG, Horvath S, van Vught PW et al (2009) Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC Genomics* 10:405. doi:[10.1186/1471-2164-10-405](https://doi.org/10.1186/1471-2164-10-405)
  93. Beyer M, Mallmann MR, Xue J et al (2012) High-resolution transcriptome of human macrophages. *PLoS One* 7, e45466. doi:[10.1371/journal.pone.0045466](https://doi.org/10.1371/journal.pone.0045466)
  94. Xue J, Schmidt SV, Sander J et al (2014) Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity* 40:274–288. doi:[10.1016/j.immuni.2014.01.006](https://doi.org/10.1016/j.immuni.2014.01.006)
  95. Raza S, Robertson KA, Lacaze PA et al (2008) A logic-based diagram of signalling pathways central to macrophage activation. *BMC Syst Biol* 2:36. doi:[10.1186/1752-0509-2-36](https://doi.org/10.1186/1752-0509-2-36)
  96. Weintz G, Olsen JV, Frühauk K et al (2010) The phosphoproteome of toll-like receptor-activated macrophages. *Mol Syst Biol* 6:371. doi:[10.1038/msb.2010.29](https://doi.org/10.1038/msb.2010.29)
  97. Jin P, Han TH, Ren J et al (2010) Molecular signatures of maturing dendritic cells: implications for testing the quality of dendritic cell therapies. *J Transl Med* 8:4. doi:[10.1186/1479-5876-8-4](https://doi.org/10.1186/1479-5876-8-4)
  98. Chaussabel D, Quinn C, Shen J et al (2008) A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* 29:150–164. doi:[10.1016/j.immuni.2008.05.012](https://doi.org/10.1016/j.immuni.2008.05.012)
  99. Chang HH, Hemberg M, Barahona M et al (2008) Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* 453:544–547. doi:[10.1038/nature06965](https://doi.org/10.1038/nature06965)
  100. Tsang JS, Schwartzberg PL, Kotliaroy Y et al (2014) Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* 157:499–513. doi:[10.1016/j.cell.2014.03.031](https://doi.org/10.1016/j.cell.2014.03.031)
  101. Querec TD, Akondy RS, Lee EK et al (2009) Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat Immunol* 10:116–125. doi:[10.1038/ni.1688](https://doi.org/10.1038/ni.1688)
  102. Vahey MT, Wang Z, Kester KE et al (2010) Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS, S malaria vaccine. *J Infect Dis* 201:580–589. doi:[10.1086/650310](https://doi.org/10.1086/650310)
  103. Pulendran B, Li S, Nakaya HI (2010) Systems vaccinology. *Immunity* 33:516–529. doi:[10.1016/j.immuni.2010.10.006](https://doi.org/10.1016/j.immuni.2010.10.006)
  104. Nakaya HI, Pulendran B (2012) Systems vaccinology: its promise and challenge for HIV vaccine development. *Curr Opin HIV AIDS* 7:24–31. doi:[10.1097/COH.0b013e32834dc37b](https://doi.org/10.1097/COH.0b013e32834dc37b)
  105. Zhang B, Li M, McDonald T et al (2013) Microenvironmental protection of CML stem and progenitor cells from tyrosine kinase inhibitors through N-cadherin and Wnt– $\beta$ -catenin signaling. *Blood* 121:1824–1838. doi:[10.1182/blood-2012-02-412890](https://doi.org/10.1182/blood-2012-02-412890)
  106. Gustafson MP, Lin Y, LaPlant B et al (2013) Immune monitoring using the predictive power of immune profiles. *J Immunother Cancer* 1:7. doi:[10.1186/2051-1426-1-7](https://doi.org/10.1186/2051-1426-1-7)
  107. Ulloa-Montoya F, Louahed J, Dizier B et al (2013) Predictive gene signature in MAGE-A3 antigen-specific cancer immunotherapy. *J Clin Oncol* 31(19):2388–2395. doi:[10.1200/JCO.2012.44.3762](https://doi.org/10.1200/JCO.2012.44.3762)



108. Consortium THMP (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214. doi:[10.1038/nature11234](https://doi.org/10.1038/nature11234)
109. Arthur JC, Perez-Chanona E, Mühlbauer M et al (2012) Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* 338:120–123. doi:[10.1126/science.1224820](https://doi.org/10.1126/science.1224820)
110. Koziel J, Maciag-Gudowska A, Mikolajczyk T et al (2009) Phagocytosis of *Staphylococcus aureus* by macrophages exerts cytoprotective effects manifested by the upregulation of anti-apoptotic factors. *PLoS ONE* 4:e5210. doi:[10.1371/journal.pone.0005210](https://doi.org/10.1371/journal.pone.0005210)
111. Favila MA, Geraci NS, Zeng E et al (2014) Human dendritic cells exhibit a pronounced type I IFN signature following leishmania major infection that is required for IL-12 induction. *J Immunol* 192:5863–5872. doi:[10.4049/jimmunol.1203230](https://doi.org/10.4049/jimmunol.1203230)
112. White JH (2008) Vitamin D signaling, infectious diseases, and regulation of innate immunity. *Infect Immun* 76:3837–3843. doi:[10.1128/IAI.00353-08](https://doi.org/10.1128/IAI.00353-08)
113. Verway M, Bouttier M, Wang T-T et al (2013) Vitamin D induces interleukin-1 $\beta$  expression: paracrine macrophage epithelial signaling controls *M. tuberculosis* infection. *PLoS Pathog* 9, e1003407. doi:[10.1371/journal.ppat.1003407](https://doi.org/10.1371/journal.ppat.1003407)
114. Franco M, Shastri AJ, Boothroyd JC (2014) Infection by *Toxoplasma gondii* specifically induces host c-Myc and the genes this pivotal transcription factor regulates. *Eukaryot Cell* 13:483–493. doi:[10.1128/EC.00316-13](https://doi.org/10.1128/EC.00316-13)
115. Huang J, Sun Z, Yan W et al (2014) Identification of MicroRNA as sepsis biomarker based on miRNAs regulatory network analysis. *Biomed Res Int* 2014:e594350. doi:[10.1155/2014/594350](https://doi.org/10.1155/2014/594350)
116. Vera J, Wolkenhauer O (2011) Mathematical tools in cancer signalling systems biology. *Cancer Syst Biol Bioinforma Med* 185–212. doi: [10.1007/978-94-007-1567-7\\_7](https://doi.org/10.1007/978-94-007-1567-7_7)
117. Chan C, Stark J, George AJT (2004) Feedback control of T-cell receptor activation. *Proc Biol Sci* 271:931–939. doi:[10.1098/rspb.2003.2587](https://doi.org/10.1098/rspb.2003.2587)
118. Altan-Bonnet G, Germain RN (2005) Modeling T cell antigen discrimination based on feedback control of digital ERK responses. *PLoS Biol* 3:e356. doi:[10.1371/journal.pbio.0030356](https://doi.org/10.1371/journal.pbio.0030356)
119. Feinerman O, Veiga J, Dorfman JR et al (2008) Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science* 321:1081–1084. doi:[10.1126/science.1158013](https://doi.org/10.1126/science.1158013)
120. Das J, Ho M, Zikherman J et al (2009) Digital signaling and hysteresis characterize ras activation in lymphoid cells. *Cell* 136:337–351. doi:[10.1016/j.cell.2008.11.051](https://doi.org/10.1016/j.cell.2008.11.051)
121. Chakraborty AK, Das J, Zikherman J, et al. (2009) Molecular origin and functional consequences of digital signaling and hysteresis during Ras activation in lymphocytes. *Sci Signal* 2: pt2. doi: [10.1126/scisignal.266pt2](https://doi.org/10.1126/scisignal.266pt2)
122. Hong T, Xing J, Li L, Tyson JJ (2012) A simple theoretical framework for understanding heterogeneous differentiation of CD4+ T cells. *BMC Syst Biol* 6:66. doi:[10.1186/1752-0509-6-66](https://doi.org/10.1186/1752-0509-6-66)
123. François P, Voisinne G, Siggia ED et al (2013) Phenotypic model for early T-cell activation displaying sensitivity, specificity, and antagonism. *Proc Natl Acad Sci U S A* 110:E888–E897. doi:[10.1073/pnas.1300752110](https://doi.org/10.1073/pnas.1300752110)
124. Mukhopadhyay H, Cordoba S-P, Maini PK et al (2013) Systems model of T cell receptor proximal signaling reveals emergent ultrasensitivity. *PLoS Comput Biol* 9:e1003004. doi:[10.1371/journal.pcbi.1003004](https://doi.org/10.1371/journal.pcbi.1003004)
125. Chylek LA, Akimov V, Dengjel J et al (2014) Phosphorylation site dynamics of early T-cell receptor signaling. *PLoS One* 9:e104240. doi:[10.1371/journal.pone.0104240](https://doi.org/10.1371/journal.pone.0104240)
126. Shahaf G, Allman D, Cancro MP, Mehr R (2004) Screening of alternative models for transitional B cell maturation. *Int Immunol* 16:1081–1090. doi:[10.1093/intimm/dxh109](https://doi.org/10.1093/intimm/dxh109)
127. Shahaf G, Johnson K, Mehr R (2006) B cell development in aging mice: lessons from mathematical modeling. *Int Immunol* 18:31–39. doi:[10.1093/intimm/dxh346](https://doi.org/10.1093/intimm/dxh346)
128. Shahaf G, Cancro MP, Mehr R (2010) Kinetic modeling reveals a common death niche for newly formed and mature B cells. *PLoS One* 5:e9497. doi:[10.1371/journal.pone.0009497](https://doi.org/10.1371/journal.pone.0009497)
129. Sciammas R, Li Y, Warmflash A et al (2011) An incoherent regulatory network architecture that orchestrates B cell diversification in response to antigen signaling. *Mol Syst Biol* 7:495. doi:[10.1038/msb.2011.25](https://doi.org/10.1038/msb.2011.25)
130. Duffy KR, Wellard CJ, Markham JF et al (2012) Activation-induced B cell fates are selected by intracellular stochastic competition. *Science* 335:338–341. doi:[10.1126/science.1213230](https://doi.org/10.1126/science.1213230)
131. Beltman JB, Allen CDC, Cyster JG, de Boer RJ (2011) B cells within germinal centers migrate preferentially from dark to light zone.

- Proc Natl Acad Sci U S A 108:8755–8760. doi:[10.1073/pnas.1101554108](https://doi.org/10.1073/pnas.1101554108)
132. Meyer-Hermann M, Figge MT, Toellner K-M (2009) Germinal centres seen through the mathematical eye: B-cell models on the catwalk. *Trends Immunol* 30:157–164. doi:[10.1016/j.it.2009.01.005](https://doi.org/10.1016/j.it.2009.01.005)
133. Meyer-Hermann M, Mohr E, Pelletier N et al (2012) A theory of germinal center B cell selection, division, and exit. *Cell Rep* 2:162–174. doi:[10.1016/j.celrep.2012.05.010](https://doi.org/10.1016/j.celrep.2012.05.010)
134. Watzl C, Sternberg-Simon M, Urlaub D, Mehr R (2012) Understanding natural killer cell regulation by mathematical approaches. *Front Immunol* 3:359. doi:[10.3389/fimmu.2012.00359](https://doi.org/10.3389/fimmu.2012.00359)
135. Kaplan A, Kotzer S, Almeida CR et al (2011) Simulations of the NK cell immune synapse reveal that activation thresholds can be established by inhibitory receptors acting locally. *J Immunol* 187:760–773. doi:[10.4049/jimmunol.1002208](https://doi.org/10.4049/jimmunol.1002208)
136. Almeida CR, Ashkenazi A, Shahaf G et al (2011) Human NK cells differ more in their KIR2DL1-dependent thresholds for HLA-Cw6-mediated inhibition than in their maximal killing capacity. *PLoS One* 6:e24927. doi:[10.1371/journal.pone.0024927](https://doi.org/10.1371/journal.pone.0024927)
137. Mesecke S, Urlaub D, Busch H et al (2011) Integration of activating and inhibitory receptor signaling by regulated phosphorylation of Vav1 in immune cells. *Sci Signal* 4:ra36. doi:[10.1126/scisignal.2001325](https://doi.org/10.1126/scisignal.2001325)
138. Yates A, Bergmann C, Van Hemmen JL et al (2000) Cytokine-modulated regulation of helper T cell populations. *J Theor Biol* 206:539–560. doi:[10.1006/jtbi.2000.2147](https://doi.org/10.1006/jtbi.2000.2147)
139. Höfer T, Nathansen H, Löhning M et al (2002) GATA-3 transcriptional imprinting in Th2 lymphocytes: a mathematical model. *Proc Natl Acad Sci U S A* 99:9364–9368. doi:[10.1073/pnas.142284699](https://doi.org/10.1073/pnas.142284699)
140. Yates A, Callard R, Stark J (2004) Combining cytokine signalling with T-bet and GATA-3 regulation in Th1 and Th2 differentiation: a model for cellular decision-making. *J Theor Biol* 231:181–196. doi:[10.1016/j.jtbi.2004.06.013](https://doi.org/10.1016/j.jtbi.2004.06.013)
141. Busse D, de la Rosa M, Hobiger K et al (2010) Competing feedback loops shape IL-2 signaling between helper and regulatory T lymphocytes in cellular microenvironments. *Proc Natl Acad Sci U S A* 107:3058–3063. doi:[10.1073/pnas.0812851107](https://doi.org/10.1073/pnas.0812851107)
142. Feinerman O, Jentsch G, Tkach KE et al (2010) Single-cell quantification of IL-2 response by effector and regulatory T cells reveals critical plasticity in immune response. *Mol Syst Biol* 6:437. doi:[10.1038/msb.2010.90](https://doi.org/10.1038/msb.2010.90)
143. Beyer T, Busse M, Hristov K et al (2011) Integrating signals from the T-cell receptor and the interleukin-2 receptor. *PLoS Comput Biol* 7:e1002121. doi:[10.1371/journal.pcbi.1002121](https://doi.org/10.1371/journal.pcbi.1002121)
144. Khailaie S, Bahrami F, Janahmadi M et al (2013) A mathematical model of immune activation with a unified self-nonself concept. *Front Immunol* 4:474. doi:[10.3389/fimmu.2013.00474](https://doi.org/10.3389/fimmu.2013.00474)
145. Tkach KE, Barik D, Voisinne G et al (2014) T cells translate individual, quantal activation into collective, analog cytokine responses via time-integrated feedbacks. *eLife* 3:e01944. doi:[10.7554/eLife.01944](https://doi.org/10.7554/eLife.01944)
146. Reynolds J, Coles M, Lythe G, Molina-París C (2013) Mathematical model of naive T cell division and survival IL-7 thresholds. *Front Immunol* 4:434. doi:[10.3389/fimmu.2013.00434](https://doi.org/10.3389/fimmu.2013.00434)
147. Hapuarachchi T, Lewis J, Callard RE (2013) A mechanistic model for naive CD4 T cell homeostasis in healthy adults and children. *Front Immunol* 4:366. doi:[10.3389/fimmu.2013.00366](https://doi.org/10.3389/fimmu.2013.00366)
148. Thiébaud R, Drylewicz J, Prague M et al (2014) Quantifying and predicting the effect of exogenous interleukin-7 on CD4+ T cells in HIV-1 infection. *PLoS Comput Biol* 10:e1003630. doi:[10.1371/journal.pcbi.1003630](https://doi.org/10.1371/journal.pcbi.1003630)
149. Owen MR, Sherratt JA (1997) Pattern formation and spatiotemporal irregularity in a model for macrophage-tumour interactions. *J Theor Biol* 189:63–80. doi:[10.1006/jtbi.1997.0494](https://doi.org/10.1006/jtbi.1997.0494)
150. Matzavinos A, Chaplain MAJ, Kuznetsov VA (2004) Mathematical modelling of the spatio-temporal response of cytotoxic T-lymphocytes to a solid tumour. *Math Med Biol* 21:1–34. doi:[10.1093/imammb/21.1.1](https://doi.org/10.1093/imammb/21.1.1)
151. Kim PS, Lee PP (2012) Modeling protective anti-tumor immunity via preventative cancer vaccines using a hybrid agent-based and delay differential equation approach. *PLoS Comput Biol* 8:e1002742. doi:[10.1371/journal.pcbi.1002742](https://doi.org/10.1371/journal.pcbi.1002742)
152. De Pillis LG, Gu W, Radunskaya AE (2006) Mixed immunotherapy and chemotherapy of tumors: modeling, applications and biological interpretations. *J Theor Biol* 238:841–862. doi:[10.1016/j.jtbi.2005.06.037](https://doi.org/10.1016/j.jtbi.2005.06.037)
153. Depillis L, Gallegos A, Radunskaya A (2013) A model of dendritic cell therapy for melanoma. *Front Oncol* 3:56. doi:[10.3389/fonc.2013.00056](https://doi.org/10.3389/fonc.2013.00056)

154. Joshi B, Wang X, Banerjee S et al (2009) On immunotherapies and cancer vaccination protocols: a mathematical modelling approach. *J Theor Biol* 259:820–827. doi:[10.1016/j.jtbi.2009.05.001](https://doi.org/10.1016/j.jtbi.2009.05.001)
155. Kogan Y, Halevi-Tobias K, Elishmereni M et al (2012) Reconsidering the paradigm of cancer immunotherapy by computationally aided real-time personalization. *Cancer Res* 72:2218–2227. doi:[10.1158/0008-5472.CAN-11-4166](https://doi.org/10.1158/0008-5472.CAN-11-4166)
156. Wilson S, Levy D (2012) A mathematical model of the enhancement of tumor vaccine efficacy by immunotherapy. *Bull Math Biol* 74:1485–1500. doi:[10.1007/s11538-012-9722-4](https://doi.org/10.1007/s11538-012-9722-4)
157. Glauche I, Horn K, Horn M et al (2012) Therapy of chronic myeloid leukaemia can benefit from the activation of stem cells: simulation studies of different treatment combinations. *Br J Cancer* 106:1742–1752. doi:[10.1038/bjc.2012.142](https://doi.org/10.1038/bjc.2012.142)
158. Parra-Guillen ZP, Berraondo P, Grenier E et al (2013) Mathematical model approach to describe tumour response in mice after vaccine administration and its applicability to immune-stimulatory cytokine-based strategies. *AAPS J* 15:797–807. doi:[10.1208/s12248-013-9483-5](https://doi.org/10.1208/s12248-013-9483-5)
159. Ciupe SM, Ribeiro RM, Nelson PW, Perelson AS (2007) Modeling the mechanisms of acute hepatitis B virus infection. *J Theor Biol* 247:23–35. doi:[10.1016/j.jtbi.2007.02.017](https://doi.org/10.1016/j.jtbi.2007.02.017)
160. Binder M, Sulaimanov N, Clausnitzer D et al (2013) Replication vesicles are load- and choke-points in the hepatitis C virus lifecycle. *PLoS Pathog* 9:e1003561. doi:[10.1371/journal.ppat.1003561](https://doi.org/10.1371/journal.ppat.1003561)
161. Althaus CL, De Boer RJ (2008) Dynamics of immune escape during HIV/SIV infection. *PLoS Comput Biol* 4:e1000103. doi:[10.1371/journal.pcbi.1000103](https://doi.org/10.1371/journal.pcbi.1000103)
162. Lee HY, Topham DJ, Park SY et al (2009) Simulation and prediction of the adaptive immune response to influenza A virus infection. *J Virol* 83:7151–7165. doi:[10.1128/JVI.00098-09](https://doi.org/10.1128/JVI.00098-09)
163. Wu H, Kumar A, Miao H et al (2011) Modeling of influenza-specific CD8+ T cells during the primary response indicates that the spleen is a major source of effectors. *J Immunol* 187:4474–4482. doi:[10.4049/jimmunol.1101443](https://doi.org/10.4049/jimmunol.1101443)
164. Wilson DP, Timms P, McElwain DLS (2003) A mathematical model for the investigation of the Th1 immune response to Chlamydia trachomatis. *Math Biosci* 182:27–44
165. Vickers DM, Zhang Q, Osgood ND (2009) Immunobiological outcomes of repeated chlamydial infection from two models of within-host population dynamics. *PLoS One* 4:e6886. doi:[10.1371/journal.pone.0006886](https://doi.org/10.1371/journal.pone.0006886)
166. Mallet DG, Bagher-Oskouei M, Farr AC et al (2013) A mathematical model of chlamydial infection incorporating movement of chlamydial particles. *Bull Math Biol* 75:2257–2270. doi:[10.1007/s11538-013-9891-9](https://doi.org/10.1007/s11538-013-9891-9)
167. Go N, Bidot C, Belloc C, Touzeau S (2014) Integrative model of the immune response to a pulmonary macrophage infection: what determines the infection duration? *PLoS One* 9:e107818. doi:[10.1371/journal.pone.0107818](https://doi.org/10.1371/journal.pone.0107818)
168. Marino S, Kirschner DE (2004) The human immune response to Mycobacterium tuberculosis in lung and lymph node. *J Theor Biol* 227:463–486. doi:[10.1016/j.jtbi.2003.11.023](https://doi.org/10.1016/j.jtbi.2003.11.023)
169. Raman K, Bhat AG, Chandra N (2010) A systems perspective of host–pathogen interactions: predicting disease outcome in tuberculosis. *Mol Biosyst* 6:516–530. doi:[10.1039/b912129c](https://doi.org/10.1039/b912129c)
170. Carbo A, Olivares-Villagómez D, Hontecillas R et al (2014) Systems modeling of the role of interleukin-21 in the maintenance of effector CD4+ T cell responses during chronic Helicobacter pylori infection. *mBio* 5:01243–01214. doi:[10.1128/mBio.01243-14](https://doi.org/10.1128/mBio.01243-14)
171. Bocharov G, Züst R, Cervantes-Barragan L et al (2010) A systems immunology approach to plasmacytoid dendritic cell function in cytopathic virus infections. *PLoS Pathog* 6:e1001017. doi:[10.1371/journal.ppat.1001017](https://doi.org/10.1371/journal.ppat.1001017)
172. Palsson S, Hickling TP, Bradshaw-Pierce EL et al (2013) The development of a fully-integrated immune response model (FIRM) simulator of the immune response through integration of multiple subset models. *BMC Syst Biol* 7:95. doi:[10.1186/1752-0509-7-95](https://doi.org/10.1186/1752-0509-7-95)
173. Fallahi-Sichani M, Schaller MA, Kirschner DE et al (2010) Identification of key processes that control tumor necrosis factor availability in a tuberculosis granuloma. *PLoS Comput Biol* 6:e1000778. doi:[10.1371/journal.pcbi.1000778](https://doi.org/10.1371/journal.pcbi.1000778)
174. Fallahi-Sichani M, El-Kebir M, Marino S et al (2011) Multiscale computational modeling reveals a critical role for TNF- $\alpha$  receptor I dynamics in tuberculosis granuloma formation. *J Immunol* 186:3472–3483. doi:[10.4049/jimmunol.1003299](https://doi.org/10.4049/jimmunol.1003299)
175. Cilfone NA, Perry CR, Kirschner DE, Linderman JJ (2013) Multi-scale modeling predicts a balance of tumor necrosis factor- $\alpha$



- and interleukin-10 controls the granuloma environment during *Mycobacterium tuberculosis* infection. *PLoS One* 8:e68680. doi:[10.1371/journal.pone.0068680](https://doi.org/10.1371/journal.pone.0068680)
176. Paiva LR, Silva HS, Ferreira SC, Martins ML (2013) Multiscale model for the effects of adaptive immunity suppression on the viral therapy of cancer. *Phys Biol* 10:025005. doi:[10.1088/1478-3975/10/2/025005](https://doi.org/10.1088/1478-3975/10/2/025005)
177. Dwivedi G, Fitz L, Hegen M et al (2014) A multiscale model of interleukin-6-mediated immune regulation in Crohn's disease and its application in drug discovery and development. *CPT Pharmacometrics Syst Pharmacol* 3:e89. doi:[10.1038/psp.2013.64](https://doi.org/10.1038/psp.2013.64)
178. Sedegah M, Kim Y, Peters B et al (2010) Identification and localization of minimal MHC-restricted CD8+ T cell epitopes within the *Plasmodium falciparum* AMA1 protein. *Malar J* 9:241. doi:[10.1186/1475-2875-9-241](https://doi.org/10.1186/1475-2875-9-241)
179. Haste Andersen P, Nielsen M, Lund O (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci* 15:2558–2567. doi:[10.1110/ps.062405906](https://doi.org/10.1110/ps.062405906)
180. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol* 15:221–231. doi:[10.1016/s0955-0674\(03\)00017-6](https://doi.org/10.1016/s0955-0674(03)00017-6)
181. Lev Bar-Or R, Maya R, Segel LA et al (2000) Generation of oscillations by the p53-Mdm2 feedback loop: a theoretical and experimental study. *Proc Natl Acad Sci* 97:11250–11255. doi:[10.1073/pnas.210171597](https://doi.org/10.1073/pnas.210171597)
182. Vera J, Schultz J, Ibrahim S et al (2009) Dynamical effects of epigenetic silencing of 14-3-3 $\sigma$  expression. *Mol BioSyst* 6:264. doi:[10.1039/b907863k](https://doi.org/10.1039/b907863k)
183. Liu Y, Liu W, Hu C et al (2011) MiR-17 modulates osteogenic differentiation through a coherent feed-forward loop in mesenchymal stem cells isolated from periodontal ligaments of patients with periodontitis. *Stem Cells* 29:1804–1816. doi:[10.1002/stem.728](https://doi.org/10.1002/stem.728)
184. Nelson DE, Ihekwaba AEC, Elliott M et al (2004) Oscillations in NF-kappaB signaling control the dynamics of gene expression. *Science* 306:704–708. doi:[10.1126/science.1099962](https://doi.org/10.1126/science.1099962)
185. Luu K, Greenhill CJ, Majoros A et al (2014) STAT1 plays a role in TLR signal transduction and inflammatory responses. *Immunol Cell Biol* 92:761–769. doi:[10.1038/icb.2014.51](https://doi.org/10.1038/icb.2014.51)

# Chapter 10

## Systems Medicine in Oncology: Signaling Network Modeling and New-Generation Decision-Support Systems

**Silvio Parodi, Giuseppe Riccardi, Nicoletta Castagnino, Lorenzo Tortolina, Massimo Maffei, Gabriele Zoppoli, Alessio Nencioni, Alberto Ballestrero, and Franco Patrone**

### Abstract

Two different perspectives are the main focus of this book chapter: (1) A perspective that looks to the future, with the goal of devising rational associations of targeted inhibitors against distinct altered signaling-network pathways. This goal implies a sufficiently in-depth molecular diagnosis of the personal cancer of a given patient. A sufficiently robust and extended dynamic modeling will suggest rational combinations of the abovementioned oncoprotein inhibitors. The work toward new selective drugs, in the field of medicinal chemistry, is very intensive. Rational associations of selective drug inhibitors will become progressively a more realistic goal within the next 3–5 years. Toward the possibility of an implementation in standard oncologic structures of technologically sufficiently advanced countries, new (legal) rules probably will have to be established through a consensus process, at the level of both diagnostic and therapeutic behaviors.

(2) The cancer patient of today is not the patient of 5–10 years from now. How to support the choice of the most convenient (and already clinically allowed) treatment for an individual cancer patient, as of today? We will consider the present level of artificial intelligence (AI) sophistication and the continuous feeding, updating, and integration of cancer-related new data, in AI systems. We will also report briefly about one of the most important projects in this field: IBM Watson US Cancer Centers. Allowing for a temporal shift, in the long term the two perspectives should move in the same direction, with a necessary time lag between them.

**Key words** Cancer genomics, Signaling-network pathways, Individual cancer patient, Oncoprotein inhibitors, Rational associations of targeted inhibitors, New clinical trial designs, Systems medicine, Decision-support systems, Artificial intelligence, IBM Watson

---

### 1 Introduction

In this book chapter we will touch two quite different perspectives that we consider however equally important, for bringing a systems medicine wavelength to the field of oncology:

1. Future perspectives toward a personalized combination therapy
2. Decision-support systems for the present-day choices of the clinical oncologist

In order to be forgiven for omitting other probably not less important perspectives that could come to the mind of an informed reader, we will start half seriously with an ancient parable from Hindu culture.

### 1.1 An Ancient Parable Adapted to Cancer

*A powerful Maharajah of old India had invited some blind people of his kingdom to define for him an unknown object, examining the thing with their hands. A successful answer would be associated with a great prize.*

*The unknown “object” was a majestic elephant.*

*Each blind man examined the “object” from his position:*

*The blind man who felt a leg said the “object” was a pillar; the one who felt the tail said the “object” was a rope; the one who felt the trunk said the “object” was like a tree branch; the one who felt the ear said the “object” was like a hand fan; the one who felt the belly said the “object” was like a wall; and the one who felt the tusk said the “object” was like a solid pipe.*

*The Maharajah explained to them: “All of you are right. The reason every one of you is telling it differently is because each one of you touched a different part of the elephant. So, actually the elephant has all the features you mentioned.”*

The parable wants to stress the need for communication, due consideration of different perspectives, and finally **the need of adequate efforts for integrating them.**

Systems medicine in oncology is typically played on a wavelength of integration of a multiplicity of aspects at different levels of expertise. The immense complexity of the cancer phenomenon and its many facets has something in common with the elephant of the blind men. Our intent is to illustrate some of the different features of the “cancer elephant.”

---

## 2 Cancer's Systems Biology

Cancer is fundamentally a genomic disease in the framework of a multi-hit evolutionary process. Systematic studies of the cancer genome have exploded in recent years. Next-generation sequencing (NGS) and other technologies, such as reverse-phase protein arrays or methylation arrays, have been successfully applied to the study of cancer disease [1]. Transcriptome-level gene expression and whole-genome copy number variation studies have also concurred to transform the way we understand and approach cancer [2, 3]. The reduction in technical costs now allows researchers to perform gene screening and transcriptome and copy number analysis for the purposes of translational and clinical oncology studies [4, 5].

These analyses are not performed at a clinical routine level, but the depth of genetic analysis will increase progressively during the incoming years even at this level.

Recently, new cancer genes affecting processes not previously known to be causal targets in cancer have been identified [6]. These new cancer genes affect processes at disparate levels: cell signaling, chromatin structure and function, epigenomic regulation, RNA splicing; protein homeostasis, metabolism, different types of regulatory noncoding RNAs, and regulatory DNA sequences. Our cancer elephant can be really multifaceted!

For instance, an altered regulation of some microRNAs (miRNAs), a class of small noncoding RNAs that function as post-transcriptional regulators of mRNA expression of oncogenes and tumor suppressors, can play a crucial role in cancer [7].

To understand the mechanisms of tumor emergence and evolution, we need to identify the genes that drive tumorigenesis. Tumor and non-tumor genomes contain thousands of somatic mutations, but only a few of them “drive” a normal cell to evolve toward a cancer cell, by affecting genes which confer selective growth advantage to tumor cells [8].

## **2.1 Major Databases in Relation to Cancer and Other Pathologies**

To support a broader, more comprehensive approach, to the complex, multifaceted features of our elephant, several available databases can be of help. Acquiring confidence in dealing with their content is potentially very important. Without being exhaustive, we touch here briefly some of the most relevant among them:

- **COSMIC**, an acronym of Catalogue of Somatic Mutations in Cancer, curates data from papers in the scientific literature and large-scale experimental screens from the Cancer Genome Project at the Sanger Institute [9]. COSMIC stores and displays information related to somatic mutations in human cancers. Cancer Genome Project links are: *Cell Line Project*, *COSMIC Whole Genomes*, *Drug Sensitivity*, *COSMIC Genome Browser*, *CONAN* (copy number analysis), *Census* (of genes causally implicated in cancer), *Trace Archive* (data generated by the Cancer Genome Project), *Systematic Screens* (a complete list of large-scale systematic screen papers and whole-genome shotgun sequencing papers curated in COSMIC), *COSMIC BioMart* (a flexible way to mine data), and *Curated Genes and Fusions* (contains curated somatic mutation data from the scientific literature, the Cancer Genome Project’s systematic screens, and data from the ICGC and TCGA databases).

At the end of 2014, the Cancer Gene Census includes almost 500 bona fide driver cancer genes (all cancers considered) [9, 10].

Large international consortia, like **The Cancer Genome Atlas** (TCGA) and the **International Cancer Genome Consortium** (ICGC), have now made accessible to researchers worldwide several omics repositories with associated clinical data [11, 12]. These efforts have the primary goal of detecting driver somatically inheritable alterations and altered molecular mechanisms present in dif-

ferent cancer types. We list here cancer genomics data portals available online, including but not limited to:

- The **TCGA Data Portal** [11] includes genomic as well as clinical information and analyses related to *individual patients cancer genomes*. TCGA sequence data are maintained by the University of California, at Santa Cruz Cancer Genomics.
- The **ICGC Data Portal** [12] is a portal for raw data from the ICGC and TCGA projects. Data can be filtered to visually present information such as top-mutated genes (in terms of frequency and alteration of function).
- The **cBioPortal for Cancer Genomics** [13] is a portal developed and maintained by the Computational Biology Center at the Memorial Sloan Kettering Cancer Center (New York). It offers curated data sets from over 50 published studies, including studies from TCGA.
- The **Cancer Genomics Hub** [14] is an online repository of the sequencing programs of the National Cancer Institute (NCI), including TCGA, the Cancer Cell Line Encyclopedia (CCLE), and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) projects.
- The **TumorPortal** [15] is developed and maintained by the Broad Institute of Harvard and MIT. This site hosts a “pan-cancer” data set from 21 tumor types. It provides visualization of computationally processed gene information. It shows which genes are mutated in many tumor samples and types.
- The **Drug–Gene Interaction database** (DGBIdb) is held at the Genome Institute, Washington University School of Medicine (St. Louis, MO, USA).

*In a broad multi-pathology perspective*, and as a help to medicinal chemistry research, this database mines resources generating hypotheses about gene products therapeutically targetable or of interest for future drug developments. Users can use this interface for searching lists of gene products against the existing compendia of known or potential interactions of drugs with targets. The online database contains over 40,000 reference genes and 10,000 drugs involved in over 15,000 drug/gene-product interactions, belonging to one of 39 potentially druggable target categories [16, 17].

- The **Genomics of Drug Sensitivity in Cancer (GDSC)** database [18] is a database curated by the Wellcome Trust Sanger Institute. The approach of GDSC is of an encyclopedic type. It is a database of recent years’ knowledge, about a variety of drugs (about 150). It includes new inhibitors of signaling proteins potentially affected by excess of function, but it also

includes traditional cytotoxic drugs (alkylating agents, antineoplastic agents, stabilized intercalating agents, topoisomerase II and I inhibitors, poisons of the mitotic spindle). We had however the impression that many older cytotoxic drugs belonging to the same families are not reported in GDSC. Among the about 150 drugs reported, we also have more or less selective inhibitors of signaling proteins or other protein targets more difficult to classify as driver–gatekeeper mutations and/or somatically inheritable alterations. The molecules reported were mostly considered of interest if clearly active in causing growth inhibition and/or cell death in a variety of cancer cell lines. More than 1200 cancer cell lines have apparently been tested, probably with different degrees of molecular characterization. Mutations in about 70 putative cancer genes have been reported for several hundreds of these cancer cell lines. It was also reported when one of these mutations affected sensitivity to a given drug in a given cancer line (only at a cell growth inhibition and/or cell death level, not at the mechanistic level of biochemical interactions and specifically altered pathways and signaling networks). Because of its significant size and encyclopedic approach, GDSC can be considered a very useful point of reference, for starting to collect information on the trinomial: (1) drugs, (2) cancer cell lines, and (3) cancer-gene mutations/somatically inheritable alterations.

The **CLARITY challenge** [19] is not an online cancer genomics data portal. It is an international effort involving thirty partners in the area of genetic disorders. We mention this effort because their conclusions highlight the need for a broader adoption of standardizations in collaborative approaches, to support clinicians in the choice of individualized treatments.

Also in the cancer field the progress of molecular research asks for standard procedures and broader collaborative networks (including collaborative efforts between academia and industry), to maximize the progress and benefits of translational oncology moving toward cancer treatment [20].

In the coming years a major challenge for cancer treatment will be the identification of highly effective targeted-drug combinations and their translation in the clinical practice. New strategies to improve preclinical and clinical drug evaluation will be needed.

This type of vast information and other multifaceted available data could be used to instruct modern decision-support systems, to generate a set of hypotheses about diagnosis and treatment of a cancer patient. As illustrated in Section 7.4, IBM Watson's data processing capabilities are already able to analyze millions of medical and scientific journal articles available through sources such as Medline and the PubMed database [21–23].

## 2.2 *Passenger versus Driver Mutations*

Fidelity of DNA replication (at DNA-base level, at chromosome level, at epigenetic levels) is intrinsically limited in normal cells (intrinsic normal fidelity limits+the presence of environmental mutagens) and is much worst in preneoplastic and cancer cells, where caretaker gene alterations (see below) can decrease fidelity in important ways [24, 25].

Any cancer cell tends to have thousands of passenger (cancer-neutral) mutations and very few (probably less than a dozen) non-neutral cancer mutations or somatically inheritable cancer alterations [8] in each individual tumor. In each specific type of cancer, driver cancer mutations are selected for, against the vast background of neutral-passenger mutations. The increased cancer-mutation (driver-mutation) frequency of a given mutation present in a large set of tumors of a given cancer type (e.g., CRC derived from the COSMIC database [9]) is already a significant alert when a given mutation frequency is high. Distinctions are delicate for low mutation frequencies and for mutations that could still be neutral even within a given cancer gene.

We will not comment here in detail the complex strategies that have been implemented to distinguish the cancer driver from passenger mutations. To make the reader aware of the relevance and complexity of this issue, we think however potentially useful to briefly mention some strategies and software for the detection of the small number of driver mutations against the vast background of passenger mutations. In the position article “*Cancer Genomes: Discerning Drivers from Passengers*” [26], a variety of driver-hunter strategies developed by different important groups are discussed. Bert Vogelstein’s group in Baltimore has proposed a “ratiometric rule”: (1) not mutation frequencies alone, but only if supported by adequate mutation patterns suggesting a functional change; (2) the rule considers separately oncogenes, which need to be hyperactive to cause cancer, and tumor suppressor genes, which cause cancer when they stop working; (3) for an excess-of-function oncogene, at least 20% of the recorded mutations in the gene should occur at the same position and cause a single switched amino acid in the protein that the gene encodes; and (4) for a tumor suppressor gene, more than 20% of the mutations in the gene should be clearly inactivating [8, 26].

The different algorithm MutSig was born in 2007 and has become an increasingly seasoned mutation signal processing tool. The algorithm evolved from MutSig to MutSigCV, MutSigCL, to MutSigFN [26–28]. (1) MutSig looked at one signal: the abundance of non-silent mutations in relation to background passenger-mutation patterns. Over time, the Broad team (Boston) changed the way the algorithm calculates gene-specific background rates and integrated covarying factors, including gene density, chromatin structure, and replication timing, related to the distance to the nearest site where DNA replication initiates; these parameters



reduce the false-positive rate among cancer genes. (2) MutSigCV abundance “is mutation abundance relative to the background level” + a preponderance of non-silent versus silent mutations. (3) MutSigCL are clustered mutations that may implicate a gene as a driver. The scientists added to the algorithm the ability to calculate the probability that the positional clustering of mutations was due to chance. MutSigCL clustering asks the question: “are the mutations clustered in hotspots?” (4) MutSigFN is an algorithm that estimates the functional impact of each mutation. The team added the capability to estimate how likely it is that a missense mutation might be deleterious to the protein encoded by the mutated gene. Evolutionary conservation is key for this signal detection: if a mutation occurs in a location conserved among multiple branches of the evolutionary tree, the algorithm knows to consider it significant. MutSigFN function asks the question: “are the mutations likely to have functional impact?” (5) The developers, heeding the fact that clustering and conservation may be correlated, calculate a joint  $P$  value for CL+FN. They also take into account all three signals (CV, CL, and FN) to deliver a global  $P$  value for a given gene, leading to the *three-signal analysis* that the team applies to cancer genomes. The Broad Institute investigators believe the catalogue of driver cancer variants is still rather incomplete.

The group of Nuria Lopez-Bigas has developed a software (IntOGen mutations) which identifies cancer drivers across tumor types [26, 29]. Also in this software, the accumulation in a given gene of mutations with a high functional impact and the clustering of non-synonymous mutations in a particular region of the protein sequence are considered suggestive of driver mutations.  $P$  values for the two parameters are computed and combined and also combined with frequency of mutation, not only at gene but also at pathway level.

This platform [30] summarizes somatic mutations, genes, and pathways involved in tumorigenesis. It identifies and visualizes cancer drivers, analyzing 4623 exomes from 13 cancer sites.

Quite different is the approach of the article by Martelotto et al. “*Benchmarking Mutation Effect Prediction-Algorithms Using Functionally Validated Cancer-Related Missense Mutations*” [31]. 849 functionally significant SNVs (single-nucleotide variants) from 15 cancer genes were used as the “golden standard” to assess the performance of 11 SNV-mutation-effect single and independent predictor algorithms + four combined meta-predictors. It is not completely clear to what extent this SNV “golden standard” derived from 15 cancer genes (six well-known dominant oncogenes containing a kinase domain, six more recent cancer genes not containing a kinase domain, a gatekeeper tumor suppressor gene—TP53—and two caretaker tumor suppressor genes, BRCA1 and BRCA2) can be really considered a large-spectrum comprehensive “golden standard,” at least for SNVs. Using this approach,

important differences of performance were observed in terms of sensitivity and specificity: some individual predictor algorithm was quite satisfactory, for instance, the “Functional Analysis Through Hidden Markov Model” (FATHMM—cancer) algorithm; FATHMM is endowed with the ability to recognize important structural and/or evolutionary constraints.

Going back to the meta-analysis of Martelotto et al. [31], the NPV (negative predictive value) was improved only by combinations of some of the individual algorithms considered.

This short subchapter is far from an in-depth comparative analysis of the performance of different software aimed at achieving the crucial distinction between passenger and driver mutations. We are far from an exhaustive treatment of this complex issue; we just wanted to arouse awareness in the reader about this critical distinction.

### **2.3 Gatekeeper Genes and Caretaker Genes**

Conceptually, it is important to offer a basic distinction between: **gatekeeper genes** and **caretaker genes** [24, 25].

**Gatekeeper genes** (mostly belonging to altered pathways) are genes directly involved in sustaining the actual frank malignancy of a tumor, at the moment it reaches our clinical observation.

Gatekeeper genes control the transmission of information along biochemical interaction pathways, interconnected among themselves within signaling-network subregions, each subregion involving perhaps few hundred network-linked molecules. An activating mutation or somatically inheritable alteration lights up, in principle irreversibly, a dormant (more precisely: physiologically regulated) pathway.

If the mutation is “druggable,” a specific correction becomes in principle possible. It will be of great interest for medicinal chemistry, the pharmaceutical industry, (pre-) clinic research, and finally the cancer patient.

A **caretaker gene** is in principle involved with the fidelity/stability of transmission of inheritable information at both the DNA level and the epigenetic level, from the parental cell to the two daughter cells, generated during cell division. Altered caretaker genes can imply a higher error frequency, both at the level of DNA bases and at the chromosome level (e.g., as altered copy numbers). In principle, they blindly increase the frequency of both passenger and driver mutations. This causes a compression of the evolutionary time from normality to cancer, and it is also equivalent to having a much higher frequency of cancers at the same age.

We can inherit an altered allele of either a gatekeeper or a caretaker gene. Both cases will give origin to a cancer predisposition.

### **2.4 Cancer as a Disease of Integrated Biochemical Signaling Networks**

Cancer systems biology has led us to an understanding of cancer primarily as a disease of integrated biochemical signaling networks of the cell [8].

It has been observed that different individual cancers tend to be more similar in terms of alterations of overall signaling pathways

than in terms of the individual mutations possibly present in a given pathway. This is because mutations within the same pathway might be mutually exclusive alternatives for the pathway alteration. Two mutations within the same pathway tend to be much less favored in terms of evolutionary pressure [32]. One alteration in a pathway can be “biologically sufficient” to achieve the alteration of that pathway. Different mutations in the same pathway remain, however distinct at the molecular level, different targets for medicinal chemistry and potentially capable of (at least partially) differentiated responses to drugs inhibiting oncoproteins affected by excess of function.

The explosion of genomic data has revealed an unexpected richness in the types of mutational processes that can be observed in individual tumors.

### **2.5 Mutational Processes in Cancer**

As an example, at the level of caretaker alterations, *kataegis* describes a pattern of localized hypermutability in some chromosomal regions at localized temporal moments [6, 33]. They have been identified in several cancer genomes. The term *kataegis* is derived from the ancient Greek word for “thunder,” *καταιγις*. Regions of *kataegis* have been shown to be co-localized with regions of somatic genome rearrangements. We typically observe cytosine to thymine mutations, in the context of a TpC dinucleotide. An enzyme of the APOBEC (apolipoprotein B-editing/catalytic) enzyme family is responsible for the process of *kataegis*. The APOBEC family is a family of amphipathic apolipoproteins, which are C to U editing enzymes. A cytidine deaminase mutagenesis (initial conversion of cytosine to uracil) was found widespread in human cancers, for instance, in breast, lung, and hematological cancers. Signatures of mutational processes in human cancer are associated with age of the patient at cancer diagnosis, known mutagenic exposures, and defects in DNA maintenance. In addition to these genome-wide mutational signatures, “kataegis,” a hypermutation localized to small genomic regions is found in many cancer types [34].

The enormous size of the potentially available information concerning mutations/alterations requires that *diagnostic priorities* are established, in the perspective of a rational (personalized) treatment of the cancer of a given patient. What this could concretely signify will be discussed in this book chapter.

---

## **3 Cancer Therapy**

The majority of antineoplastic agents currently used for the treatment of patients are (in terms of drug families and basic mechanisms of action) 50–60 years old. On a didactic wavelength, let us recall to our minds the following major families of traditional anti-

neoplastic agents: alkylating agents, antimetabolites, stabilized intercalating agents, topoisomerase II and I inhibitors, and poisons of the mitotic spindle. These traditional families of antineoplastic agents in fact all belong to a unique superfamily. They are rather blind poisons of the cell-replication machinery, at some point or another. In principle, they are *poorly capable of distinguishing between dividing normal cells and dividing cancer cells*. Perhaps for this reason, in solid tumors, surgery still saves many more lives than antineoplastic pharmacology.

### 3.1 Targeted Therapy

Targeted agents are intended to exploit the phenomenon of *oncogene addiction* [35]. Cancer cells, along their evolutionary process, can become dependent for their survival on the constitutive activity of a defined mutated (or functionally drastically altered) oncogene. Therefore, the consequence of its inhibition can in principle become lethal (or more severe) for the malignant cell than for the corresponding healthy cell. We do not know if this phenomenon is always taking place, but several relevant examples have been published in literature over the recent years. Without any intention of being exhaustive, and without entering in details, we mention here the role of imatinib in Bcr–Abl fusion protein in chronic myelogenous leukemia [36], crizotinib in AML4–ALK fusion protein in lung cancer [37], and vemurafenib in BRAFV600 mutant melanoma [38].

The strategy of a personalized cancer medicine will require more and more a detailed/personalized stratification of patients in small subsets, based on shared altered-gene features, concerning especially driver–gatekeeper mutations. Only such a molecular diagnostic progress will open the way to rational combination treatments with selective inhibitors of altered signaling proteins (oncoproteins) affected by excess of function, signaling proteins at the same time favorable in terms of “druggability” at the medicinal chemistry level.

We feel that a major goal of the next 3–5–10 years will be to achieve a personalized selection of *rational multidrug combinations of signaling-protein (oncoprotein) inhibitors*, in the perspective of a long-time due, seriously innovative, anticancer therapy.

Systems medicine perspectives in oncology have to interplay with much needed further advances in the broad field of medicinal chemistry.

Selective small-molecule kinase inhibitors have emerged over the past decade as an important (not unique) class of new anticancer agents. They can antagonize specific oncogenic signaling processes. In the late 1990s, imatinib (Gleevec), an inhibitor designed to target the BCR–ABL fusion complex in chronic myelogenous leukemia (CML), was the first very successful drug of this kind. An uncontrolled high level of the ABL protein function in the chimera had conferred to the CML cancer cells addiction to the ABL excess of function [36].

In solid tumors, kinase inhibitors such as gefitinib (Iressa) are effective in cancers with an appropriately sensitive and overactive EGFR mutation [39]. Downstream mutations could however impair this activity. In addition to the small-molecule kinase inhibitors, kinase-targeted antibodies have also demonstrated clinical efficacy in solid tumors. Cetuximab and panitumumab (anti-EGFR monoclonal antibodies) are currently used in clinical practice for the treatment of metastatic forms of colorectal cancer, in the absence of a downstream KRAS mutation conferring resistance [40, 41].

At present they are used either as single agents or in combination with traditional anticancer agents. Sorafenib (an inhibitor of VEGFR, PDGFR, and other kinases) in combination with traditional anticancer irinotecan showed some activity as second- or later-line treatment in pretreated metastatic CRC cancer patients [42]. Bevacizumab, a monoclonal antibody directed against the vascular endothelial growth factor (VEGF), which promotes neoangiogenesis, was also clinically employed [43]. A kinase inhibitor of uncertain specificity (regorafenib), studied in a phase III trial, demonstrated some activity in a standard-CT treatment-refractory population [44].

The survival gains (progression free survival, PFS; overall survival, OS) reported in these studies [42–44] on metastatic CRC were in general less than four months, in comparison with other types of previous chemotherapeutic treatments or a placebo. We consider these studies substantially old types of trials. They were referred to patients genetically/epigenetically not deeply characterized (molecularly inhomogeneous). Statistically significant results were prevalently related to a sufficiently large size of the study, but usually the real clinical gain for the patient was intrinsically small. They cannot be considered relevant advances in all cases, at least in a long-term multi-decade historical perspective.

MEK inhibitors, for instance, selumetinib (AZD6244), in association with afatinib (BIBW2992—an ERBB2 irreversible inhibitor), can synergize in KRAS-mutant lung and colon cancers [45]. In this last case, clinical trials have just started.

In a phase II study, everolimus (an mTOR inhibitor) was well tolerated but did not confer meaningful efficacy in heavily pretreated patients with metastatic CRC [46]. Targeted sequencing in bladder cancer revealed that TSC1 mutations are correlated with everolimus sensitivity [47]. A mutated TSC1 acts as an inactive GTPase toward Rheb, which in turn phosphorylates/activates mTORC1. Again, a pathway alteration seems important.

We and others [48] tend to consider a design of the past a poorly targeted large clinical trial, showing a small but statistically significant treatment effect. These types of clinical reports (mostly referred to variations of associations and schedules of cytotoxic agents) have been going on for decades.

The range of new-generation agents directed against specific signaling proteins, currently evaluated in preclinical studies, is much wider than the number of drugs already approved for clinical use. New ethical and legal approaches to speed up a possible transition should be found. Too slow a process, at the end, could impair/delay the implementation of more effective patient therapies.

Several agents, whose targets are aberrantly activated, are being investigated at the preclinical level, and some compounds have been advanced into early phase clinical trials: for instance, PI3K or AKT inhibitors (along the PI3K pathway) [49] and MDM2 inhibitors (in the TP53 pathway) [50]. In addition, new agents are emerging, such as tankyrase inhibitors [51], which, by inhibiting axin degradation, increase its intracellular concentration and the function of the APC complex, ultimately promoting  $\beta$ -catenin degradation and inhibiting  $\beta$ -catenin co-operation with transcription factors like TCF7L2. In which specific constellation of molecular alterations a tankyrase inhibitor could become relevant is not clear yet.

Probably only a fraction of the kinases directly affected by an excess of function during the process of malignant transformation have turned out to be practically “druggable,” giving origin to selective inhibitors, usable at a preclinical level and in perspective also at a clinical level.

Malignant transformation is an evolutionary process from the bad to the worst. Late mutations/inheritable alterations could play significant roles in terms of favored recurrence and metastasis. These late events could obviously be of interest as potential drug targets.

New selective molecules, capable of moving from a biochemical level (lead molecules) to a cellular level, to molecularly appropriated tumors transplanted in experimental animals, and to phase 1, 2, and 3 clinical studies, are part of an incremental process, for which we expect important advances (hopefully also in terms of modified/accelerated procedural strategies) during the next 3–5–10 years. We and others [52] would suggest the treatment of smaller tumor subsets, characterized at a higher analytical depth and molecularly more homogeneous. However, in a future perspective, suppose we had achieved an advanced/predictive dynamic modeling at the biochemical interaction level, of a sufficiently well-parameterized signaling-network subregion, with input in the model of a sufficiently detailed genetic pathology of major driver-gatekeeper somatically inheritable alterations: this could be conducive to a rational form of “n of 1” clinical trials [53]. A statistical evaluation of such an approach could be based on a sort of “meta-analysis” of the performance of a sufficiently large set of “n of 1” clinical trials. In each “n of 1” clinical trial, combination therapy

could have been suggested by a homogeneous procedure of dynamic modeling, finely tuned according to altered pathways and the best-performing oncoprotein–inhibitor combination, in a personalized view for each patient case. This procedure would be compared with the outcomes of more traditional treatments implemented in correspondingly more routinely characterized cancer patients. A well-established dynamic model invariant both in its basic and more detailed procedures would be a sufficiently homogeneous point of reference for meta-analytic studies. The ability to implement a meta-analytic review of many “n of 1” clinical trials united by the type of homogeneity indicated above would be an innovative way of shortening drastically the time to implementation in regular clinical practice of targeted combination therapies and perhaps at the end to save lives.

Additional factors, such as circumvention of a secondarily acquired drug resistance and/or intra-tumor heterogeneity, should also be considered in the development and use of new inhibitors.

To rationalize therapies and improve clinical responses, we need *to go beyond a set of statistically significant biomarkers*. In our opinion even a consistently altered landscape of mRNAs expressions can be considered a sort of snapshot “biomarker” of a given cancer condition. It is difficult to recognize which mRNA products (e.g., co-operating transcription factors) are the controller of downstream mRNAs, especially in the presence of mutated gene products.

We will discuss below to a significant extent a strategy of integrated understanding of the role of driver–gatekeeper mutations/somatically inheritable alterations, based on the reconstruction and mathematical dynamic modeling of a signaling-network subregion, a strategy capable of suggesting more rationally new associations of targeted inhibitors.

Artificial intelligence approaches integrated with vast amount of data generated by NGS and other techniques can be used to develop better treatments also for the cancer patient of the present day. The support of artificial intelligence and its computational approaches can help to suggest the most convenient treatment for an individual cancer patient of today [21, 54].

In the framework of the IBM Watson’s cognitive computing capabilities illustrated in Section 8.4, the partnership with the Mayo Clinic in Rochester, Minnesota [55], started on August 2014, launched a project to help oncologists to match patients with the right clinical trials. This pilot approach tries to develop an individualized treatment, making use of the speed and accuracy that Watson offers, in synergism with all clinical trials available at the Mayo Clinic, the patient records, and the public database ClinicalTrials.gov [56].



---

## 4 Combined Targeted Therapy: The Future Is Now

### **4.1 A Research Report Moving toward an Appropriate Selection of Oncoprotein Inhibitor Combinations, via a Pharmacologic Screening with Targeted Agents**

In a recent paper, Crystal and colleagues combined genetic analysis of resistant tumors with a pharmacologic screening with targeted agents [57].

In their innovative approach, they were able to establish 24 viable lines from patients' resistant lung tumors. Studying the genetics of tumor-resistant biopsies has the advantage that the discovered genetic alterations actually occurred clinically. They integrated the genetics of cancers with acquired resistance, with pharmacologic interrogation of cell lines systematically developed from those same resistant patients' tumors—a pharmacology discovery platform. Each cell line was tested against a panel of 76 drugs, mostly inhibitors of signaling proteins (cancer genes), 47 drugs approved for clinical use or already tested in clinical trials and 29 at a preclinical stage. Primary tumors of patients had been genetically characterized. In some cases a given gatekeeper mutation (i.e., EGFR or ALK) was of a “druggable” type. The patient initially responded to the first inhibitor. The authors worked with patient-derived resistance models. Some additional new gatekeeper mutation conferred resistance. A set of 76 drugs (mostly oncoprotein inhibitors) that could overcome resistance mutations were used in association with the initial TKI, in a pharmacologic combination screen.

The patient-derived resistance models were also analyzed by NGS to identify potential genetic causes of resistance. NGS and other omics-based strategies are extremely useful to guide treatment.

The authors highlighted how the addition of an MEK inhibitor was active in an ALK-positive resistant tumor that had developed an MAP2K1 (MEK) activating mutation. Genetic analysis of that cell line revealed both a mutation known to activate the MEK pathway and another mutation affecting an enzyme called JAK3. Only the pharmacologic screen was able to determine that resistance was conferred by the MEK mutation, since JAK inhibitors did not resensitize that cell line to ALK inhibition.

In other cases the association of EGFR and FGFR inhibitors was active in an EGFR mutant resistant cancer carrying a novel mutation in FGFR3. Simultaneous inhibition of both interconnected pathways, suppressed downstream signaling, resulting in growth arrest and cell death. In some cases tumor resistance was not the consequence of the appearance of a new resistance mutation but rather the consequence of a signaling-network modified feedback after administration of the first oncoprotein inhibitor.

Combined ALK and SRC inhibition was effective in several ALK-driven patient-derived models, a result not predicted by

genetic analysis (mutations and somatically inheritable alterations) alone. In fact, SRC goes up as consequence of a positive feedback of ALK inhibition. This was a network effect, not the consequence of an SRC mutation!

Acquired resistance may be mediated not only by a secondary resistance mutation but also by “a compensatory signaling pathway or bypass track,” stimulated by the first inhibitor drug and resulting in a secondary activation of an interconnected signaling pathway.

The “pharmacologic platform strategy” proposed by the authors could help to direct therapeutic choices for individual patients; however, for the moment the adopted approach remains technically cumbersome, probably beyond the practical technological potential of an average oncologic facility. With their present technology, cultures from lung biopsies or tumor effusions were successful only in a fraction of cases (~50%). It was technically possible to obtain such lines only in the presence of a feeder layer of irradiated fibroblasts. It usually takes 2–6 months to get a stable cancer line from a patient’s tumor. As a second drug inhibitor against a newly acquired resistance, they tried, after the initial inhibitor, their set of 76 antineoplastic drugs: a lot of experimental work was required to find a working association, in terms of cell growth inhibition. This seems an important limitation (in terms of required technology, costs, and time), especially for the clinical practice of an average facility for cancer treatment.

#### ***4.2 Mathematical Modeling of Signaling Networks: An Example of a Rationalized Approach to Targeted-Therapy Combinations***

A dynamic simulation approach at the level of protein–protein interactions within a signaling network is potentially valuable for predicting synergistic combination therapies (testing in the model the behavior of distinct drugs affecting distinct altered pathways). It represents a complementary approach to the previous one, but of much more practical implementation. Mathematical dynamic models can guide decisions regarding individualized options for the treatment of patients, treated simultaneously with more than one drug, inhibiting distinct specific oncoprotein targets, belonging to distinct pathways.

Driver–gatekeeper alterations have to be incorporated within altered pathways and signaling-network subregions. In this kind of approach, we can tune our model to the pathway alterations of a patient’s individual tumor. Pathway reconstructions will carry private cancer genes of that given tumor. Dynamic mathematical modeling will be able to detect signaling-network feedbacks (bypass track mechanisms) involved in specific cases of acquired resistance. In part, these may be suggested by previous works. They could also emerge directly from MIM (Molecular Interaction Map) reconstruction, dynamic modeling, and independent experimental validations. To understand the role of cancer-relevant gatekeeper inheritable alterations, the reconstruction of a

signaling-network subregion at the biochemical interaction level, following an extensive parameterization and pre-training of a corresponding dynamic model, is a natural complement to a deeper genetic characterization of the cancer of an individual patient. We will show below an example of modeling concerning CRC.

The central goal of this type of approach is predicting on a rational basis the behavior of associations of inhibitors of signaling proteins affected by excess of function in multiple specifically altered pathways.

A modeling approach can help to restrict the priority and the number of associations to be later tested/validated experimentally also at a cellular level. Against the molecular pathology of an individual patient, this containment of wet lab experimentations could represent a very significant practical advantage. As a vicarious substitute of a cancer line obtained by the patient's tumor, we could perhaps also employ an already existent and characterized cancer line, carrying the same type of driver mutations or somatically inheritable alterations. About 1000 cancer lines have already been deeply characterized at a genetic/epigenetic level, and this work continues to go on [58].

---

## 5 A Colorectal Cancer Dynamic Simulation Example

What follows is a modified synthetic report from a recent paper published by our group [59].

In the “didactic” dimension (overall setting) of a book chapter, we must keep in mind that, from a basic general perspective, chemical interactions/reactions are at the very basis of the life of any cell and therefore of life for short. In our opinion this dimension includes also prebiological and biological evolution. Even the formation of lipid bilayers and trafficking and spatial arrangements of molecules in organelles obey the same chemical-interaction laws, on the background of diffusion laws. Biochemical interactions/reactions will be the level at which we will treat and discuss our signaling networks.

The variety and complexity of the cancer facets in front of us force the choice of a specific type of solid tumor (we selected colorectal cancer—CRC) and, in the perspective of discussing possible rationalizations of targeted therapies, also the choice of a specific signaling-network subregion.

As we will show below, we focused our attention on a signaling-network subregion mostly involved in the G0–G1 cell-cycle transition. At this stage a cell takes crucial decisions about entering the DNA-synthesis S phase and a chain of subsequent events.

At a cancer level, the importance of this phase is underlined by the high frequency of driver–gatekeeper mutations and altered pathways, involved in this cell-cycle phase.

Not all but a significant portion of the G0–G1 signaling-network subregion involves pathways downstream of the TGF $\beta$  family, WNT family, and EGF family of growth-factor proteins. For reasons of practical confinement of our attention to a tractable dimension, we concentrate our attention on this signaling-network subregion.

Many preclinical and clinical treatments use inhibitors of onco-proteins affected by excess of function, alterations often present in these pathways in CRC (and other cancers as well).

Our dynamic modeling of the G0–G1 cell-cycle phase ends with the activation of transcription factors required for many protein neo-syntheses taking place during the S phase. We remained temporally at the transit through the activity of co-operating transcription factors and cofactors, because the events downstream of this transition imply a drastic increase in complexity and at the same time many “decisions” (accompanied with driver mutations) take place already during the cell-cycle phase we have explored in depth.

### **5.1 Dynamic Model Development**

We studied and implemented dynamic simulations of multiple interconnected downstream pathways [60–62]. Following a sort of partially subjective fuzzy logic, we summarize them according to ten pathway fragments.

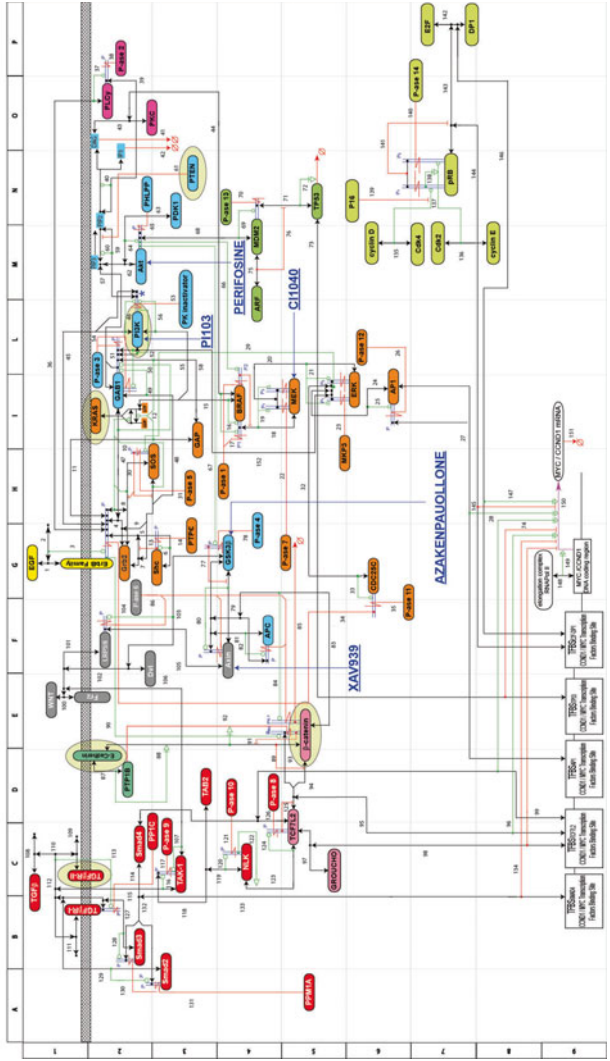
We described the section of the signaling network considered as an MIM (*see* Fig. 1) [63–65]. The example of dynamic modeling presented here [59] used ordinary differential equations (ODEs), and it involved 447 reactants (basic species, modified species, complexes, and inhibitors), 348 protein–protein interactions, and 174 catalytic reactions. The ODE models were developed and simulated with the SimBiology toolbox in Matlab (Mathworks) [66].

In our simulations, we assumed (as a logic default) that all reactions followed a mass action kinetic law (a consequence of the second law of thermodynamics). According to this kinetic law, the velocity of the reaction is directly proportional to the concentration of the reactants multiplied by the reaction rate. This is the main logical background of all our dynamic simulations.

In Fig. 1 (reproduced from [59], with permission of 2008–2015 Impact Journals, LLC), we show the MIM we have simulated.

Applying a sort of fuzzy logic, our MIM can be summarized in the following pathway fragments:

1. Pathway [ErbB family receptors–PI3K–PTEN–Akt–GSK3 $\beta$ –APC– $\beta$ -catenin–TCF7L2–DNA binding site 2, transcription agonist].
2. Pathway [ErbB family receptors–Grb2–Shc–SOS–GAP–KRAS–BRAF–MEK–ERK–AP1–DNA binding site 3, transcription agonist].



**Fig. 1** Our reconstructed and simulated MIM. The interconnected pathways of the TGFβ family, WNT family, and EGF family of growth-factor proteins. The cartouches of mutated/altered signaling proteins (HCT116 line) have been surrounded by an oval in light yellow. Names and sites of inhibitors' activity are also indicated. Slightly modified from Fig. 1 of [59], with permission of 2008–2015 Impact Journals, LLC

3. Pathway [ErbB family receptors–E-cadherin (cadherin/cadherin adhesive complex)].
4. Pathway [ErbB family receptors–PLC $\gamma$ –PIP2–PKC–BRAF–MEK–ERK–API–DNA binding site 3, transcription agonist]; the terminal parts of pathways 2 and 4 are the same.
5. Pathway [WNT–Frz/LRP5/6–Dvl–AXIN–APC–GSK3 $\beta$ – $\beta$ -catenin].
6. Pathway [TGF $\beta$  receptors–SMAD2/3–SMAD4–DNA binding site 1, transcription antagonist].
7. Pathway [TGF $\beta$  receptors–TAK-1–TAB2–NLK–TCF7L2, converging with 8].
8. Pathway [WNT–Frz/LRP5/6–TAK-1–TAB2–NLK–TCF7L2, converging with 7].
9. Pathway [Akt–MDM2–TP53].
10. Pathway [cyclin (D/E)/CDK (2/4)–pRB–E2F:DP1].

We disturbed the behavior of our model by introducing virtual inhibitors (perturbation approach). We validated our predictions with wet experiments in two CRC lines: HCT116 and HT29, respectively. We used the following inhibitors, alone or in combination:

- MEK inhibitor CI-1040 (also known as PD184352)
- PI3K inhibitor PI-103
- AKT inhibitor perifosine
- GSK3 $\beta$  inhibitor azakenpaullone
- Tankyrase inhibitor XAV939
- Pan-transcriptional inhibitor actinomycin D

## **5.2 Implementation of Individual Cancer Patients in the Model**

Our model initially simulates a “physiologic state.” The model can be subsequently adapted to simulate individual pathologic CRC conditions. In perspective, we move toward personalized cancer models, by the implementation of patient-specific alterations and/or mutations in relevant oncoproteins.

We generated personalized models for two CRC cell lines, intended in perspective as a proxy for individual tumor data:

### **Mutations/alterations considered in our HCT116 cancer line:**

PTEN (60% of HT-29 level—our experiments), KRAS, PI3K,  $\beta$ -catenin, TGF $\beta$  receptor II, E-cadherin

### **Mutations/alterations considered in our HT29 cancer line:**

ErbB2 ~2X of HCT116 level—our experiments), BRAF, PI3K, APC, SMAD4

Having established our model, we verified experimentally some salient model predictions using the mutated colorectal cancer lines HCT116 and HT29 (from ATCC). The mutations present in our ATCC CRC lines are described in the CCLE database [58].

Notice that some mutations involve different proteins but target the same pathway.

Analyzing the spectrum of mutations present in our two colon cancer lines, it is possible to observe four dominant mutations in terms of pathway activation: KRAS (HCT116),  $\beta$ -catenin (HCT116), BRAF (HT29), and PI3K (both lines). In the dynamic simulations, these mutations have been implemented according to the simplifying rule that the mutated protein will remain in a phosphorylated or in an “active” form. According to our experimental results, HT29 cells overexpressed ErbB2 ~twofold, and this variation was also inserted in our simulations.

In our CRC lines we also had five alterations involving loss-of-function mutations, resulting in an alteration of the corresponding pathway: PTEN (HCT116), E-cadherin (HCT116), TGF $\beta$  receptor II (HCT116), SMAD4 (HT29), and APC (HT29). Notice that the final consequence of these loss-of-function mutations is always in favor of the G0–G1 transition.

To model these loss-of-function mutations, we set at a zero level their corresponding protein concentrations, thereby simulating the absence of a functional protein.

In the case of PTEN in the HCT116 cell line, a more complex situation was reported [67–69].

Our direct experimental assessment also suggested only a partial PTEN inactivation. For HCT116 cell line, PTEN homozygous mutations consist of a single-nucleotide deletion at the first part of the mRNA 3'-UTR that could interfere with the seed region of a regulatory miRNA. In fact PTEN expression was recently described to be fine-tuned by miRNAs that have seed sites in the mRNA 3'-UTR region, in equilibrium with the mRNA 3'-UTR region of pseudogenes that show a “sponge effect” [67]. We therefore put the protein concentration of PTEN at 60% of its physiological value, according to our experimental results in HCT116 cells.

We were also able to simulate the presence of a single functional allele (data not reported).

During the long pre-training phase of our dynamic modeling (about 3 years), we implemented a long and painstaking patchwork approach, a work of continuous parameter readjustments aimed at interpolating/adjusting multiple parameters, in order to fit the results reported in several dozens of pertinent molecular, preclinical, but sometimes also clinical, oncology papers published in reputable journals.

Our very long and patient tuning of the model during the training phase was a process somehow describable as a parallel “in silico evolution,” where likely deficiencies at the level of network components and parameters inputs have been somehow compensated during the gradual, step-by-step, 3-year-long evolution of our dynamic model.



### **5.3 Experimental Verification of the Dynamic Model through P-Protein Levels and mRNAs Regulation: Statistical Analysis of the Results**

After the pre-training phase of our model, we submitted the model to a posteriori experimental verifications, to compare simulated and experimental data.

We performed two classes of experiments on HCT116 and HT29 cancer cell lines, in the presence or absence of different signaling-protein inhibitors, alone or in combination. Only inhibitors which acted at or downstream of mutated genes were considered.

We implemented a semiquantitative assessment of pp-ERK (Thr202/Tyr204) and pAKT (Ser473) protein levels, using western blots and measuring blot intensity. We mostly considered [P-protein/total-protein] ratios. We also evaluated mRNA levels for MYC and CCND1, by qPCR. Changes in protein phosphorylation before transcription (spatially located in the upper and larger part of our MIM (Fig. 1) above of the transcription region) take place rather rapidly, and they were experimentally assessed at 30 min (and 60 min with substantially similar results). After preliminary time-series experiments, we performed our wet lab experiments and simulations for mRNA levels at conveniently longer times (4–8 h).

The fundamental aim of this phase was to examine to what extent the level of our partial information was already generating sensible predictions. We wanted to verify if our general approach seemed tendentially correct, requiring only possible and achievable gradual incremental improvements, in order to produce models with increasing clinical utility.

To illustrate our point, we present below the observed relationships between the predictions of our mathematical dynamic modeling and the subsequent results of wet lab experiments.

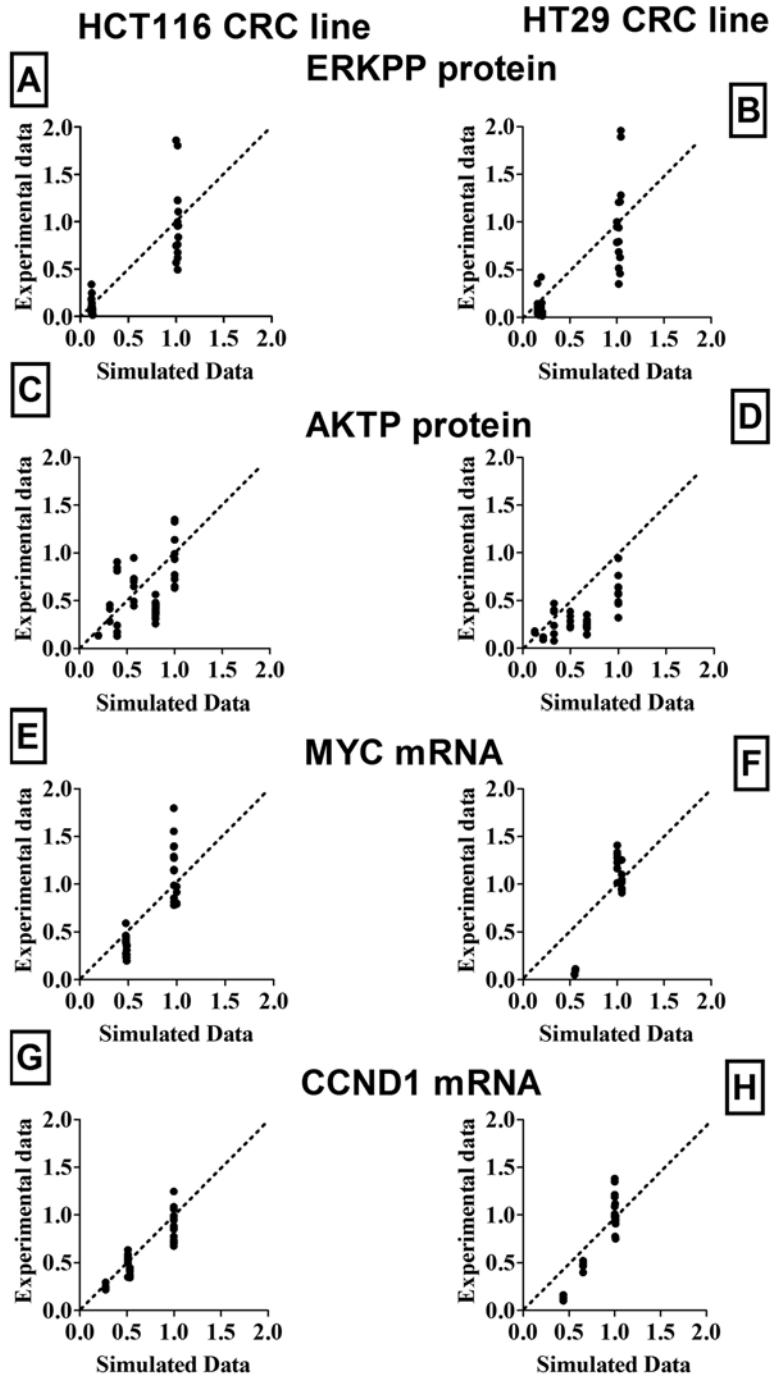
In the first two plots, we report the [P-protein/total-protein] ratios for the signaling proteins ERK and AKT, respectively, in the cancer lines HCT116 and HT29, respectively.

The next two plots refer to MYC and CCND1 mRNAs levels, always in the two cancer lines.

In Fig. 2 (modified from Figs. 5 and 7 of Ref. 59, with permission of 2008–2015 Impact Journals, LLC) for ERKPP and AKTP protein phosphorylation, c-MYC and CCND1 mRNA levels (in HCT116 and HT29 cell lines, respectively), we drew a 45° diagonal line between the  $X$  axis (simulated data) and the  $Y$  axis (experimental data), reflecting a theoretic 1:1 match. Both in the case of strong inhibitions and weak or absent inhibitions, simulated and experimental data tended to go together reasonably well.

Notice that, as expected, an experimental variability is present in the wet lab experiments, but not in the mathematical dynamic simulations, where one input condition generates just one prediction.

A statistical analysis of the data confirmed our qualitative impression of an already reasonable behavior of our mathematical



**Fig. 2** Scatter plots of experimental (Y axis) versus simulated (X axis) values in response to different inhibitor treatments. (a, b): ERKPP protein. (c, d): AKTP protein. (e, f): MYC mRNA. (g, h): CCND1 mRNA. Reproduced, in a modified form, from Figs. 5 and 7 of Ref. 59, with permission of 2008–2015 Impact Journals, LLC

model. Considering the different distributions of the data on the  $X$  axis (simulated values) and the  $Y$  axis (experimental values), we started our statistical analysis with a nonparametric approach (the **Spearman’s rho** test).

However, we also implemented an analysis of the **R2** coefficient (coefficient of determination), to get an idea of which fraction of the wet lab experiments ( $Y$  variable) was predicted by the dynamic simulation ( $X$  variable) [70]. Notice that there is no evidence of non-normality in  $y$  experimental values, the  $x$  values (simulated values) are obviously not subjected to experimental errors, and we are not actually performing statistical inference linked to the normality assumption [71].

Most predictions were already reasonably satisfactory; however, some modeling refinement seems to be suggested at the level of AKTP related pathways (statistical Table 1, modified from Tables 1 and 2 of Ref. 59, with permission of 2008–2015 Impact Journals, LLC).

Our statistical analysis does not want to convey a misleading message about our stage of MIM reconstruction, parameterization, and mathematical modeling, as if it was already close to a working instrument. What we have shown instead is that the multiple (and inevitable at this stage) deficiencies in input are definitely

**Table 1**  
**Spearman’s rho (first three columns); R<sup>2</sup> coefficient for a linear regression (fourth column)**

	<b>Spearman’s rho</b>	<b>95% confid. interv.</b>	<b>p-val (two tailed)</b>	<b>R<sup>2</sup> (all individual experim. results consider.)</b>
<i>HTC 116 CRC line</i>				
ERKPP (protein)	0.56	(0.28, 0.73)	0.0002	0.77
AKTP (protein)	0.55	(0.23, 0.77)	0.0004	0.35
MYC (mRNA)	0.59	(0.34, 0.68)	0.0002	0.81
CCND1 (mRNA)	0.76	(0.58, 0.86)	6E-07	0.82
<i>HT29 CRC line</i>				
ERKPP (protein)	0.66	(0.39, 0.80)	9E-06	0.75
AKTP (protein)	0.61	(0.30, 0.80)	5E-05	0.49
MYC (mRNA)	0.54	(0.03, 0.84)	0.007	0.96
CCND1 (mRNA)	0.53	(0.05, 0.82)	0.009	0.91

Data were computed for the HTC116 and HT29 cell lines, respectively, in all experimental conditions (two protein and two mRNAs end points for each cancer line). Reproduced, in a modified form, from Tables 1 and 2 of Ref. 59, with permission of 2008–2015 Impact Journals, LLC

not generating a noise obscuring any connection between model and experiments. Approximations, simplifications, and incompleteness, in the MIM reconstruction of our signaling-network subregion (G0–G1 cell-cycle transition), are obviously present (and inevitable). However, we have already a prevalence of information over noise, and this could be further reduced in the future. Along this road, future incremental progress seems quite possible.

**5.4 Recent  
Independently  
Published Results  
Have Been Predicted  
by Our CRC Model**

Well after our model finalization and our own experimental verifications, the predictions of our model were tested against preclinical results obtained by independent investigators, in DiFi, LIM1215, HCA-46, and OXCO-2 CRC lines, before and after induction of panErb resistance through a subsequent KRAS mutation [72]. These lines were initially sensitive to the panErb inhibitors cetuximab or panitumumab, but resistance emerged through subsequent new KRAS mutations. Misale et al. [72] demonstrated that, in CRC lines that had become resistant to panErb inhibitors, the addition of MEK inhibitors could partially overcome resistance. This resensitization was more complete when the panErb inhibitor continued to be given with the MEK inhibitor.

We examined in our model the simulated behavior before and after the emergence of a (post-KRAS mutation) resistance to panErb inhibitors, implementing an ad hoc MIM modeling, in the absence or presence of KRAS mutations.

We simulated the presence of a panErb inhibitor, an MEK inhibitor, or both and observed that the behavior of [P-protein/total protein] for EGFR, ERK, and AKT (at 30 min–1 h) was well compatible with the authors' observations.

Moreover, c-MYC and CCND1 mRNAs (both playing a crucial role for cell replication) were completely normalized (at 4–8 h), only by the combination of panErb and MEK inhibitors in the presence of a mutated KRAS. Our simulations suggest that this behavior is due to a synergic effect of the two inhibitors, which target two different pathways: MEK inhibitor downstream of the KRAS mutated pathway and the PanErb inhibitor along the PI3K-AKT pathway. This synergism at a protein and mRNA level seems in line with the one observed by the authors at a cellular level. Our dynamic modeling seems to offer a mechanistic explanation of what is going on in an altered signaling network.

The important message (reinforced by statistical analysis) is that an incremental path forward seems now open, justifying more long-term future cooperative efforts in the direction of building larger MIMs, supported by more input parameters at the level of molecular (signaling-protein) concentrations and reaction rates.

Our present model is already intriguing and encouraging, but we work toward future more advanced models, as operative instruments for a rationalization of the treatment of individual cancer patients, with the intended future goal of implementing rational

associations of inhibitors of specifically altered pathways in a given specific cancer of a specific patient.

---

## 6 Cancer Therapy in the Next 3–5–10 Years

It is perhaps true that major innovations in cancer therapy could require an approach similar to a table with three legs:

1. A more personalized, more detailed, molecular characterization of individual tumors, for instance, through next-generation sequencing (NGS) targeted toward the most frequent driver mutations for a given type of tumor (DNA analysis on tumor samples and liquid biopsies?) plus the analysis of other important driver somatically inheritable alterations.
2. A more comprehensive reconstruction and modeling of the behavior of a larger signaling-network subregion. In the example illustrated above (Subheading 5), we have focused our attention on the G0–G1 cell-cycle transition. Both physiologically and pathologically important cell-replication decisions are taken at this stage. In fact, the signaling network concerning this cell-cycle phase is especially rich of driver and gatekeeper mutations.
3. A larger and better assortment of signaling-protein-targeting selective drugs. Here the multifaceted aspects of medicinal chemistry come into play. New molecules available for clinical trials will definitely come into play during the next 3–5–10 years. Always with reference to the field of oncology, they are at the research focus of most pharmaceutical companies, large, medium, and small, often in synergism with the academic world.

---

## 7 Toward a New Clinical Trial Strategy

Suppose that the modeling example we have illustrated above (Subheading 5) is followed by more extended MIM reconstructions, accompanied by much more systematic parameterization (both for involved signaling proteins and reaction rates). At this point the inhibitor combinations recommended by the model could become unsatisfactory quite infrequently. In addition, validations at the cellular level, concentrated on a small number of drug combinations, would be reasonably easy to perform, considering also the high number of genetically extensively characterized cancer lines [58, 73], a number steadily growing. It is in principle not difficult to find one or more cancer lines of the same histological type, carrying the same (or quasi the same) driver–gatekeeper genetic alterations than the tumor of a given patient, and to implement on them experimental validations of the proposals coming from the mathematical model.

A deep genetic/epigenetic characterization of patients' cancers creates recruiting difficulties in building homogeneous subsets, both in terms of subset size and required time for patients enrolling. These types of difficulties have also been discussed in a recent FDA Breast Cancer Workshop [52]. Several speakers touched the issue that when one, or more than one, driver-gatekeeper mutation(s) important in an individual patient's cancer corresponds to "low-prevalence biomarkers" (using their terminology), then few thousands of BC women would already be needed to get a homogeneous subset of 40–50 women (convenient for a phase II study involving a compound with a large effect size compared to controls). A phase III study could involve numbers about ten times larger [52].

Considering the inevitable slow pace of the above procedures, much quicker suggestions of rational drug combinations for each personal cancer case could come by more advanced versions of the example of mathematical modeling shown above.

**7.1 The  
"Homogeneity  
of a Treatment  
Strategy"  
versus Genetic/  
Epigenetic  
Homogeneity  
in a Patient Subset**

We have suggested (Subheading 3.1) that the "homogeneity of a treatment strategy" can come from the "homogeneous procedure" adopted for selecting the personal treatment combination, downstream of an efficient network reconstruction and modeling, rather than a traditional genetic/epigenetic homogeneity within a small subset of patients, requiring long recruitment times.

If regulative and ethic procedures could be adapted to this different new context, we could have *treatment sets homogeneous for the procedural decision strategy adopted*, not for each personal genetic analysis of each individual patient.

In the presence of a sufficiently deep patient-by-patient genetic characterization, the more traditional genetic/epigenetic homogeneity is problematic especially in terms of recruitment times and also times required to reach a clear assessment of a given trial.

On the contrary, *at a meta-analysis statistical level*, it would take a relatively short time to discover if this different personalized approach, based on a homogeneous decisional strategy for treatment, rather than very small homogeneous patients' subsets, works better and faster, in respect to past more traditional diagnostic and treatment approaches.

To achieve the implementation of this new approach in standard oncologic structures of technologically sufficiently advanced countries, new ethical considerations and new legal rules will have to be established, through a progressive consensus process, at the level of both diagnostic and therapeutic behaviors.

The genetic complexity and the heterogeneity of each individual tumor lead to a variability in patients' response to a more traditional approach to therapy: only a small sub-cohort of cancer patients will have improvement in response to a given treatment [20]. On the side of a more stringent genetic/epigenetic characterization, initiatives like the I-SPY 2 trial [74], the MATCH trial

[75], and the SAFIR02\_Breast trial [76] imply the recruitment of small homogeneous subsets that, as a consequence of their homogeneity, could display (after a targeted treatment) a statistically significant behavior even in the presence of small patient numbers.

The identification of sufficient numbers of patients with individual (rare) genetic aberrations for enrolment into homogeneous clinical trials is an important obstacle to a rational introduction of combinations of targeted agents.

This more “orthodox” approach, and our more “heretic” approach, could turn out to be complementary and convergent on the long run, but our approach could save precious time.

Our proposal could improve the average rationality of an association of targeted drugs.

We are well aware that our proposal is just at the beginning of a new strategic road that could turn out to be a road on which we can walk on to advance further. We don’t exclude unexpected difficulties.

## **7.2 Precision Medicine in Oncology According to a Recent Perspective**

Francis S. Collins and Harold Varmus at NIH–NCI recently commented an important point of President Barack Obama’s State of the Union address, of January 20, 2015: “Tonight, I’m launching a new **Precision Medicine** Initiative to bring us closer to curing diseases like cancer and diabetes—and to give all of us access to the personalized information we need to keep ourselves and our families healthier” [77].

Among the many points already touched in the previous sections of this book chapter, let us recall synthetically some of them through the phrasing of Collins and Varmus: “Research has already revealed many of the molecular lesions that drive cancers, showing that each cancer has its own genomic signature, with some tumor-specific features and some features common to multiple types”. And later on they speak of: “therapeutic strategies, with increasing use of drugs and antibodies designed to counter the influence of specific molecular drivers.” They also mention the need of knowing more about combinations of oncoprotein inhibitors: “These features make efforts to improve the ways we anticipate, prevent, diagnose, and treat cancers both urgent and promising.” They even touch the need of future “clinical trials with novel designs,” a point to which we dedicated a specific proposal (Subheadings 3.1,7, and 7.1).

Optimal international utilizations of large open databases of multiple kinds are also an important issue in their mind. We stressed the importance of some of them in Subheading 2.1.

Future, open-access, very large longitudinal cohorts of voluntary citizens followed for many years will increase the role of these very large databases. Every bright scientific and clinical mind all over the world will be able to give a contribution in understanding.



The authors are aware that it is important to start, but “this medicine initiative will probably yield its greatest benefits **years down the road.**”

The authors also touch other relevant points, but it was a pleasure to acknowledge a communality of wavelength between this book chapter and their perception of the future of precision medicine in oncology.

---

## 8 Next-Generation Decision-Support Systems in Oncology

### 8.1 *Decision-Making in Oncology*

In a medical field like oncology, decisions are taken when formulating a diagnosis or recommending a therapy or reviewing therapy outcomes. Such decisions are complex with respect to the dimensionality of the problem itself and the large set of variables and facts that need to be taken into account.

Tumors have been originally classified at a morphologic microscopic level (histology). As the name itself suggests, the classical distinction was according to the tissue of origin. Within a given tissue of origin (e.g., the breast) we can have, as the starting point for a malignant transformation, different cellular types, for instance, the ductal and lobular types. Nowadays, however, the derivation of “ductal” and “lobular” breast cancer from “ductal” and “lobular” normal cells is strongly disputed [78]. In both cases we can have more initial breast cancer (BC) lesions (carcinoma in situ) or more advanced infiltrating carcinoma. At a molecular level (e.g., [1, 3]), a better characterization has been based, already for some years, on biomarkers, like the presence of estrogen and progesterone receptors, the level of the HER2 tyrosine kinase membrane receptor, and other biomarkers. A clear positivity of the estrogen receptors or the HER2 receptor opened the way to treatments with corresponding drug inhibitors. Usually a triple-negative BC (for the three biomarkers just mentioned) has a less favorable prognosis.

In recent years cancer has been viewed as a multi-hit alteration in multiple interconnected pathways at the level of biochemical networks among signaling proteins. This perspective and the therapeutic possibilities it opens have been discussed at length in Subheadings 2–7. Using rational associations of sufficiently selective inhibitors of signaling proteins (oncoproteins) affected by excess of function, clinical treatments are going to change dramatically probably within 3–5 years (Subheadings 5, 6, and 7).

Hospital oncologists usually are broadly aware of the therapeutic strategies already widely accepted, also because they can easily find on the Web a variety of practical updating and comments. However, as we mentioned in the introduction of this chapter, the complexity and variety of the issues involved is very high, and an individual decision-maker (e.g., an oncologist) could easily miss some relevant detail. A decision-support system could therefore be

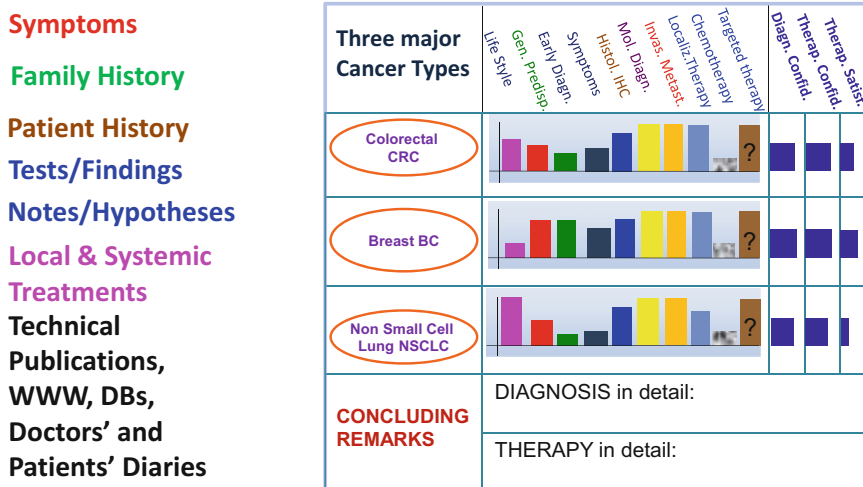
very useful, even more in a relatively small hospital facility. In a small facility awareness about very infrequent outcomes of treatment or new trends could be less well known.

In Europe and many industrialized countries, in terms of incidence, cancer is more and more a disease of elderly patients. We make reference to this subset of cancer patients, to give a non-irrelevant example of the multiplicity of issues and complications involved in cancer therapy.

The most commonly administered chemotherapy (CT) regimens, but even more modern molecularly targeted agents, confer survival benefits but cause significant toxicity. In an elderly cancer patient, the toxic side effects of a pharmacologic treatment tend to be amplified, and CT by itself is thought to accelerate the process of aging. In addition, elderly patients are often affected by a variety of comorbidities, which make even more delicate the overall pharmacological treatment. This is not the place for discussing more in detail the issue of the elderly cancer patients. It is however rather evident that this could be a natural area of expansion of a decision-support system, in a systems medicine perspective of broad integration and elaboration of inputs.

**8.2 Modern Artificial Intelligence for Systems Medicine**

Systems medicine should benefit from the latest advances on what we may collectively call modern artificial intelligence (AI). Modern AI is strictly related and influenced by disciplines such as computer science, machine learning, and signal/natural language processing, to design the new generation of decision-support systems (DSS). In Fig. 3 we describe the functional architecture of such DSSs in an oncology context. Computers may be able to process data that are relevant for a patient case: from his/her symptoms, patient/family



**Fig. 3** Decision-making with the support of confidence parameters. Modeling from diverse sources of information

track records, test results, effectiveness of past treatments, as well as updated technical publications (journals, Web), doctors' and patients' forums, and personal doctor/patient diaries.

Modern computational systems are able to provide an important component of a decision: especially the degree of confidence we can have in a given decision or an alternative decision.

Each cancer-type hypothesis may be weighed in by its conditional confidence measure:  $P(\textit{breast BC} | F)$ , where *breast BC* is the cancer type and *F* is the multidimensional distribution over symptoms, family records, etc. (see Fig. 3). Today we have available to us large amounts of preclinical, as reported in Subheading 2, and clinical information (e.g., [1–20]), which however has to be integrated and interconnected, in a functional way. In addition, this type of software has to be frequently and regularly updated. This is definitely a very important opportunity in a systems medicine perspective. The overall size of the potentially available and usable global information is so large that it goes over the capabilities of a single clinical oncologist and an average oncological team. An artificial intelligence DSS (decision-support system) can support oncologic clinical decisions, both at the diagnostic and therapeutic levels. In synergism with a deeper molecular diagnostic level, and the availability of a more rational panel of choices in terms of more rational drug combinations [57], a DSS can significantly support a more performing personalized medicine treatment.

The diversity of input variables and time series makes the problem highly complex from the engineering point of view of the data/process integration as well as from a data analysis/prediction point view. However, the current state of the art in the aforementioned disciplines is at a point where the challenge of designing computational architectures is not only achievable in the midterm but more importantly can have a midterm disruptive effect in the medical domain.

In the next sections we review the scientific approach of modern AI and the early experiments on the clinical trials of modern DSS technology and close with the future perspectives of AI-based decision-support systems.

### **8.3 The Modern AI Approach**

The new generation of decision-support systems (DSS) in medicine should be able to:

1. Read and interpret automatically Web-scale collections of medical publications from around the world, including the very large and continuously updated databases we have mentioned above. Notes and diaries from doctors and patients can also be useful. Create knowledge bases from this process that can be used for inference generation (e.g., deductions about diagnosis) or decision recommendations (e.g., suggestions about therapy).

2. Analyze and interpret doctor–patient vis-à-vis conversations, physiological signals generated from continuous streams of brain–computer interfaces and wearable computers.

These processes are summarizable as “natural language processing” capabilities and “pattern recognition” capabilities.

3. Integrate large-scale-structure data from the above **step 1**, with available on-time patient’s data and local (hospital) or global (health networks and professional communities) knowledge bases, for inference generation and decision recommendations.
4. Support the decision of medical doctors, who will use facts, hypotheses, inferences, and recommendations generated automatically, to select decisions using appropriate risk models.
5. Decisions should be generated via modeling the confidence and risk based on the best evidences derived from all sources (*see* Fig. 3):

For instance, we may envision that the next-generation DSS would be able to generate the following decisions, given the inputs that can be any of the type in Fig. 3:

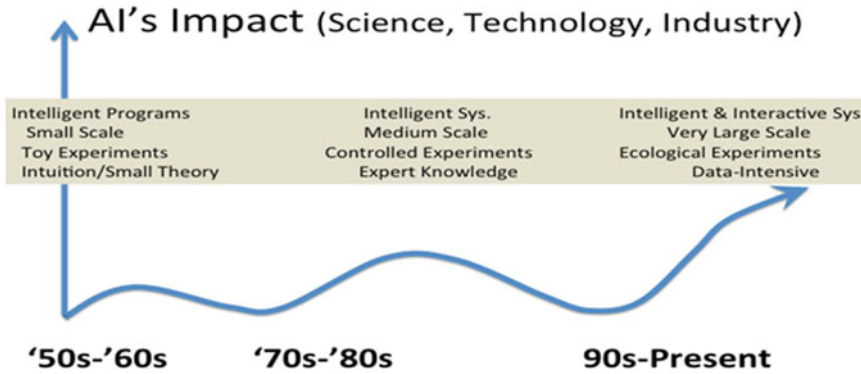
- The best diagnosis that has a confidence 0.8 is  $X$  (roughly this would mean that the DSS expects  $X$  to be the correct diagnosis with 80% probability).
- The expected success, with confidence 0.5, is  $\mathcal{Y}$ .
- The expected quality of life (for the next  $N$  months), with confidence 0.7, is  $Z$ .

The cognitive abilities of modern DSSs in **steps 1–3** are one of the most important and exciting achievements of an ensemble of disciplines, including computer science, computational linguistics, machine learning, and electrical engineering [79].

We will collectively use the term “modern AI” to refer to a machine that is able to exploit the collective components of the aforementioned disciplines.

In the past AI has had oscillating successes and failures [80]. The original plan and vision is part of the current vision. In Fig. 4 we characterize the type of intelligent systems based on AI as function of the time from the 1950s through the present time and the impact it had on science and industry development. In the early 1960s John McCarthy [81] was planning to program a machine with common sense and generating new knowledge without being reprogrammed. Marvin Minsky would envision and model micro-worlds where machines (robots) would operate and arrange autonomously blocks (block world).

Almost 20 years later, the MYCIN [82] research project was expected to provide types of AI systems to the end users, namely, the health professional community. MYCIN was supposed to diagnose blood infections. The MYCIN expert system included 450



**Fig. 4** Intelligent systems evolution over time

rules designed by the domain engineers and acquired through interviews with domain experts. MYCIN incorporated a calculus of uncertainty called “certainty factors,” which attempted to mimic how doctors assessed impact of evidence on the diagnosis.

The expectation on the impact and performance of such types of systems rose to the point that by the mid-1980s, most corporations, in diverse industry sectors, had assembled an AI group.

The transition from the research lab and in vitro experiments to the real-world problems, data and end users, was not smooth.

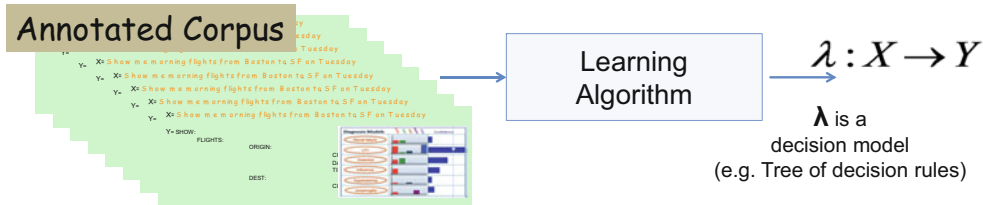
Within a few years, industry stopped investing into AI technology, and researchers entered the so-called AI winter. At the same time, independent events, such as the availability of storage and computing across research communities, favored an empiricist approach to scientific investigation and engineering models. For instance, hidden Markov models (HMMs) [83], although not very accurate from a modeling point of view, learned extremely well from examples (observations) when applied to speech technology. HMMs and other types of statistical models quickly were adopted and found successful exploitations in a wide range of disciplines and applications.

One of the most important events in the AI disciplines was the adoption of experimental scientific methodology. It required a *modus operandi* expecting scientists to prove or falsify hypotheses on the basis of objective data and sound and repeatable experimental design.

In Fig. 5, we outline the prototypical process of machine learning [84] from examples: large amounts of structured and unstructured data are annotated by experts. Such annotations are analyzed and then fed into automatic learning algorithms that will model the decision-making in a statistical sense. There have been great empirical validations of mathematical models (e.g., decision trees). They have been applied to diverse fields from engineering to neuroscience and to economics [85–87]. Last but not least, human–

# Learning from Examples

$(X, Y) = (\text{Publications,}$   
 $\text{Medical Records}$   
 $\text{Patient Diaries}$   
 $\dots\dots\dots,$   
 $\text{Human Decision ( Diagnosis, etc.. )}$



**Fig. 5** Machine learning from examples. Examples are pairs of annotations  $(X, Y)$ , where  $X$  may include the medical records, medical tests, excerpts from decision-relevant publications, etc., and  $Y$  is the annotation (decision) from a domain expert. The machine-learning algorithm elaborates the observations  $(X, Y)$  and infers a decision model such as a decision tree (see Ref. 84 for a more exhaustive coverage)

computer interaction engineers started to design such systems for end users that were looking to benefit them and not to burden them with the machine-intelligence-world-specific aspects.

In the last few years, intelligent systems have been designed to elaborate very large amounts of data, generated from diverse sources and with different features (e.g., scholarly publications, including very large databases, but also patient social media networks).

The ability to take advantage of very large amounts of observations generated in realistic environments (e.g., robots in a factory plant) is both an ongoing research challenge and opportunity. At present most information technology companies are using modern AI as part of their operation infrastructure. The goal in the next years is to expand and reach domains with high societal and economical impacts, including the health domain.

## 8.4 Modern AI in the Medical Domain

In the last 10 years there have been various initiatives to engage doctors and the medical community with modern AI technology. There are established conferences/journals and companies addressing the *automatic processing of medical publications* and *design of very large knowledge bases*. An increasing number of companies provide services to create medical knowledge bases from scientific publications. An important initiative to design modern DSS based on the latest advances in intelligent systems is the *Watson project at IBM* [21–23] and its applications to the medical domain.

The IBM Watson group has started research collaborations for the deployment of their questioning/answering system within hospitals' operations. In particular early experimentation results have been published by the Memorial Sloan Kettering (MSK) Cancer Center in the oncological domain [54].

Medical oncologists, working on a variety of cancer types, from MSK's regional network, were invited to evaluate the IBM intelligent system. These physicians were managing patients affected by different types of cancer (e.g., NSCLCs, colorectal and breast cancers). The results are encouraging in terms of expected useful decision support for a clinical oncologist, but there is still a need for further progress, in terms of systems performance and support to the end-user decisions. IBM Watson is in partnership (among others) with the MD Anderson Cancer Center, where Watson helps oncologists to create individualized treatment for leukemia patients, and with the New York Genome Center, where Watson helps the experts to examine the molecular profile of each patient to identify drugs for individual treatment [20].

A method of analyzing complex patient history data into a summary of information most important to a physician is one of the possible additional applications underway for the IBM Watson technology [55, 88].

Several approaches are developing to aggregate and evaluate data elements currently captured in disparate health-care settings and information systems to inform the development of "omics" clinical practice guidelines.

CancerLinQ is a data informatics system of the American Society of Clinical Oncology (ASCO) and its Institute for Quality, developed to monitor, coordinate, and improve the quality of care provided to patients with cancer through the collection, aggregation, and analysis of data extracted from the EHRs (electronic health records) and practice management systems. CancerLinQ will rapidly analyze information and provide clinical practice guidelines to physicians and real-time clinical decision support to facilitate treatment planning for specific patients [89, 90].

These types of initiatives, given the societal impact of such intelligent computers, should be extended across different medical domains, hospitals, and countries.

At a technological level, an IBM team [91] developed an efficient, scalable, and flexible non-von Neumann architecture broadly inspired by the brain's structure (TrueNorth/SyNAPSE technology). They built a 5.4-billion transistor chip with 4096 neurosynaptic cores interconnected via an intra-chip network integrating one million programmable spiking neurons and 256 million configurable synapses. Chips can be tiled in two dimensions via an inter-chip communication interface, seamlessly scaling the architecture to a cortex-like sheet of arbitrary size (excerpts from the summary of their article). According to the above authors, the



IBM Watson project could derive important advantages from this new technology. In addition, they consider “TrueNorth” a direction and not a destination.

### **8.5 Future Perspectives**

The vision for a modern DSS is that it will be a crucial tool (in terms of performance enhancement) supporting the decision-making process of health professionals. For legal reasons, AI systems shall be intended as a support to the final decisions of an expert field specialist, not as a replacement.

The interaction of a doctor with these advanced and interactive expert systems can decrease the risk of omitting or forgetting some important parameter or set of facts. At the same time, these systems could improve the decision-making process of clinicians, by providing them with decisions based on up-to-date and validated suggestions, by a multiplicity of different professionals, carrying a multiplicity of complementary expertise. This help could contribute to both a wiser diagnostic assessment and a better therapeutic decision. Modern DSS could assume an even more important decision-support role, for small medical centers, which are dealing not only with cancer but also with a multiplicity of internal-medicine pathologies.

We are still at the infancy stage of training DSS for medical applications. However, the research and development in this field has advanced to a tipping point where realistic experimentations can be carried out in work environments (e.g., hospitals) with real users (e.g., medical doctors). The current state of AI science and engineering is a good predictor for an upcoming paradigm shift in systems medicine.

---

## **9 Conclusions**

Subheading 5 focused its attention mostly on the possibility of rationalizing the detection and implementation of appropriate combinations of oncoprotein inhibitors, on the background of a more advanced reconstruction, pre-training, and parameterization of altered signaling-network pathways. These models can incorporate (and take advantage of) a deeper genetic characterization of the tumor of any individual patient, a diagnostic progress which is presently progressively taking place (Subheading 6). In addition, these models can easily take advantage of the discovery of new selective inhibitors (Subheading 3), because the biochemical interference of these types of molecules can be “naturally” included in a dynamic model of biochemical interactions, of the type we have illustrated.

Subheadings 4 and 5 were therefore intended to facilitate the choice of new rational drug-treatment combinations. Clinical out-

puts have already started, and we can reasonably expect an explosion of these more rational targeted associations within the next 3–5–10 years (Subheadings 6 and 7).

Subheading 8 focused its attention on suggesting the most convenient treatment for an individual cancer patient, as of today. The support by AI and computational approaches facilitates enormously the continuous feeding, updating, and integration of cancer-related new data, overcoming the obstacles posed by their disparate levels, the consequent problems of normalization and linkage of heterogeneous data, and often their enormous database size.

Advances in AI and computing technologies *will shrink the gap* between:

- What could be done based on integrating all possible knowledge about the personal cancer of a given patient, including combination therapy suggestions coming from more advanced computational models (future developments of the subject touched in Subheading 5)
- What a clinical oncologist can do today in an average hospital

---

## Acknowledgments

This work was supported by the Italian Ministry of Economic Development “Industry 2015—Made in Italy” (MI01\_00424) (A.B., F.P., S.P.) (2011–2014); Compagnia di San Paolo (1471 SD/CC N.2009.1822) (2011–2013) (2013.0927 ID ROL 4195); fellowship to M. M., Carige Foundation (2012); Start-Up grant AIRC #6108, Italian Ministry of Health grant GR-2008-1135635 (G.Z.); and FP7 project PANACREAS #256986 (A.N.).

## References

1. Hoadley KA, Yau C, Wolf DM et al (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158:929–944
2. Perou CM, Sørlie T, Eisen MB et al (2000) Molecular portraits of human breast tumors. *Nature* 406:747–752
3. Zack TI, Schumacher SE, Carter SL et al (2013) Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 45:1134–1140
4. Lacombe D, Tejpar S, Salgado R et al (2014) European perspective for effective cancer drug development. *Nat Rev Clin Oncol* 11:492–498
5. Zardavas D, Maetens M, Irrthum A et al (2014) The AURORA initiative for metastatic breast cancer. *Br J Cancer* 111:1881–1887
6. Garraway LA, Lander ES (2013) Lessons from the cancer genome. *Cell* 153:17–37
7. Amirkhah R, Schmitz U, Linnebacher M et al (2014) MicroRNA-mRNA interactions in colorectal cancer and their role in tumor progression. *Genes Chromosom Cancer* 54:129–141
8. Vogelstein B, Papadopoulos N, Velculescu VE et al (2013) Cancer genome landscapes. *Science* 339:1546–1558
9. Catalogue of Somatic Mutation in Cancer (COSMIC). <http://www.sanger.ac.uk/genetics/CPG/cosmic>
10. Futreal PA, Coin L, Marshall M et al (2004) A census of human cancer genes. *Nat Rev Cancer* 4:177–183
11. The Cancer Genome Atlas (TCGA). <http://cancergenome.nih.gov/>

12. International Cancer Genome Consortium (ICGC). The ICGC Data Portal. <https://icgc.org/>
13. The cBioPortal for Cancer Genomics. <http://www.cbioportal.org/public-portal/>
14. The Cancer Genomics Hub (CGHub). <https://cghub.ucsc.edu/>
15. The Tumor Portal. <http://cancergenome.broadinstitute.org/>
16. Griffith M, Griffith OL, Coffman AC et al (2013) DGIdb: mining the druggable genome. *Nat Methods* 10:1209–1210
17. The Drug Gene Interaction Database. <http://dgidb.genome.wustl.edu/>
18. The Genomics of Drug Sensitivity in Cancer (GDSC). <http://www.cancerrxgene.org/>
19. Brownstein CA, Beggs AH, Homer N et al (2014) An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. *Genome Biol* 15:R53
20. Stahel R, Bogaerts J, Ciardiello F, de Ruyscher D et al. (2014) Optimising translational oncology in clinical practice: strategies to accelerate progress in drug development. *Cancer Treat Rev.* pii: S0305-7372 (14) 00209-6
21. Watson IBM. <http://www.ibm.com/smarterplanet/us/en/ibmwatson/20>
22. Ferrucci D, Brown E, Chu-Carroll J et al (2010) Building Watson: an overview of the DeepQA project. *AI Mag* 31:59–79
23. Moschitti A, Chu-Carroll J, Patwardhan S et al. (2011) Using syntactic and semantic structural kernels for classifying definition questions in jeopardy!. *Proceedings of the conference on empirical methods in natural language processing.* pp 712–724
24. Kinzler KW, Vogelstein B (1997) Cancer-susceptibility genes Gatekeepers and caretakers. *Nature* 386:761–763
25. Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. *Nat Med* 10:789–799
26. Marx V (2014) Cancer genomes: discerning drivers from passengers. *Nat Methods* 11:375–379
27. Lawrence MS, Lawrence MS, Stojanov P et al (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499:214–218
28. Lawrence MS, Stojanov P, Mermel CH et al (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505:495–501
29. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J et al (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* 10:1081–1082
30. IntOGen-mutations platform. <http://www.intogen.org/mutations/>
31. Martelotto LG, Ng C, De Filippo MR et al (2014) Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol* 15:484
32. Yeang CH, McCormick F, Levine A (2008) Combinatorial patterns of somatic gene mutations in cancer. *FASEB J* 22:2605–2622
33. Nik-Zainal S, Alexandrov LB, Wedge DC et al (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* 149:979–993
34. Alexandrov LB, Nik-Zainal S, Wedge DC et al (2013) Signatures of mutational processes in human cancer. *Nature* 500:415–421
35. Weinstein IB, Joe AK (2006) Mechanisms of disease: oncogene addiction – a rationale for molecular targeting in cancer therapy. *Nat Clin Pract Oncol* 3:448–457
36. Druker BJ, Talpaz M, Resta DJ et al (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N Engl J Med* 344:1031–1037
37. Shaw AT, Kim DW, Nakagawa K et al (2013) Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N Engl J Med* 368:2385–2394
38. Chapman PB, Hauschild A, Robert C et al (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* 364:2507–2516
39. Mok TS, Wu YL, Thongprasert S et al (2009) Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* 361:947–957
40. De Roock W, Claes B, Bernasconi D et al (2010) Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol* 11:753–762
41. Grothey A, Lenz HJ (2012) Explaining the unexplainable: EGFR antibodies in colorectal cancer. *J Clin Oncol* 30:1735–1737
42. Samalin E, Bouché O, Thézenas S et al (2014) Sorafenib and irinotecan (NEXIRI) as second- or later-line treatment for patients with metastatic colorectal cancer and KRAS-mutated

- tumours: a multicentre Phase I/II trial. *Br J Cancer* 110:1148–1154
43. Shih T, Lindley C (2006) Bevacizumab: an angiogenesis inhibitor for the treatment of solid malignancies. *Clin Ther* 28:1779–1802
  44. Grothey A, Van Cutsem E, Sobrero A et al (2013) Regorafenib monotherapy for previously treated metastatic colorectal cancer (CORRECT): an international, multicentre, randomised, placebo-controlled, phase 3 trial. *Lancet* 381:303–312
  45. Sun C, Hobor S, Bertotti A et al (2014) Intrinsic resistance to MEK inhibition in KRAS mutant lung and colon cancer through transcriptional induction of ERBB3. *Cell Rep* 7:86–93
  46. Ng K, Tabernero J, Hwang J et al (2013) Phase II study of everolimus in patients with metastatic colorectal adenocarcinoma previously treated with bevacizumab-, fluoropyrimidine-, oxaliplatin-, and irinotecan-based regimens. *Clin Cancer Res* 19:3987–3995
  47. Iyer G, Hanrahan AJ, Milowsky MI et al (2012) Genome sequencing identifies a basis for everolimus sensitivity. *Science* 338:221
  48. Integrating personalised medicine into EU strategy. EAPM annual conference report Bibliothèque Solvay and the European Parliament, Brussels 9–10 September, 2014. <http://euapm.eu/wp-content/uploads/2012/07/EAPM-Annual-Conf-Report-Integrating-Personalised-Medicine-into-the-EU-Health-Strategy.pdf>
  49. Pal I, Mandal M (2012) PI3K and Akt as molecular targets for cancer therapy: current clinical outcomes. *Acta Pharmacol Sin* 33:1441–1458
  50. Zhao Y, Aguilar A, Bernard D et al (2015) (2014) Small-molecule inhibitors of the MDM2–p53 protein-protein interaction (MDM2 inhibitors) in clinical trials for cancer treatment. *J Med Chem* 8(3):1038–52
  51. Huang SM, Mishina YM, Liu S et al (2009) Tankyrase inhibition stabilizes axin and antagonizes Wnt signalling. *Nature* 461:614–620
  52. FDA Public Workshop. Innovations in breast cancer drug development – next generation oncology trials. Breast Cancer Workshop. October 21, 2014. Session 1 improving targeted drug development for “small” populations with genomic. <http://www.fda.gov/Drugs/NewsEvents/ucm410332.htm>
  53. Lillie EO, Patay B, Diamant J et al (2011) The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Per Med* 8:161–173
  54. Zauderer MG, Gucalp A, Epstein AS et al (2014) Piloting IBM Watson Oncology within Memorial Sloan Kettering’s regional network. *Journal of Clinical Oncology* 32(15 suppl):e17653, 2014 ASCO Annual Meeting Abstracts
  55. Rodin M. IBM Watson: Transforming expertise in the new era of computing. Presented at Mayo Clinic Transform 2014, Washington, DC/San Francisco, Sept 7–9, 2014. [www.mayo.edu/transform/talks/2014/ibm-watson-transforming-expertise-in-the-new-era-of-computing](http://www.mayo.edu/transform/talks/2014/ibm-watson-transforming-expertise-in-the-new-era-of-computing)
  56. ClinicalTrials.gov. <https://clinicaltrials.gov/>
  57. Crystal AS, Shaw AT, Sequist LV et al (2014) Patient-derived models of acquired resistance can identify effective drug combinations for cancer. *Science* 346:1480–1486
  58. Cancer Cell Line Encyclopedia. <http://www.broadinstitute.org/ccle/home>
  59. Zauderer MG, Gucalp A, Epstein AS, Seidman AD, Caroline A, Granovsky S, Julia F, Keesing J, Lewis S, Co H, Petri J, Megerian M, Eggebraaten T, Bach P, Kris MG, Tortolina L, Duffy DJ, Maffei M et al (2015) Advances in dynamic modeling of colorectal cancer signaling-network regions, a path toward targeted therapies. *Oncotarget* 10:5041–5058
  60. Castagnino N, Tortolina L, Balbi A et al (2010) Dynamic simulations of pathways downstream of ERBB-family, including mutations and treatments. Concordance with experimental results. *Curr Cancer Drug Targets* 10:737–757
  61. Tortolina L, Castagnino N, De Ambrosi C et al (2012) A multi-scale approach to colorectal cancer: from a biochemical-interaction signaling-network level, to multi-cellular dynamics of malignant transformation. Interplay with mutations and onco-protein inhibitor drugs. *Curr Cancer Drug Targets* 12:339–355
  62. De Ambrosi C, Barla A, Tortolina L et al (2013) Parameter space exploration within dynamic simulations of signaling networks. *Math Biosci Eng* 10:103–120
  63. Kohn KW (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol Biol Cell* 10:2703–2734
  64. Aladjem M.I., Pasa S., Parodi S. et al. (2004) Molecular interaction maps--a diagrammatic graphical language for bioregulatory networks. *Sci STKE* 2004(222):pe8.
  65. Kohn KW, Aladjem MI, Weinstein JN et al (2006) Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. *Mol Biol Cell* 17:1–13
  66. MATLAB. <http://www.mathworks.com/products/simbiology/?BB=1>
  67. Polisenio L, Salmena L, Zhang J et al (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465:1033–1038

68. Tay Y, Kats L, Salmena L et al (2011) Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* 147:344–357
69. Song MS, Salmena L, Pandolfi PP (2012) The functions and regulation of the PTEN tumour suppressor. *Nat Rev Mol Cell Biol* 13:283–296
70. Snedecor GW, Cochran WG (1967) *Statistical methods* 1967. Blackwell, Ames, IA
71. Statistical inference. [http://en.wikipedia.org/wiki/Statistical\\_inference](http://en.wikipedia.org/wiki/Statistical_inference)
72. Misale S, Arena S, Lamba S et al (2014) Blockade of EGFR and MEK intercepts heterogeneous mechanisms of acquired resistance to anti-EGFR therapies in colorectal cancer. *Sci Transl Med* 6(224):224ra26
73. Cell Miner. <http://discover.nci.nih.gov/cellminer/home.do>
74. The I-SPY 2 TRIAL – Investigation of serial studies to predict your therapeutic response with imaging and molecular analysis 2. <http://ispy2.org/>
75. The NCI Molecular Analysis for Therapy Choice (MATCH) program. <http://www.cancer.gov/clinicaltrials/noteworthy-trials/match>
76. SAFIR02\_Breast. <https://clinicaltrials.gov/ct2/show/NCT02299999?term=Safir02&rank=2>
77. Collins FS, Varmus H (2015) A new initiative on precision medicine. *N Engl J Med* 372(9):793–795
78. Blanpain C (2013) Tracing the cellular origin of cancer. *Nat Cell Biol* 15:126–134
79. Schmitz U, Wolkenhauer O (eds) (2016) *Systems medicine methods and protocols: methods in molecular biology*, vol 1386. Springer, New York
80. Russell S, Norvig P (1995) *Artificial intelligence: a modern approach*. Prentice-Hall, Englewood Cliffs, NJ
81. McCarthy J (1963) *Programming with common sense*. Defense Technical Information Center, Washington, DC
82. Shortliffe EH (1974) MYCIN: a rule based computer program for advising physicians regarding antimicrobial therapy selection. PhD dissertation in Medical Information Sciences. Stanford University
83. Rabiner L (1989) A tutorial on hidden Markov Models and selected applications in speech recognition. *Proc IEEE* 77:257–286
84. Bishop CM (2006) *Pattern recognition and machine learning*. Springer, New York
85. Jurafsky D, James H (2000) *Speech and language processing an introduction to natural language processing, computational linguistics, and speech*. Prentice-Hall, Englewood Cliffs, NJ
86. Gokhan T, De Mori R (2011) *Spoken language understanding: systems for extracting semantic information from speech*. John Wiley, New York
87. Narayanan S, Panayiotis GG (2013) Behavioral signal processing: deriving human behavioral informatics from speech and language. *Proc IEEE* 101:1203–1233
88. Mayo Clinic partners with IBM’s Watson to improve clinical trial patient selection. <http://www.healio.com/endocrinology/practice-management/news/online/%7B193f1642-342d-492f-9be3-0e447becbf02%7D/mayo-clinic-partners-with-ibms-watson-to-improve-clinical-trial-patient-selection>
89. Sledge GW Jr, Miller RS, Hauser R (2013) CancerLinQ and the future of cancer care. *Am Soc Clin Oncol Educ Book*. pp 430–434
90. Schilsky RL, Michels DL, Kearbey AH et al (2014) Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ. *J Clin Oncol* 32: 2373–2379
91. Merolla PA, Arthur JV, Alvarez-Icaza R et al (2014) Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345:668–673

# Chapter 11

## Neurological Diseases from a Systems Medicine Point of View

Marek Ostaszewski, Alexander Skupin, and Rudi Balling

### Abstract

The difficulty to understand, diagnose, and treat neurological disorders stems from the great complexity of the central nervous system on different levels of physiological granularity. The individual components, their interactions, and dynamics involved in brain development and function can be represented as molecular, cellular, or functional networks, where diseases are perturbations of networks. These networks can become a useful research tool in investigating neurological disorders if they are properly tailored to reflect corresponding mechanisms. Here, we review approaches to construct networks specific for neurological disorders describing disease-related pathology on different scales: the molecular, cellular, and brain level. We also briefly discuss cross-scale network analysis as a necessary integrator of these scales.

**Key words** Systems medicine, Multiscale brain networks, Network reconstruction, Molecular networks, Cellular networks, Connectome, Cross-scale analysis, Neurodegenerative diseases, Epilepsies

---

### 1 Introduction

The human brain is an organ of an extraordinary complexity, where high-level processes emerge from simultaneous and continuous interaction of mechanisms on molecular, cellular, and anatomical scales. We need to properly analyze this complexity to be able to address the question of how brain disorders should be diagnosed and treated. A systems approach is a proper paradigm to address the challenges of brain pathophysiology.

The dynamics and close coupling between different scales of brain physiology are already clearly seen in the development of the nervous system. Importantly, the molecular and cellular processes observed in acute and chronic diseases often reflect and reuse mechanisms of embryogenesis. Many of the canonical pathways such as sonic hedgehog, Wnt, FGF, BMP, and their underlying

---

*If you look at the anatomy, the structure, the function, there's nothing in the universe that's more beautiful, that's more complex, than the human brain.* Keith Black



transcriptional regulatory networks are highly conserved during evolution. The homologs or paralogs of certain genes are expressed at different times and in different pre- and postnatal cell lineages [1]. For instance, genes required in the formation of the embryonic vascular system are re-expressed during wound healing in adults. Of course the context of these evolutionary conserved modules within the circuitry of the adult organism differs greatly, which might lead to a different outcome after activation of their expression in an embryonic versus an adult environment [2, 3]. Nevertheless developmental biology can very well inform and support the generation of hypotheses about the disease pathogenesis, especially considering that the developmental processes integrate control on molecular, cellular, and anatomical levels, and the perturbation of this control may reflect the pathogenesis on the later stages [4, 5].

### ***1.1 The Molecular, Cellular, and Anatomical Regulation of the Development of the Nervous System***

One of the earliest events during embryogenesis is the determination of the principal body axis. Following the development of the primitive streak and the formation of the notochord, different cell layers, mesoderm, ectoderm, and endoderm are formed. The ectoderm located immediately dorsally to the notochord is induced to form neuroectoderm, the precursor of the nervous system. The neuroectoderm, initially a flat sheet, then folds into the neural tube, which in itself differentiates further into a number of neuronal different cell types dependent on their anterior-posterior, dorsal-ventral, and lateral position [6]. The development of the specific neuronal cell types within the spinal cord and the brain including the formation of the peripheral motor and sensory system can be traced back to the induction and programming during these specific early developmental phases [7].

The most anterior part of the neuroectoderm and the neural tube develop into the brain as a result of highly complex folding, proliferation, and migration events [8]. In this period the segmental nature of the brain becomes masked by region-specific migration and outgrowth of specific brain regions, for instance the cortex. Newly developing neurons contain cell-autonomous positional identity information and in addition receive spatial intracellular and extracellular cues guiding their migration and homing within the developing embryo in an anterior-posterior as well as a dorsal-ventral direction. These events are overlaid by the expression and activity of intricate cell survival, apoptotic, proliferation, and differentiation signals, leading to the final formation of the different neuronal and glial cell types and their wiring into the final connectome of the brain [9]. Already during embryogenesis electrical activity of neurons starts and is an important factor in the development of the nervous system [10].

Similarly to neural tube formation, specific molecular and cellular processes govern the segmentation into specific



components of nervous system. One of them is the formation of the midbrain-hindbrain boundary, reflecting the pronounced segmentation of the developing brain [11]. One of the key pathways driving this process is the Wnt pathway. A series of hierarchically organized transcription factors (e.g., EN1, PAX2, PAX8) and secreted morphogens (WNT1, SHH, FGF8) set up an asymmetry at a precise anterior-posterior boundary [12]. It is at this interface where an “organizing center” forms, leading to the differentiation and outgrowth of various neuronal precursors, for instance of the dopaminergic neurons. Tracing specific neuronal subtypes and assigning them specific gene expression signatures have greatly facilitated our understanding of the underlying wiring principles of the nervous system.

Brain development is a process integrating different layers of biological complexity, from genetic programs and molecular mechanisms, through cellular interactions and migration, to the development of functional anatomical regions. It is expected that perturbation of these networks may result in pathogenic states of the brain. Indeed, recent findings suggest that molecular, cellular, and anatomical dysfunction during brain development are for example resulting in epilepsies [13]. Interestingly components of the Wnt pathway are affected in diseases associated with the specific brain regions or neurons later on [14, 15]. These findings fuel intensive efforts under way to develop systems-based computational models for many of the molecular events of nervous system development.

The integration of mechanisms on different layers takes place across the whole life-span of the brain, through its homeostasis to degeneration. Similarly, it is needed to integrate spatial and temporal scales of representation of brain disease to be able to grasp the full picture of neuropathogenesis.

## **1.2 A Systems Approach is Required for Neurological Disorders**

The processes of brain development, homeostasis and function, and neurodegeneration are complex. Elaborated architecture and functionality on molecular, tissue, and anatomical levels are constantly changing due to intrinsic brain functions and interactions with the environment. Disorders of such a complex system affect different aspects of its function, ranging from molecular structure, through dysfunction of neuronal subpopulations, to alteration of anatomical or functional brain connectivity. To be able to properly address the challenge of neurological disorders, we need to understand key processes implicated in brain function. For this purpose, existing knowledge is being combined with experimental readouts to construct networks describing pathological processes on molecular, cellular, and anatomical levels in the brain. Constantly improving analytical methods are applied to dissect structure and dynamics of these networks in an attempt to understand the pathology behind.

However, this systems approach is not sufficient to fully answer challenges of neurological disorders. The networks of dynamic topology, responsible for emergence of coherent function of the brain, should be considered along with their relation to other scales of brain organization. For instance, frequency and amplitude of neuronal firing that maps onto interactions of neurons and other cells should be considered in the context of the function of anatomical location containing the neurons, as well as molecular processes responsible for the firing.

In the following sections we review the three physiological scales to consider when approaching neurological disorders: molecular, cellular, and the whole brain. We discuss recent approaches to characterize components of networks on these scales, and to construct and refine networks specific to each scale. Finally, we emphasize the need for cross-scale network analysis to gain further understanding of the complexity of neurological disorders.

---

## 2 Molecular Interaction Networks

### 2.1 *Components*

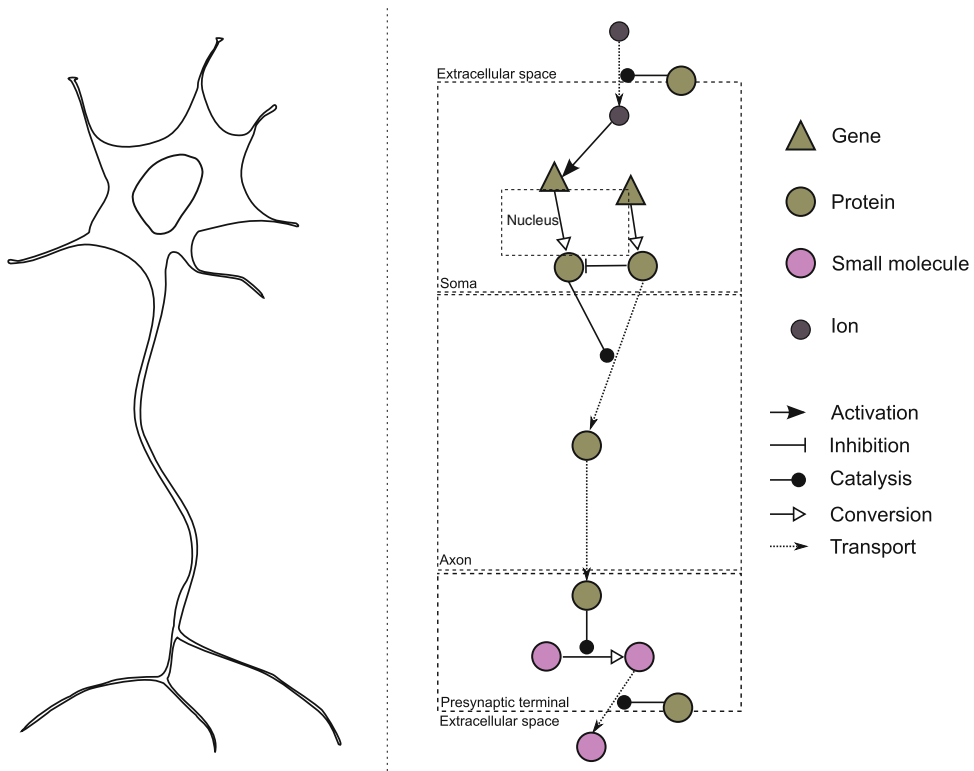
Characterization of disease-related mechanisms on the level of molecular neurobiology is both necessary and extremely challenging. Our knowledge of the physiology of neuronal and glial cells is still limited, mostly because the brain tissue is both heterogeneous and difficult to access. In effect, molecular networks usually represent only a reduced view of the molecular biology of nervous cells. Figure 1 illustrates this reduced network view next to a cell it represents.

This reduced, but not reductionist, view is the essence of disease-oriented molecular networks. The network has to model the processes implicated in the pathogenesis; thus it has to focus only on relevant components and interactions. However, taking into account the multitude of processes implicated in neurological disorders, identifying which elements of the molecular networks are relevant is not a trivial task [16].

Typical components and corresponding interaction types of molecular networks are listed in Table 1, with an indication of potential interactions between different components. It should be emphasized that the network representation describes dynamical processes and the abovementioned interaction types have various temporal and spatial resolutions. For instance, the axonal transport of substrates of synaptic activity is quite different from the calcium transport across the neuronal membrane.

### 2.2 *Disease-Specific Molecular Networks*

The focus of disease-specific molecular networks depends on the nature of the pathogenesis. This scope ranges from well-defined mechanisms through a set of implicated pathways to a number of



**Fig. 1** A network representing molecular processes in a neuronal cell, illustrating the reduced network view of complex cell physiology. Activation and inhibition interactions describe regulatory events within a neuronal cell. Catalysis interaction denotes a catalyst role of an element. Conversion refers to change of state of molecules, be it biochemical reaction, or protein complex assembly. Transport describes translocation of molecules within the neuronal cell, or across its boundaries

**Table 1**  
**Elements and interaction types typically used for constructing molecular networks. See Fig. 1 for description of interaction types**

Element types	Interaction types			
	Regulatory	Catalytic	Conversion	Transport
Gene/mRNA	•		•	
MicroRNA	•			
Protein	•	•	•	•
Small molecule	•	•	•	
Pathway	•	•	•	•

involved molecules. The focus of the molecular network is in effect closely related to questions that systems-level analyses should answer. The more focused the network model, the more precise

questions may be asked, up to the level of high-quality, computable metabolic models [17].

In the case of prion-like diseases, the causative mechanism is a misfolding prion protein, inducing neurodegeneration and spreading across the nervous system [18]. Regardless of our insight into the structural properties of prions [19], the knowledge about the pathology of this single molecular mechanism is insufficient to propose a cure.

Huntington's disease (HD) is a genetic disorder caused by excessive glutamine repeats in the gene encoding the huntingtin protein [20]. Such mutated huntingtin induces formation of pathogenic inclusions in neuronal cells and is supposed to burden their protein degradation systems. Although the genetic factor is convincingly identified, the exact mechanism of molecular neuropathology remains elusive.

Chronic neurodegenerative disorders, like Alzheimer's disease (AD) or Parkinson's disease (PD), are influenced by a combination of genetic and environmental factors [21, 22]. A number of familial genes and disease-inducing toxins indicate a range of molecular pathways affected in the course of these diseases. Nevertheless, causative factors remain unclear.

Epilepsies are neurological disorders where genetic components are known, or become a risk factor. Here, dysfunction of molecular mechanisms leads to the emergence of pathology on higher levels of organization of the central nervous system [23, 24]. The utility of molecular networks in studying this class of disorder seems to be limited, as existing approaches are reductionistic, not able to apprehend the complexity of the pathology [25].

### **2.3 Construction of Molecular Networks**

Construction of networks reflecting molecular level of neuropathology usually follows a number of well-defined steps. In general, these are (a) identification of candidate molecules, (b) connection of the molecules by querying databases or manually curating the interactions, and (c) refinement and evaluation of the network.

#### **2.3.1 Identification of Candidate Molecules**

Identification of candidate molecules to construct a molecular network is often supported by high-throughput screens in experimental disease models. In many cases associative networks are established using the underlying data, i.e., networks, where interactions do not represent any mechanistic link between connected elements. In the end, these associative networks support candidate prioritization for assembling disease-related, mechanistic models [26–28].

For prion diseases, mouse models [29], cellular models [30], and yeast genetic screens [31] supported construction of molecular networks. Similarly for HD-related pathology, yeast screens helped to prioritize candidates for network construction [32]. Whenever available, human *postmortem* tissue is used for omics

profiling allowing, for instance, to pinpoint genes involved in PD [33–35] and AD [36, 37] pathogenesis. Interestingly, candidates for molecular interaction networks in case of certain epilepsies base on reconstruction from brain tissue biopsies collected during surgical procedures [38].

### 2.3.2 Connection of the Molecules

Establishing interactions between candidate molecules can involve querying databases of molecular mechanisms [39], or manual curation [40]. Profiling of the transcriptome combined with literature-based network reconstruction has been proposed for instance for prion diseases [29, 41] or AD [42] as a method to indicate pathways affected during the disease progression. Network reconstruction based on genomic data, i.e., focused on genetic risk factors of neurological diseases, was proposed for prion [43], epilepsies [44], and PD [45].

The construction of molecular networks may require manual curation either to de novo assemble the interactions between candidate molecules or to review an automatically constructed network. Development of a large-scale, disease-focused network is a challenging task. In the fields of AD [46] and PD [47], heterogeneous molecular interaction maps were established. A more focused approach resulted in the curation of existing metabolic pathways into a brain-specific network [40]. Finally, in the field of PD, even more focused network-based models were constructed, representing in detail processes related to cellular stress of neuronal metabolism and to protein misfolding [48, 49].

### 2.3.3 Refinement and Evaluation

Networks constructed on the basis of analytically identified candidate molecules and interactions are prone to bias. Evaluation and refinement of constructed networks should be performed to ensure their proper focus. The quality of established disease-related networks may be evaluated using relevant experimental datasets mapped on the network structure. Fujita et al. proposed to visualize brain tissue transcriptomics data on their manually curated PD-relevant network [47], what allows to assess its relevance. In case of epilepsies such an evaluation helps to tailor networks for different disease subtypes, for instance focusing on specific neuronal receptors [50] or filtering using gene expression profiles from brain tissue of pharmacoresistant cases [51]. In addition to human brain samples, datasets from disease-related experimental models can be similarly applied [52, 53]. Especially experimental setups focused on detailed analysis of specific pathways are useful in such a network evaluation. For instance, recent work on the mechanisms of the Wnt signaling pathway [54] produced time series of gene expression following Wnt stimulation. Network analysis of these series confirmed known mechanisms governing canonical and noncanonical activation of the Wnt-pathway, and shed light on molecular mechanisms relevant for AD. Although the constructed

network was associative, such gene expression time series data can also be mapped on curated, mechanistic models of the disease to validate their accuracy in reflecting crucial mechanisms.

Besides using tissue-specific experimental datasets, additional sources of information can be applied to refine the shape of developed molecular networks. Recently, microRNAs gained attention as potential modulators of neurological disorders [55–57]. These regulators of mRNA are especially relevant when constructing brain-focused, gene regulatory networks. Similarly, DNA methylation, or protein acetylation, emerges as a potent regulator of a large number of genes in neurological disorders [58–60], which can affect entire functional modules of molecular networks.

#### **2.4 Summary: Molecular Networks**

Molecular networks of brain disorders are very heterogeneous, such as the data sources used for their construction. Molecular mechanisms of the brain are studied using experimental models, postmortem tissue, and, in particular cases, brain biopsies. When constructing these networks, a trade-off has to be made between network breadth and depth. Large-scale networks provide an overview of disease processes, allowing limited analytical approaches [47]. Moreover, they enable studies on molecular cross-disease comparison, aiming to elucidate overlapping mechanisms between diseases like AD or PD, and diabetes or autoimmune diseases [61–63]. In turn small, focused networks can describe disease-related processes with high quality and using established mathematical frameworks. Simulation of dynamics in such networks allows predictions on causality and temporal resolution of represented processes [48, 64].

Importantly, molecular networks should not be considered as stand-alone structures. The cellular machinery of brain cells works in the context of its embedding tissue, which in turn forms functional areas of the brain. Thus, although prion pathology has a molecular basis, the disease has to be considered also from the perspective of higher order networks. Recent findings on prion interactions with GABA receptors and their influence on excitotoxicity allow forming a link with cellular networks [65]. This link is further reinforced by the findings on the modulatory role of prion protein in the dopaminergic system [66]. It might be necessary to bridge molecular and brain layers to explain symptomatic biomarkers of prion disease [67].

Our knowledge on the molecular basis of HD is insufficient to explain its pathogenesis. This fact suggests broadening the scope of systems analysis beyond the molecular interaction networks. Studies correlating genetics of early HD with neuroimaging studies form a bridge between molecular and brain-level networks [68].

Importantly, higher levels of network representation should be considered when analyzing molecular mechanisms. In PD, degeneration of a particular neuronal populations is observed,

suggesting that cellular interactions [69, 70] play an important role in the pathology. Moreover, growing body of evidence points towards pathological spreading of synuclein aggregates across brain areas [71, 72] as the key mechanism of PD. Higher levels of network organization may provide further understanding in PD pathology. Recent studies in AD and PD follow this concept by analyzing omics of different brain areas affected in PD and AD [37, 73], or genetic factors affecting functioning of brain-level networks [74, 75].

The molecular pathogenesis of epilepsies contributes significantly to the pathology of networks of higher order [76]. Therefore, the need for systems biology is pressing, as their emergent properties span not only over many elements of molecular networks, but also over different network layers.

---

### 3 Cellular Interaction Networks

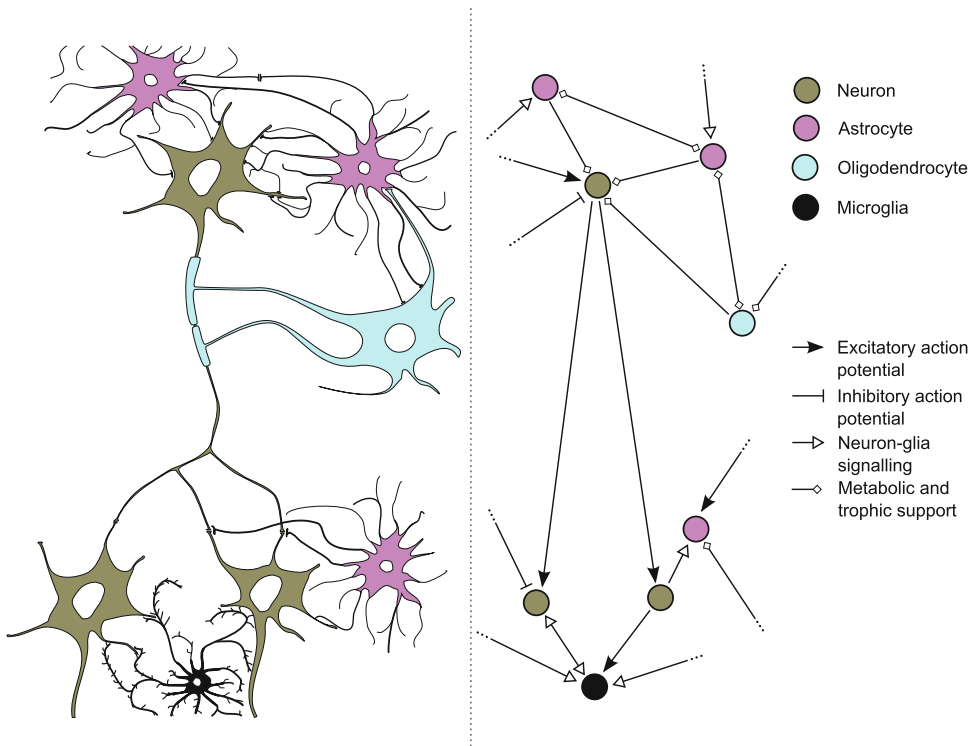
#### 3.1 Components

The human brain consists of approximately  $10^9$  neurons, each of which has on average 100,000 synaptic connections to other neuronal cells [77, 78]. This plethora of neuronal interactions reveals a well-defined network structure, established already during the developmental stage. This network has varying spatiotemporal characteristics as some cells are more locally connected whereas others project to distant regions within the brain and the body, with some connections being longer than a meter [79, 80]. Moreover, the interaction modes between neuronal cells are diverse, being either excitatory or inhibitory in dependence on neurotransmitters and corresponding receptors of their synapses.

While the main brain structure remains stable over lifetime, the brain demonstrates a huge local plasticity compared to all other organs, enabling learning and memory. This plasticity is achieved by an input-dependent rewiring of the neuronal network topology in specific brain regions like the hippocampus and the cortex. The main mechanisms for this rewiring are long-term potentiation and long-term depression that alter synaptic connection between neurons in an activity-dependent manner according to Hebb's learning rule [81, 82].

Importantly, the human brain consists of more than 50 % of glial cells that play an important role in the activity of brain cellular networks. Among these cells astrocytes are the majority, complemented by oligodendrocytes and microglia [83]. Astrocytes translate neuronal activity and the related energy demands to blood flow regulation and corresponding uptake of glucose and oxygen to facilitate neuronal metabolism [84, 85]. Oligodendrocytes insulate axons of the neurons by myelin sheets that allows for fast signal transduction and protects the fragile structure from the exterior [86]. Microglia represent the macrophages of the brain.





**Fig. 2** The diverse cell types within the brain generate a complex interaction network with different classes of interaction edges that also exhibit distinguished dynamic properties. Excitable connections (excitatory or inhibitory) denote action potentials of the neurons. Metabolic and trophic support interactions represent exchange of substrates required for cellular network homeostasis. Topology-altering interactions denote cellular mechanisms leading to changes in the local network structure

They sense pathogens and damaged cells, migrate to the specific areas, and clean them by phagocytosis [87]. Figure 2 gives a schematic overview of these interactions including a corresponding network representation.

The brain exhibits a wide and heterogeneous spectrum of cellular interactions that covers many spatial and temporal scales. The fastest intercellular signaling occurs between neurons on a millisecond time scale [88]. Thereby electric impulses of axon potentials are transmitted at synapses to connected cells. This fast communication and typical feedback loops enable fast perception, appropriate responses, and refinements of actions [89]. Importantly, the signaling within the neuronal network is also influencing the surrounding glia, which in turn can modulate the neuronal communication.

Astrocytes seal up the synaptic cleft to facilitate chemical information transmission and are responsible for clearing the neurotransmitters from the cleft and recycling them back to the pre-synaptic terminal [90]. Importantly, astrocytes express receptors for diverse neurotransmitters. For instance, the glutamate release at

glutamatergic synapses activates not only the postsynaptic neuron but also the surrounding astrocytes. Activated astrocytes increase their cytosolic  $\text{Ca}^{2+}$  on the time scale of seconds, which triggers downstream signaling processes including potential release of ATP and glutamate [91]. This release induces the regulation of the blood flow but also a local amplification process [92]. Subsequently, the signal can propagate within the astrocytic network by intercellular  $\text{Ca}^{2+}$  waves activating tens of cells and spreading hundreds of micrometer [93], where it may induce or modulate neuronal activity including synapse genesis by long-term potentiation and long-term depression [94].

Oligodendrocytes support neuronal functionality by myelination that occurs on the time scale of minutes to hours [87]. Moreover, recent findings point to their role in metabolic supporting and regulation of neuronal function [95, 96]. Similarly, microglia have a long-term influence on neuronal dynamics. Besides removing pathogens and cell debris, including damaged neurons from the brain, microglia are responsible for synapse pruning [97]. The resulting changes in the neuronal network topology occur on the time scale of hours and are essential for brain function. Interestingly, a similar role was recently reported for astrocytes [98]. Table 2 gives an overview on the different cell types and their role in brain dynamics.

Overall, the huge neuronal connectivity of neurons leads to dense network structures that translate the nonlinear dynamics of the single entities into a mesoscopically more ordered behavior. The resulting fine-tuned activity patterns often exhibit locally synchronized firing of neurons that correspond to specific representations of information such as visual memory [99] or movement controls [100]. The underlying neuronal microcircuits are embedded in and modulated by a number of regulatory cellular interactions that allow for their plasticity and adaptation by changing the network topology [89]. To integrate the different involved levels and scales, we need to rely on systems approaches.

**Table 2**  
**Elements and interaction types of cellular networks. The temporal resolution of each interaction is explicitly indicated**

Element types	Interaction types		
	Excitable	Metabolic	Topology altering
Neuron	Milliseconds	Seconds	Hours
Astrocyte	Seconds	Seconds/minutes	Hours
Oligodendrocyte		Minutes/hours	
Microglia	Minutes		Hours

### **3.2 Disease-Specific Network Topology**

The challenge to identify disease-specific network topologies and dynamics is to distinguish between primary and secondary effects. Within this context, the general question arises as to how single-cell properties reflecting individual entities are translated to the cellular network behavior that may cause the pathology.

A direct link between molecular modification and impaired network dynamics is observed in epilepsy. In this case, a single mutation of a channel protein can lead to increased excitability on a single-cell level, inducing more frequent spiking [101, 102]. Nodes with such modified properties within the cellular network can induce drastic changes in the mesoscopic dynamics. Higher excitability of single cells can cause globally synchronized activity of many neurons, inducing seizures. At the same time, the affected cellular network is often capable to compensate for synchronized firing. In effect, both time and brain area of seizure occurrence are difficult to predict [103]. Interestingly, for cases where antiepileptic medication does not exhibit seizure-suppressing effects, a possible therapy is to remove parts of the temporal lobe or to disconnect specific projections that allow for seizure spreading [104].

In case of HD, the pathogenic genetic factor is well correlated with the cellular phenotype. Resulting neurodegeneration predominantly takes place in the striatum; however the mechanistic relation between the single-cell characteristic and the pathogenesis on the cellular network level is still not understood. Recent reports suggest an increased neuronal activity that induces larger energy demands and facilitates aging in the corresponding brain areas, leading to earlier cell death [105].

Similarly, current evidence on PD points to an unbalanced energy budget of dopaminergic neurons in the substantia nigra and their corresponding intercellular interactions [106, 107]. Selective vulnerability of these neurons comes from their extra energy demand due to dopamine synthesis and homeostasis of long projections. Disturbances in the energy balance prime these neurons for an early cell death [108]. Moreover, the proportion of glia, and their resulting metabolic support, within affected regions of dopamine synthesis is lower compared to other brain areas [109]. Another factor of dopaminergic degeneration may be the intracellular spreading of misfolded  $\alpha$ -synuclein protein [110, 111]. In consequence to the tissue-level stress excessive degeneration of dopaminergic neurons in the substantia nigra depletes the pool of striatal dopamine, affecting the basal ganglia feedback loop coordinating signals from the peripheral nervous system and the sensorimotor cortex [112]. Consequently, thalamic neurons fire synchronously, inducing the stereotypic tremor. The activity of these neurons can be targeted by deep brain stimulation (DBS) that de-synchronizes the neuronal activity and suppresses the tremor [113].

The molecular basis of deregulation of cellular networks can be observed in prion diseases. The misfolding chain reaction leads to

intra- or extracellular aggregates and eventually to neuronal death [114]. The associated changes of the neuronal network structure and corresponding dynamics subsequently evoke neurological symptoms such as dementia. Compared with the described direct dynamical impairments in epilepsy or PD that are observable by highly synchronized neuronal activity, the consequences of the modified network topology in most prion diseases are less understood [115]. A promising approach for a unifying perspective on neurodegeneration is brain energy metabolism linking many phenotypic traits and symptoms across several diseases [116, 117].

### **3.3 Construction of Cellular Networks**

#### *3.3.1 Identification of Cellular Interactions*

Cellular networks are difficult to construct due to the huge structural and dynamical complexity of represented interactions, in particular the specialized neuronal morphology and the extraordinary synaptic connectivity of the neuronal network. The architecture of glial cells, although less elaborate than of neurons, also features processes and multicellular interactions. Determination of a cellular network topology from such a heterogeneous and interconnected mosaic of cells is a nontrivial task for neurohistology.

A first approach to this problem, proposed by Golgi, was the low-efficiency plasma membrane staining with silver chromate. The resulting single-neuron stains revealed the ramified morphology of neurons and the layer-like organization of the cortex [118]. However, this approach is not suited for identification of cellular networks, as only a subset of cells is labeled in a region of interest.

Currently, electron microscopy is applied to track the neuronal interconnections, down to their fine substructures [119, 120]. The resulting large data sets have to be subsequently analyzed, mainly manually, because the variety of synaptic topologies limits automated segmentation approaches. More recent developments of synapse-specific dyes [121] enable more high-throughput investigations and functional regulation studies [122, 123]. A general limitation of all these approaches is that they can identify individual synapses but are unable to allocate these to specific neuron-neuron connections. In the context of network reconstruction, this means that only the edges of the neuronal network are identified without the necessary node associations.

#### *3.3.2 Connecting the Elements*

This intrinsic difficulty of studying cellular networks in human brain tissue brought a significant focus to animal model studies. In 2007, Lichtman and coworkers established a landmark invention addressing the neuronal connectivity challenge with their transgenic Brainbow mouse model. The approach is based on the random expression of fluorescent proteins of different colors [124] producing cell-specific color mixtures that enable to discern single neurons and identify their connections. In effect, it became possible to describe whole neuronal microcircuits [125]. Although the

Brainbow method allows for identification of neuronal connections, it provides no information on network dynamics.

In the cellular networks of the brain, dynamics is of crucial importance with respect to synaptic signaling and plasticity. Moreover, interactions between glia and neuron-glia cross talk cannot be inferred from histological information as they take place in the extracellular space without direct cell-to-cell connections [126]. Currently, the dynamics of the cellular networks [127–129] are studied using in vitro approaches [130], which provide a good footing to understand disease-specific modulation of network dynamics in vivo [131]. Recent developments combining genetics and optics enable well-controllable optogenetic experimental model systems for neuronal microcircuits [132, 133]. Application of two-photon microscopy on a brain with genetically modified reporters allows imaging of brain areas and optical control of neuronal activity [134]. Imaging and control of neuronal microcircuits are especially plausible to study disorders featuring acute neuronal misfiring, like epilepsy. For other neurological diseases, with a chronic impairment of neuronal network dynamics and associated topology, cellular network reconstruction requires more input information concerning in particular the modulatory effect of neuron-glia interactions [135, 136]. However, establishing these interactions in the cellular networks will require approaches allowing simultaneous molecular and activity profiling.

### 3.3.3 Refinement and Evaluation

The original Brainbow method is restricted to optically accessible regions. As two-photon microscopy allows for penetration of tissues to the range of a few mm [137], this approach is still beyond typical mammalian brain size. To overcome these limitations, new methods have been developed, allowing to remove the lipids of the tissue by electrophoresis, leading to transparent organs [138, 139]. Clearing a brain and applying specific fluorescent antibody staining enable imaging of a whole brain on single-cell resolution without any sectioning. The resulting brain maps do not only include all neuronal connections [9] but can also provide spatial information on glia localization. When imaging such a treated brain of a patient with autism, Deisseroth and colleagues found abnormal neuronal projections that exhibit closed loops within individual cells [138]. This finding demonstrates how a modified structure may influence brain dynamics and behavior.

Despite the substantial progress, the clearing methods are still restricted to the analysis of a fixed tissue and are unable to monitor the intricate interplay of neuronal network dynamics and structural development. This challenge may be addressed in near future in zebrafish experiments [140]. The transparency of the fish and the availability of genetically encoded  $\text{Ca}^{2+}$  dyes combined with appropriate image analysis tools [141] will allow for system-wide data acquisition that has to be inferred with computational modeling

[129] for a mechanistic understanding of brain dynamics. Another strategy is to optimize noninvasive diffusion MRI and fMRI techniques (*see* Subheading 4) to single-cell resolution that could reveal microcircuit dynamics in patients and provide bottom-up understanding in neurological pathogenesis. The ambitious Human Brain Project [142] may become an integrative initiative for these approaches.

The evaluation of constructed cellular networks is possible *in vivo* with available methods. The optogenetics approach allowed Tønnesen and colleagues to achieve light-induced hyperpolarization of neurons in an animal model of epilepsy. Hyperpolarization of certain neuronal populations was found to suppress neuronal bursting, demonstrating new targets for epilepsy treatments [143]. In human brain, the technique of magnetoencephalography (MEG) allows to measure oscillatory activity of neuronal populations in given brain areas [144]. Although MEG lacks single-cell resolution, it allows to track disease-specific frequency patterns, which in turn may validate analytical outcomes of cellular network analyses.

### **3.4 Summary: Cellular Networks**

The cellular network level bridges between molecular pathogenesis and the resulting neurological phenotype of the brain. The diversity of the intercellular interactions and their rich spatiotemporal spectrum (*see* Table 2) render this level exceptionally complex to model. At the same time, cellular network analysis may indicate promising candidates for therapeutic interventions, as pathogenic cell properties may be altered by drugs targeting key molecular pathways, or corrected by tissue-level interventions like DBS.

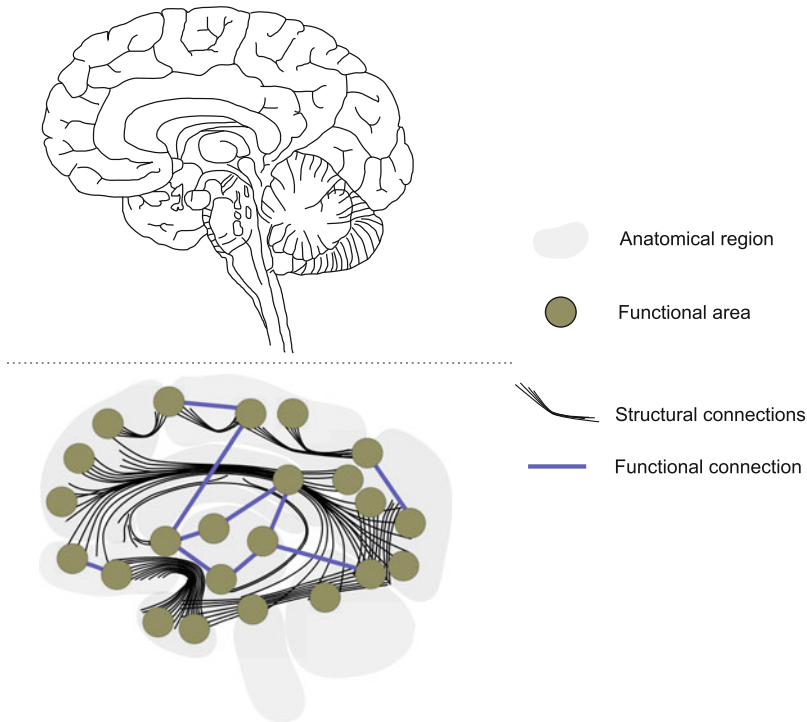
The major challenge for a mechanistic understanding of these intercellular interactions is the high connectivity of the neuronal network and dynamics covering many spatiotemporal scales. This complexity permits experimental methods to focus only on a subset of phenomena and integrative systems approaches are needed to understand underlying signaling mechanisms and support development of novel therapeutic strategies.

---

## **4 Brain-Level Networks**

### **4.1 Components**

Network representation is very appropriate to describe brain-level activity. Connectivity of different anatomical and functional brain areas suggests efficient network structure, optimized to provide high-level cognitive functions at a relatively low cost [145]. Disruption of this network is associated with pathological states of the brain. On the other hand, changes in the brain wiring may happen also due to compensatory mechanisms [146]. Similarly to network representations on other levels, we face certain simplification of extremely complex structure of the brain to a set of elements and interactions. Figure 3 illustrates this situation.



**Fig. 3** Brain-level networks represent connections between anatomical and functional brain areas, representing structural connections of directly interacting groups of neuronal cells, or functional associations between areas co-activated during a given type of brain activity

The number of components of such a network is quite limited, mostly due to the narrow scope of current neuroimaging approaches. In general, it is possible to measure functional areas of the brain by assessing oxygen consumption (fMRI-BOLD), or measure distribution of radiolabeled tracers (PET and SPECT). In turn, structural measurements are achieved assessing diffusion rates in asymmetric neuronal cells (DTI-MRI, or dMRI). Finally, our knowledge on brain topology provides us with certain mapped brain areas and their associated neurotransmitter signaling. It needs to be emphasized that the construction of brain-level networks depends heavily on proper labeling of brain areas. The task of brain parcellation is challenging, and can heavily influence the properties of obtained networks [147]. Table 3 summarizes components and interactions of brain networks.

#### **4.2 Disease-Specific Brain Networks**

The goal of a disease-oriented network approach on the brain level is primarily to synthesize experimental readouts from neuroimaging studies into a coherent picture of changes in brain function and structure caused by specific pathogenesis [148]. Because brain networks are inferred directly from neuroimaging readouts, or from established brain topology, they are usually more homogeneous



**Table 3**  
**Elements and interactions typically used for constructing brain-level networks**

Element types	Interaction types		
	Structural connectivity	Functional association	Mapped connectivity
Functional area of metabolic activity	•	•	•
Functional area of neuronal activity		•	•
Anatomical area	•		•

than molecular or cellular networks. Moreover, the neuroimaging framework is similar for all neurological diseases, and network construction efforts aim to identify brain areas associated with specific pathogeneses. Thus, in contrary to molecular networks, mechanisms specific for different diseases will not affect the focus of the network being constructed.

### 4.3 Construction of Brain-Level Networks

#### 4.3.1 Identification and Connecting of Candidate Areas

One of the most widely used methods to construct disease-relevant brain networks is functional MRI (magnetic resonance imaging), recording BOLD (blood-oxygen-level dependent) contrast signal in the brain during rest or while performing various tasks. Analysis of an fMRI signal allows to identify activated candidate areas, but is a nontrivial task, often requiring advanced data exploratory techniques [149]. Identified activated areas become elements in a brain-level network, while interactions are established on the basis of correlation of their co-activation [150, 151]. The approaches based on fMRI are numerous in the field of neurological research. In the context of this review it is important to highlight research, where a systems approach is followed to obtain a global picture of the disease. An important example is the work of Baggio et al. [152], who analyzed resting-state fMRI data from PD, mild cognitive impairment, and healthy subjects to reconstruct a global brain network associated with cognitive deficits in PD. Another interesting example of an fMRI study is the construction of a brain network for epilepsy by Toussaint and colleagues [153], aiming to highlight disruption in the functional network of the brain following epileptic discharges. Network approaches to epilepsies are reviewed in [154].

Another approach to identify and connect brain networks is neuroimaging of radiolabeled tracers. This approach highlights specific metabolic processes in the brain, involving the chosen radiotracer. The processes can be general, like glucose metabolism, or disease specific, like circulation of synaptic vesicles. The so-called metabolic networks of the brain are acquired in a similar manner to fMRI-derived networks, namely by analyzing temporal

correlations of radiotracer expression between different brain areas [155]. Such metabolic networks were recently constructed for PD [156] and AD [157].

Structural brain networks represent physical connections between brain regions by white matter fiber tracts. These connections are calculated on the basis of so-called diffusion MRI (dMRI) measures by tractography approaches. Disease-associated alterations, either pathogenic or compensatory, may be reflected in the topology of these structural connections [158]. Interestingly, these structural networks were recently shown to reflect brain response to treatment of PD [159] and epilepsy [160].

Finally, besides neuroimaging-based brain networks, prior knowledge on brain anatomy and function is used to construct networks of disease-specific dysregulation of established brain circuits. One of such circuits is the default mode network, the brain circuit active when the brain performs no particular cognitive task. This network, for instance, was found distorted in AD [153, 161, 162] and in PD [152]. One of the very-well-explored disease-related circuits is the model of basal ganglia dysfunction in PD [163, 164]. The architecture of corticobasal ganglia–cortical loops [165] is in fact a reconstructed network, with interactions being projections of different neuronal subtypes to basal ganglia and cortical regions [166]. Here, the disturbance of these mapped circuits may be assessed using the technique of recording neuronal activity in local brain areas, called electroencephalography (EEG). EEG allows to obtain a good temporal resolution when measuring brain activity during epileptic seizures [167], or permits longitudinal tracking of the disruption of disease-relevant brain circuits, as shown for PD [168].

#### 4.3.2 Evaluation and Refinement

Brain-level networks, whether neuroimaging based, or derived from prior knowledge, need to be evaluated concerning their relevance and refined. One possible approach to reach this goal is correlation with available clinical data. In their study Morales and colleagues [169] performed a cognitive assessment of PD patients along with recording of the fMRI data and constructed non-overlapping subgroups of patients with different cognitive impairments. This allowed them to improve the interpretation of neuroimaging data. Similarly to clinical assessment, drug therapy-related information can improve the quality of obtained networks. In a recent study, Cole and coworkers demonstrated that connectivity among a number of well-defined brain circuits is influenced by dopamine therapy [170]. Finally, longitudinal neuroimaging can greatly help to refine and improve the quality of brain networks. This approach was considered by Seibyl et al. [171] to help stratifying subgroups of subjects and better approach the evolution of the disease.

Certain neurological disorders carry a significant genetic burden. This information can also be used to better tailor the

constructed networks. In their work, Rao and colleagues performed fMRI on HD patients taking into account the number of glutamine repeats in the huntingtin sequence. This stratification, together with a list of HD-associated brain circuits, allowed them to identify networks specific to the forecasted severity of the disease [172]. Genetic stratification was coupled with fMRI measurements of cognitive tasks in PD [173]. In this work Nombella and coworkers demonstrated that three PD-associated alleles influence cognitive systems in PD, although no direct network construction attempt was made.

#### 4.4 Summary

Disease-specific brain-level networks are quite homogeneous concerning their composition. Their common denominators are brain anatomy and function. Their construction usually heavily depends on supplementing neuroimaging data, where the most important factors are the choice of subjects and the neuroimaging approach. While dMRI supports construction of networks with fixed topology, fMRI and metabolic imaging can produce dynamic networks. Importantly, networks obtained with the latter approaches are correlation based and represent patterns of temporal associations. Thus, their interpretation needs to be performed with care and prior information on mapped relevant brain circuits has to be taken into account.

Importantly, brain networks are the highest order representation of pathogenic processes in neurological diseases. This aggregated view allows for convincingly linking the network disturbance to clinical endophenotypes. At the same time it is difficult to assess the emergence of the disturbance from the pathology on cellular and molecular levels. Here, improvements in metabolic imaging [156] and genetic stratification of neuroimaging subjects [173] should allow to correlate molecular and brain-level networks.

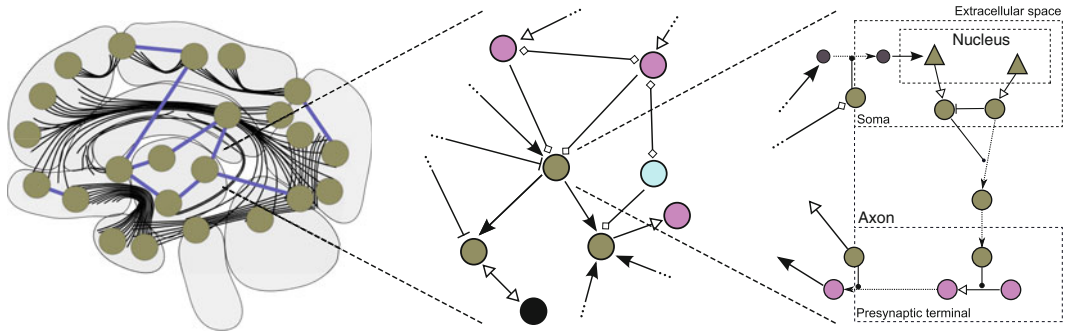
---

## 5 Synthesis

The brain is an extremely complex structure and this complexity can be observed on the level of the whole organ, cellular, and molecular organization. This nested network architecture, illustrated in Fig. 4, increases the difficulty in studying pathogenesis of neurological disorders. Nevertheless, systems approaches applied on each of these levels independently start to bring better understanding of the nature of these disorders.

### 5.1 Cross-Scale Network Analysis

Regardless of the level of brain organization, construction of networks for the purpose of systems analysis involves a similar trade-off between the scope and the depth. Broad-scope networks, constructed on the basis of omics screens (molecular) [41], micro-electrode arrays (cellular) [130], or MRI data (brain level)



**Fig. 4** Networks representing brain disorder may be deeply nested, with each of the levels contributing to the phenotype

[174, 175], have usually a broad scope, but limited depth. In effect, network- and systems-level analysis provides general large-scale insights into disruption of disease-associated brain function [33], microcircuits [138], and pathways [176]. On the other side, carefully constructed, focused networks offer detailed analyses and conclusions concerning the dynamics of the analyzed system. These focused networks require manual curation on the basis of known molecular interactions (molecular) [49], known or monitored activity of cell populations (cellular) [177], or well-mapped brain circuits (whole brain) [178]. Currently, the size-scope trade-off is inevitable. However, as the efforts towards high-quality network curation gain community-scale attention [47, 179], and new high-content screening approaches are proposed for network construction [180], we may expect that disease-specific networks will grow in size without sacrificing their quality.

Concerning the nested network architecture, it is important not only to analyze in detail the behavior of a system on a given level of complexity—molecular, cellular, or whole brain—but also to cross the scales with the systems analysis. Experimental approaches allowing to achieve this cross-scale analysis are topic of intensive research. For instance, novel imaging techniques [181, 182] allow us to gain deep insights into the molecular basis of cellular dysfunction, as well as bridge between cellular and brain scales [139].

What remains a challenge is a proper analytical approach to draw meaningful conclusions on the level of nested networks that will ultimately lead to better understanding of the disease. Currently, a number of computational approaches have been proposed that bridge the cellular and brain-level networks, focusing on modeling neuronal activity from specific brain areas to gain insight into whole-brain network dynamics. The granularity of these approaches varies from simulation of spiking behavior of single neuron [183] or neuronal population [184] models based on Hodgkin–Huxley equations to analysis of neurotransmitter release by specific brain circuits using reinforced learning models [185].

A framework bringing together molecular and higher scales still remains to be proposed; however efforts in this direction can be seen in brain network reconstructions concerning molecular profiles of subjects [172, 173].

---

## 6 Perspectives

Our current understanding of developmental processes, including brain development, encompasses the emergence of organ-level structure and function from elaborate cellular and molecular interactions. We appreciate the importance of temporal and spatial dynamics of these processes, and associate their perturbations with pathogenic states. Similarly, when approaching diseases of an adult brain and analyzing associated pathological processes, we should consider molecular, cellular, and organ-level dysfunction simultaneously. Emerging evidence on close coupling between developmental processes and the condition of specific neuronal subpopulations affected by neurodegeneration [186, 187] reinforces this perspective.

Systems biomedicine in neurology is expected to gain insight into the complex nature of human brain and its disorders, facilitating accurate diagnosis, suggesting efficient treatment, and, finally, allowing for preventing the pathogenesis. An important step to achieve these goals is to consider the brain as architecture of nested networks: molecular, cellular, and brain level. These networks become substrates of various mathematical and computational approaches, with an assumption that a given disorder is a dysfunction on a network level. Assembly of such a multi-layer network will require integration of data, but also of expertise. Community-driven approaches, like the Allen brain atlas [188] or the Parkinson's disease map [47], extending well past the molecular layer, are needed to address this challenge.

Diseases with prominent molecular components, like prion diseases or HD, will primarily benefit from detailed analysis of molecular networks. For these disorders to be pharmacologically treated, it is necessary to accurately identify mechanisms to target. Similarly, in the field of chronic neurodegenerative diseases like PD or AD only symptomatic treatment is available. Consistent failure of drug design [189] and of gene-therapeutic approaches [190] reveals a pressing need for an insightful methodology to identify causal factors of these diseases, which should be likely sought on the molecular level. Importantly, concurrent or integrated analysis of cellular or brain level networks may allow to interpret how both molecular pathogenesis and drug treatment influence higher order brain networks [191].

Disorders like epilepsies will primarily benefit from insights from cellular and brain network analysis. Here, the pathogenic

state emerges in much shorter time frame than in chronic diseases. Distorted patterns of neuronal firing need to be stratified for different epilepsy subtypes and analyzed for their response to treatment [192]. Recent advances in multi-scale [193, 194] and multi-modal neuroimaging [195] come forward to meet the needs of an integrative network analysis approach. The assessment of treatment outcomes by cellular and brain network analysis is especially important concerning increasing application of DBS in the field of PD [196], but also epilepsies [197]. Currently, DBS electrodes deliver pacemaking stimulation in single site, and in an open loop. However, the possibility to read out the firing frequency of the neurons at the stimulation site allows designing feedback systems, or considering multi-site stimulation [198]. In both cases, systems analysis of cellular and brain networks is indispensable to properly design the therapeutic approaches.

Finally, advanced therapeutic and preventive approaches can benefit from network analysis integrating molecular, cellular, and brain layers of complexity. One of them is regenerative medicine using stem cells [199, 200]. The main challenge in stem cell grafting is the question where to place them. This, in turn, requires insights integrating information on molecular function of the cell, its role within the targeted tissue, and the impact of the grafted area on the whole-brain network structure and function.

## References

1. Parker MH, Seale P, Rudnicki MA (2003) Looking back to the embryo: defining transcriptional networks in adult myogenesis. *Nat Rev Genet* 4:497–507. doi:10.1038/nrg1109
2. Drake CJ (2003) Embryonic and adult vasculogenesis. *Birth Defects Res C Embryo Today* 69:73–82. doi:10.1002/bdrc.10003
3. Urbán N, Guillemot F (2014) Neurogenesis in the embryonic and adult brain: same regulators, different roles. *Front Cell Neurosci*. doi:10.3389/fncel.2014.00396
4. Sharan R, Ideker T (2006) Modeling cellular machinery through biological network comparison. *Nat Biotechnol* 24:427–433. doi:10.1038/nbt1196
5. Edelman LB, Chandrasekaran S, Price ND (2010) Systems biology of embryogenesis. *Reprod Fertil Dev* 22:98–105. doi:10.1071/RD09215
6. Stiles J, Jernigan TL (2010) The basics of brain development. *Neuropsychol Rev* 20:327–348. doi:10.1007/s11065-010-9148-4
7. Kintner C (2002) Neurogenesis in embryos and in adult neural stem cells. *J Neurosci* 22:639–643
8. Gilbert SF (2000) *Developmental biology*, 6th edn. Sinauer Associates, Sunderland, MA
9. Sporns O, Tononi G, Kötter R (2005) The human connectome: a structural description of the human brain. *PLoS Comput Biol* 1:e42. doi:10.1371/journal.pcbi.0010042
10. Spitzer NC (2006) Electrical activity in early neuronal development. *Nature* 444:707–712. doi:10.1038/nature05300
11. Rhinn M, Brand M (2001) The midbrain-hindbrain boundary organizer. *Curr Opin Neurobiol* 11:34–42
12. Joyner AL (1996) Engrailed, Wnt and Pax genes regulate midbrain-hindbrain development. *Trends Genet* 12:15–20
13. Bozzi Y, Casarosa S, Caleo M (2012) Epilepsy as a neurodevelopmental disorder. *Front Psychiatry* 3:19. doi:10.3389/fpsy.2012.00019
14. Nissim-Eliraz E, Zisman S, Schatz O, Ben-Arie N (2013) *Nato3* integrates with the *Shh-Foxa2* transcriptional network regulating the differentiation of midbrain dopaminergic neurons. *J Mol Neurosci* 51:13–27. doi:10.1007/s12031-012-9939-6



15. Rogers D, Schor NF (2010) The child is father to the man: developmental roles for proteins of importance for neurodegenerative disease. *Ann Neurol* 67:151–158. doi:[10.1002/ana.21841](https://doi.org/10.1002/ana.21841)
16. Kolodkin A, Simeonidis E, Balling R, Westerhoff HV (2012) Understanding complexity in neurodegenerative diseases: in silico reconstruction of emergence. *Front Physiol* 3:291. doi:[10.3389/fphys.2012.00291](https://doi.org/10.3389/fphys.2012.00291)
17. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 5:93–121
18. Head MW (2013) Human prion diseases: molecular, cellular and population biology. *Neuropathology* 33:221–236. doi:[10.1111/neup.12016](https://doi.org/10.1111/neup.12016)
19. Tycko R, Wickner RB (2013) Molecular structures of amyloid and prion fibrils: consensus versus controversy. *Acc Chem Res* 46:1487–1496. doi:[10.1021/ar300282r](https://doi.org/10.1021/ar300282r)
20. Ortega Z, Lucas JJ (2014) Ubiquitin-proteasome system involvement in Huntington's disease. *Front Mol Neurosci* 7:77. doi:[10.3389/fnmol.2014.00077](https://doi.org/10.3389/fnmol.2014.00077)
21. Nacmias B, Piaceri I, Bagnoli S et al (2014) Genetics of Alzheimer's disease and frontotemporal dementia. *Curr Mol Med* 14:993–1000
22. Petrucci S, Consoli F, Valente EM (2014) Parkinson disease genetics: a “Continuum” from mendelian to multifactorial inheritance. *Curr Mol Med* 14:1079–1088
23. Lignani G, Raimondi A, Ferrea E et al (2013) Epileptogenic Q555X SYN1 mutant triggers imbalances in release dynamics and short-term plasticity. *Hum Mol Genet* 22:2186–2199. doi:[10.1093/hmg/ddt071](https://doi.org/10.1093/hmg/ddt071)
24. Oliva M, Berkovic SF, Petrou S (2012) Sodium channels and the neurobiology of epilepsy. *Epilepsia* 53:1849–1859. doi:[10.1111/j.1528-1167.2012.03631.x](https://doi.org/10.1111/j.1528-1167.2012.03631.x)
25. Margineanu DG (2013) Systems biology, complexity, and the impact on antiepileptic drug discovery. *Epilepsy Behav*. doi:[10.1016/j.yebeh.2013.08.029](https://doi.org/10.1016/j.yebeh.2013.08.029)
26. Cui S, Sun H, Gu X et al (2014) Gene expression profiling analysis of locus coeruleus in idiopathic Parkinson's disease by bioinformatics. *Neurol Sci*. doi:[10.1007/s10072-014-1889-z](https://doi.org/10.1007/s10072-014-1889-z)
27. Winden KD, Karsten SL, Bragin A et al (2011) A systems level, functional genomics analysis of chronic epilepsy. *PLoS One* 6:e20763. doi:[10.1371/journal.pone.0020763](https://doi.org/10.1371/journal.pone.0020763)
28. Tauber E, Miller-Fleming L, Mason RP et al (2011) Functional gene expression profiling in yeast implicates translational dysfunction in mutant huntingtin toxicity. *J Biol Chem* 286:410–419. doi:[10.1074/jbc.M110.101527](https://doi.org/10.1074/jbc.M110.101527)
29. Hwang D, Lee IY, Yoo H et al (2009) A systems approach to prion disease. *Mol Syst Biol* 5:252. doi:[10.1038/msb.2009.10](https://doi.org/10.1038/msb.2009.10)
30. Marbiah MM, Harvey A, West BT et al (2014) Identification of a gene regulatory network associated with prion replication. *EMBO J* 33:1527–1547. doi:[10.15252/embj.201387150](https://doi.org/10.15252/embj.201387150)
31. Manogaran AL, Hong JY, Hufana J et al (2011) Prion formation and polyglutamine aggregation are controlled by two classes of genes. *PLoS Genet* 7:e1001386. doi:[10.1371/journal.pgen.1001386](https://doi.org/10.1371/journal.pgen.1001386)
32. Tourette C, Li B, Bell R et al (2014) A large scale Huntingtin protein interaction network implicates Rho GTPase signaling pathways in Huntington disease. *J Biol Chem* 289:6709–6726. doi:[10.1074/jbc.M113.523696](https://doi.org/10.1074/jbc.M113.523696)
33. Rakshit H, Rathi N, Roy D (2014) Construction and analysis of the protein-protein interaction networks based on gene expression profiles of Parkinson's disease. *PLoS One* 9:e103047. doi:[10.1371/journal.pone.0103047](https://doi.org/10.1371/journal.pone.0103047)
34. Dusonchet J, Li H, Guillily M et al (2014) A Parkinson's disease gene regulatory network identifies the signaling protein RGS2 as a modulator of LRRK2 activity and neuronal toxicity. *Hum Mol Genet* 23:4887–4905. doi:[10.1093/hmg/ddu202](https://doi.org/10.1093/hmg/ddu202)
35. Chandrasekaran S, Bonchev D (2013) A network view on Parkinson's disease. *Comput Struct Biotechnol J* 7:e201304004. doi:[10.5936/csbi.201304004](https://doi.org/10.5936/csbi.201304004)
36. Armananzas R, Larranaga P, Bielza C (2012) Ensemble transcript interaction networks: a case study on Alzheimer's disease. *Comput Methods Programs Biomed* 108:442–450. doi:[10.1016/j.cmpb.2011.11.011](https://doi.org/10.1016/j.cmpb.2011.11.011)
37. Liu Z-P, Wang Y, Zhang X-S et al (2011) Detecting and analyzing differentially activated pathways in brain regions of Alzheimer's disease patients. *Mol Biosyst* 7:1441–1452. doi:[10.1039/c0mb00325e](https://doi.org/10.1039/c0mb00325e)
38. Bando SY, Alegro MC, Amaro EJ et al (2011) Hippocampal CA3 transcriptome signature correlates with initial precipitating injury in refractory mesial temporal lobe epilepsy. *PLoS One* 6:e26268. doi:[10.1371/journal.pone.0026268](https://doi.org/10.1371/journal.pone.0026268)
39. Kanehisa M, Limviphuvadh V, Tanabe M (2010) Chapter 9 knowledge-based analysis of protein interaction networks in



- neurodegenerative diseases. In: Alzate O (ed) Neuroproteomics. CRC Press, Boca Raton, pp 1–17
40. Sertbas M, Ulgen K, Cakir T (2014) Systematic analysis of transcription-level effects of neurodegenerative diseases on human brain metabolism by a newly reconstructed brain-specific metabolic network. *FEBS Open Bio* 4:542–553. doi:10.1016/j.fob.2014.05.006
  41. Crespo I, Roomp K, Jurkowski W et al (2012) Gene regulatory network analysis supports inflammation as a key neurodegeneration process in prion disease. *BMC Syst Biol* 6:132. doi:10.1186/1752-0509-6-132
  42. Mayburd A, Baranova A (2013) Knowledge-based compact disease models identify new molecular players contributing to early-stage Alzheimer's disease. *BMC Syst Biol* 7:121. doi:10.1186/1752-0509-7-121
  43. Lee SM, Chung M, Hwang KJ et al (2014) Biological network inferences for a protection mechanism against familial Creutzfeldt-Jakob disease with E200K pathogenic mutation. *BMC Med Genomics* 7:52. doi:10.1186/1755-8794-7-52
  44. Bakir-Gungor B, Baykan B, Ugur Iseri S et al (2013) Identifying SNP targeted pathways in partial epilepsies with genome-wide association study data. *Epilepsy Res* 105:92–102. doi:10.1016/j.epilepsyres.2013.02.008
  45. Edwards YJK, Beecham GW, Scott WK et al (2011) Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach. *PLoS One* 6:e16917. doi:10.1371/journal.pone.0016917
  46. Ogishima S, Mizuno S, Kikuchi M et al (2013) A map of Alzheimer's disease-signaling pathways: a hope for drug target discovery. *Clin Pharmacol Ther* 93:399–401. doi:10.1038/clpt.2013.37
  47. Fujita KA, Ostaszewski M, Matsuoka Y et al (2014) Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol Neurobiol* 49:88–102. doi:10.1007/s12035-013-8489-4
  48. Ouzounoglou E, Kalamatianos D, Emmanouilidou E et al (2014) In silico modeling of the effects of alpha-synuclein oligomerization on dopaminergic neuronal homeostasis. *BMC Syst Biol* 8:54. doi:10.1186/1752-0509-8-54
  49. Buchel F, Saliger S, Drager A et al (2013) Parkinson's disease: dopaminergic nerve cell model is consistent with experimental finding of increased extracellular transport of alpha-synuclein. *BMC Neurosci* 14:136. doi:10.1186/1471-2202-14-136
  50. Palomero-Gallagher N, Schleicher A, Bidmon H-J et al (2012) Multireceptor analysis in human neocortex reveals complex alterations of receptor ligand binding in focal epilepsies. *Epilepsia* 53:1987–1997. doi:10.1111/j.1528-1167.2012.03634.x
  51. Mirza N, Vasieva O, Marson AG, Pirmohamed M (2011) Exploring the genomic basis of pharmacoresistance in epilepsy: an integrative analysis of large-scale gene expression profiling studies on brain tissue from epilepsy surgery. *Hum Mol Genet* 20:4381–4394. doi:10.1093/hmg/ddr365
  52. Qi Z, Miller GW, Voit EO (2014) Rotenone and paraquat perturb dopamine metabolism: a computational analysis of pesticide toxicity. *Toxicology* 315:92–101. doi:10.1016/j.tox.2013.11.003
  53. Rhodes SL, Buchanan DD, Ahmed I et al (2014) Pooled analysis of iron-related genes in Parkinson's disease: association with transferrin. *Neurobiol Dis* 62:172–178. doi:10.1016/j.nbd.2013.09.019
  54. Wexler EM, Rosen E, Lu D et al (2011) Genome-wide analysis of a Wnt1-regulated transcriptional network implicates neurodegenerative pathways. *Sci Signal* 4:ra65. doi:10.1126/scisignal.2002282
  55. Chatterjee P, Bhattacharyya M, Bandyopadhyay S, Roy D (2014) Studying the system-level involvement of microRNAs in Parkinson's disease. *PLoS One* 9:e93751. doi:10.1371/journal.pone.0093751
  56. Satoh J (2012) Molecular network of microRNA targets in Alzheimer's disease brains. *Exp Neurol* 235:436–446. doi:10.1016/j.expneurol.2011.09.003
  57. Jimenez-Mateos EM, Henshall DC (2013) Epilepsy and microRNA. *Neuroscience* 238:218–229. doi:10.1016/j.neuroscience.2013.02.027
  58. Stilling RM, Ronicke R, Benito E et al (2014) K-Lysine acetyltransferase 2a regulates a hippocampal gene expression network linked to memory formation. *EMBO J* 33:1912–1927. doi:10.15252/emboj.201487870
  59. Li Y, Chen JA, Sears RL et al (2014) An epigenetic signature in peripheral blood associated with the haplotype on 17q21.31, a risk factor for neurodegenerative tauopathy. *PLoS Genet* 10:e1004211. doi:10.1371/journal.pgen.1004211
  60. Li G, Jiang H, Chang M et al (2011) HDAC6 alpha-tubulin deacetylase: a potential therapeutic target in neurodegenerative diseases. *J Neurol Sci* 304:1–8. doi:10.1016/j.jns.2011.02.017

61. Santiago JA, Potashkin JA (2014) System-based approaches to decode the molecular links in Parkinson's disease and diabetes. *Neurobiol Dis.* doi:[10.1016/j.nbd.2014.03.019](https://doi.org/10.1016/j.nbd.2014.03.019)
62. Menon R, Farina C (2011) Shared molecular and functional frameworks among five complex human disorders: a comparative study on interactomes linked to susceptibility genes. *PLoS One* 6:e18660. doi:[10.1371/journal.pone.0018660](https://doi.org/10.1371/journal.pone.0018660)
63. Tu Z, Keller MP, Zhang C et al (2012) Integrative analysis of a cross-loci regulation network identifies App as a gene regulating insulin secretion from pancreatic islets. *PLoS Genet* 8:e1003107. doi:[10.1371/journal.pgen.1003107](https://doi.org/10.1371/journal.pgen.1003107)
64. Schluesener JK, Zhu X, Schluesener HJ et al (2014) Key network approach reveals new insight into Alzheimer's disease. *IET Syst Biol* 8:169–175. doi:[10.1049/iet-syb.2013.0047](https://doi.org/10.1049/iet-syb.2013.0047)
65. Llorens F, Del Rio JA (2012) Unraveling the neuroprotective mechanisms of PrP (C) in excitotoxicity. *Prion* 6:245–251. doi:[10.4161/pri.19639](https://doi.org/10.4161/pri.19639)
66. Rial D, Pamplona FA, Moreira ELG et al (2014) Cellular prion protein is present in dopaminergic neurons and modulates the dopaminergic system. *Eur J Neurosci* 40:2479–2486. doi:[10.1111/ejn.12600](https://doi.org/10.1111/ejn.12600)
67. Cramm M, Schmitz M, Karch A et al (2014) Characteristic CSF prion seeding efficiency in humans with prion diseases. *Mol Neurobiol.* doi:[10.1007/s12035-014-8709-6](https://doi.org/10.1007/s12035-014-8709-6)
68. Koenig KA, Lowe MJ, Harrington DL et al (2014) Functional connectivity of primary motor cortex is dependent on genetic burden in prodromal Huntington disease. *Brain Connect* 4:535–546. doi:[10.1089/brain.2014.0271](https://doi.org/10.1089/brain.2014.0271)
69. Szabadi E (2013) Functional neuroanatomy of the central noradrenergic system. *J Psychopharmacol* 27:659–693. doi:[10.1177/0269881113490326](https://doi.org/10.1177/0269881113490326)
70. Elstner M, Morris CM, Heim K et al (2011) Expression analysis of dopaminergic neurons in Parkinson's disease and aging links transcriptional dysregulation of energy metabolism to cell death. *Acta Neuropathol* 122:75–86. doi:[10.1007/s00401-011-0828-9](https://doi.org/10.1007/s00401-011-0828-9)
71. Bae E-J, Yang N-Y, Song M et al (2014) Glucocerebrosidase depletion enhances cell-to-cell transmission of alpha-synuclein. *Nat Commun* 5:4755. doi:[10.1038/ncomms5755](https://doi.org/10.1038/ncomms5755)
72. Ubeda-Banon I, Saiz-Sanchez D, de la Rosa-Prieto C, Martinez-Marcos A (2014) alpha-Synuclein in the olfactory system in Parkinson's disease: role of neural connections on spreading pathology. *Brain Struct Funct* 219:1513–1526. doi:[10.1007/s00429-013-0651-2](https://doi.org/10.1007/s00429-013-0651-2)
73. Riley BE, Gardai SJ, Emig-Agius D et al (2014) Systems-based analyses of brain regions functionally impacted in Parkinson's disease reveals underlying causal mechanisms. *PLoS One* 9:e102909. doi:[10.1371/journal.pone.0102909](https://doi.org/10.1371/journal.pone.0102909)
74. Thaler A, Mirelman A, Helmich RC et al (2013) Neural correlates of executive functions in healthy G2019S LRRK2 mutation carriers. *Cortex* 49:2501–2511. doi:[10.1016/j.cortex.2012.12.017](https://doi.org/10.1016/j.cortex.2012.12.017)
75. Yan J, Du L, Kim S et al (2014) Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics* 30:i564–i571. doi:[10.1093/bioinformatics/btu465](https://doi.org/10.1093/bioinformatics/btu465)
76. Campbell IM, Rao M, Arredondo SD et al (2013) Fusion of large-scale genomic knowledge and frequency data computationally prioritizes variants in epilepsy. *PLoS Genet* 9:e1003797. doi:[10.1371/journal.pgen.1003797](https://doi.org/10.1371/journal.pgen.1003797)
77. Cherniak C (1990) The bounded brain: toward quantitative neuroanatomy. *J Cogn Neurosci* 2:58–68. doi:[10.1162/jocn.1990.2.1.58](https://doi.org/10.1162/jocn.1990.2.1.58)
78. Azevedo FAC, Carvalho LRB, Grinberg LT et al (2009) Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J Comp Neurol* 513:532–541. doi:[10.1002/cne.21974](https://doi.org/10.1002/cne.21974)
79. Swanson LW (2000) What is the brain? *Trends Neurosci* 23:519–527
80. Swanson LW (2011) *Brain architecture: understanding the basic plan*, 2nd edn. Oxford University Press, New York
81. Stanton PK, Sejnowski TJ (1989) Associative long-term depression in the hippocampus induced by hebbian covariance. *Nature* 339:215–218. doi:[10.1038/339215a0](https://doi.org/10.1038/339215a0)
82. Bliss TV, Collingridge GL (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* 361:31–39. doi:[10.1038/361031a0](https://doi.org/10.1038/361031a0)
83. Allen NJ, Barres BA (2009) Neuroscience: Glia - more than just brain glue. *Nature* 457:675–677. doi:[10.1038/457675a](https://doi.org/10.1038/457675a)
84. Takano T, Tian G-F, Peng W et al (2006) Astrocyte-mediated control of cerebral blood flow. *Nat Neurosci* 9:260–267. doi:[10.1038/nn1623](https://doi.org/10.1038/nn1623)

85. Attwell D, Buchan AM, Charpak S et al (2010) Glial and neuronal control of brain blood flow. *Nature* 468:232–243. doi:[10.1038/nature09613](https://doi.org/10.1038/nature09613)
86. Baumann N, Pham-Dinh D (2001) Biology of oligodendrocyte and myelin in the mammalian central nervous system. *Physiol Rev* 81:871–927
87. Kreutzberg GW (1996) Microglia: a sensor for pathological events in the CNS. *Trends Neurosci* 19:312–318
88. Boyden ES, Zhang F, Bamberg E et al (2005) Millisecond-timescale, genetically targeted optical control of neural activity. *Nat Neurosci* 8:1263–1268. doi:[10.1038/nn1525](https://doi.org/10.1038/nn1525)
89. Sporns O, Tononi G, Edelman GM (2000) Connectivity and complexity: the relationship between neuroanatomy and brain dynamics. *Neural Netw* 13:909–922
90. Allen NJ (2014) Synaptic plasticity: astrocytes wrap it up. *Curr Biol* 24:R697–R699. doi:[10.1016/j.cub.2014.06.030](https://doi.org/10.1016/j.cub.2014.06.030)
91. Zhang J, Wang H, Ye C et al (2003) ATP released by astrocytes mediates glutamatergic activity-dependent heterosynaptic suppression. *Neuron* 40:971–982
92. Halassa MM, Fellin T, Haydon PG (2007) The tripartite synapse: roles for gliotransmission in health and disease. *Trends Mol Med* 13:54–63. doi:[10.1016/j.molmed.2006.12.005](https://doi.org/10.1016/j.molmed.2006.12.005)
93. Höfer T, Venance L, Giaume C (2002) Control and plasticity of intercellular calcium waves in astrocytes: a modeling approach. *J Neurosci* 22:4850–4859
94. Barker AJ, Ullian EM (2010) Astrocytes and synaptic plasticity. *Neuroscientist* 16:40–50. doi:[10.1177/1073858409339215](https://doi.org/10.1177/1073858409339215)
95. Morrison BM, Lee Y, Rothstein JD (2013) Oligodendroglia: metabolic supporters of axons. *Trends Cell Biol* 23:644–651. doi:[10.1016/j.tcb.2013.07.007](https://doi.org/10.1016/j.tcb.2013.07.007)
96. Fünfschilling U, Supplie LM, Mahad D et al (2012) Glycolytic oligodendrocytes maintain myelin and long-term axonal integrity. *Nature* 485:517–521. doi:[10.1038/nature11007](https://doi.org/10.1038/nature11007)
97. Paolicelli RC, Bolasco G, Pagani F et al (2011) Synaptic pruning by microglia is necessary for normal brain development. *Science* 333:1456–1458. doi:[10.1126/science.1202529](https://doi.org/10.1126/science.1202529)
98. Chung W-S, Clarke LE, Wang GX et al (2013) Astrocytes mediate synapse elimination through MEGF10 and MERTK pathways. *Nature* 504:394–400. doi:[10.1038/nature12776](https://doi.org/10.1038/nature12776)
99. Kreiter AK, Singer W (1996) Stimulus-dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. *J Neurosci* 16:2381–2396
100. Cassidy M, Mazzone P, Oliviero A et al (2002) Movement-related changes in synchronization in the human basal ganglia. *Brain* 125:1235–1246
101. Lerche H, Shah M, Beck H et al (2013) Ion channels in genetic and acquired forms of epilepsy. *J Physiol* 591:753–764. doi:[10.1113/jphysiol.2012.240606](https://doi.org/10.1113/jphysiol.2012.240606)
102. Ragsdale DS (2008) How do mutant Nav1.1 sodium channels cause epilepsy? *Brain Res Rev* 58:149–159. doi:[10.1016/j.brainresrev.2008.01.003](https://doi.org/10.1016/j.brainresrev.2008.01.003)
103. Lopes da Silva FH, Blanes W, Kalitzin SN et al (2003) Dynamical diseases of brain systems: different routes to epileptic seizures. *IEEE Trans Biomed Eng* 50:540–548. doi:[10.1109/TBME.2003.810703](https://doi.org/10.1109/TBME.2003.810703)
104. Jaha LE, Najm I, Bingaman W et al (2007) Surgical outcome and prognostic factors of frontal lobe epilepsy surgery. *Brain* 130:574–584. doi:[10.1093/brain/awl364](https://doi.org/10.1093/brain/awl364)
105. Hong SL, Cossyleon D, Hussain WA et al (2012) Dysfunctional behavioral modulation of corticostriatal communication in the R6/2 mouse model of Huntington's disease. *PLoS One* 7:e47026. doi:[10.1371/journal.pone.0047026](https://doi.org/10.1371/journal.pone.0047026)
106. Wellstead P, Cloutier M (2011) An energy systems approach to Parkinson's disease. *Wiley Interdiscip Rev Syst Biol Med* 3:1–6. doi:[10.1002/wsbm.107](https://doi.org/10.1002/wsbm.107)
107. Pissadaki EK, Bolam JP (2013) The energy cost of action potential propagation in dopamine neurons: clues to susceptibility in Parkinson's disease. *Front Comput Neurosci* 7:13. doi:[10.3389/fncom.2013.00013](https://doi.org/10.3389/fncom.2013.00013)
108. Uhl GR (1998) Hypothesis: the role of dopaminergic transporters in selective vulnerability of cells in Parkinson's disease. *Ann Neurol* 43:555–560. doi:[10.1002/ana.410430503](https://doi.org/10.1002/ana.410430503)
109. Chauhan NB, Siegel GJ, Lee JM (2001) Depletion of glial cell line-derived neurotrophic factor in substantia nigra neurons of Parkinson's disease brain. *J Chem Neuroanat* 21:277–288
110. Goedert M (2001) Alpha-synuclein and neurodegenerative diseases. *Nat Rev Neurosci* 2:492–501. doi:[10.1038/35081564](https://doi.org/10.1038/35081564)
111. Goedert M, Spillantini MG, Del Tredici K, Braak H (2013) 100 years of Lewy pathology. *Nat Rev Neurol* 9:13–24. doi:[10.1038/nrneurol.2012.242](https://doi.org/10.1038/nrneurol.2012.242)
112. Brown P (2003) Oscillatory nature of human basal ganglia activity: relationship to the pathophysiology of Parkinson's disease.

- Mov Disord 18:357–363. doi:[10.1002/mds.10358](https://doi.org/10.1002/mds.10358)
113. Benazzouz A, Hallett M (2000) Mechanism of action of deep brain stimulation. *Neurology* 55:S13–S16
  114. Ross CA, Poirier MA (2004) Protein aggregation and neurodegenerative disease. *Nat Med* 10(Suppl):S10–S17. doi:[10.1038/nml1066](https://doi.org/10.1038/nml1066)
  115. Raj A, Kuceyeski A, Weiner M (2012) A network diffusion model of disease progression in dementia. *Neuron* 73:1204–1215. doi:[10.1016/j.neuron.2011.12.040](https://doi.org/10.1016/j.neuron.2011.12.040)
  116. Beal MF (1995) Aging, energy, and oxidative stress in neurodegenerative diseases. *Ann Neurol* 38:357–366. doi:[10.1002/ana.410380304](https://doi.org/10.1002/ana.410380304)
  117. Van Praag H, Fleshner M, Schwartz MW, Mattson MP (2014) Exercise, energy intake, glucose homeostasis, and the brain. *J Neurosci* 34:15139–15149. doi:[10.1523/JNEUROSCI.2814-14.2014](https://doi.org/10.1523/JNEUROSCI.2814-14.2014)
  118. Somogyi P, Hodgson AJ, Smith AD (1979) An approach to tracing neuron networks in the cerebral cortex and basal ganglia. Combination of Golgi staining, retrograde transport of horseradish peroxidase and anterograde degeneration of synaptic boutons in the same material. *Neuroscience* 4:1805–1852
  119. Gray EG (1959) Axo-somatic and axo-dendritic synapses of the cerebral cortex: an electron microscope study. *J Anat* 93:420–433
  120. Colonnier M (1968) Synaptic patterns on different cell types in the different laminae of the cat visual cortex. An electron microscope study. *Brain Res* 9:268–287
  121. Dhawale A, Bhalla US (2008) The network and the synapse: 100 years after Cajal. *HFSP J* 2:12–16. doi:[10.2976/1.2835214](https://doi.org/10.2976/1.2835214)
  122. Roberts RC, Gaiher LA, Peretti FJ et al (1996) Synaptic organization of the human striatum: a postmortem ultrastructural study. *J Comp Neurol* 374:523–534. doi:[10.1002/\(SICI\)1096-9861\(19961028\)374:4<523::AID-CNE4>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1096-9861(19961028)374:4<523::AID-CNE4>3.0.CO;2-3)
  123. Biederer T, Sara Y, Mozhayeva M et al (2002) SynCAM, a synaptic adhesion molecule that drives synapse assembly. *Science* 297:1525–1531. doi:[10.1126/science.1072356](https://doi.org/10.1126/science.1072356)
  124. Livet J, Weissman TA, Kang H et al (2007) Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* 450:56–62. doi:[10.1038/nature06293](https://doi.org/10.1038/nature06293)
  125. Lichtman JW, Livet J, Sanes JR (2008) A technicolour approach to the connectome. *Nat Rev Neurosci* 9:417–422. doi:[10.1038/nrn2391](https://doi.org/10.1038/nrn2391)
  126. Gomes FC, Spohr TC, Martinez R, Moura Neto V (2001) Cross-talk between neurons and glia: highlights on soluble factors. *Braz J Med Biol Res* 34:611–620
  127. Piet R, Vargová L, Syková E et al (2004) Physiological contribution of the astrocytic environment of neurons to intersynaptic crosstalk. *Proc Natl Acad Sci USA* 101:2151–2155. doi:[10.1073/pnas.0308408100](https://doi.org/10.1073/pnas.0308408100)
  128. Soriano J, Rodríguez Martínez M, Tlustý T, Moses E (2008) Development of input connections in neural cultures. *Proc Natl Acad Sci U S A* 105:13758–13763. doi:[10.1073/pnas.0707492105](https://doi.org/10.1073/pnas.0707492105)
  129. Van Bussel F, Kriener B, Timme M (2011) Inferring synaptic connectivity from spatio-temporal spike patterns. *Front Comput Neurosci* 5:3. doi:[10.3389/fncom.2011.00003](https://doi.org/10.3389/fncom.2011.00003)
  130. Napoli A, Xie J, Obeid I (2014) Understanding the temporal evolution of neuronal connectivity in cultured networks using statistical analysis. *BMC Neurosci* 15:17. doi:[10.1186/1471-2202-15-17](https://doi.org/10.1186/1471-2202-15-17)
  131. Tye KM, Deisseroth K (2012) Optogenetic investigation of neural circuits underlying brain disease in animal models. *Nat Rev Neurosci* 13:251–266. doi:[10.1038/nrn3171](https://doi.org/10.1038/nrn3171)
  132. Zhang F, Gradinaru V, Adamantidis AR et al (2010) Optogenetic interrogation of neural circuits: technology for probing mammalian brain structures. *Nat Protoc* 5:439–456. doi:[10.1038/nprot.2009.226](https://doi.org/10.1038/nprot.2009.226)
  133. Fenno LE, Mattis J, Ramakrishnan C et al (2014) Targeting cells with single vectors using multiple-feature Boolean logic. *Nat Methods* 11:763–772. doi:[10.1038/nmeth.2996](https://doi.org/10.1038/nmeth.2996)
  134. Deisseroth K, Feng G, Majewska AK et al (2006) Next-generation optical technologies for illuminating genetically targeted brain circuits. *J Neurosci* 26:10380–10386. doi:[10.1523/JNEUROSCI.3863-06.2006](https://doi.org/10.1523/JNEUROSCI.3863-06.2006)
  135. Papa M, De Luca C, Petta F et al (2014) Astrocyte-neuron interplay in maladaptive plasticity. *Neurosci Biobehav Rev* 42:35–54. doi:[10.1016/j.neubiorev.2014.01.010](https://doi.org/10.1016/j.neubiorev.2014.01.010)
  136. López-Hidalgo M, Schummers J (2014) Cortical maps: a role for astrocytes? *Curr Opin Neurobiol* 24:176–189. doi:[10.1016/j.conb.2013.11.001](https://doi.org/10.1016/j.conb.2013.11.001)
  137. Gullo F, Maffezzoli A, Dossi E et al (2012) Classifying heterogeneity of spontaneous



- up-states: a method for revealing variations in firing probability, engaged neurons and Fano factor. *J Neurosci Methods* 203:407–417. doi:[10.1016/j.jneumeth.2011.10.014](https://doi.org/10.1016/j.jneumeth.2011.10.014)
138. Chung K, Wallace J, Kim S-Y et al (2013) Structural and molecular interrogation of intact biological systems. *Nature* 497:332–337. doi:[10.1038/nature12107](https://doi.org/10.1038/nature12107)
  139. Ke M-T, Fujimoto S, Imai T (2013) SeeDB: a simple and morphology-preserving optical clearing agent for neuronal circuit reconstruction. *Nat Neurosci* 16:1154–1161. doi:[10.1038/nn.3447](https://doi.org/10.1038/nn.3447)
  140. Ahrens MB, Li JM, Orger MB et al (2012) Brain-wide neuronal dynamics during motor adaptation in zebrafish. *Nature* 485:471–477. doi:[10.1038/nature11057](https://doi.org/10.1038/nature11057)
  141. Freeman J, Vladimirov N, Kawashima T et al (2014) Mapping brain activity at scale with cluster computing. *Nat Methods* 11:941–950. doi:[10.1038/nmeth.3041](https://doi.org/10.1038/nmeth.3041)
  142. Calimera A, Macii E, Poncino M (2013) The Human Brain Project and neuromorphic computing. *Funct Neurol* 28:191–196
  143. Tønnesen J, Sørensen AT, Deisseroth K et al (2009) Optogenetic control of epileptiform activity. *Proc Natl Acad Sci U S A* 106:12162–12167. doi:[10.1073/pnas.0901915106](https://doi.org/10.1073/pnas.0901915106)
  144. Heinrichs-Graham E, Wilson TW, Santamaria PM et al (2014) Neuromagnetic evidence of abnormal movement-related beta desynchronization in Parkinson's disease. *Cereb Cortex* 24:2669–2678. doi:[10.1093/cercor/bht121](https://doi.org/10.1093/cercor/bht121)
  145. Bullmore E, Sporns O (2012) The economy of brain network organization. *Nat Rev Neurosci* 13:336–349. doi:[10.1038/nrn3214](https://doi.org/10.1038/nrn3214)
  146. Appel-Cresswell S, de la Fuente-Fernandez R, Galley S, McKeown MJ (2010) Imaging of compensatory mechanisms in Parkinson's disease. *Curr Opin Neurol* 23:407–412. doi:[10.1097/WCO.0b013e32833b6019](https://doi.org/10.1097/WCO.0b013e32833b6019)
  147. De Reus MA, van den Heuvel MP (2013) The parcellation-based connectome: limitations and extensions. *Neuroimage* 80:397–404. doi:[10.1016/j.neuroimage.2013.03.053](https://doi.org/10.1016/j.neuroimage.2013.03.053)
  148. Wen W, He Y, Sachdev P (2011) Structural brain networks and neuropsychiatric disorders. *Curr Opin Psychiatry* 24:219–225. doi:[10.1097/YCO.0b013e32834591f8](https://doi.org/10.1097/YCO.0b013e32834591f8)
  149. Rombouts SARB, Damoiseaux JS, Goekoop R et al (2009) Model-free group analysis shows altered BOLD fMRI networks in dementia. *Hum Brain Mapp* 30:256–266. doi:[10.1002/hbm.20505](https://doi.org/10.1002/hbm.20505)
  150. Onias H, Viol A, Palhano-Fontes F et al (2013) Brain complex network analysis by means of resting state fMRI and graph analysis: will it be helpful in clinical epilepsy? *Epilepsy Behav.* doi:[10.1016/j.yebeh.2013.11.019](https://doi.org/10.1016/j.yebeh.2013.11.019)
  151. Huang S, Li J, Ye J et al (2013) A sparse structure learning algorithm for Gaussian Bayesian Network identification from high-dimensional data. *IEEE Trans Pattern Anal Mach Intell* 35:1328–1342. doi:[10.1109/TPAMI.2012.129](https://doi.org/10.1109/TPAMI.2012.129)
  152. Baggio H-C, Sala-Llonch R, Segura B et al (2014) Functional brain networks and cognitive deficits in Parkinson's disease. *Hum Brain Mapp* 35:4620–4634. doi:[10.1002/hbm.22499](https://doi.org/10.1002/hbm.22499)
  153. Toussaint P-J, Maiz S, Coynel D et al (2014) Characteristics of the default mode functional connectivity in normal ageing and Alzheimer's disease using resting state fMRI with a combined approach of entropy-based and graph theoretical measurements. *Neuroimage* 101:778–786. doi:[10.1016/j.neuroimage.2014.08.003](https://doi.org/10.1016/j.neuroimage.2014.08.003)
  154. Abela E, Rummel C, Hauf M et al (2014) Neuroimaging of epilepsy: lesions, networks, oscillations. *Clin Neuroradiol* 24:5–15. doi:[10.1007/s00062-014-0284-8](https://doi.org/10.1007/s00062-014-0284-8)
  155. Tang CC, Eidelberg D (2010) Abnormal metabolic brain networks in Parkinson's disease from blackboard to bedside. *Prog Brain Res* 184:161–176. doi:[10.1016/S0079-6123\(10\)84008-7](https://doi.org/10.1016/S0079-6123(10)84008-7)
  156. Niethammer M, Tang CC, Ma Y et al (2013) Parkinson's disease cognitive network correlates with caudate dopamine. *Neuroimage* 78:204–209. doi:[10.1016/j.neuroimage.2013.03.070](https://doi.org/10.1016/j.neuroimage.2013.03.070)
  157. Teune LK, Strijkert F, Renken RJ et al (2014) The Alzheimer's disease-related glucose metabolic brain pattern. *Curr Alzheimer Res* 11:725–732
  158. Crossley NA, Mechelli A, Scott J et al (2014) The hubs of the human connectome are generally implicated in the anatomy of brain disorders. *Brain* 137:2382–2395. doi:[10.1093/brain/awu132](https://doi.org/10.1093/brain/awu132)
  159. Coenen VA, Allert N, Paus S et al (2014) Modulation of the cerebello-thalamo-cortical network in thalamic deep brain stimulation for tremor: a diffusion tensor imaging study. *Neurosurgery.* doi:[10.1227/NEU.0000000000000540](https://doi.org/10.1227/NEU.0000000000000540)
  160. Park KM, Shin KJ, Ha SY et al (2014) Response to antiepileptic drugs in partial epilepsy with structural lesions on MRI. *Clin Neurol Neurosurg* 123:64–68. doi:[10.1016/j.clineuro.2014.04.029](https://doi.org/10.1016/j.clineuro.2014.04.029)
  161. Liu Y, Yu C, Zhang X et al (2014) Impaired long distance functional connectivity and weighted network architecture in Alzheimer's

- disease. *Cereb Cortex* 24:1422–1435. doi:[10.1093/cercor/bhs410](https://doi.org/10.1093/cercor/bhs410)
162. Zhong Y, Huang L, Cai S et al (2014) Altered effective connectivity patterns of the default mode network in Alzheimer's disease: an fMRI study. *Neurosci Lett* 578:171–175. doi:[10.1016/j.neulet.2014.06.043](https://doi.org/10.1016/j.neulet.2014.06.043)
163. Schiff SJ (2010) Towards model-based control of Parkinson's disease. *Philos Trans A Math Phys Eng Sci* 368:2269–2308. doi:[10.1098/rsta.2010.0050](https://doi.org/10.1098/rsta.2010.0050)
164. Stocco A, Lebiere C, Anderson JR (2010) Conditional routing of information to the cortex: a model of the basal ganglia's role in cognitive coordination. *Psychol Rev* 117:541–574. doi:[10.1037/a0019077](https://doi.org/10.1037/a0019077)
165. Redgrave P, Rodriguez M, Smith Y et al (2010) Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nat Rev Neurosci* 11:760–772. doi:[10.1038/nrn2915](https://doi.org/10.1038/nrn2915)
166. Thibeault CM, Srinivasa N (2013) Using a hybrid neuron in physiologically inspired models of the basal ganglia. *Front Comput Neurosci* 7:88. doi:[10.3389/fncom.2013.00088](https://doi.org/10.3389/fncom.2013.00088)
167. Rummel C, Goodfellow M, Gast H et al (2013) A systems-level approach to human epileptic seizures. *Neuroinformatics* 11:159–173. doi:[10.1007/s12021-012-9161-2](https://doi.org/10.1007/s12021-012-9161-2)
168. Olde Dubbelink KTE, Hillebrand A, Stoffers D et al (2014) Disrupted brain network topology in Parkinson's disease: a longitudinal magnetoencephalography study. *Brain* 137:197–207. doi:[10.1093/brain/awt316](https://doi.org/10.1093/brain/awt316)
169. Morales DA, Vives-Gilabert Y, Gomez-Anson B et al (2013) Predicting dementia development in Parkinson's disease using Bayesian network classifiers. *Psychiatry Res* 213:92–98. doi:[10.1016/j.psychres.2012.06.001](https://doi.org/10.1016/j.psychres.2012.06.001)
170. Cole DM, Oei NYL, Soeter RP et al (2013) Dopamine-dependent architecture of cortico-subcortical network connectivity. *Cereb Cortex* 23:1509–1516. doi:[10.1093/cercor/bhs136](https://doi.org/10.1093/cercor/bhs136)
171. Seibyl J, Russell D, Jennings D, Marek K (2012) Neuroimaging over the course of Parkinson's disease: from early detection of the at-risk patient to improving pharmacotherapy of later-stage disease. *Semin Nucl Med* 42:406–414. doi:[10.1053/j.semnuclmed.2012.06.003](https://doi.org/10.1053/j.semnuclmed.2012.06.003)
172. Rao JA, Harrington DL, Durgerian S et al (2014) Disruption of response inhibition circuits in prodromal Huntington disease. *Cortex* 58:72–85. doi:[10.1016/j.cortex.2014.04.018](https://doi.org/10.1016/j.cortex.2014.04.018)
173. Nombela C, Rowe JB, Winder-Rhodes SE et al (2014) Genetic impact on cognition and brain function in newly diagnosed Parkinson's disease: ICICLE-PD study. *Brain* 137:2743–2758. doi:[10.1093/brain/awu201](https://doi.org/10.1093/brain/awu201)
174. Van Diessen E, Diederer SJH, Braun KPJ et al (2013) Functional and structural brain networks in epilepsy: what have we learned? *Epilepsia* 54:1855–1865. doi:[10.1111/epi.12350](https://doi.org/10.1111/epi.12350)
175. Stam CJ, Tewarie P, Van Dellen E et al (2014) The trees and the forest: characterization of complex brain networks with minimum spanning trees. *Int J Psychophysiol* 92:129–138. doi:[10.1016/j.ijpsycho.2014.04.001](https://doi.org/10.1016/j.ijpsycho.2014.04.001)
176. Otte WM, Dijkhuizen RM, van Meer MPA et al (2012) Characterization of functional and structural integrity in experimental focal epilepsy: reduced network efficiency coincides with white matter changes. *PLoS One* 7:e39078. doi:[10.1371/journal.pone.0039078](https://doi.org/10.1371/journal.pone.0039078)
177. Kalitzin S, Koppert M, Petkov G, da Silva FL (2014) Multiple oscillatory states in models of collective neuronal dynamics. *Int J Neural Syst* 24:1450020. doi:[10.1142/S0129065714500208](https://doi.org/10.1142/S0129065714500208)
178. Holt AB, Netoff TI (2014) Origins and suppression of oscillations in a computational model of Parkinson's disease. *J Comput Neurosci* 37:505–521. doi:[10.1007/s10827-014-0523-7](https://doi.org/10.1007/s10827-014-0523-7)
179. Thiele I, Swainston N, Fleming RMT et al (2013) A community-driven global reconstruction of human metabolism. *Nat Biotechnol*. doi:[10.1038/nbt.2488](https://doi.org/10.1038/nbt.2488)
180. Jain S, van Kesteren RE, Heutink P (2012) High content screening in neurodegenerative diseases. *J Vis Exp*:e3452. doi:[10.3791/3452](https://doi.org/10.3791/3452)
181. Herrera F, Goncalves S, Outeiro TF (2012) Imaging protein oligomerization in neurodegeneration using bimolecular fluorescence complementation. *Methods Enzymol* 506:157–174. doi:[10.1016/B978-0-12-391856-7.00033-0](https://doi.org/10.1016/B978-0-12-391856-7.00033-0)
182. Boassa D, Berlanga ML, Yang MA et al (2013) Mapping the subcellular distribution of alpha-synuclein in neurons using genetically encoded probes for correlated light and electron microscopy: implications for Parkinson's disease pathogenesis. *J Neurosci* 33:2605–2615. doi:[10.1523/JNEUROSCI.2898-12.2013](https://doi.org/10.1523/JNEUROSCI.2898-12.2013)
183. Meisel C, Kuehn C (2012) Scaling effects and spatio-temporal multilevel dynamics in epileptic seizures. *PLoS One* 7:e30371. doi:[10.1371/journal.pone.0030371](https://doi.org/10.1371/journal.pone.0030371)

184. Abuhassan K, Coyle D, Maguire LP (2012) Investigating the neural correlates of pathological cortical networks in Alzheimer's disease using heterogeneous neuronal models. *IEEE Trans Biomed Eng* 59:890–896. doi:[10.1109/TBME.2011.2181843](https://doi.org/10.1109/TBME.2011.2181843)
185. Piray P, Keramati MM, Dezfouli A et al (2010) Individual differences in nucleus accumbens dopamine receptors predict development of addiction-like behavior: a computational approach. *Neural Comput* 22:2334–2368. doi:[10.1162/NECO\\_a\\_00009](https://doi.org/10.1162/NECO_a_00009)
186. Garcia-Reitboeck P, Anichtchik O, Dalley JW et al (2013) Endogenous alpha-synuclein influences the number of dopaminergic neurons in mouse substantia nigra. *Exp Neurol* 248:541–545. doi:[10.1016/j.expneurol.2013.07.015](https://doi.org/10.1016/j.expneurol.2013.07.015)
187. McKinstry SU, Karadeniz YB, Worthington AK et al (2014) Huntingtin is required for normal excitatory synapse development in cortical and striatal circuits. *J Neurosci* 34:9455–9472. doi:[10.1523/JNEUROSCI.4699-13.2014](https://doi.org/10.1523/JNEUROSCI.4699-13.2014)
188. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL et al (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489:391–399. doi:[10.1038/nature11405](https://doi.org/10.1038/nature11405)
189. Lew MF (2011) The evidence for disease modification in Parkinson's disease. *Int J Neurosci* 121(Suppl):18–26. doi:[10.3109/00207454.2011.620194](https://doi.org/10.3109/00207454.2011.620194)
190. Bartus RT, Weinberg MS, Samulski RJ (2014) Parkinson's disease gene therapy: success by design meets failure by efficacy. *Mol Ther* 22:487–497. doi:[10.1038/mt.2013.281](https://doi.org/10.1038/mt.2013.281)
191. Lorenzi M, Beltramello A, Mercuri NB et al (2011) Effect of memantine on resting state default mode network activity in Alzheimer's disease. *Drugs Aging* 28:205–217. doi:[10.2165/11586440-000000000-00000](https://doi.org/10.2165/11586440-000000000-00000)
192. Fraschini M, Demuru M, Puligheddu M et al (2014) The re-organization of functional brain networks in pharmaco-resistant epileptic patients who respond to VNS. *Neurosci Lett* 580:153–157. doi:[10.1016/j.neulet.2014.08.010](https://doi.org/10.1016/j.neulet.2014.08.010)
193. Ojemann GA, Ojemann J, Ramsey NF (2013) Relation between functional magnetic resonance imaging (fMRI) and single neuron, local field potential (LFP) and electrocorticography (ECoG) activity in human cortex. *Front Hum Neurosci* 7:34. doi:[10.3389/fnhum.2013.00034](https://doi.org/10.3389/fnhum.2013.00034)
194. Hill NJ, Gupta D, Brunner P et al (2012) Recording human electrocorticographic (ECoG) signals for neuroscientific research and real-time functional cortical mapping. *J Vis Exp*. doi:[10.3791/3993](https://doi.org/10.3791/3993)
195. Kemmotsu N, Kucukboyaci NE, Leyden KM et al (2014) Frontolimbic brain networks predict depressive symptoms in temporal lobe epilepsy. *Epilepsy Res* 108:1554–1563. doi:[10.1016/j.eplepsyres.2014.08.018](https://doi.org/10.1016/j.eplepsyres.2014.08.018)
196. Kahan J, Urner M, Moran R et al (2014) Resting state functional MRI in Parkinson's disease: the impact of deep brain stimulation on “effective” connectivity. *Brain* 137:1130–1144. doi:[10.1093/brain/awu027](https://doi.org/10.1093/brain/awu027)
197. Stypulkowski PH, Stanslaski SR, Jensen RM et al (2014) Brain stimulation for epilepsy—local and remote modulation of network excitability. *Brain Stimul* 7:350–358. doi:[10.1016/j.brs.2014.02.002](https://doi.org/10.1016/j.brs.2014.02.002)
198. Guo Y, Rubin JE (2011) Multi-site stimulation of subthalamic nucleus diminishes thalamocortical relay errors in a biophysical network model. *Neural Netw* 24:602–616. doi:[10.1016/j.neunet.2011.03.010](https://doi.org/10.1016/j.neunet.2011.03.010)
199. De Munter JPJM, Melamed E, Wolters EC (2014) Stem cell grafting in parkinsonism—why, how and when. *Parkinsonism Relat Disord* 20(Suppl 1):S150–S153. doi:[10.1016/S1353-8020\(13\)70036-1](https://doi.org/10.1016/S1353-8020(13)70036-1)
200. Ben-Yehudah A, Easley CA 4th, Hermann BP et al (2010) Systems biology discoveries using non-human primate pluripotent stem and germ cells: novel gene and genomic imprinting interactions as well as unique expression patterns. *Stem Cell Res Ther* 1:24. doi:[10.1186/scrt24](https://doi.org/10.1186/scrt24)



# **Part III**

## **Systems Medicine Projects and Case Studies**

## Computational Modeling of Human Metabolism and Its Application to Systems Biomedicine

Maike K. Aurich and Ines Thiele

### Abstract

Modern high-throughput techniques offer immense opportunities to investigate whole-systems behavior, such as those underlying human diseases. However, the complexity of the data presents challenges in interpretation, and new avenues are needed to address the complexity of both diseases and data. Constraint-based modeling is one formalism applied in systems biology. It relies on a genome-scale reconstruction that captures extensive biochemical knowledge regarding an organism. The human genome-scale metabolic reconstruction is increasingly used to understand normal cellular and disease states because metabolism is an important factor in many human diseases. The application of human genome-scale reconstruction ranges from mere querying of the model as a knowledge base to studies that take advantage of the model's topology and, most notably, to functional predictions based on cell- and condition-specific metabolic models built based on omics data.

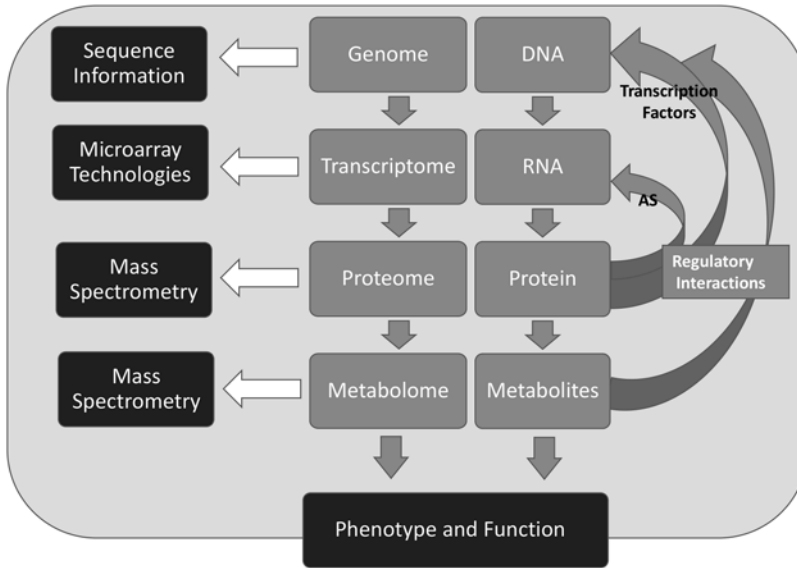
An increasing number and diversity of biomedical questions are being addressed using constraint-based modeling and metabolic models. One of the most successful biomedical applications to date is cancer metabolism, but constraint-based modeling also holds great potential for inborn errors of metabolism or obesity. In addition, it offers great prospects for individualized approaches to diagnostics and the design of disease prevention and intervention strategies. Metabolic models support this endeavor by providing easy access to complex high-throughput datasets. Personalized metabolic models have been introduced. Finally, constraint-based modeling can be used to model whole-body metabolism, which will enable the elucidation of metabolic interactions between organs and disturbances of these interactions as either causes or consequence of metabolic diseases. This chapter introduces constraint-based modeling and describes some of its contributions to systems biomedicine.

**Key words** Systems biology, Constraint-based modeling, Personalized health, Metabolomics, OMICS, COBRA, Flux balance analysis, Cancer metabolism, Human disease, Personalized models

---

### 1 From Systems Biology to Constrained-Based Modeling

Technical developments in molecular biology have facilitated the emergence of systems biology, which seeks to investigate the behavior of whole systems [1]. High-throughput techniques enable the simultaneous measurement of thousands of cellular components (Fig. 1). Detailed dynamic models of metabolism remain limited in scope due to a lack of kinetic parameters to describe each reaction in



**Fig. 1** High-throughput techniques allow the simultaneous assessment of many thousands of cellular variables. These datasets provide a snapshot of the sampled cell. Sequencing techniques provide comprehensive insights into the genomes of organisms. The cellular RNA content can be assessed through RNA sequencing and microarray techniques. The metabolome and the proteome can be defined through mass spectrometry. The cellular components are highly regulated and influence the readout at different stages of the omics cascade. Metabolomic data are the endpoint of the cascade and are the closest to the actual phenotype, thereby allowing the connection of a genotype to a phenotype. *AS* alternative splicing

complex systems and the amount of data needed to obtain these parameters [2, 3]. Constraint-based reconstruction and analysis circumvent this parameter bottleneck by assuming steady-state conditions of the modeled biological system [4], permitting the investigation of biological systems on a large scale.

## 2 Brief Introduction to Constraint-Based Modeling

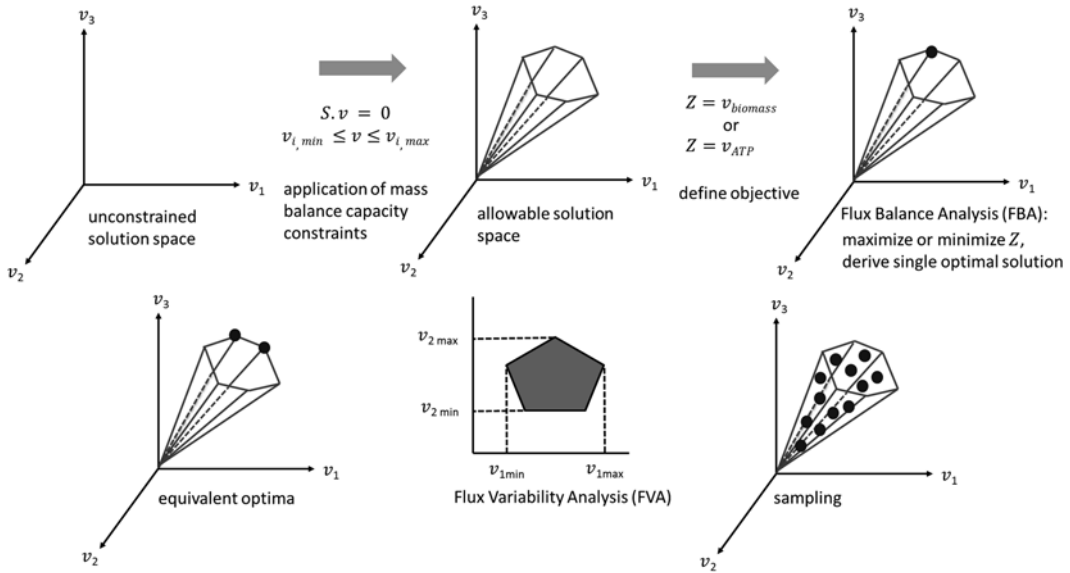
The constraint-based modeling and analysis approach uses the stoichiometry of biochemical reactions to mathematically represent biochemical networks of cellular processes, i.e., metabolism, signaling, or transcriptional/translational networks (Fig. 2) [5–8]. Such a genome-scale network reconstruction is assembled in a bottom-up reconstruction process, according to standardized operating procedures and using ready-to-use reconstruction and analysis tools [9, 10]. This process is accomplished through the manual consideration of extensive amounts of organism-specific literature [11]. Consequently, the result of the reconstruction process is a biochemically, genetically, and genomically structured knowledge base of the target organism. This knowledge base is curated and validated to ensure the correct prediction of biological functions by the resulting biochemical network model [4, 11, 12].

Vocabulary	
<p><b>Constraint-based modeling</b> – A computational modeling approach that uses reaction stoichiometry and constraints to predict steady-states of large-scale metabolic network models independent of kinetic parameters.</p>	<p><b>Solution space</b> – The set of feasible metabolic states of the model is defined by the imposed constraints. Together, all feasible states of the metabolic model make up the solution space.</p>
<p><b>Steady-state</b> – Constraint-based modeling assumes a steady-state of the system in which the change in metabolite concentration over time is equal to zero.</p>	<p><b>Flux balance analysis</b> – A computational method for interrogating the metabolic model. The linear programming problem returns a solution of how the metabolic model can be used to optimize a specified objective function, e.g., maximization of growth. However, there are usually many alternate optimal solutions. These alternative solutions result from the reaction and pathway redundancy of the metabolic network.</p>
<p><b>Genome scale metabolic reconstruction</b> – Contains the entire set of metabolic reactions that appear throughout an organism. The reactions are further associated with the genes encoding proteins that catalyze reactions. The reconstruction is compiled based on genome annotation, large amounts of biochemical literature, and experimental data.</p>	<p><b>Flux variability analysis</b> – A variation of flux balance analysis that calculates for each reaction the minimally and maximally possible reaction flux.</p>
<p><b>Gene-protein-reaction associations (GPRs)</b> – Define through Boolean formalism, which gene products catalyze respective network reactions, and consider the assembly of protein complexes (AND) from multiple subunits and isozymes (OR).</p>	<p><b>Objective function</b> – The objective function defines the goal of the predicted metabolic state. In microorganisms, the objective function is thought to be the outcome of an evolutionary process that drove these organisms to optimize their metabolic network for growth.</p>
<p><b>Metabolic model</b> – A reconstruction can be converted into a metabolic model through the conversion of the reaction list into a computable matrix format, the imposition of constraints, and the definition of the exchange medium. While there is only one metabolic reconstruction, many condition-specific metabolic models can be derived from the reconstruction.</p>	<p><b>Recon (1 and 2)</b> – The human genome-scale metabolic reconstruction is referred to as Recon followed by the version number. It is subject to constant improvement and extension.</p>
<p><b>Stoichiometric matrix (S)</b> – The computable matrix that contains the reaction stoichiometry of metabolites that participate in each reaction. It contains one row for each metabolite and a column for each reaction.</p>	<p><b>Tissue-(Cell-type)-specific metabolic model</b> – Compared to genome-scale metabolic reconstructions, a tissue-specific metabolic model contains only a cell type- or tissue-specific subset of metabolic pathways and functions. However, many different condition-specific models can be derived from tissue (cell type)-specific metabolic models through the integration of additional constraints.</p>
<p><b>Constraints</b> – Limit the behavior of the metabolic model to the set of feasible network states (e.g., reaction directionality and reaction rates (bounds) or the accumulation of metabolites (mass-balance)).</p>	

**Fig. 2** Explanations of the technical terms used

Genome-scale network reconstructions contain a hierarchical structure. Genes are connected to proteins and enzymes and the catalyzed reactions. These gene–protein–reaction associations are formulated as Boolean rules considering isozymes (OR) and all subunits of protein complexes (AND). The gene–protein–reaction associations are the entry points for the integration of transcriptomic and proteomic data into the network context. In addition, these associations are used to model the impact of gene loss of functions [13] and drug interventions [14]. Hence, the correct formulation of gene–protein–reaction associations is an important prerequisite for any network contextualization.

The reconstruction can be converted into a mathematical model after the compilation of a comprehensive reaction list and association of all known genes with those reactions. First, the reaction list is converted into a computable matrix format (stoichiometric matrix). The stoichiometric matrix contains a row for each metabolite and a column for each reaction [3, 12, 15]. The non-zero entries of the stoichiometric matrix describe which metabolites participate in each reaction. A negative entry identifies a substrate and a positive entry defines a product [3]. Second, systems boundaries are formulated equivalent to those occurring in



**Fig. 3** Definitions and methods for the functional analysis of the feasible solution space. Figure redrawn based on [15, 114]

in vivo or in vitro systems (Fig. 3, [12, 15]). In the model, constraints are applied as either balances or bounds.

According to the physical law of mass conservation, the net production and consumption of a metabolite are balanced in the steady state. The steady-state assumption, including all mass balance equations ( $dx/dt=0$ ), is contained in the equation  $S \times V=0$ .  $S$  is the stoichiometric matrix and  $V$  is the flux vector containing all reaction fluxes of one of the entire set of optimal states of the system [16]. The biological justification of the steady state is that biochemical reactions occur at a much faster timescale (milliseconds to seconds) than other cellular events, e.g., cellular growth (minutes to hours), and environmental and regulatory changes (hours or days) [16].

Bounds constitute upper and lower limits ( $v_{\min}, i \leq v_i \leq v_{\max}, i$ ) on each reaction  $i$  ( $i \in N$ ). The bounds can be set to  $v_{\min}, i=0$  for forward reactions and  $v_{\max}, i=0$  for reverse reactions (Fig. 3) [3, 17]. These upper and lower bounds can be further adjusted based on experimental data (e.g., metabolite uptake and secretion fluxes or enzyme capacity) [3, 17] for a more realistic definition of the steady-state solution space (Fig. 2). By integrating experimental data, a condition-specific subset of feasible flux distributions from the entity of possible network states is selected (Fig. 3). Bounds on the exchange reactions provide the entry point for the integration of extracellular metabolomic data [18]. After establishing the model format, the set of feasible network states can be interrogated (Fig. 3), e.g., using MATLAB as a programming and simulation environment and the COBRA toolbox [10, 19].

### 3 Methods to Explore the Solution Space

Methods to interrogate constraint-based models [10, 19] can be classified as biased and unbiased. Biased methods rely on the optimality principle and require a user-defined objective function that biologically translates into the (assumed) cellular goal (Fig. 3). By contrast, unbiased methods do not require such an a priori optimality assumption. Thus, network analysis methods have different a priori requirements.

Biomass generation and ATP production are commonly used as objective functions [20, 21]. A biomass objective function accounts for all known precursors required to create a new cell [22]. This function is used to identify the subset of network states that support optimal biomass generation. This optimality principle is, at least in microorganisms, thought to be the outcome of an evolutionary process driving the organism to maximal proliferation rates and the optimal use of available, usually limited, resources. Similarly, the metabolism of highly proliferating human cancer cells might be optimized for biomass production [23], and such biomass objective functions can be applied to model cancer cell metabolism.

However, the definition of an objective function is more difficult for differentiated, non-proliferating human cells and tissues. Modeling the metabolism of cells that are not optimized for biomass production might lead to false or biologically irrelevant predictions. To avoid this outcome, different objectives have been formulated. One alternative to the biomass objective function is to investigate organ-, cell-, and tissue-specific metabolic functions, e.g., the uptake and secretion of bile acids or the synthesis of polyamines in enterocytes [13, 24, 25]. These functions optimize the model for the utilization of particular pathways or for metabolite production.

Further, the classical biomass objective function has been modified to better represent human tissues or cells. Biomass maintenance describes a cell's or tissue's capability to synthesize all biomass components except nuclear deoxynucleotides [13, 26]. The model is no longer forced to produce the building blocks for DNA replication. For cells or conditions, in which no transcription or translation occurs (e.g., platelets or red blood cells), a corresponding maintenance biomass objective function can be formulated that does not include deoxyNTP, NTP, or amino acid requirements for replication, transcription, and translation. Thus, modifications to the biomass objective function could be used to model human cells.

Another means of investigating metabolic models is the use of unbiased network interrogation methods. These unbiased methods permit the interrogation of the allowable steady-state solution

space without any optimality assumption. In recent years, sampling methods have been increasingly applied to investigate cell type-specific metabolic networks [3, 18, 27–30]. Because these methods permit an unbiased assessment of the solution space, they are well suited for biomedical applications. Below, a subset of biased and unbiased interrogation methods are briefly introduced.

### **3.1 Flux Balance Analysis**

Flux balance analysis is used to predict network states subject to a given objective and has been widely applied to investigate metabolic networks. Flux balance analysis is used to predict a single flux distribution through the formulation of a linear programming problem that either minimizes or maximizes the flux through the objective function subject to all imposed constraints (Fig. 3) [15, 29]. The output of this analysis is a single flux distribution that lies at the extremity of the allowable solution space (Fig. 3) [15]. This flux distribution describes the contribution of each reaction in the network to the computed phenotype [15].

Because the stoichiometric matrix is underdetermined, i.e., the number of reactions (variables) exceeds the number of metabolites (equations), an infinite number of flux distributions exist for the same maximal objective value (alternative objective solutions) [16, 31]. The multitude of alternate optimal solutions reflects the system's flexibility. This redundancy contributes to the robustness of the metabolic network (Fig. 3) [31]. The actual cellular state depends on additional factors, such as the interplay of enzymatic and genetic regulatory events [3, 16]. In the absence of sufficiently detailed constraints to exclude physiologically irrelevant network states, alternate optimal solutions should be investigated to obtain a comprehensive overview of the model's metabolic capabilities (Fig. 3).

### **3.2 Flux Variability Analysis**

Flux variability analysis can provide insights into these alternate optimal solutions. Flux variability analysis is a variation of flux balance analysis that calculates the minimal and maximal allowable flux for each reaction in the model [31]. The analysis returns the range of allowable fluxes for each reaction and can identify reactions that are never used or differently used under the applied sets of condition-specific constraints representing environmental or genetic conditions [3, 31]. Depending on the optional input, flux variability analysis can be used without relying on the optimality principle [10, 32] and is thus well suited to the study of human metabolism.

### **3.3 Sampling Analysis**

A more comprehensive resolution of alternate flux distributions can be obtained through sampling. During the sampling process, randomly distributed points (each comprising a flux distribution) are selected from the feasible solution space as an unbiased representation of the entire solution space (Fig. 3) [33]. For example, the artificial centering hit-and-run sampler has been implemented



in the COBRA toolbox and used to study the solution space of larger networks [10, 27, 33, 34]. The procedure starts from an initial point moving through the space with a randomly chosen direction and step length [27]. Only every  $i$ -th point is collected to ensure a random selection of the sampling points. However, in large-scale networks, the high dimensionality and size of the solution space render the coverage of the entire solution space in a finite time uncertain; this uncertainty has also been referred to as the slow mixing problem [33]. The outcome of the sampling analysis is commonly illustrated using histograms counting how often a flux value was computed for a reaction, i.e., the probability distribution of flux values. Note that flux variability analysis computes the extreme values of this probability distribution [29].

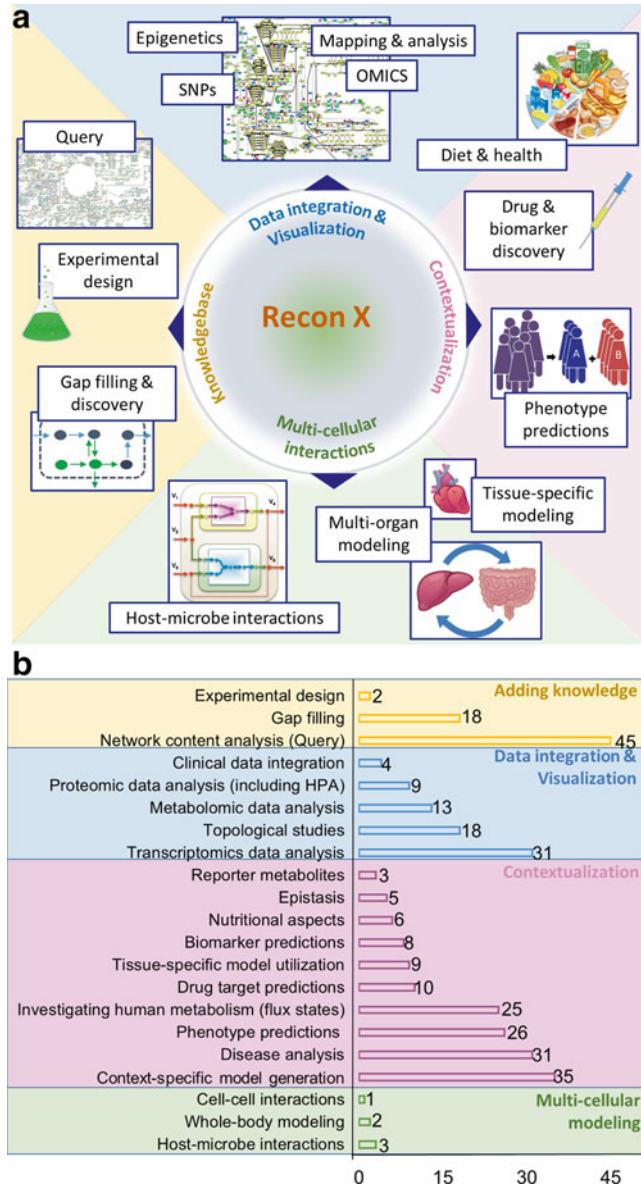
---

## 4 Human Metabolic Genome-Scale Reconstructions

Published in 2007, Homo Sapiens Recon 1 was the first genome-scale reconstruction of human metabolism [24]. Recon 1 captured the functions of 2004 proteins, 2766 metabolites, and 3311 metabolic and transport reactions, which were assembled in a bottom-up reconstruction process based on extensive literature reports (more than 1500 primary and review publications). The pathways of Recon 1 are distributed over eight cellular compartments (cytoplasm, mitochondria, nucleus, endoplasmic reticulum, Golgi apparatus, lysosome, peroxisome, and extracellular environment). Recon 1 was validated for 288 metabolic functions known to occur in cells throughout the human body, enabling its use as a predictive model of human metabolism (Fig. 4). This global human metabolic model represents most known biochemical transformations that can occur in at least one cell in the human body, rather than the metabolic functions of a particular cell type.

Since its publication, Recon 1 has been extensively used as a knowledge base of human metabolism, to investigate general and cell type-specific metabolism, to address knowledge and network gaps in human metabolism, to map experimental data, and to generate tissue-specific models to investigate human disease processes (Fig. 4) [14, 35–40]. Further, Recon 1 and the tissue-specific networks derived from it have been used to investigate host–pathogen and host–gut microbial interactions (Fig. 4) [30, 41, 42]. The broad applicability of Recon 1 makes it a pivotal resource for human metabolic research.

Other human metabolic reconstructions have been published in addition to Recon 1 [24]. The Edinburgh human metabolic network comprises 2823 reactions, 2671 metabolites, and 2322 genes [43]. Cellular compartments were not distinguished in the initial network, but a compartmentalized version was subsequently added [44]. A manually curated liver-specific metabolic network



**Fig. 4** Four categories of applications of human metabolic models. The figure does not distinguish between Recon 1 and Recon 2 and hence refers to Recon X. (a) Recon X can be used as a knowledge base of human metabolism, for data integration and visualization, and for the generation of cell- and condition-specific metabolic models. Recon X and its contextualized metabolic models can be combined to create cell–cell and host–microbe interaction networks. (b) Frequency distribution of the four major applications of Recon X

has also been published [45]. HepatoNet1 comprised 2539 reactions and 777 metabolites [45]. Finally, Recon 1 was used as an input for a human metabolic reactions database (HMR and HMR2.0) [46, 47]. This human metabolic reactions database has

been used to generate cell- and condition-specific models and to investigate the molecular mechanisms of cancer as well as obesity-related diseases, along with the RAVEN toolbox and tINIT [46–50]. Thus, different large-scale network resources of human metabolism exist and have been successfully applied to biomedical questions.

The process of network reconstruction is a laborious task. To improve predictability and to increase the applications of the human metabolic model (Fig. 4, [24]), expansion and correction of network content based on emerging biochemical knowledge are an ongoing, iterative process. As part of this iterative process, Recon 2 was published [25]. Recon 2 was created as a community-driven effort to combine Recon 1 with four other resources on human metabolism, namely, EHMN [44], HepatoNet1 [45], an acylcarnitine/fatty acid oxidation module [51], and the human small intestinal enterocyte reconstruction [13]. Recon 2 encompasses 1789 genes, 7440 reactions, and 2626 unique metabolites distributed over the eight aforementioned cellular compartments. The reaction content was doubled compared to Recon 1, and 62 % of the Recon 1 pathways were expanded in Recon 2. The improvement in pathway content was particularly evident in xenobiotic metabolism, cholesterol metabolism, and intracellular transporters of the endoplasmic reticulum, mitochondrial, and peroxisomal compartments, among others [25]. In addition, Recon 2 incorporates nine metabolic pathways completely absent in Recon 1 [25]. Thus, combining different resources on human metabolism improved pathway coverage in Recon 2 compared to Recon 1.

Furthermore, many dead-end metabolites and blocked reactions in Recon 1 were resolved in Recon 2. Dead-end metabolites are metabolites that can only be produced or consumed by the model. Blocked reactions cannot carry nonzero flux and are associated directly or indirectly with dead-end metabolites. Recon 2 resolved 307 dead-end metabolites and 827 blocked reactions from Recon 1. Moreover, Recon 2 was tested against 354 defined metabolic tasks [25]. Hence, Recon 2 closed many of the knowledge gaps present in Recon 1.

Recon 2's predictive capability has been demonstrated, e.g., through mapping of inborn errors of metabolism (IEMs) [25]. The model better predicted changes in biomarker metabolites of IEMs compared to Recon 1 (49 compared to 29) and with higher accuracy [25]. This higher predictive capability of Recon 2 further demonstrates its superiority over Recon 1 [25].

Similar to Recon 1, Recon 2 can be used for data mapping. To demonstrate this application, different datasets were mapped to Recon 2. Recon 2 captured the majority of the metabolites detected in the spent medium of the NCI-60 cancer cell line collection [25, 52]. The direction of exchange of most of the metabolites consumed or released by the cancer cells was unconstrained in Recon

2, indicating that Recon 2 could both uptake and secrete metabolites, whereas the cells could only secrete a metabolite. However, this mismatching was intuitive because the model comprises metabolic pathways found in cells throughout the human body and is not limited to a subset found in a particular cell type.

This universalism enables Recon 2 to function as a blueprint for the generation of cell- and condition-specific metabolic models. These subnetworks comprise cell type- and condition-specific subsets of metabolic pathways present in an individual cell or used under specific sets of environmental or (epi)genetic conditions. To demonstrate this template function, 65 cell type-specific draft reconstructions were generated from protein expression data from the human protein atlas [53]. Thus, Recon 2 constitutes a resource for the integration of various data types.

#### **4.1 Module Approach: Adding Missing Parts Sequentially, Piece by Piece**

Although Recon 2 is an extensive and detailed representation of biochemical transformations occurring in humans, knowledge gaps remain, which will be tackled using a combined computational and experimental approach [40, 54, 55]. To include novel evidence, we envision regular future updates of Recon in the form of “modules.” A functional module captures a specific part of the global human metabolic network, adds new reactions, and expands existing pathways (e.g., adding newly identified gene functions). It also corrects preexisting content (e.g., correction of enzyme localization, gene–protein–reaction associations, or substrate and cofactor utilization) as new knowledge becomes available. Through this modular approach, the original model is retained, while advancements are incorporated. Hence, both the universal and specific scope of the reconstruction are preserved.

The acylcarnitine/fatty acid oxidation module [51] expanded the lipid metabolism of Recon 1 [24] by accounting for the acylcarnitines that are used as biomarkers in newborn screening programs worldwide. This module captured the alpha, beta, and omega oxidation pathways of fatty acids in the form of 352 reactions, 139 metabolites, and 14 genes that were not only new but also compatible with Recon 1. Combining the acylcarnitine/fatty acid oxidation module with Recon 1 enabled the mapping of newborn screening data onto this network [51].

In another case, an intestinal transport/absorption module was assembled using manual curation of the scientific literature [13]. The intestinal transport/absorption module comprised 371 transport and exchange reactions, 27 metabolites, and 29 genes [13]. This module was combined with Recon 2 and the recently built enterocyte reconstruction to expand their transporter content [13].

Only 44 % of the metabolites present in the Human Metabolome Database [56] were present in the extracellular compartment of Recon 2, illustrating that a large portion of the transport systems appearing in humans was not captured [25]. In fact, corrections

and additions to extracellular metabolite transporters in both Recon 1 and Recon 2 have recently been summarized [57]. This assessment of the extracellular metabolite transporters highlighted general knowledge gaps in this field of metabolism [57] and revealed corrections that affect the representation of pH, drug resistance, and proliferative energy metabolism in cancer cells. In addition, it was reported that almost half of IEM genes associated with transport-associated IEMs were not captured in Recon 2 [57]. This work resulted in a transport module comprising 78 reactions, 87 metabolites, and 33 new genes that can be added to Recon 2, e.g., to better capture metabolite transport by cancer cells [57].

Moreover, a drug module capturing the metabolism of the 18 most highly prescribed drug groups has been reconstructed [58]. This module captured 187 metabolic reactions, 386 transport reactions, 210 metabolites, and 57 genes and can be combined with Recon 2. This module enabled the investigation of the influence of metabolism and factors, such as diet and genetics, on drug metabolism. Thus, this example illustrates how the modular approach can enable novel applications of Recon 2.

Finally, extensions can be driven by the attempt to integrate experimental datasets with the reconstruction. The mapping of extracellular metabolomic datasets led to the identification of metabolite transporters that were not captured in Recon [25]. After identifying transport proteins and transport mechanisms from the literature, 34 transport and modeling reactions were added to the model to improve mapping of the extracellular metabolomic profiles of two lymphoblastic leukemia cell lines [18]. The updated human metabolic reconstruction was subsequently used to generate condition-specific cell line models. Thus, another incentive for additions is improved representation of experimental and omics data by the model.

---

## 5 Analysis of Omics Data in the Context of COBRA Models

The development of methods to analyze omics datasets, reduce their complexity, and make their content accessible to non-biochemistry researchers is of major interest to the broader research community. Metabolic reconstructions can provide a context for the analysis of omics datasets because they provide a mechanistic and biologically well-defined framework for these data [4].

Cellular phenotypes differ, although all human cells carry the same genetic information. A cell type-specific set of genes is transcribed and mRNAs are subsequently translated into proteins that once activated, can perform their (catalytic) function (Fig. 1). Proteins are subject to degradation processes or may be inactivated because the abundance and activity of enzymes and proteins within the cell are tightly regulated [59]. The phenotypic and functional

differences among the approximately 200 human cell types arise from the regulation of gene expression and active cellular protein contents, pathways, and reaction fluxes [60].

High-throughput technologies offer snapshots of the cellular network on the level of RNA, protein, and metabolites covering thousands of cellular components. Differences in metabolic states can be read from these high-throughput data, e.g., to distinguish healthy states from disease states. However, the complexity of these datasets makes it challenging to obtain a clear picture of differences that can appear throughout the metabolic network. By providing a large-scale template to place the data into perspective, the human metabolic reconstruction can be used to aid the interpretation of these omics datasets and of cell type- and condition-specific metabolism.

There are multiple ways to combine omics data and metabolic models (Fig. 4). First, the topology of the reconstruction can be used for structured visualization of the data, e.g., by pathways, and to facilitate interpretation. As an example, Recon 1 was used for the interpretation of gene expression data to reveal the effects of gastric bypass surgery on skeletal muscle metabolism [24]. Comparison of size and pathway topology has further been used to verify a reduced metabolic network in clear cell renal cell carcinoma [61].

In addition, functional and topological aspects can be combined and used to identify correlations that could not be derived from the data alone. The analysis of single nucleotide polymorphisms in the network context revealed examples, in which single nucleotide polymorphisms with similar pathological impact were mapped to reactions that belonged to the same functionally correlated reaction set (reactions with 100 % correlated flux) [62]. Thus, the topology of the network in combination with functional predictions can aid the interpretation of these data.

The predictive nature of the human metabolic model can be exploited to investigate metabolic phenotypes. Omics data were used to formulate additional constraints to reduce or alter the solution space [63]. One example is the integration of extracellular metabolomic profiles of yeast into the network context [64]. First, metabolic models were tailored to different environmental conditions. Subsequently, the allowable network states were compared based on sampling analysis, and differences in intracellular metabolic flux states were identified along with the regions of the metabolic network that were distinct [64]. Thus, the metabolic model can predict phenotypic differences between cell type- and condition-specific differences.

Finally, omics data can be used to generate cell type-specific metabolic models. Such models constitute subnetworks of the global human models, and the decision to retain or discard reactions is made based on the experimental data and various criteria (Table 1). To date, cell type- and condition-specific metabolic

**Table 1**  
**Methods for network contextualization**

Method	Description	Objective function required	Data	Input format	Solver	Examples for models derived	Reference
Binary	Fluxes through reactions associated with absent genes are constrained to zero	No	Transcriptomic	P/A calls based on user-defined threshold	No	<i>S. cerevisiae</i> batch cultures [115]	[115]
GIMME	Functional flux is defined through FBA, and active and suppressed reactions are defined based on the GPR associations. Subsequently, inactive reactions are removed. Removed reactions required for the predetermined functional flux are reinserted to produce a functional submodel. Incorporates the idea of posttranscriptional regulation but relies on arbitrarily chosen flux distribution (FBA)	Yes	Transcriptomic or proteomic	P/A calls	LP	Macrophage [70], brain cells [37]	[35]
GIMMEp	Creates models, each satisfying a proteome-based objective, which are combined into one final GIMMEp model	Proteome-based OF	Transcriptomic and proteomic	P/A calls	LP	Murine macrophage [70]	[70]
iMAT	Maximizes the number of enzymes whose flux activity is consistent with their measured expression level (high flux to high expression and low flux to low expression) along with pathway length. Enzyme expression levels are considered as cues rather than fixed determinants of enzyme activity and connected flux. Posttranscriptional regulation is assumed to be the difference between the measured expression level and the predicted flux. Predicts tissue-specific metabolite exchanges	No	Transcriptomic or proteomic	High, medium, and low expression	MILP	Macrophage [70], brain cells [37], cardiomyocyte [67]	[65, 116]

(continued)



**Table 1**  
(continued)

Method	Description	Objective function required	Data	Input format	Solver	Examples for models derived	Reference
E-flux	Predicts metabolic states by setting maximum flux constraints as a function of measured gene expression. Reactions associated with low-expression genes are tightly constrained, and reactions associated with highly expressed genes are subject to loose constraints	Yes	Transcriptomic	Gene expression levels	LP	<i>M. tuberculosis</i> bacterium [59]	[59]
PROM	Generates integrated metabolic regulatory networks. Requires a metabolic network, a regulatory network, gene expression data from different conditions, and additional regulatory interactions. It uses probabilities estimated from the expression data and different conditions to represent gene states and gene-transcription factor interactions	Yes	Transcriptomic	P/A calls	LP	<i>E. coli</i> and <i>M. tuberculosis</i> [117]	[117]
MBA	The functional output model comprises all user-defined, high-probability reactions, a maximum number of medium-probability reactions, and a minimal number of reactions. This minimal but consistent output model is compiled based on confidence values assigned to each reaction. The confidence values are based on the frequency of appearance of reactions in the 1000 candidate models, each generated with a random pruning order	No	Literature-based knowledge, transcriptomic, proteomic, metabolomic, and phenotypic data	High- and medium-probability reactions	MILP	Human heart liver [68], cancer model [14, 91]	[68]

<p><b>FASTCORE</b></p>	<p>Searches for a consistent flux subnetwork that contains all user-defined core reactions and a minimal set of additional reactions</p>	<p>Flux consistency</p>	<p>Literature-based knowledge, multi-omics</p>	<p>Core set of reactions</p>	<p>LP</p>	<p>Liver and macrophage [76]</p>	<p>[76]</p>
<p><b>INIT</b></p>	<p>Finds a subnetwork by maximizing the sum of evidence scores and provides a connected and functional model. All included reactions should be able to carry flux. In addition, the production of specified metabolites by the output model is ensured</p>	<p>Flux consistency</p>	<p>Tailored for the use of HPA (proteomic), transcriptomic data, observed metabolites</p>	<p>Scores for high, medium, low, and absent proteins</p>	<p>MILP</p>	<p>69 normal and 16 cancer cell types [47]</p>	<p>[47]</p>
<p><b>tINIT</b></p>	<p>In contrast to the preceding version (INIT), which delivered a connected and consistent network, tINIT generates functional networks based on user-defined, cell type-specific set of metabolic functions. The algorithm defines the reaction set necessary for the realization of the specified metabolic tasks; in case the resulting model misses a task in the test phase, gap filling is applied to ensure the functionality of the output model. In addition, the output model has only irreversible reactions. Compared to INIT, it is optional whether net production of metabolites is allowed</p>	<p>Metabolic tasks</p>	<p>Proteomic</p>	<p>Personalized HCC models [49]</p>	<p>MILP</p>	<p>Personalized HCC models [49]</p>	<p>[49]</p>
<p><b>mCADRE</b></p>	<p>Generates a subnetwork by removing reactions that are not part of the high-confidence core reaction set, which is defined based on gene expression data and connectivity clues, while preserving the flux capacity of core reactions and defined metabolic functionality</p>	<p>Flux capacity of core reactions and metabolic functions</p>	<p>Transcriptomic</p>	<p>A high-confidence core reaction set based on expression evidence</p>	<p>LP</p>	<p>126 tumor and normal tissue and cell types [73]</p>	<p>[73]</p>

models have been reconstructed for many human tissues and cell types. Most of these networks have been generated using omics datasets, particularly transcriptomic and proteomic data (Table 1) [35, 65]. The reconstructed cell types include the brain [37], heart [66], cardiomyocyte [67], liver [45, 68], kidney [69], macrophage [30, 70], red blood cell [71], and enterocyte [13]. Among these models, the enterocyte has been entirely reconstructed in a bottom-up process without considering omics datasets [13]. Moreover, reconstruction efforts have generated high numbers of metabolic cell line models of normal and cancer tissues [47, 72, 73]. Together, these cell type models have been applied to interrogate the metabolic aspects of diverse human disease conditions, such as cancer [47], neurodegeneration [37], and diabetes [26]. Thus, a great variety of metabolic models are readily available for the investigation of diverse biomedical topics.

Constraint-based modeling has further been used to investigate interactions between different cell types. A computational model comprising different cell types is available for human brain cells [37]. A multi-tissue model connecting adipocyte, hepatocyte, and myocyte has been applied to study whole-body systems physiology [26]. Host–pathogen [30] and host–gut microbe [41, 42] interactions have also been investigated. These multicell models contribute further to the scope of biomedical applications.

### **5.1 Methods for Network Contextualization Based on Transcriptomic and Proteomic Data**

A number of algorithms have been published for the integration of omics datasets with genome-scale reconstructions (Table 1) and summarized [74, 75]. These methods emphasize the integration of transcriptomic and proteomic data but differ in whether they depend on a defined objective function, whether exchange profiles need to be predefined for the target cell type or are predicted as part of the output model [35, 65], and the form in which input data are compiled and incorporated.

Algorithms, such as GIMME [35] and iMAT [65], have been designed to support completely automated subnetwork generation (Table 1) despite the general noisiness of transcriptomic data and posttranscriptional and posttranslational regulatory impacts (Fig. 1) [35, 65]. These algorithms predict posttranscriptional or posttranslational regulation based on the model context. In the case of GIMME, a functional model is obtained with regard to a stated objective function [35]. More recent algorithms, such as MBA [68] and FASTCORE [76], allow the manual definition of high-confidence reaction sets around which the functional model is built. This development emphasizes that manual work and biological insight in addition to the datasets are required to ensure the quality and biological relevance of the generated models [68, 76].

Methods that depend on a user-defined threshold to distinguish reaction activity from inactivity (incorporated through gene–protein–reaction associations) can be applied even if only data from

one condition are available. Other methods require data from multiple conditions (Table 1). The selection of the algorithms depends on the dataset at hand.

It is important to distinguish cell type- and condition-specific models. Cell type-specific models should be able to perform the entire range of metabolic functions regardless of a particular environmental or genetic condition, and they can only be built using a compilation of datasets measuring the expression of cellular components under many different conditions. Cell type-specific models require substantial manual curation based on the literature to ensure cell type-specific functionality [5] because these models should capture all metabolic functions of the target cells. Consequently, cell type-specific models can be further tailored to create condition-specific cell type models. Condition-specific models capture only a subset of cell type-specific metabolic functions, namely, those active under a particular environmental condition. Based on their limited scope, these subnetworks can be built from a single (e.g., extracellular metabolomic or transcriptomic) dataset, either starting from the global human metabolic model or from a cell type-specific metabolic model [18]. Thus, cell type- and condition-specific models differ in scope.

## **5.2 Integration of Metabolomic Datasets**

Metabolomic datasets have become increasingly comprehensive. The metabolome comprises low-molecular-weight molecules and is measured by profiling metabolites in biofluids or in extracts (intracellular) or supernatants (extracellular) of human cell cultures [77]. Metabolomics is the youngest of the omics disciplines. These measurements are stable, relatively inexpensive, and highly reproducible [77]. In addition, metabolomics defines the actual cellular phenotype, in contrast to transcriptomic or proteomic data [77], which makes it the most straightforward resource for integration with metabolic networks. These datasets can also be obtained easily from patients using noninvasive methods. Together, these factors make metabolomics particularly interesting for biomedical applications.

A number of the algorithms introduced above consider the use of metabolomic data (Table 1). For mCADRE, metabolomic data are discussed as potential clues and to define metabolic functions that are checked for during network pruning [73]. (t)INIT allows the inclusion of metabolomic data as clues for model building and to enable the final model to produce detected metabolites [47, 49]. In addition, MBA and FASTCORE depend on the a priori definition of core reaction sets, which can include the consideration of metabolomic data [68, 76]. Gene Inactivation Moderated by Metabolism, Metabolomics, and Expression (GIM(3)E) has recently become available. GIM(3)E enforces minimum turnover of detected metabolites [78]. Metabolomic data constitute one source of evidence during automated, top-down cell type-specific

model generation. Metabolomic data can be integrated with metabolic networks as qualitative, quantitative, and thermodynamic constraints [18, 64, 79–81] to increase the accuracy of model predictions [63].

The integration of metabolomic data was also used to drive metabolic discovery. Recon 1, in combination with urine metabolomic profiles and transcriptomic data, was used to predict novel putative endogenous substrates of the OAT1 transporter [82]. Target substrates were identified by comparing predictions from models constrained based on metabolomic or transcriptomic data from wild-type and mOAT1 knockout mice. Intermediates of the polyamine pathway were subsequently experimentally confirmed as putative substrates of mOAT1 [82]. Recon 1, flux balance analysis, and published metabolite uptake and secretion rates [52] were further used to support findings derived from liquid chromatography–mass spectrometry-based isotope tracer studies and a metabolic flux model and congruently highlighted oxidative phosphorylation as a major contributor to ATP production (on average 84 % across the NCI-60 cell lines) in cancer cells [83]. This example illustrates how metabolic knowledge can be expanded by the integration of metabolomic data into the network context.

Many recent research endeavors generate multiple datasets. Consequently, approaches are needed that allow the integration of different datasets into one coherent picture [18]. An example using a multi-omics approach including metabolomic data was published by Cakir et al. [84]. The authors used a small set of metabolites to constrain a yeast metabolic model, which they subsequently used to identify reporter reactions associated with changes in metabolite levels as a consequence of environmental or genetic perturbations. The reporter reactions were subsequently related to transcriptomic data to infer different forms of regulation [84].

Furthermore, an integrative analysis of semiquantitative extracellular metabolomic profiles and transcriptomic data was performed for two lymphoblastic leukemia cell lines (CCRF-CEM and Molt-4) [18]. This approach added to earlier studies that inferred internal metabolic states from extracellular metabolomic data [64, 85, 86]. Differences in metabolite exchange evoked differences in the metabolic phenotypes of the two cell line models, which were predicted using sampling analysis. The metabolic CCRF-CEM model displayed a more glycolytic phenotype, whereas the Molt-4 model exhibited an oxidative phenotype. These computational predictions were supported by experimental data. Moreover, differential gene expression and alternative splicing analysis revealed a high incidence of regulated genes of rate-limiting and commitment steps in central metabolic pathways, thereby supporting the predicted metabolic differences between the two cell lines [18]. In summary, the context of the metabolic model can be used to analyze metabolomic data alone as well as congruently with other omics datasets.

---

## 6 Biomedical Applications

The number of biomedical applications has expanded continuously since the human genome-scale reconstruction was published. This development is closely linked to the growing number of omics datasets. In the following sections, some of these applications will be discussed.

### **6.1 Using COBRA to Investigate Cancer Metabolism**

Metabolism has been recognized as an important aspect of cancer [87]. Consequently, the metabolic model constitutes an ideal framework for the investigation of cancer. Since 2010, a growing number of constraint-based modeling studies have investigated the high utilization of glycolysis connected to lactate secretion under aerobic conditions (i.e., the Warburg effect) and various other aspects of cancer metabolism [14, 47, 88–93]. These studies have been reviewed extensively [23, 36, 94–96].

Metabolic models have been applied to identify drug targets [14]. One study used a small cancer metabolic model that captured most of the experimentally studied pathways related to cancer, namely, glycolysis, citric acid cycle, pentose phosphate pathway, glutaminolysis, and oxidative phosphorylation. This model represented the physiological conditions measured in HeLa cells and predicted lactate dehydrogenase and pyruvate dehydrogenase as candidate metabolic drug targets [88].

The first genome-scale model of cancer metabolism was derived from Recon 1 [24] using a version of the MBA approach [14, 68]. The same algorithm was further applied to generate a non-small cell lung cancer model using multiple gene expression datasets, which demonstrated predictive superiority for cell line-specific, growth-supporting genes compared to a generic cancer model [14]. The generic cancer model was used to predict synthetic lethal gene pairs as potential drug targets, of which a subset was nontoxic to Recon 1. Succinate dehydrogenase and fumarate hydratase, both frequently mutated in different cancers, were both predicted to be synthetically lethal with pyruvate carboxylase. Pyruvate carboxylase was validated to specifically target succinate dehydrogenase and fumarate hydratase-deficient cancer cells [14]. In a follow-up study, lethal synergy between fumarate hydratase and enzymes of the heme pathway was experimentally validated to provide insight into the unresolved mechanism by which fumarate hydratase-deficient cells, e.g., in renal cell cancer (HLRCC), survive with a defective TCA cycle [91]. Thus, both small-scale and genome-scale metabolic cancer models have been used to predict therapeutic intervention strategies.

Moreover, metabolic models have been used to investigate the causes of the metabolic alterations observed in cancer cells [97]. The role of solvent capacity constraints, i.e., the upper limit of mitochondrial density in the cytoplasm, in the aberrant metabo-

lism of cancer cells has been investigated. The application of solvent capacity constraints evoked a glucose uptake-dependent dichotomy of metabolic states in a reduced flux balance model of ATP production. This dichotomy consisted of a switch from oxidative phosphorylation to aerobic glycolysis [97], thus supporting the induction of Warburg metabolism through cytosolic capacity restrictions in these cells.

Another study predicted that aerobic glycolysis arises equally from solvent capacity limits in cancer and proliferating normal muscle cells in both a small-scale and a large-scale model [98]. Aerobic glycolysis provided a higher ATP yield per volume density than mitochondrial oxidative phosphorylation when assuming glucose (cancer) and/or fatty acid (muscle tissue) utilization [98]. The large-scale model even predicted glutaminolysis [98], which is well known to support cancer cell proliferation. This important role of enzyme mass restrictions at high proliferation rates and the potential emergence of the Warburg effect were consolidated by another group using Recon 1 [93]. Thus, the metabolic models supported molecular crowding as a potential contributor to metabolic alterations in cancer cells.

The metabolic specificities of cancer cells have been predicted using metabolic models. Whereas the aforementioned results were obtained by making use of normal glycolysis, further work suggested the existence of an alternative pathway for ATP production in cancer cells [90]. Metabolic flux is diverted into an alternative glycolytic pathway (with net zero ATP production) that involves reactions in the serine biosynthesis, one-carbon metabolism, and glycine cleavage system (SOG pathway) [90]. Tedeschi et al. further investigated ATP generation through the SOG pathway [89] and predicted that the SOG pathway supports cancer proliferation through ATP, NADPH, and purines [89]. First, gene expression data revealed high expression of SOG-associated genes in certain cancer cell lines that coincided with increased expression of Myc, Myc target genes, and genes associated with poor prognosis or metastasis [89]. Second, the human metabolic model was constrained by the cell line-specific proliferation rate, cell volume, DNA content, and specific metabolic fluxes derived from the NCI-60 tumor-derived cell lines [52]. Subsequently, a correlation between serine synthesis from 3-phosphoglycerate and proliferation rate was observed [89]. Reaction flux through the SOG pathway was determined to support the production of ATP, NADPH, and purines [89]. Third, SOG pathway inhibition in prostate cancer PC-3 cells caused a decrease in ATP that was not associated with decreasing flux through glycolysis or oxidative phosphorylation but through inhibition of MTHFD1 in the initial response phase (0–4 h) [89]. Thus, metabolic models predicted the use of the SOG pathway for energy generation.



Metabolic models were further applied to investigate the topology of the divergent gene expression signature observed in renal carcinoma [61]. Gatto et al. used cancer type-specific metabolic models as an estimator to confirm the reduced metabolic network of clear cell renal cell carcinoma [61]. The authors observed unique metabolic reprogramming in clear cell renal cell carcinoma based on transcriptomic and proteomic data that were not shared by any other tumor tissue [61]. The renal cancer model captured approximately 35 % fewer genes than a normal kidney model [47] and models of other cancer types reconstructed in the same manner [61]. This reduced model size was consistent with the down-regulation of genes across metabolic pathways observed in clear cell renal cell carcinoma [61]. Thus, contextualized models revealed the reduced redundancy of central metabolic genes unique to clear cell renal cell carcinoma compared to other cancers [61]. Taken together, these examples illustrate the broad range of cancer metabolism that can be studied using metabolic models.

**6.2 Using Metabolic Modeling to Investigate IEMs: Nutritional Therapeutic Approaches for Inherited Enzymopathies**

IEMs are individually rare but collectively have a relatively high incidence rate of 1:800 live births [99]. Numerous reviews have been published reviewing IEMs from various aspects, including newborn screening programs and systems biology [100–103]. Most IEMs are recessive inherited disorders. IEMs can be present at any age, from fetal life to old age. The numerous IEMs and their wide range of symptoms can be grouped into three diagnostically useful groups [104]: (1) disorders giving rise to intoxication via accumulation of intracellular compounds over time, (2) disorders involving energy metabolism, and (3) disorders involving the metabolism of complex molecules. The symptoms vary among groups as well as within groups. Not all IEMs are easily diagnosed and treatments may not be available. Disorders of the first group can be treated by changing the patient's diet [104]. Current IEM treatments can be broadly classified as (1) reducing load on the affected pathway via substrate restriction, (2) correcting product deficiency, (3) decreasing metabolite toxicity, (4) stimulating residual enzymes, and (5) enzyme replacement [104].

The use of comprehensive mathematical metabolic models for the analysis of IEMs is promising, particularly for those IEMs that cannot be diagnosed because no pattern in biofluid composition can be directly correlated to the IEM (either due to the absence of such a clear pattern or due to a small number of patients, which limits traditional statistical approaches). Computational modeling may be particularly valuable for very rare IEMs and IEMs for which insufficient data exists for diagnosis as well as for therapies. The human metabolic reconstruction was expanded to account for all metabolic pathways of the metabolites (20 amino acids and 35 acylcarnitines) routinely measured in many newborn screening programs worldwide [25]. This expanded human metabolic recon-

struction captured 235 distinct IEMs, of which 144 affect the nervous system [51]. Of these IEMs, 37 are currently treated through nutritional adjustment (e.g., phenylketonuria, OMIM: 261260) or nutritional supplementation (e.g., l-carnitine supplement for systemic carnitine deficiency, OMIM: 212140).

A recent study of a mouse metabolic network associated with the gut microbe *Bacteroides thetaiotaomicron* proposed that either the association with the microbe or nutritional supplementation of metabolites could rescue the otherwise lethal growth phenotype of mice with different IEMs [42]. Orotic aciduria (OMIM: 258900) could be rescued by uracil or pyrimidine nucleosides (uridine, deoxyuridine, cytidine, or deoxycytidine). AICA-ribosiduria (OMIM: 608688) could be rescued by purine nucleosides (guanosine, inosine, deoxyadenosine), and ribose-5-phosphate isomerase deficiency (OMIM: 608611) could be rescued by ribose. These predictions were supported by literature evidence, as uridine supplementation is an established treatment strategy for individuals with orotic aciduria [105, 106]. However, nutritional supplementation has not been reported for the other two IEMs. These initial results emphasize the prospects of applying constraint-based modeling to investigate inherited human diseases.

### **6.3 Personalized Metabolic Models**

Preliminary examples of personalized metabolic models have been generated. Jamshidi et al. analyzed differences in the serum metabolomic profiles of a single hereditary hemorrhagic telangiectasia (HHT) patient versus controls using Recon 1 [107]. Recon 1 thereby functioned as a whole-body metabolic network (including all organs), and the differences in the plasma were interpreted as net metabolite changes (uptake and secretion) mediated by cells throughout the human body [107]. The differences in the metabolic profiles were integrated through differential scaling of the coefficients of a non-growth-associated biomass objective to distinguish the HHT patient and the controls [107]. Subsequently, the authors used the flux span ratio (comparison of the difference between the minimal and the maximal flux values per reaction and between conditions). The flux span ratio was defined by flux variability analysis and used to compare the HHT patient and control models. It identified decreased energy production and increased flux potentials in nitrogen handling and disposition pathways in the HHT patients that were linked to an anti-VEGF drug (bevacizumab (Avastin)) [107]. After treatment, the HHT patient's metabolomic profile and the metabolomic profiles of the controls were more similar compared to the pretreatment HHT sample [107]. Hence, the study demonstrated the successful prediction of the metabolic mechanisms underlying HHT based on a personalized metabolic model and exploitation of those predictions to shift the metabolism of a patient toward a healthier state.

Other studies have combined multiple omics datasets to generate and analyze a metabolic adipocyte model [50]. The authors used the model to investigate metabolic alternations in adipocytes that would allow the stratification of obese patients [50]. The sets of genes that were differentially expressed between lean and obese males and females were used for model building and additionally correlated with predicted reporter metabolites [50]. Their model predictions coincided with the differential transcriptional (down) regulation of mitochondrial pathways in obese individuals. Consequently, the authors proposed to increase mitochondrial acetyl-CoA as a potential therapeutic target to decrease fat in patients [50]. In addition, plasma androsterone levels were suggested as a biomarker of metabolic alterations in these patients [50]. Personalized models and differentially expressed genes were correlated to predict intervention strategies for obese patients.

Another study from the same group predicted new potential drugs (antimetabolites) to specifically target hepatocellular carcinoma while sparing normal cells [49]. Six hepatocellular carcinoma patient-specific metabolic models, a generic hepatocellular carcinoma model, and 83 normal tissue models were generated by integrating proteomic data with the human metabolic reaction database (HMR) 2.0 and tINIT [46, 49]. Different yet overlapping sets of antimetabolites were predicted for the individual personalized cancer models as well as for the generic cancer model. Among the predicted antimetabolites was l-carnitine, which selectively inhibits growth in HepG2 cells [49]. The results illustrate that personalized models can be successfully applied to reveal individual as well as common intervention targets and to predict the response of individual patients to metabolic drugs.

#### **6.4 Toward Whole-Body Modeling**

Cells in the human body do not operate in isolation. Cellular metabolism depends on, e.g., the supply of nutrients dependent on the cells lining the intestinal tract or the exploitation of fat storage during periods of starvation. Thus, interactions between cells in the human body are an important factor to understand biomedical questions in their entirety. Bordbar et al. published a first multi-tissue model of human metabolism [26]. The authors employed the human metabolic reconstruction, Recon 1 [24], tissue-specific information from UniProt [108], and information from the literature to generate metabolic models for the liver (hepatocytes), skeletal muscle (myocytes), and fat (adipocytes). When integrating these three tissue-specific models into one body metabolic model, the authors connected them by a blood compartment to permit nutrient uptake and byproduct secretion for each tissue. Interestingly, the authors found that the inter-tissue transport of metabolites was not correctly balanced and required the addition of a bicarbonate buffer system. The authors used flux variability analysis to determine the interdependency of the three tissue-

specific models in the body metabolic context. This work represents a first milestone toward whole-body modeling but also highlights the need for appropriate, powerful computational methods to investigate metabolic dependencies and crosstalk among different cells and organs.

### **6.5 Opportunities for Modeling in Predictive Medicine**

The importance of computational modeling in biomedicine will continue to increase substantially, and the medical field will continue to develop toward the individualization of health maintenance, diagnosis, and treatment design. Constraint-based modeling has the potential to contribute to personalized medicine. Herein, we have discussed some of the biomedical success stories of metabolic modeling. However, a multitude of computational and modeling challenges remain to be addressed.

The use of metabolomic data is particularly promising for personalized health and diagnostics because these data are closest to the actual phenotype and can easily be derived from biofluids. One future goal would therefore be the use of metabolic models as tools to predict disease risk based on the integration of biofluid metabolomics. To achieve this goal, methods must be established to integrate patient-derived body fluid samples (e.g., blood, urine, interstitial fluid) into the context of metabolic reconstruction.

The identities of the cells in the body that produce and consume the metabolites detected in human body fluids are not known. Changes in biofluid metabolic profiles observed as a consequence of disease may be caused by metabolic changes in individual cells or tissues or by changes in whole-body metabolism. Metabolic models could be used to understand the biochemistry behind these detected changes in body fluid metabolic profiles. The application of Recon to explain changes in the biofluid metabolome, as performed to investigate hereditary hemorrhagic telangiectasia [107], could be a good starting point. Furthermore, a comprehensive organ-resolved model would be essential for the interpretation of (multiple) biofluid metabolomic profiles.

Such prospective whole-body metabolic reconstruction would need to be benchmarked using standard criteria developed by the systems biology community [109] and available experimental data, including biofluid and tissue metabolomic data, to assess its accuracy and predictive value. Metabolomic data for large cohorts are becoming available and could be used to test, refine, and benchmark such a whole-body model (e.g., the KORA study [110, 111]).

One of the major challenges associated with a whole-body model is to define the metabolic functions of the different organs as well as those of the entire body. Different metabolic function formulations will need to be investigated, similar to the work performed for single cells [112, 113]. Such formulations will require in-depth knowledge of human physiology and comprehensive experimental organ- and cell type-resolved data.

---

## 7 Conclusions

Applications of constraint-based modeling of human metabolism are expanding. Existing studies have already provided novel insights into various human disease processes. These studies demonstrate that constraint-based modeling can contribute significantly to systems biomedicine and personalized health, particularly if progress in the integration and analysis of the various omics data types in the network context continues at its current pace.

---

## Acknowledgment

This study was supported by an ATTRACT program grant (FNR/A12/01) from the Luxembourg National Research Fund (FNR).

## References

1. Kitano H (2001) Foundations of systems biology. MIT Press, Cambridge, MA
2. Machado D, Costa R, Rocha M et al (2011) Modeling formalisms in systems biology. *AMB Express* 1:45
3. Durot M, Bourguignon PY, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev* 33:164–190
4. Palsson BØ (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, Cambridge
5. Aurich MK, Thiele I (2012) Contextualization procedure and modeling of monocyte specific TLR signaling. *PLoS One* 7:e49978
6. Li F, Thiele I, Jamshidi N, Palsson BØ (2009) Identification of potential pathway mediation targets in toll-like receptor signaling. *PLoS Comput Biol* 5:e1000292
7. Papin JA, Palsson BØ (2004) The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys J* 87:37–46
8. Thiele I, Jamshidi N, Fleming RMT et al (2009) Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol* 5:e1000312
9. Thorleifsson SG, Thiele I (2011) rBioNet: a COBRA toolbox extension for reconstructing high-quality biochemical networks. *Bioinformatics* 27:2009–2010
10. Schellenberger J, Que R, Fleming RMT et al (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 6:1290–1307
11. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 5:93–121
12. Reed JL, Famili I, Thiele I et al (2006) Towards multidimensional genome annotation. *Nat Rev Genet* 7:130–141
13. Sahoo S, Thiele I (2013) Predicting the impact of diet and enzymopathies on human small intestinal epithelial cells. *Hum Mol Genet* 22:2705–2722
14. Folger O, Jerby L, Frezza C et al (2011) Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol* 7:501
15. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nat Biotech* 28:245–248
16. Varma A, Palsson BØ (1994) Metabolic flux balancing: basic concepts, scientific and practical use. *Nat Biotech* 12:994–998
17. Terzer M, Maynard ND, Covert MW et al (2009) Genome-scale metabolic networks. *Wiley Interdiscip Rev Syst Biol Med* 1:285–297
18. Aurich M, Paglia G, Rolfsson Ó et al (2015) Prediction of intracellular metabolic states from extracellular metabolomic data. *Metabolomics* 11:603–619

19. Lewis NE, Nagarajan H, Palsson BØ (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* 10:291–305
20. Savinell JM, Palsson BØ (1992) Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *J Theor Biol* 154:421–454
21. Vo TD, Greenberg HJ, Palsson BØ (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J Biol Chem* 279:39532–39540
22. Feist AM, Palsson BØ (2010) The biomass objective function. *Curr Opin Microbiol* 1:344–349
23. Hernández Patiño CE, Jaime-Muñoz G, Resendis-Antonio O (2013) Systems biology of cancer: moving toward the integrative study of the metabolic alterations in cancer cells. *Front Physiol* 3:481
24. Duarte NC, Becker SA, Jamshidi N et al (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *PNAS* 104:1777–1782
25. Thiele I, Swainston N, Fleming RMT et al (2013) A community-driven global reconstruction of human metabolism. *Nat Biotechnol* 31:419–425
26. Bordbar A, Feist AM, Usaite-Black R et al (2011) A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst Biol* 5:180
27. Thiele I, Price ND, Vo TD et al (2005) Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. *J Biol Chem* 280:11683–11695
28. Bordel S, Agren R, Nielsen J (2010) Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput Biol* 6:e1000859
29. Lewis NE, Jamshidi N, Thiele I et al (2009) Metabolic systems biology: a constraint-based approach. In: *Encyclopedia of complexity and system science*. Chapter 329, 5535–5552, Springer, New York, ISBN 978-0-387-75888-6
30. Bordbar A, Lewis NE, Schellenberger J et al (2010) Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. *Mol Syst Biol* 6:422
31. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 5:264–276
32. Gudmundsson S, Thiele I (2010) Computationally efficient flux variability analysis. *BMC Bioinformatics* 11:489
33. Schellenberger J, Palsson BØ (2009) Use of randomized sampling for analysis of metabolic networks. *J Biol Chem* 284:5457–5461
34. Kaufman DE, Smith RL (1998) Direction choice for accelerated convergence in hit-and-run sampling. *Oper Res* 46:84–95
35. Becker SA, Palsson BØ (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol* 4:e1000082
36. Jerby L, Ruppin E (2012) Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling. *Clin Cancer Res* 18:5572–5584
37. Lewis NE, Schramm G, Bordbar A et al (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat Biotechnol* 28:1279–1285
38. Bordbar A, Palsson BØ (2012) Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *J Intern Med* 271:131–141
39. Shlomi T, Cabili MN, Ruppin E (2009) Predicting metabolic biomarkers of human inborn errors of metabolism. *Mol Syst Biol* 5:263
40. Rolfsson O, Palsson BØ, Thiele I (2011) The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC Syst Biol* 5:155
41. Heinken A, Thiele I (2015) Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. *Gut Microbes*. doi:10.1080/19490976.2015.1023494
42. Heinken A, Sahoo S, Fleming RMT et al (2013) Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes* 4:28–40
43. Ma H, Sorokin A, Mazein A et al (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* 3:135
44. Hao T, Ma HW, Zhao XM et al (2010) Compartmentalization of the Edinburgh human metabolic network. *BMC Bioinformatics* 11:393
45. Gille C, Bolling C, Hoppe A et al (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol Syst Biol* 6:411

46. Mardinoglu A, Agren R, Kampf C et al (2014) Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat Commun* 5:3083
47. Agren R, Bordel S, Mardinoglu A et al (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput Biol* 8:e1002518
48. Agren R, Liu L, Shoaie S et al (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Comput Biol* 9:e1002980
49. Agren R, Mardinoglu A, Asplund A et al (2014) Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol Syst Biol* 10:721
50. Mardinoglu A, Agren R, Kampf C et al (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol Syst Biol* 9:649
51. Sahoo S, Franzson L, Jonsson JJ et al (2012) A compendium of inborn errors of metabolism mapped onto the human metabolic network. *Mol Biosyst* 8:2545–2558
52. Jain M, Nilsson R, Sharma S et al (2012) Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science* 336:1040–1044
53. Uhlen M, Oksvold P, Fagerberg L et al (2010) Towards a knowledge-based human protein atlas. *Nat Biotech* 28:1248–1250
54. Orth JD, Palsson B (2012) Gap-filling analysis of the iJO1366 *Escherichia coli* metabolic network reconstruction for discovery of metabolic functions. *BMC Syst Biol* 6:30
55. Thiele I, Vlassis N, Fleming RMT (2014) fastGapFill: efficient gap filling in metabolic networks. *Bioinformatics* 30:2529–2531
56. Wishart DS, Knox C, Guo AC et al (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* 37: D603–D610
57. Sahoo S, Aurich MK, Jonsson JJ et al (2014) Membrane transporters in a human genome-scale metabolic knowledgebase and their implications for disease. *Front Physiol* 5:91
58. Sahoo S, Haraldsdottir HS, Fleming RMT et al (2014) Modeling the effects of commonly used drugs on human metabolism. *FEBS J* 282:297–317
59. Colijn C, Brandes A, Zucker J et al (2009) Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Comput Biol* 5:e1000489
60. Cox J, Mann M (2007) Is proteomics the new genomics? *Cell* 130:395–398
61. Gatto F, Nookaew I, Nielsen J (2014) Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. *PNAS* 111: E866–E875
62. Jamshidi N, Palsson BØ (2006) Systems biology of SNPs. *Mol Syst Biol* 2:38
63. Reed JL (2012) Shrinking the metabolic solution space using experimental datasets. *PLoS Comput Biol* 8:e1002662
64. Mo ML, Palsson BØ, Herrgard MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* 3:37
65. Shlomi T, Cabili MN, Herrgard MJ et al (2008) Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol* 26:1003–1010
66. Zhao Y, Huang J (2011) Reconstruction and analysis of human heart-specific metabolic network based on transcriptome and proteome data. *Biochem Biophys Res Commun* 415:450–454
67. Karlstadt A, Fliegner D, Kararigas G et al (2012) CardioNet: a human metabolic network suited for the study of cardiomyocyte metabolism. *BMC Syst Biol* 6:114
68. Jerby L, Shlomi T, Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol* 6:401
69. Chang RL, Xie L, Xie L et al (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput Biol* 6:e1000938
70. Bordbar A, Mo ML, Nakayasu ES et al (2012) Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation. *Mol Syst Biol* 8:558
71. Bordbar A, Jamshidi N, Palsson BØ (2011) iAB-RBC-283: a proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states. *BMC Syst Biol* 5:110
72. Yizhak K, Gaude E, Le Devedec S et al (2014) Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *eLife* 3:e03641
73. Wang Y, Eddy JA, Price ND (2012) Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst Biol* 6:153



74. Blazier AS, Papin JA (2012) Integration of expression data in genome-scale metabolic network reconstructions. *Front Physiol* 3:299
75. Shlomi T (2010) Metabolic network-based interpretation of gene expression data elucidates human cellular metabolism. *Biotechnol Genet Eng Rev* 26:281–296
76. Vlassis N, Pacheco MP, Sauter T (2014) Fast reconstruction of compact context-specific metabolic network models. *PLoS Comput Biol* 10:e1003424
77. Antonucci R, Pilloni MD, Atzori L et al (2012) Pharmaceutical research and metabolomics in the newborn. *J Matern Fetal Neonatal Med* 25:22–26
78. Schmidt BJ, Ebrahim A, Metz TO et al (2013) GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics* 29:2900–2908
79. Fleming RMT, Thiele I, Nasheuer HP (2009) Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to *Escherichia coli*. *Biophys Chem* 145:47–56
80. Yizhak K, Benyamini T, Liebermeister W et al (2010) Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* 26:i255–i260
81. Kummel A, Panke S, Heinemann M (2006) Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* 7:512
82. Ahn SY, Jamshidi N, Mo ML et al (2011) Linkage of organic anion transporter-1 to metabolic pathways through integrated “omics”-driven network and functional analysis. *J Biol Chem* 286:31522–31531
83. Fan J, Kamphorst JJ, Mathew R et al (2013) Glutamine-driven oxidative phosphorylation is a major ATP source in transformed mammalian cells in both normoxia and hypoxia. *Mol Syst Biol* 9:712
84. Cakir T, Patil KR, Onsan Z et al (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. *Mol Syst Biol* 2:50
85. Allen J, Davey HM, Broadhurst D et al (2004) Discrimination of modes of action of antifungal substances by use of metabolic footprinting. *Appl Environ Microbiol* 70:6157–6165
86. Allen J, Davey HM, Broadhurst D et al (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat Biotechnol* 21:692–696
87. Warburg O (1956) On the origin of cancer cells. *Science* 123:309–314
88. Resendis-Antonio O, Checa A, Encarnacion S (2010) Modeling core metabolism in cancer cells: surveying the topology underlying the Warburg effect. *PLoS One* 5:e12383
89. Tedeschi PM, Markert EK, Gounder M et al (2013) Contribution of serine, folate and glycine metabolism to the ATP, NADPH and purine requirements of cancer cells. *Cell Death Dis* 4:e877
90. Vazquez A, Markert EK, Oltvai ZN (2011) Serine biosynthesis with one carbon catabolism and the glycine cleavage system represents a novel pathway for ATP generation. *PLoS One* 6:e25881
91. Frezza C, Zheng L, Folger O et al (2011) Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature* 477:225–228
92. Jerby L, Wolf L, Denkert C et al (2012) Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer. *Cancer Res* 72:5712–5720
93. Shlomi T, Benyamini T, Gottlieb E et al (2011) Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect. *PLoS Comput Biol* 7:e1002018
94. Bordbar A, Monk JM, King ZA et al (2014) Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* 15:107–120
95. Lewis NE, Abdel-Haleem AM (2013) The evolution of genome-scale models of cancer metabolism. *Front Physiol* 4:237
96. Masoudi-Nejad A, Asgari Y (2014) Metabolic cancer biology: structural-based analysis of cancer as a metabolic disease, new sights and opportunities for disease treatment. *Semin Cancer Biol* 30C:21–29
97. Vazquez A, Liu J, Zhou Y et al (2010) Catabolic efficiency of aerobic glycolysis: the Warburg effect revisited. *BMC Syst Biol* 4:58
98. Vazquez A, Oltvai ZN (2011) Molecular crowding defines a common origin for the Warburg effect in proliferating cells and the lactate threshold in muscle physiology. *PLoS One* 6:e19538
99. Pampols T (2010) Inherited metabolic rare disease. *Adv Exp Med Biol* 686:397–431
100. Levy HL (2010) Newborn screening conditions: what we know, what we do not know, and how we will know it. *Genet Med* 12:S213–S214

101. Seymour CA, Thomason MJ, Chalmers RA et al (1997) Newborn screening for inborn errors of metabolism: a systematic review. *Health Technol Assess* 1:84–95
102. Lanpher B, Brunetti-Pierri N, Lee B (2006) Inborn errors of metabolism: the flux from Mendelian to complex diseases. *Nat Rev Genet* 7:449–460
103. Vockley J (2008) Metabolism as a complex genetic trait, a systems biology approach: implications for inborn errors of metabolism and clinical diseases. *J Inherit Metab Dis* 31:619–629
104. Fernandes J (2006) Inborn metabolic diseases: diagnosis and treatment, 4th edn. Springer, Heidelberg
105. Becroft DM, Phillips LI (1965) Hereditary orotic aciduria and megaloblastic anaemia: a second case, with response to uridine. *Br Med J* 1:547–552
106. Becroft DM, Phillips LI, Simmonds A (1969) Hereditary orotic aciduria: long-term therapy with uridine and a trial of uracil. *J Pediatr* 75:885–891
107. Jamshidi N, Miller FJ, Mandel J et al (2011) Individualized therapy of HHT driven by network analysis of metabolomic profiles. *BMC Syst Biol* 5:200
108. Bairoch A, Apweiler R, Wu CH et al (2005) The universal protein resource (UniProt). *Nucleic Acids Res* 33:D154–D159
109. Thiele I, Palsson BØ (2010) Reconstruction annotation jamborees: a community approach to systems biology. *Mol Syst Biol* 6:361
110. Suhre K, Wallaschofski H, Raffler J et al (2011) A genome-wide association study of metabolic traits in human urine. *Nat Genet* 43:565–569
111. Krug S, Kastenmuller G, Stuckler F et al (2012) The dynamic range of the human metabolome revealed by challenges. *FASEB J* 26:2607–2619
112. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol* 3:1–15
113. Gianchandani EP, Oberhardt MA, Burgard AP et al (2008) Predicting biological system objectives de novo from internal state measurements. *BMC Bioinformatics* 9:43
114. Price ND, Schellenberger J, Palsson BØ (2004) Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. *Biophys J* 87: 2172–2186
115. Akesson M, Forster J, Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. *Metab Eng* 6:285–293
116. Zur H, Ruppin E, Shlomi T (2010) iMAT: an integrative metabolic analysis tool. *Bioinformatics* 26:3140–3142
117. Chandrasekaran S, Price ND (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *PNAS* 107: 17845–17850

# Chapter 13

## From Systems Understanding to Personalized Medicine: Lessons and Recommendations Based on a Multidisciplinary and Translational Analysis of COPD

Josep Roca, Isaac Cano, David Gomez-Cabrero, and Jesper Tegnér

### Abstract

Systems medicine, using and adapting methods and approaches as developed within systems biology, promises to be essential in ongoing efforts of realizing and implementing personalized medicine in clinical practice and research. Here we review and critically assess these opportunities and challenges using our work on COPD as a case study. We find that there are significant unresolved biomedical challenges in how to unravel complex multifactorial components in disease initiation and progression producing different clinical phenotypes. Yet, while such a systems understanding of COPD is necessary, there are other auxiliary challenges that need to be addressed in concert with a systems analysis of COPD. These include information and communication technology (ICT)-related issues such as data harmonization, systematic handling of knowledge, computational modeling, and importantly their translation and support of clinical practice. For example, clinical decision-support systems need a seamless integration with new models and knowledge as systems analysis of COPD continues to develop. Our experience with clinical implementation of systems medicine targeting COPD highlights the need for a change of management including design of appropriate business models and adoption of ICT providing and supporting organizational interoperability among professional teams across healthcare tiers, working around the patient. In conclusion, in our hands the scope and efforts of systems medicine need to concurrently consider these aspects of clinical implementation, which inherently drives the selection of the most relevant and urgent issues and methods that need further development in a systems analysis of disease.

**Key words** Clinical decision support, Integrated care, Comorbidity, Disease modeling, Knowledge management

---

### 1 Introduction

The ultimate aim of personalized medicine [1] is to design and deliver healthcare interventions adjusted to the needs of the individual patient. In practice, this translates into the process followed to establish an individual longitudinal health plan with well-identified objectives for each patient. Such an approach aims at fostering optimization of health outcomes, preventing both useless and/or harmful effects provoked by some medical interventions,

and enhancing healthcare value generation and cost containment provided that care delivery is done in the appropriate setting. It should be highlighted that generalization of the practice of personalized medicine remains a vision still far away from the characteristics of current healthcare practice.

A systems approach to health, understood as a holistic analysis of health determinants including multilevel integration of information and data analytics using computational modeling, constitutes a fundamental requirement to pave the way toward personalized medicine. But, systems medicine represents only a key methodological orientation needed to achieve a medical practice based on the principles that define personalized medicine as one of the components of a 4P medicine strategy (personalized, predictive, preventive, and participatory) [1]. Here we review challenges and opportunities within the area of systems medicine as well as issues related to the implementation of personalized medicine in the clinic based on a systems medicine approach.

### **1.1 Drivers of the Changing Landscape of Medicine and Clinical Practice**

The interplay between three driving forces pushing toward a radical change in the health paradigm are major epidemiological changes [2], the urgent need for increasing healthcare efficiencies to ensure sustainability of current health systems [3–5], and a novel approach to practice based on a network medicine analysis facilitating an understanding of disease mechanisms for different subgroups of patients within and between comorbid diseases [6–8]. Overall, these three driving forces are significantly contributing to shape the concept of personalized medicine, as well as to the design of strategies to make that concept progressively a reality in specific medical areas.

Over the last years, the still-increasing epidemics of noncommunicable diseases (NCDs) [2] has been the principal triggering factor for a profound reshaping of the way we approach delivery of care for chronic patients [9, 10]. This has been so, mainly because of the interplay of two main factors: population aging and unhealthy lifestyles [2] leading to a high burden worldwide on both healthcare and societal aspects. Major disorders responsible for such a burden are cardiovascular conditions; cancer; chronic respiratory diseases, such as chronic obstructive pulmonary disease (COPD); type II diabetes mellitus; and mental illnesses [5].

Integrated care, following the chronic care model [9–11], is widely accepted as a conceptual approach to profoundly redesign future health systems to face the challenge generated by NCDs and to pave the way for personalized medicine for chronic patients. In this new scenario, conventional disease-oriented approaches, centered on the management of clinical episodes, are being and ought to be replaced by articulation of novel patient-centered integrated care services. Such a transition has proven successful in areas wherein one organization is providing care [12–14], but extensive

deployment of integrated care services in settings with heterogeneous healthcare providers remains a challenge [11].

The three major barriers for adoption of integrated care [11] are (1) change of management, (2) implementation of appropriate business models, and (3) adoption of information and communication technologies (ICT) providing organizational interoperability among professional teams across healthcare tiers, working around the patient. Different ongoing initiatives aiming at enforcing the transition toward adoption of the novel healthcare model, such as the EIP-AHA (European Innovation Partnership on Active and Healthy Ageing) [9] as well as the program currently being shaped by the European Institute of Technology for Health (EIT-Health), are generating and disseminating specific proposals to foster extensive deployment of integrated care.

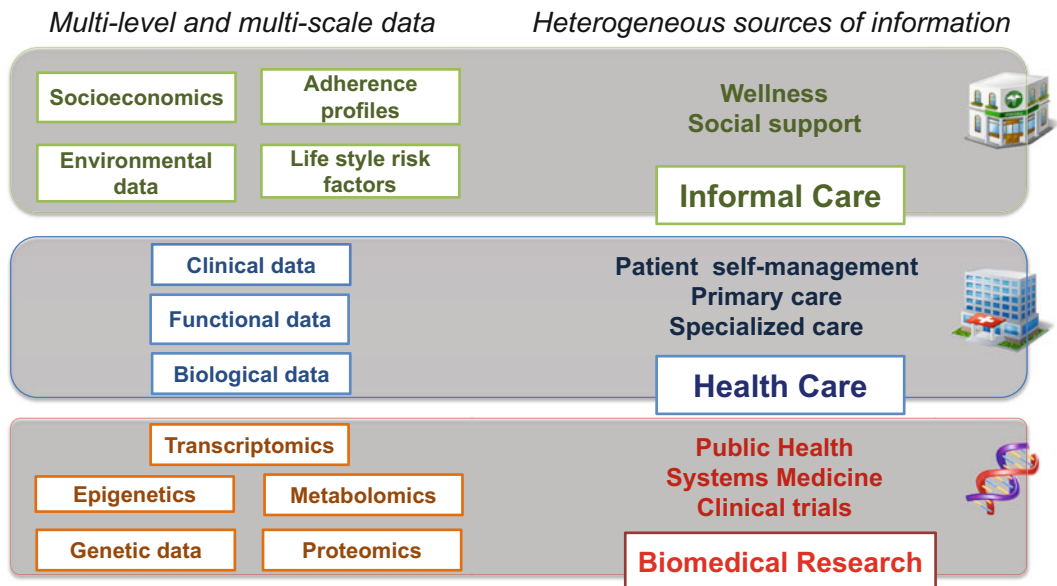
It is currently well accepted that extensive deployment of ICT-supported integrated care services may contribute to enhance health outcomes without increasing overall costs of the health system. One of the main factors generating healthcare efficiencies is individualized health risk prediction and stratification fostering delivery of care in the most appropriate setting. Targeting stratification and thereby improving individualized predictions is therefore not only a major challenge in realizing personalized medicine but is also significant opportunity using new techniques for multi-dimensional data collection and analysis [15]. Cost savings are partly achieved by the transfer of service complexities from specialized to primary care. Moreover, it is generally hypothesized that the generation of health efficiencies can be markedly boosted by promoting a more active role of both citizens and patients allowing implementation of novel cost-effective preventive strategies aiming at modulating disease progress.

As alluded above, adoption of proper strategies for patient's health risk assessment and stratification constitute a key element for large-scale deployment of integrated care. However, current stratification tools [4] rely on population-based analyses [16–18] of past use of healthcare resources. They are useful to support interventions and/or to define health policies at group level, but show limitations for clinical applicability at patient level. It is of note that these predictive tools have proven potentially useful for case-finding purposes, that is, for detection of citizens showing high-risk occurrence of major undesired health events such as unplanned hospitalizations, fast functional decline, and/or risk of death. However, scalability of the existing predicting tools for case finding and their transferability into the clinical area still requires further developments. It is acknowledged that this level of stratification is currently not sufficiently sophisticated, and it is very far from the concept of personalized medicine. Current changes in the landscape of risk assessment are driven largely by the convergence of two trends: (1) phenomenal advances in molecular and

systems biology leading to a progressively mature network medicine [9, 19–21] and (2) ICT-supported integrated care facilitating novel scenarios for data analytics, including longitudinal analyses combining biological and nonbiological phenomena [22]. These two drivers are prompting adoption of the emerging methods in systems medicine as tools to inform risk assessment and decision-making in the clinical arena that, ultimately, should contribute to shape personalized medicine to its full extent.

**1.2 Convergent Strategies Toward Personalized Medicine**

The adoption of the novel health paradigm involves convergence between adoption of the integrated care approach [10, 11] and a holistic (systems medicine) orientation aiming at generating knowledge on different dimensions (scales) in space and time influencing disease that cannot be achieved otherwise. It is currently accepted that only a small proportion of disease susceptibility (~10%) is explained by genetic variants identified to date [23]. As it has been noted [23], moving forward requires greater awareness and inclusion of what is referred to as the exposome paradigm. The concept of exposome, which consists of all of the internal and external exposures an individual incurs over a lifetime, is dynamic and variable and changes with age. The concept offers an expansive view of environmental exposures over the life course and is likely to contribute to clarify disease etiology and mechanisms. Efforts should be made to combine information from different levels (Fig. 1) in order to identify possible causal pathways and opportunities



**Fig. 1** Integration of heterogeneous multilevel and multi-scale data is needed to encompass all dimensions modulating patient health status. To this end, communication among informal care, healthcare, and biomedical research constitutes a key functional requirement that was addressed in the Synergy-COPD project through the concept of Digital Health Framework (DHF)

for intervention. Consequently, additional research is needed to clarify biological, as well as nonbiological, mechanisms and their causality in exposure-disease associations. It has become increasingly evident that such a program requires new methods and approaches enabling integrative bioinformatics to bridge between these different levels of description, ranging from molecules to clinical phenotypes [24]. Moreover, to dissect causal relationships operating at different scales, it is necessary to use different types of mathematical modeling facilitating an analysis of the causal effect of different types of interventions [15]. Ongoing collaborative efforts to decode the human epigenome [25] are likely to be key in combining different “omics” levels, as well as clinical data, with information about environmental exposures, behavioral profiles, and socioeconomic traits that individuals incur over a lifetime. Recently, emerging evidence suggests that the effect of environmental and lifestyle-related factors is mediated through the epigenome [26].

In this regard, the epigenome may serve as the bridge for traditional healthcare delivery (i.e., formal care) and informal care (e.g., patient self-management, wellness programs, social care) through adoption of citizen’s (patient’s) personal health records as management tools. In this new scenario, the appropriate articulation of patient gateways and mobile devices, also known as mHealth [27], is promising to empower for the first time an efficient channel enhancing accessibility to the health system, facilitating monitoring, and including patient’s behavioral and environmental factors into health management. The ultimate goal of patient gateways is to support cost-effective preventive interventions to modulate the evolution of the disease, which might represent tremendous sources of efficiencies if in place.

### **1.3 COPD as an Instructive Use Case**

Chronic obstructive pulmonary disease (COPD) is a prevalent chronic respiratory disease that is currently the fourth leading cause of mortality [2]. It is caused by inhalation of irritants, mainly tobacco smoking, in susceptible patients. However, only approximately 15–20% of all tobacco smokers are prone to develop the disease, and there is marked individual variability of both clinical manifestations and COPD progression [28–30] with relevant implications in terms of health risk assessment and patient management [31]. Moreover, COPD patients can also show systemic effects of the disease [31, 32] and comorbid conditions [33]. Highly prevalent chronic conditions such as cardiovascular disorders (CVD) and type 2 diabetes mellitus (T2DM)-metabolic syndrome (MS) and anxiety-depression often occur as a comorbidity cluster in COPD patients [31, 34]. Likewise, the risk of lung cancer is increased in these patients such that it can conceptually also be considered as a comorbidity of COPD [35]. There is evidence suggesting that systemic effects of the disease and comorbidity clustering are independently associated with poor prognosis [31].



Since COPD is a highly heterogeneous disorder and that comorbidities are one of the most relevant phenomena that modulates patient prognosis, the disease constitutes an optimal use case to address complexity of chronic conditions in general. There is a strong rationale for further research on subject-specific health risk prediction and stratification aiming at enhancing cost-effective management of COPD patients. The ability to better understand heterogeneity of COPD [36] should permit the development and implementation of therapeutic strategies that are specific for subgroups of patients, as well as the development of new therapies [37]. From the strategic standpoint, the approach will likely show transferability to other complex chronic conditions.

#### **1.4 Synergy-COPD**

Synergy-COPD (2011–2014) [38] was a European Union project, within the Virtual Physiological Human 7th Framework Program, conceived to explore the potential of systems medicine to generate knowledge on underlying mechanisms of chronic obstructive pulmonary disease (COPD) heterogeneities observed in the patients both in terms of clinical manifestations and disease progression [28, 31]. A core component of the project was the transfer of the acquired knowledge into the clinical arena with a twofold purpose. Firstly, analysis of the role of a systems approach to COPD heterogeneity to enhance individual health risk assessment and stratification leading to innovative patient management strategies. The second purpose was to identify novel modalities for the interplay between healthcare and biomedical research aiming at fostering deployment of 4P medicine for patients with chronic disorders [39–41]. Ultimately, Synergy-COPD was designed to generate outcomes in three different dimensions: (1) biomedical area, (2) information and communication technologies (ICT), and (3) transfer into healthcare.

The central biomedical hypothesis of the project was that heterogeneities observed in COPD patients cannot be explained by the activity of pulmonary disease only, as suggested by an organ-centric vision of the disease [42]. Alternatively, it is hypothesized that abnormalities in co-regulation of core metabolic pathways (bioenergetics, inflammation, tissue remodeling) at systemic level seem to play a central role on both systemic effects of COPD and comorbidity clustering often seen in these patients. In this scenario, overlap among certain modules of the interactome could be expected in complex COPD patients [6]. Moreover, there is evidence that oxidative stress is a characteristic feature of the disease [43] likely playing a central causal role in complex COPD. To this end, relationships among cell oxygenation, bioenergetics, and abnormal reactive oxygen species (ROS) generation were analyzed as a relevant part of the project.

The current chapter describes how a systems-oriented research on COPD heterogeneity generated novel knowledge, not achievable

through classical methods. It indicates the elevated potential for generalization of the research findings to other prevalent chronic disorders. Moreover, it describes relevant bottlenecks encountered during the project's development as well as recommends effective strategies to overcome the barriers to pave the way for a stepwise implementation of personalized medicine for chronic patients.

---

## 2 Project Outcomes

Lessons learned during the Synergy-COPD life span are grouped in three main fields: (1) biomedical outcomes, (2) ICT-related achievements, and (3) strategies for transfer of the project results into healthcare. Under the current subheading, description of project outcomes combines analysis of well-defined achievements with identification of bottlenecks that precluded further progress during the EU project.

### 2.1 Biomedical Challenges

The biomedical rationale of the entire project was based on the results of an unbiased clustering analysis of clinically stable COPD patients, the PAC\_COPD (*phenotypic characterization and course of COPD patients*) study [44], assessed after their first hospitalization and followed up during a 5-year period. The study identified and prospectively validated three COPD subtypes: (1) *group I*, severe respiratory COPD; (2) *group II*, moderate respiratory COPD patients in whom the most distinctive trait was a dissociation between severe emphysema score together with mild to moderate airway remodeling leading and moderate airflow limitation, as expressed by forced expiratory volume during the first second (FEV<sub>1</sub>); and (3) *group III*, including COPD patients in whom the most characteristic trait was comorbidity clustering, mainly cardiovascular disorders (CVD) and type II diabetes mellitus (T2D) often accompanied by metabolic syndrome (MS). It is of note that skeletal muscle dysfunction was a transversal characteristic with patients distributed in all three PAC\_COPD groups [45]. The findings of the PAC\_COPD study prompted the need for tackling COPD heterogeneity with a systems approach and prompted the four main biomedical challenges described below.

#### 2.1.1 Abnormal Regulation of Relevant Skeletal Muscle Biological Pathways

Integrative multilevel analyses of skeletal muscle of healthy subjects and COPD patients [46] including different “omics” layers (transcriptomics, epigenetics, proteomics, and metabolomics), physiological characteristics, and clinical information generated strong evidence of abnormal regulation of muscle bioenergetics both at baseline and after the perturbation of the biological system by a standard endurance training protocol. Abnormal training-induced adaptations were observed at several different levels of the mitochondrial respiratory chain, but also in the interplay between

oxidative and glycolytic pathways, as well as in fatty acid metabolism. Network analysis of metabolic pathways indicated abnormalities in key mechanisms governing skeletal muscle bioenergetics and ribosome biogenesis [8], such as mTOR and its interplay with the insulin signaling pathway. Interestingly, the different analyses carried out in COPD patients consistently showed abnormal relationships between cytokines and tissue remodeling at baseline and after training. These results were supported by experimental animal studies in guinea pigs and mice wherein it was shown that combined effects of tobacco smoking and cellular hypoxia may generate abnormal inflammatory responses [47].

Acknowledged limitations for the multilevel analysis of the interactome in the skeletal muscle in COPD patients with and without systemic effects and in healthy subjects, studied before and after endurance training, were both the reduced sample size and the unbalanced number of subjects in each study group.

### *2.1.2 Increased Risk of Comorbid Conditions in COPD Patients*

Using 13 million health records from US Medicare [48, 49], the project identified 27 disease groups (DG) with significantly elevated risk to co-occur with COPD; in all cases, the risk increased with aging. These groups included both well-established associations like CVD or lung cancer, but also unexpected ones, like digestive track disorders, that could be interesting candidates for more focused follow-up investigations. For each DG, we constructed a comprehensive list of known associated genes from the literature, and by performing a pathway enrichment analysis, a number of pathways that are shared between different disease groups were identified, suggesting that the observed comorbidities are indeed rooted in shared molecular mechanisms. By further inspecting the characteristics of the interactome, the project was able to identify a number of genes with the potential to characterize COPD comorbidities. Ongoing analyses on potential biomarkers predicting the level of comorbidity remain to be validated in further studies.

### *2.1.3 Identification of COPD Candidates for Lung Cancer Case Finding*

An ancillary aim of the project was the analysis of group II of the PAC\_COPD study [44] using the mechanistic model of spatial pulmonary heterogeneities as described in [50] in order to generate rules for identification of this subset of COPD patients in primary care. The rationale behind this approach is that this subset of COPD patients could be a candidate for screening programs for early diagnosis of lung cancer, which is one of the priorities in respiratory medicine. The literature appears to indicate that dissociation between high emphysema score and mild airway remodeling is associated with a higher probability of developing lung cancer [35]. Unfortunately, the maturity of the modeling development did not allow completion of the analyses as initially planned.

### *2.1.4 Computational Modeling for Better Understanding Biological Mechanisms of Disease*

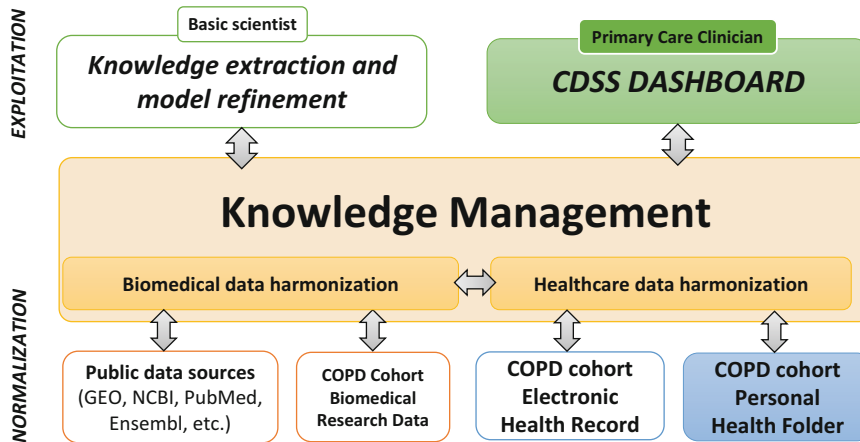
The project combined several system-based modeling approaches, probabilistic and mechanistic [51], to further explore underlying responsible mechanisms of the three biomedical areas alluded above, namely, (1) skeletal muscle dysfunction, (2) comorbidity clustering, as well as (3) group II from the PAC\_COPD study [44]. Moreover, a novel application of existing modeling techniques, Bayesian analyses [52] and Thomas network formalism [53], assessing the interplay between probabilistic and mechanistic modeling, was used aiming at expanding the potential of future systems-oriented analyses of biological phenomena, as explained in detail in [51].

Overall, both the modeling developments and the strategies adopted showed to be useful to explore the biomedical challenges of the project and to identify potential biomarkers. Moreover, the interplay between the two main modeling strategies indicated the potential of probabilistic modeling approaches to contribute to parameter refinement in mechanistic modeling (e.g., enhanced parameter estimation for mitochondrial function in the integrated model).

Overall, the biomedical results seem to support the central hypothesis of the project indicating that abnormal regulation of pivotal pathways at systemic level can contribute to both comorbidity clustering and systemic effects in COPD patients. Moreover, the analyses support a causal role for nitroso-redox disequilibrium [54] as contributor to the abnormal pathway regulations observed in COPD. Accordingly, individual susceptibility to deregulation of metabolic pathways, together with epigenetic mechanisms, may play a role modulating both systemic effects and comorbidity clustering in these patients, beyond well-known risk disease factors, such as tobacco smoking. Despite that the results from different animal experiments carried out during the project lifetime seem to provide support to our interpretations, we fully acknowledge that further validation of the current speculations is required before we move to transfer of knowledge into the clinical scenario, as discussed in the following sections of the current chapter.

## **2.2 ICT Challenges: Digital Health Framework**

The formulation of the concept, detailed characteristics [41], and road map for deployment of a Digital Health Framework (DHF) was one of the most relevant achievements of the project (Fig. 2). The DHF aims to embrace the emerging requirements—data and tools—for applying systems medicine into healthcare with a three-tier strategy articulating formal healthcare, informal care, and biomedical research. Accordingly, it has been constructed based on three key building blocks, namely, novel integrated care services with the support of information and communication technologies, a personal health folder (PHF), and a biomedical research environment (DHF-research). Details on the functional requirements and necessary components of the DHF-research were extensively



**Fig. 2** Diagram describing the core elements of a Digital Health Framework fostering communication among: (1) informal care, (2) healthcare, and (3) biomedical research. The color-filled areas are the areas prioritized for development in the DHF deployment road map

presented in [41]. The specifics of the building block strategy for deployment of the DHF and the steps toward adoption are analyzed during the project lifetime, and recommendations for implementation at local level have been formulated. The proposed architectural solutions and implementation steps constitute a pivotal strategy to foster and enable 4P medicine (predictive, preventive, personalized, and participatory) in practice and should provide a head start to any community and institution currently considering to implement a biomedical research platform.

### 2.2.1 Data Harmonization, Data Analytics, and Knowledge Generation

The COPD knowledge base (COPDKB) developed [55] in Synergy-COPD provides interoperability and integration between multiple data sources and tools commonly used in biomedical research. The COPDKB is based on the concept of “knowledge as network” and bridges multiple sources and scales of knowledge by abstracting commonly used concepts to communicate disease-specific knowledge into objects and their relations. Structuring explicit and implicit knowledge into these formal concepts enables the use of existing, well-defined vocabularies (e.g., GO [56], ICD10 [57]) and standards (e.g., SBML [58], HL7 [59]) to represent molecular, biochemical, and clinical processes. The COPDKB plays a key role facilitating the interplay with public datasets for omics analyses, as displayed Table 1.

The use of computation models in biomedical research poses the challenge of the integration of models at different scales as well as the mapping to corresponding clinical, physiological, or molecular data. We defined standard operating procedures for model documentation and developed a concept of orthogonal ontology use to create semantic descriptions for models, model parameters, and clinical parameters. These included standards for the definition of spatiotemporal compartments to allow ontology-based

**Table 1**  
**List of public datasets for omics analyses included in the COPDKB**

Bioassays	DrugBank	ICD-9/10	miRBase	Prosite
BIND	EMBL	ITFP	miRTarBase	Pubchem
BioGrid	Ensembl	IntAct	miRWalk	Reactome
BRENDA	Enzyme	IPI	OMIM	REBASE
CATH	EPD	KEGG	PDB	RefSeq
ChEBI	FASTA	LIGAND	PDQ clinical trials	SBML
CHEMBL	GenBank	LIPID maps	Pfam	SCOP
ChemIDplus	GEO	Medline	PharmGKB	SMART
ClinicalTrials.gov	GenPept	MEROPS	PLACE	Taxonomy
COG	GOA	MeSH	Plant-QTL	TransFac
COSMIC	Gramene	MiMI	Prints	TransPath
CTD comparative	HSSP	MINT	ProDom	Unigene
dbSNP	Human metabolome	MGI Phen	Prolink	UniProt

model-model and model-data connection. The orthogonal ontology use allows generating semantic descriptions of complex statements such as “partial arterial oxygen pressure” which are not represented in any current ontology.

A network search enables the use of interconnecting information and the generation of disease-specific subnetworks from general knowledge. Integration with a clinical decision-support system allows delivery into clinical practice.

The COPDKB is the only publicly available knowledge resource dedicated to COPD and combining genetic information with molecular, physiological, and clinical data as well as mathematical modeling. Its integrated analysis functions provide overviews about clinical trends and connections, while its semantically mapped content enables complex analysis approaches. The COPDKB is freely available after registration at [www.copdknowledgebase.eu](http://www.copdknowledgebase.eu).

### 2.2.2 User Profiled Interfaces

Figure 2 identifies two well-defined user profiles: (1) practicing clinician and (2) scientist performing basic and/or translational research. Practicing physicians, as described below, will require clinical decision-support systems (CDSS) with an adaptive visualization interface responsible for presenting a meaningful view of all relevant patient-specific data as well as dynamic predictions and recommendations generated by the reasoning systems component of the CDSS. It is of note, however, that beyond formulation of the two basic user profiles, no further progress was done within the project life span.

**Table 2**  
**Clinical decision-support systems (CDSS) for COPD management in an integrated care scenario**

<p>1. Early diagnosis—COPD case-finding program</p> <p>The suite of CDSS supports the regional deployment of a program of early COPD diagnosis targeting citizens at risk examined in pharmacy offices and non-diagnosed patients studied in primary care. Additional objectives of the program are to ensure high-quality forced spirometry accessible across healthcare tiers, as well as prevention of overdiagnosis of COPD in the elderly (73)</p>
<p>2. Enhanced stratification of COPD patients</p> <p>It should include three families of CDSS with well-differentiated objectives: (1) enhance applicability of the 2011 GOLD Update criteria for COPD staging; (2) facilitate offline comparisons with other COPD staging criteria, namely, BODE, DOSE, ADO, etc.; and (3) enhanced patient-based health risk assessment and stratification</p>
<p>3. Community-based integrated care program</p> <p>The suite of CDSS aims at supporting different integrated care services fostering the transfer of complexity from specialized care to the community with an active role of patients. The two programs being deployed are (1) sustainability of training-induced effects and promotion of physical activity in clinically stable moderate to severe COPD and (2) management of patients under long-term oxygen therapy (LTOT). The two programs were assessed within NEXES [69], as part of the deployment of integrated care services in the health district of Hospital Clinic</p>

### **2.3 Transfer to Healthcare**

As part of the strategies for transferring novel biomedical knowledge into the clinical arena, the three families of clinical decision-support systems (Table 2) were conceived to be embedded into the clinical processes at primary care level using an ICT platform supporting integrated care services [60]. The three CDSS families displayed in Table 2 show heterogeneous degrees of deployment: (1) early COPD diagnosis is ready for deployment at regional level in Catalonia (ES) within 2015–2016; (2) enhanced COPD stratification was only formulated conceptually without real deployment so far; and (3) community-based COPD management encompasses two programs being prepared for deployment at healthcare sector level (urban area of 540,000 inhabitants in Barcelona).

#### **2.3.1 Early Diagnosis of COPD**

The program has a twofold aim: (1) achievement of high-quality spirometry in primary care and (2) COPD case-finding program in both primary care [61] and pharmacy offices [62]. It encompasses different aspects: (1) remote support to automatic assessment of quality of forced spirometry in the community including offline support of specialized professionals if needed [63], (2) standardization of forced spirometry information and accessibility to testing across levels of care and providers, and (3) enhanced communication and support to coordination between informal care (pharmacy offices) and formal care (primary care and specialists). Accomplishment of regional deployment of the program should generate the following outcomes: (1) enhanced quality of testing;



**Table 3**  
**Risk classification of COPD patients according to the 2011 Gold Update**

Risk GOLD classification	3–4 (C) High risk, less symptoms	(D) High risk, more symptoms	$\geq 2$ Risk exacerbation history
	1–2 (A) Low risk, less symptoms	(B) Low risk, more symptoms	0–1
mMRC 0–1	CAT < 10		mMRC $\geq 2$ CAT $\geq 10$

The 2011 COPD Update [28] defines four risk categories for COPD patients (A to D) depending upon: (1) *symptoms* (modified dyspnea score from the Medical Research Council, mMRC) or CAT questionnaire; (2) *spirometric classification*: GOLD I:  $FEV_1 \geq 80\%$  pred; GOLD II:  $50\% \leq FEV_1 < 80\%$  pred; GOLD III:  $30\% \leq FEV_1 < 50\%$  pred; and GOLD IV:  $FEV_1 < 30\%$  and/or  $PaO_2 < 60$  mmHg breathing  $F_iO_2$  0.21); and (3) *frequency of exacerbations per year*. Recent reports have assessed the predictive value of this classification

(2) early COPD diagnosis, (3) enhanced case management, and (4) open new avenues for early detection of patients with abnormally fast lung function decline and/or those with abnormal biological variability of testing suggesting bronchial hyperresponsiveness.

### 2.3.2 Enhanced Stratification of COPD Patients

Patient-based health risk assessment and stratification for COPD patients is an unmet need. Appropriate patient stratification including various aspects of COPD heterogeneity, namely, (1) pulmonary disease severity, (2) disease activity, (3) systemic effects, and (4) comorbidities, is still a challenge for COPD patients. The Global Initiative for Chronic Obstructive Pulmonary Disease (GOLD) (Table 3) has represented one step forward in terms of assessment of expert-based knowledge in the field, but the proposed approach based on (1) lung function impairment, (2) symptoms, and (3) exacerbations has not yet been fully validated [64]. Moreover, several composite indices of COPD severity with proven prognostic accuracy have been developed in single studies (i.e., various BODE indices: ADO, DOSE, CODEX [65–68]), but no comprehensive comparisons are available to support evidence-based strategies for patient-based stratification in COPD. A better understanding of COPD heterogeneity should permit the development and implementation of both therapeutic strategies for subgroups of patients aiming at generating cost-effective preventive interventions fostering synergies between pharmacological and non-pharmacological approaches. Yet, Synergy-COPD was not able to develop a consistent strategy to approach the problem, as discussed below.

Future developments should be likely based on CDSS combining expert-based knowledge and outcomes from patient-based risk prediction modeling taking into account holistic approaches that consider all the elements included in Fig. 1. Unfortunately, refined strategies to achieve this goal are not in place yet.

*2.3.3 Community-Based  
Integrated Care  
Management of COPD  
Patients*

Deployment experiences of integrated care services [69] developed in parallel with Synergy-COPD have demonstrated positive health outcomes together with cost containment through the transfer of healthcare complexity from specialized care to the community fostering an active and participatory role of both citizens at risk, patients and carers. In this scenario, the use of CDSS to support health professionals for chronic care management appears as an effective approach to transfer novel biomedical knowledge into healthcare. Such an approach was successfully addressed through qualitative assessment approaches in the validation work package of the project. Moreover, the parallel deployment experiences [69] carried out during the lifetime of the project identified the high potential of the personal health folder (PHF) [70] for transferring different types of nonmedical patient information, namely, lifestyles, social frailty, adherence profile, etc., into formal healthcare, as detailed in [41].

*2.3.4 CDSS Design*

Key factors that contribute to successful CDSS outcomes in terms of impact in healthcare are (1) decision support integrated into the clinical workflow, (2) decision support delivered at the time and place of decision-making, and (3) actionable recommendations [71]. Thus, one of the important aspects that should be taken into account in the design of the CDSS is the ability to interface to existing health information systems that are already used by the intended target users of the CDSS. The implementation challenge is to design a modular CDSS framework that is portable enough to be deployed in various site clinical environments and be able to enhance the day-to-day workflow of the target clinical user with minimal impact on additional overhead.

Major factors to be taken into account in the design and implementation of CDSS are (1) the need to interface to existing health information systems in place in each of the sites, (2) a modular CDSS framework that is flexible enough to be deployed in various pilot clinical environments, and, finally, (3) the capacity to enhance the day-to-day workflow of the target clinical users. CDSS should comprise three main components: (1) an adaptive visualization interface responsible for presenting a meaningful view of all relevant patient-specific data as well as dynamic predictions and recommendations generated by the reasoning systems component, (2) a reasoning system operating on clinical rules from expert knowledge-based models and health risk predictive modeling tools, and (3) a patient data exchange module that should implement one or more interoperable clinical information standards (such as HL7, EN/ISO 13606) for receiving and uploading patient-specific data to the existing health information system.

*2.3.5 Logistics for 4P  
Medicine*

The accepted limitations in terms of subject-specific predictive modeling did not preclude other relevant technological and organizational outcomes such as the described developments of

CDSS [71], as well as formulation of the Digital Health Framework (DHF) [41]. We believe that the deployment of these tools within an integrated care scenario paves the way toward predictive, preventive, participatory, and personalized (4P) medicine for these patients preventing fragmentation of care. It is important to note that the entire DHF still requires a proof-of-concept validation before considering specific strategies for its scale-up.

The transition toward a novel biomedical research scenario fostering 4P medicine has two major biomedical research goals, namely, (1) to speed up the transfer of biomedical knowledge, including novel therapies, into healthcare and (2) to generate operational feedback from healthcare and informal care into biomedical research. The last step shall produce two main benefits. Firstly, biological knowledge will be enriched with information on different dimensions of the patient (adherence profile, frailty, lifestyles, socioeconomic and environmental factors, etc.), and secondly, it will facilitate an iterative process that shall result in progressive refinement of subject-specific predictive modeling. In this regard, the interoperability among the PHF, the healthcare through ICT-supported services [60], and the novel biomedical research platform proposed in [41], within the concept of the DHF, constitute a major achievement of the project toward the consolidation of innovative biomedical research scenario that overcomes current limitations due to fragmentation of the information.

---

### 3 Discussion

The Synergy-COPD project has demonstrated that embracing a systems-oriented research targeting COPD heterogeneity generated novel knowledge, not achievable through classical methods. In the project, COPD was chosen as a use case because of the high prevalence and impact of the disease, as well as the relevance of COPD heterogeneity for subject-based health risk assessment and stratification in the clinical arena. Several of the biomedical and ICT-related challenges are in our hands generic to several other chronic conditions. Hence, COPD provides an opportunity to address these core challenges while also having an elevated potential for generalization of the research findings to other prevalent chronic disorders.

The concept of Digital Health Framework developed in the project and the road map for its implementation involves an overall strategy for the transition from current healthcare practice to a novel scenario fostering cross talk between informal care, healthcare, and systems-oriented biomedical research that shall facilitate implementation of 4P medicine for chronic patients.

#### 3.1 *Priorities Beyond Synergy-COPD*

Both the outcomes of the project and the limitations faced during the project life span are key pieces to delineate the priorities beyond Synergy-COPD, as discussed below.

### 3.1.1 *Datasets Availability to Facilitate Patient-Based Health Risk Predictive Modeling*

Despite the current exponential generation of large amounts of biomedical data of different natures, several factors associated to availability of appropriate datasets have determined two major limitations of Synergy-COPD outcomes. Firstly, the project has generated insufficient consolidation of knowledge on underlying mechanisms of systemic effects of COPD and comorbidity clustering to bring the new knowledge closer to clinical application. A second limitation is availability of data to adequately generate clinically applicable patient-based health risk predictive modeling.

The following limiting aspects were identified: (1) the fragmented nature of the available datasets, (2) insufficient context-specific information, and (3) the lack of large datasets with proper experimental designs including multilevel “omics” information and clinically driven hypotheses.

Additional problems encountered throughout the project lifetime have been: (1) insufficient harmonization of medical coding across countries and within large longitudinal datasets, (2) gaps in semantic interoperability with large variations in disease definitions and coding, (3) publicly available information biased toward well-established and expected diseases and their underlying mechanisms, (4) lack of multilevel “omics” information bridging between GWAS information and phenotypic characterization, and (5) lack of accompanying nonclinical information (environmental, lifestyle, socioeconomic factors) in biobanking data. In summary, the characteristics of the available datasets had a negative impact on the project precluding the generation of subject-specific predictive modeling. However, they also constituted a limitation to validate the explored novel modeling approaches (e.g., Bayesian analysis and Thomas formalism) that should facilitate the interplay between probabilistic and mechanistic modeling for further characterization of complex biological processes. In this regard, policies promoting data sharing are highly recommended. In addition, generation of smart strategies linking population-based health risk assessment and subject-specific predictive modeling to enhance patient stratification and to generate real progress toward predictive and personalized medicine for chronic patients is needed.

### 3.1.2 *Maturity of the Field*

Mechanistic modeling techniques have shown usefulness to characterize biological mechanisms and to provide quantitative assessment of the phenomena analyzed, but they have serious limitations to address complex biological phenomena. In contrast, network medicine approaches based on statistical models seem suitable to address complex biomedical phenomena when large amount of data are available. Moreover, high-throughput analysis shows that canonical analysis of biological pathways is too simplistic not reflecting the real complexity of interconnectedness of biological networks [72]. It is of note, however, that the high expectations generated by emerging high-throughput methods are not yet balanced by a sufficient degree of applications in the clinical field.

### 3.1.3 Societal Changes

The project clearly identified that major organizational and technological changes are required to pave the way for a credible transition toward 4P medicine. Some of the key requirements for such a transition are described in [41] within the concept of Digital Health Framework. But, cultural factors such as (1) workforce preparation, (2) evolving concepts in terms of ethical factors relative to privacy of information transfer and information sharing, and (3) development of novel business environments fulfilling the requirements of the novel scenario are relevant elements to be taken into account in the definition of strategies leading to a successful implementation of the change. It must be emphasized that the identification of the limiting elements alluded above does not define at all a negative landscape for systems-oriented research in the biomedical area. On the contrary, one of the most important outcomes of Synergy-COPD has been the identification of the challenges to be faced and the definition of innovative strategies to adequately overcome the limiting factors alluded above that should lead to unprecedented developments in the medical practice.

### 3.2 Opportunities Identified During the Project Lifetime

The profound change in the health paradigm is leading to major healthcare transformations. Overall, the emerging scenario is exceedingly favorable for the convergence between integrated care and systems medicine as an efficient way to accelerate a mature deployment of 4P medicine for chronic patients. The outcomes of the Synergy-COPD project clearly reinforce such an orientation for future developments in the field.

The acknowledgment of the complexities faced during the project lifespan delineates the need for planning a building-block strategy in future endeavors designed to achieve further progress in the area. Moreover, the concept of Digital Health Framework provides the rationale for prioritization of the ICT developments, as identified in the proposed road map for the deployment.

It is currently well accepted that the chronic care model through deployment of integrated care services supported by information and communication technologies can contribute to enhance health outcomes without increasing overall costs of the health system. Such a generation of healthcare efficiencies is partly achieved by the transfer of complexities from specialized to primary care and to the community. It is reasonable to hypothesize that the generation of health efficiencies can be markedly boosted by: (1) promoting a more active role of citizens, patients, and carers in self-management and codesign of the services and (2) fostering cost-effective preventive strategies aiming at modulating disease progress. These two strategic proposals require adoption of the novel health paradigm that involves bridging traditional healthcare delivery (i.e., formal care) and informal care (e.g., patient self-management, wellness programs, social care, etc.) through adoption of citizen's (patient's) personal health records as

management tools. In this new scenario, the correct articulation of patient gateways and mobile devices (mHealth) [27] is promising to empower, for the first time, an efficient channel enhancing accessibility to the health system, facilitating monitoring, and including patient's behavioral and environmental factors into health management. The ultimate goal of patient gateways is to support cost-effective preventive interventions to modulate the evolution of the disease, which might represent tremendous sources of efficiencies if in place.

Moreover, the analysis of the lessons learned during the Synergy-COPD project facilitates the identification of specific challenge-driven opportunities in all the areas described above. A proper prioritization of future actions following the general recommendations generated by the project should contribute to make 4P medicine for chronic patients a successful reality.

### 3.3 Conclusions

The chapter summarizes main outcomes and lessons learned from the Synergy-COPD project. The characteristics of the disease (COPD), the inherent challenges, and our actions toward mitigating the gap between research and clinical practice on the one hand and personalized medicine on the other reinforce the high potential for generalization of the results to other chronic conditions. Overall, the project showed that convergence between a systems medicine approach and integrated care may generate substantial healthcare efficiencies for the management of complex chronic patients.

---

## Acknowledgments

This work was supported by the Swedish Research Council, Stockholm County Council, Torsten Söderberg Foundation, and Karolinska Institutet. The authors thank PITES PI12/01241, Synergy-COPD (FP7, Id:270086), and Comissionat per a Universitats i Recerca de la Generalitat de Catalunya (2009SGR1308, 2009SGR911, and 2009-SGR-393).

## References

1. Auffray C, Charron D, Hood L (2010) Predictive, preventive, personalized and participatory medicine: back to the future. *Genome Med* 2(8):57
2. Murray C, Lopez A (2013) Measuring the global burden of disease. *New Engl J Med* 369:448–457
3. WHO (2002) Innovative care for chronic conditions: building blocks for action. World Health Organization (WHO/MNC/CCH/02.01), Geneva, <http://www.who.int/chp/knowledge/publications/iccreport/en/>
4. EU C (2010) Innovative approaches for chronic diseases in public health and healthcare systems. Council of the EU 3053rd Employment, social policy health and consumer affairs
5. WHO (2008) 2008–2013 Action plan for the global strategy for the prevention and control of noncommunicable diseases. <http://www.who.int/nmh/publications/9789241597418/en/>
6. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J et al (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347(6224):1257601



7. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H et al (2014) Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun* 5:4022
8. Roca J, Vargas C, Cano I, Selivanov V, Barreiro E, Maier D et al (2014) Chronic obstructive pulmonary disease heterogeneity: challenges for health risk assessment, stratification and management. *J Transl Med* 12(Suppl 2):S3
9. EIP-AHA (2014) European scaling-up strategy in active and healthy ageing. [http://ec.europa.eu/research/innovation-union/pdf/active-healthy-ageing/scaling\\_up\\_strategy.pdf](http://ec.europa.eu/research/innovation-union/pdf/active-healthy-ageing/scaling_up_strategy.pdf)
10. ACT (2015) The Advancing Care Coordination & Telehealth Deployment (ACT) Programme. <http://www.act-programme.eu/>
11. Hernandez C, Alonso A, Garcia-Aymerich J, Grimsmo A, Vontetsianos T, García Cuyàs F et al (2015) Integrated care services: lessons learned from the deployment of the NEXES project. *Int J Integr Care* 15:e006
12. Atkins D, Kupersmith J, Eisen S (2010) The veterans affairs experience: comparative effectiveness research in a large health system. *Health Affair* 29(10):1906–1912
13. McCreary L (2010) Kaiser Permanente's innovation on the front lines. *Harvard Bus Rev* 88(9):92, 4–7, 126
14. True G, Butler AE, Lamparska BG, Lempa ML, Shea JA, Asch DA et al (2013) Open access in the patient-centered medical home: lessons from the Veterans Health Administration. *J Gen Intern Med* 28(4):539–545
15. Tegner JN, Compte A, Auffray C, An G, Cedersund G, Clermont G et al (2009) Computational disease modeling – fact or fiction? *BMC Syst Biol* 3:56
16. Wharam JF, Weiner JP (2012) The promise and peril of healthcare forecasting. *Am J Manag Care* 18(3):e82–e85
17. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M et al (2011) Risk prediction models for hospital readmission: a systematic review. *JAMA* 306(15):1688–1698
18. Moharra M, Vela E et al (2015) Comparison of predictive risk modeling among 5 European regions in the ACT project. International Foundation for Integrated Care (IFIC), 25 Març, Edinburg (abstract)
19. Loscalzo J, Barabasi AL (2011) Systems biology and the future of medicine. *Wiley Interdiscip Rev Syst Biol Med* 3(6):619–627
20. Gomez-Cabrero D, Lluch-Ariet M, Tegner J, Cascante M, Miralles F, Roca J et al (2014) Synergy-COPD: a systems approach for understanding and managing chronic diseases. *J Transl Med* 12(Suppl 2):S2
21. Topol EJ (2014) Individualized medicine from prewomb to tomb. *Cell* 157(1):241–253
22. Abugessaisa I, Saevarsdottir S, Tsipras G, Lindblad S, Sandin C, Nikamo P et al (2014) Accelerating translational research by clinically driven development of an informatics platform – a case study. *PloS One* 9(9):e104382
23. Coughlin SS (2014) Toward a road map for global -omics: a primer on -omic technologies. *Am J Epidemiol* 180(12):1188–1195
24. Tegner J, Abugessaisa I (2013) Pediatric systems medicine: evaluating needs and opportunities using congenital heart block as a case study. *Pediatr Res* 73(4 Pt 2):508–513
25. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A et al (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotech* 28(10):1045–1048
26. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A et al (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotech* 31(2):142–147
27. EU C (2014) eHealth Action Plan 2012-2020. Green Paper on mobile health (“mHealth”). <https://ec.europa.eu/digital-agenda/en/news/green-paper-mobile-health-mhealth>
28. Vestbo J, Hurd SS, Agusti AG, Jones PW, Vogelmeier C, Anzueto A et al (2013) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med* 187(4):347–365
29. Agusti A, Calverley PM, Celli B, Coxson HO, Edwards LD, Lomas DA et al (2010) Characterisation of COPD heterogeneity in the ECLIPSE cohort. *Respir Res* 11:122
30. Casanova C, de Torres JP, Aguirre-Jaime A, Pinto-Plata V, Marin JM, Cordoba E et al (2011) The progression of chronic obstructive pulmonary disease is heterogeneous: the experience of the BODE cohort. *Am J Respir Crit Care Med* 184(9):1015–1021
31. Vestbo J, Agusti A, Wouters EF, Bakke P, Calverley PM, Celli B et al (2014) Should we view chronic obstructive pulmonary disease differently after ECLIPSE? A clinical perspective from the study team. *Am J Respir Crit Care Med* 189(9):1022–1030
32. Maltais F, Decramer M, Casaburi R, Barreiro E, Burelle Y, Debigare R et al (2014) An official American Thoracic Society/European Respiratory Society statement: update on limb muscle dysfunction in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 189(9):e15–e62



33. Divo M, Cote C, de Torres JP, Casanova C, Marin JM, Pinto-Plata V et al (2012) Comorbidities and risk of mortality in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 186(2):155–161
34. Van Hoyweghen I, Horstman K (2008) European practices of genetic information and insurance: lessons for the Genetic Information Nondiscrimination Act. *JAMA* 300(3):326–327
35. Celli BR, Decramer M, Wedzicha JA, Wilson KC, Agusti A, Criner GJ et al (2015) An official American Thoracic Society/European Respiratory Society statement: research questions in COPD. *Eur Respir J* 45(4):879–905
36. Rennard SI (2011) COPD heterogeneity: what this will mean in practice. *Respir Care* 56(8):1181–1187
37. Chen X, Xu X, Xiao F (2013) Heterogeneity of chronic obstructive pulmonary disease: from phenotype to genotype. *Front Med* 7(4):425–432
38. Synergy-COPD (2011/2013) Modelling and simulation environment for systems medicine: chronic obstructive pulmonary disease (COPD) as a use case. FP7-ICT-270086
39. Bousquet J, Anto J, Sterk P, Adcock I, Chung K, Roca J et al (2011) Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Med* 3(7):43
40. Hood L, Auffray C (2013) Participatory medicine: a driving force for revolutionizing health-care. *Genome Med* 5(12):110
41. Cano I, Lluh-Ariet M, Gomez-Cabrero D, Maier D, Kalko S, Cascante M et al (2014) Biomedical research in a digital health framework. *J Transl Med* 12(Suppl 2):S10
42. Barnes PJ, Celli BR (2009) Systemic manifestations and comorbidities of COPD. *Eur Respir J* 33(5):1165–1185
43. Barnes PJ (2015) Mechanisms of development of multimorbidity in the elderly. *Eur Respir J* 45(3):790–806
44. Garcia-Aymerich J, Gómez FP, Antó JM (2009) Phenotypic characterization and course of chronic obstructive pulmonary disease in the PAC-COPD study: design and methods. *Arch Bronconeumol (English Version)* 45(01):4–11
45. Garcia-Aymerich J, Gomez FP, Benet M, Ferrero E, Basagana X, Gayete A et al (2011) Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax* 66(5):430–437
46. Turan N, Kalko S, Stincone A, Clarke K, Sabah A, Howlett K et al (2011) A systems biology approach identifies molecular networks defining skeletal muscle abnormalities in chronic obstructive pulmonary disease. *PLoS Comput Biol* 7:e1002129
47. Davidsen PK, Herbert JM, Antczak P, Clarke K, Ferrer E, Peinado VI et al (2014) A systems biology approach reveals a link between systemic cytokines and skeletal muscle energy metabolism in a rodent smoking model and human COPD. *Genome Med* 6(8):59
48. Barabasi A, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12:56–68
49. Lee D, Park J, Kay K, Christakis N, Oltvai Z, Barabasi A (2008) The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci U S A* 105:9880–9885
50. Burrowes KSD, Brightling C (2014) Computational modeling of the obstructive lung diseases asthma and COPD. *J Transl Med* 12(Suppl 2):S5
51. Gomez-Cabrero DM, Cano I, Abugessaisa I, Huertas-Miguelanez M, Tenyi A, Marin de Mas I et al (2014) Systems medicine: from molecular features and models to the clinic in COPD. *J Transl Med* 12(Suppl 2):S4
52. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR (2007) A primer on learning in Bayesian networks for computational biology. *PLoS Comput Biol* 3(8):e129
53. Thomas R, Kaufman M (2001) Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behavior. *Chaos* 11(1):170–179
54. Poole D, Hirai D, Copp S, Musch T (2012) Muscle oxygen transport and utilization in heart failure: implications for exercise (in)tolerance. *Am J Physiol Heart Circ Physiol* 302:H1050–H1063
55. Cano I, Tenyi A, Schueller C, Wolff M, Huertas Miguelanez MM, Gomez-Cabrero D et al (2014) The COPD knowledge base: enabling data analysis and computational simulation in translational COPD research. *J Transl Med* 12(Suppl 2):S6
56. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1):25–29
57. WHO (1992) World Health Organisation. International statistical classification of diseases and related health problems, 10th Revision (ICD-10), Geneva
58. Hucka M, Finney A, Sauro H, Bolouri H, Doyle J, Kitano H et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19:524–531

59. Van Hentenryck K (1997) Health level seven. Shedding light on HL7's version 2.3 standard. *Healthc Inform* 14(3):74
60. Cano I, Alonso A, Hernandez C, Burgos F, Barberan-Garcia A, Roldan J et al (2015) An adaptive case management system to support integrated care services: lessons learned from the NEXES project. *J Biomed Inform* 55:11–22
61. Burgos F, Disdier C, de Santamaria EL, Galdiz B, Roger N, Rivera ML et al (2012) Telemedicine enhances quality of forced spirometry in primary care. *Eur Respir J* 39(6):1313–1318
62. Castillo D, Burgos F, Guayta R, Giner J, Soriano JB, Lozano P et al (2012) Effect of a Community Pharmacy COPD case finding program: a novel approach to reduce COPD underdiagnosis. *Congreso Nacional SEPAR, Oviedo 2011. Arch Bronconeumol* 47:91
63. Burgos F, Melia U (2014) Clinical decision support system to enhance quality control of spirometry using information and communication technologies. *JMIR Med Inform* 2(2):e29
64. Calverley P (2013) The ABCD of GOLD made clear. *Eur Respir J* 42:1163–1165
65. Puhan M, Garcia-Aymerich J, Frey M, ter Riet G, Anto J, Agustí A et al (2009) Expansion of the prognostic assessment of patients with chronic obstructive pulmonary disease: the updated BODE index and the ADO index. *Lancet* 374:704–711
66. Puhan M, Hansel N, Sobradillo P, Enright P, Lange P, Hickson D et al (2012) Large-scale international validation of the ADO index in subjects with COPD: an individual subject data analysis of 10 cohorts. *BMJ Open* 2:e002152. doi:10.1136/bmjopen-2012-002152
67. Jones R, Donaldson G, Chavannes N, Kida K, Dickson-Spillmann M, Harding S et al (2009) Derivation and validation of a composite index of severity in chronic obstructive pulmonary disease: the DOSE Index. *Am J Respir Crit Care Med* 180:1189–1195
68. Motegi T, Jones R, Ishii T, Hattori K, Kusunoki Y, Furutate R et al (2013) A comparison of three multidimensional indices of COPD severity as predictors of future exacerbations. *Int J Chron Obstruct Pulmon Dis* 8:259–271
69. Roca JGH, Grimsmo A, Meya M, Alonso A, Gorman J et al (2013) NEXES: supporting healthier and independent living for chronic patients and elderly: final report. [http://www.nexeshealth.eu/media/pdf/nexes\\_final\\_report.pdf](http://www.nexeshealth.eu/media/pdf/nexes_final_report.pdf)
70. Barberan-Garcia A, Vogiatzis I, Solberg HS, Vilaro J, Rodriguez DA, Garasen HM et al (2014) Effects and barriers to deployment of telehealth wellness programs for chronic patients across 3 European countries. *Resp Med* 108(4):628–637
71. Velickovski F, Roca J, Burgos F, Galdiz J, Nueria M, Lluch Ariet M (2014) Clinical decision support systems (CDSS) for preventive management of COPD patients. *J Transl Med* 12(Suppl 2):S9
72. Silverman EK, Loscalzo J (2012) Network medicine approaches to the genetics of complex diseases. *Discov Med* 14(75):143–152

# Chapter 14

## RNA Systems Biology for Cancer: From Diagnosis to Therapy

Raheleh Amirkhah, Ali Farazmand, Olaf Wolkenhauer, and Ulf Schmitz

### Abstract

It is due to the advances in high-throughput omics data generation that RNA species have re-entered the focus of biomedical research. International collaborate efforts, like the ENCODE and GENCODE projects, have spawned thousands of previously unknown functional non-coding RNAs (ncRNAs) with various but primarily regulatory roles. Many of these are linked to the emergence and progression of human diseases. In particular, interdisciplinary studies integrating bioinformatics, systems biology, and biotechnological approaches have successfully characterized the role of ncRNAs in different human cancers. These efforts led to the identification of a new tool-kit for cancer diagnosis, monitoring, and treatment, which is now starting to enter and impact on clinical practice. This chapter is to elaborate on the state of the art in RNA systems biology, including a review and perspective on clinical applications toward an integrative RNA systems medicine approach. The focus is on the role of ncRNAs in cancer.

**Key words** Non-coding RNA, microRNA, Integrative workflows, Bioinformatics tools, Systems biology methods, Biomarker prediction, Therapeutic target identification

---

### 1 Introduction

In this chapter, we highlight the importance of an integration of bioinformatics and system biology in handling cancer genomics and transcriptomics data. More specifically, we discuss RNA expression profiling approaches and how they can be used and analyzed to functionally characterize sets of differentially expressed ncRNAs. Computational and experimental approaches are introduced for the prediction and validation of miRNA targets, and modeling approaches used to investigate the structure and dynamics of cancer-related ncRNAs are surveyed. Furthermore, we introduce web resources developed to share data about ncRNAs in human diseases. Thereafter, miRNA-based therapeutic strategies for controlling cancer and treatment responses to cancer are discussed. Additionally, we introduce some prognostic models of cancer

progression and patient survival. Finally, we illustrate a workflow for the use in RNA systems medicine. However, we start with elaborating on the role of ncRNAs in the hallmarks of cancer.

### **1.1 Non-coding RNAs**

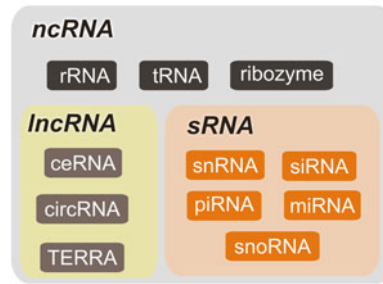
There are two conceptual roles of RNA molecules. One is linked to messenger RNAs (mRNA) which are blueprints of genetic code and are used to carry the construction plan for proteins from the nucleus to the cytoplasm, where mRNA is translated by the ribosome into amino acid chains.

The second role is regarding RNAs that do not serve as templates for proteins. These ncRNAs are functional and their stable structural fold is the basis for their biological function [1]. Most prominent representatives of this domain are ribosomal RNAs (rRNA), which are involved in protein synthesis and form a part of the ribosome, and transfer RNAs (tRNA), known as carrier of amino acids, used by the ribosome in the translation of ribonucleic acid messages (mRNAs) into peptides composed of amino acids. The discovery of another class of functional RNAs, the catalytic RNAs (a.k.a. ribonucleic acid enzymes or ribozymes) that have been found to be involved in RNA processing reactions, such as RNA splicing, viral replication, and transfer RNA biosynthesis, stimulated the discussion surrounding the RNA world hypothesis, which suggests RNA-based organisms as an essential step in evolution [2–4]. Two other classes of functional RNAs are small nuclear RNAs (snRNA) and small nucleolar RNAs (snoRNAs), which are involved in splicing and modification of rRNA, respectively. Regarding their size, ncRNA can be categorized into two groups (1) small ncRNAs, like microRNAs (miRNAs), small interfering RNAs (siRNAs) or PIWI-interacting RNAs (piRNAs), and (2) long ncRNAs (lncRNA). siRNAs which are active molecules in RNAi-dependent silencing pathways regulate various processes, including regulation of gene expression, protection of genome integrity, and innate immune responses against viruses. In the animal germline, however, PIWI-interacting RNAs (piRNAs) are in charge of silencing mutagenic transposable elements. It was then evident that the RNA molecule domain offers many more functionalities than just the conveyance of the genetic message. However, miRNAs and lncRNAs are among the most recently discovered ncRNAs and received most attention because of their importance in the cell and in diseases such as cancer. In the two following sections, the role of miRNAs and lncRNAs in human cancers is explained.

A classification scheme for ncRNA molecules that includes the classes introduced above is provided in Fig. 1.

### **1.2 MiRNAs and Cancer**

MiRNAs are an abundant class of short single-stranded RNA molecules that can regulate the expression of genes at the post-transcriptional level in most cell-biological processes [5].



**Fig. 1** Non-coding RNA classification scheme. The scheme is composed of three boxes, each representing a ncRNA domain containing several classes or subdomains. In the case of lncRNAs, more classes can be expected to be found and characterized in the near future

The biogenesis of miRNAs is a multistep enzymatic process which starts in the nucleus where a precursor miRNA (pre-miRNA) is produced through the function of Drosha, and then in the cytoplasm where Dicer processes pre-miRNA to the mature miRNA with the ability to regulate target mRNA.

The potential of miRNA to regulate the expression of a large portion of the human genome makes miRNAs a versatile tool for a context-specific regulation of most cellular processes. Deregulation of miRNA expression, however, can lead to the emergence and also the progression of human diseases including cancer [6]. Proximity of miRNAs to chromosomal breakpoints [7] and their dysregulated expression in many malignancies link them to tumorigenesis [8, 9].

With respect to cancer, some miRNAs are classified as tumor suppressors because they inhibit expression of proto-oncogenes under normal conditions. Conversely, those that negatively regulate tumor suppressor genes are called *oncomirs*. Recently, some miRNAs have also been associated with the progression of tumors toward the formation of metastasis and are typically referred to as *metastamir*. Therefore, this class of ncRNAs forms an integral part of present and future biochemical, cell biological, biomedical, and clinical investigations [10–13].

Although a wealth of knowledge about regulatory mechanisms in concerted miRNA–target interactions already exist, the details of many regulatory scenarios remain obscure. For example, target genes being regulated by two or more miRNAs or the involvement of miRNAs in regulatory networks that include transcriptional, post-transcriptional, and post-translational regulation have yet to be understood.

### 1.3 lncRNAs and Cancer

lncRNAs are RNA transcripts ranging from 200 to 100,000 nucleotides in length, located in nuclear or cytosolic fractions and commonly defined as RNA polymerase II (RNAP II) transcripts.

LncRNAs are emerging as a new layer of regulation in the cancer paradigm. Over the past several years, multiple lines of evidence demonstrate that dysregulated lncRNA expression can be associated with various human diseases including cancer.

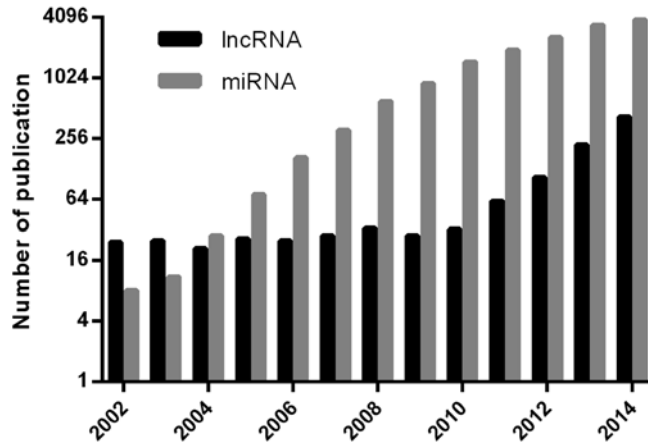
Dysregulation of lncRNAs in cancers can be due to genetic and epigenetic changes. The reason is that many lncRNAs are located in regions of the chromosome associated with chromosomal translocations, copy-number alterations, and genomic imprinting. Additionally, several studies reported that some specific signals or regulatory factors are involved in misexpression of lncRNAs. For example, Yuan and co-authors have shown that lncRNA-ATB (lncRNA-activated by TGF- $\beta$ ) triggered by TGF- $\beta$  signaling promotes the invasion-metastasis cascade in hepatocellular carcinoma (HCC) [14].

LncRNAs have been reported to regulate gene expression at both transcriptional and post-transcriptional levels by interacting with regulatory proteins [15–17], miRNAs [18–20], mRNAs, and the DNA [21]. They have also been implicated in chromatin remodeling and the integrity of subcellular compartments [22]. Misexpression of lncRNAs and consequently their malfunction as scaffold, and/or molecular guide for specific regulatory modules, and decoy for sequestering RNAs or proteins can result in cancer development [23].

Interestingly, a growing volume of literature has recently highlighted the application of lncRNAs as therapeutic targets and novel diagnostic markers [24, 25]. The use of lncRNA expression levels for diagnosis has recently been reported for several types of cancer. For example, Zheng et al. [26] reported that higher expression levels of metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) might serve as a negative prognostic marker in stage II/III colorectal cancer patients. MALAT1 has been described as a regulator of metastasis and motility, and its expression is associated with metastasis in multiple types of human malignancies. Other studies suggested a prognostic role for HOX antisense intergenic RNA (HOTAIR) in a variety of human cancers, including breast cancer, colorectal cancer, laryngeal squamous cell carcinoma, and liver cancer [27]. HOTAIR has been implicated in cancer invasion and metastasis through redirecting chromatin remodeling complexes.

Recent studies describe the role of lncRNAs, such as SPRY4-IT1, BANCR, and HOTAIR, in melanomagenesis [28, 29]. Khaitan et al. [30], for example, showed that dysregulation of SPRY4-IT1 can function as an early biomarker and key regulatory event for melanoma pathogenesis in humans.

Figure 2 illustrates the ever growing number of publications that link miRNAs and lncRNAs to cancer, focusing on the involvement in tumorigenesis, metastagenesis, and their role as biomarkers and therapeutic targets.



**Fig. 2** Growing number of publications investigating the role of miRNAs and lncRNAs in cancer

Long non-coding RNAs, as a relatively new class of transcripts, provide opportunities to identify both functional drivers and cancer-type-specific biomarkers. System biology approaches, in combination with bioinformatics tools, present a promising way to disclose the complex network of regulatory genes and their targets [31].

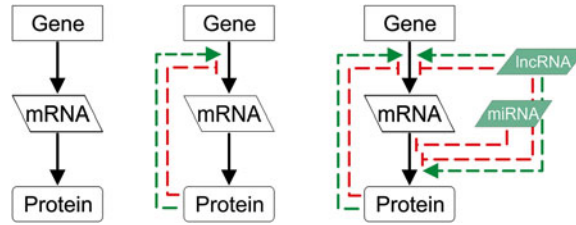
## 2 Functional RNAs from a Systems Perspective

With the discovery of post-transcriptional gene regulation by RNA interference mechanisms (e.g. miRNA and lncRNA regulation), the long-time established central dogma of molecular biology, which describes gene expression as a linear process, had to be revised (Fig. 3).

Since the critical function of miRNAs in cancer development has been discovered, bioinformatics and systems biology approaches have paved the path for advancements in the field [32]. Several computational methods have been developed to identify miRNA genes and their potential mRNA targets [33]. These approaches help with the functional categorization of target sets, and for building statistical and mathematical models which can interpret miRNA regulation in complex biological networks, e.g. those associated with cancer [32].

Accumulating evidence shows that the study of single miRNA–target interactions is not informative [12, 34]. Instead, the joint repression of mRNA targets by multiple miRNAs and the multiplicity of targets affected by single miRNAs need to be considered [35]. Or understanding the role of miRNAs in cellular functions and cancer development, one needs to consider them as part of





**Fig. 3** Evolution of the central dogma of molecular biology. Our view on gene expression has evolved from the assumption of a simple and linear sequence of events to a complex and sophisticated regulatory structure. An important piece in this puzzle is the recently discovered miRNA which can regulate gene expression at the post-transcriptional level

complex regulatory networks [36]. Mathematical modeling has played a valuable role in this [31].

Furthermore, recent studies have shown the value of an integrated bioinformatics and systems biology approach as a promising avenue toward RNA systems medicine. This approach involves (1) the integration and analysis of biomedical omics data, (2) the reconstruction and analysis of molecular interaction networks, and (3) the analysis and prediction of complex behavior and regulatory pattern emerging from biochemical networks [37].

## 2.1 Models of miRNA Regulation in Cancer

Accumulative evidence indicates that miRNAs have a critical role in controlling the hallmarks of cancer. For example, many miRNAs regulate cell proliferation pathways by direct interaction with critical regulators such as Ras, Myc, PTEN, and Cyclin-CDK complexes or cell cycle inhibitors. Bhattacharya et al., for example, showed that the miR-195 regulates cell proliferation in malignant melanoma by the cell cycle checkpoint kinase WEE1 [38].

MiRNAs can also break the balance between pro- and anti-apoptotic factors to keep tumor cell survival. In this line, Alla et al. [39] and Vera et al. [40] found that miR-205 can mediate anticancer drug resistance and facilitate E2F1 accumulation which leads to apoptosis resistance through suppression of DNp73. Bhattacharya et al. [41] also reported miR-638, as an oncogenic miRNA, which is capable of protecting melanoma cells from apoptosis and autophagy.

Some studies highlight the role of miRNAs as mediators of tumor invasion and metastasis. For instance, the miR-224/miR-452 cluster targets the metastasis suppressor TXNIP, according to Knoll and co-authors, and thereby mediates E2F1-induced EMT and invasion in malignant melanoma [42]. Additionally, there are many other studies that demonstrate the role of miRNAs in the regulation of angiogenesis, immune responses in cancer, genomic instability, and limitless replicative potential of cancer cells [43, 44].

It has been proven that miRNAs affect cancer hallmarks by regulating the activity and the stability of specific mRNAs; however, miRNAs are also embedded in complex regulatory networks that involve gene activation, post-translational regulation, and protein–protein interactions.

To better understand properties of miRNA-mediated gene regulation and the role of miRNAs in large regulatory networks or signaling pathways, it is necessary to employ a systems biology approach. In line with this, several approaches and methods have been developed for discovering miRNA–phenotype associations and for the construction and analysis of network models integrating miRNA regulation [45]. For example, Liu et al. [46] based on differentially expressed miRNAs (DEMs) in prostate cancer constructed a regulatory network by integrating information on predicted targets of these miRNAs and their interactions with other biomolecules based on the STRING (Search Tool for the Retrieval of Interacting Genes) and KEGG databases. Their results from a network and pathway enrichment analysis suggest miR-20 as an important player in the regulation of prostate cancer onset.

Modeling of gene regulatory networks has become a widely used computational approach in systems biology. However, only a few publications employed mathematical modeling for investigating the structure and dynamics of miRNA-involving networks in the context of cancer. For example, Khanin and Vinciotti [47] constructed a model based on time-series transcriptomics data that describes target gene expression dynamics based on transfection experiments with the tumor suppressor miR-124a, which is epigenetically silenced in hepatocellular carcinoma and acute lymphoblastic leukemia. A network model characterizing miR-204 as a tumor suppressor was constructed by Lee et al. [48] and was used to detect and functionally characterize 18 gene targets related to tumor progression.

Nikolov et al. [49] by using a systems biology approach analyzed the dynamics of a time delay model of the interaction between miRNA-17-92 and the transcription factors Myc and E2F and demonstrated how dynamic modeling of miRNA regulation can enhance the understanding of a specific biological process and lead to the discovery of new regulatory interactions.

In another study, Weber et al. [50] used NetGenerator, which is based on linear ordinary differential equations (ODE), to construct a dynamical regulatory model based on mRNA and miRNA time series data and prior knowledge about potential regulatory interactions between the components to discover the role of miRNAs in the chondrogenic regulatory network. Their inferred network identified a regulatory effect of miR-524-5p on the expression of the transcription factor SOX9 and the chondrogenic marker genes COL2A1, ACAN, and COL10A1.

Röhr et al. [51] modeled the individual response of four colorectal cancer patients to either miRNA-1 downregulation or miRNA-1 drug treatment using a Monte Carlo-based computational cancer model to simulate the behavior of the signaling network influenced by miRNA-1.

See Schmitz et al. [32] for a comprehensive overview about the role of miRNAs in human cancer and mathematical approaches to model miRNA-mediated tumorigenesis and progression.

In addition to miRNAs, some studies highlighted the regulatory role of lncRNAs for the hallmarks of cancer. However, there are very few published reports in which a mathematical modeling approach is used to investigate the structure and dynamical impact of lncRNAs on cancer pathways.

One example is Batagov et al. [52] which studied expression of multiple lncRNAs in a 120-h time course of differentiation of human neuroblastoma SH-SY5Y cells into neurons upon treatment with retinoic acid (RA), the compound used for the treatment of neuroblastoma. The aim of the study was to consider the role of genomic architecture in the expression dynamics of lncRNAs using differentiating human neuroblastoma cells. They could show that lncRNAs have a specific behavior in different genomic architectures by integrating genomic and transcriptomic levels of information.

## **2.2 Approaches for the Prediction and Analysis of miRNA Target Genes**

It is evident that miRNAs play significant roles in gene expression regulation in animals and plants. In order to understand their function, identification of their targets is indispensable. Today, the computational identification of miRNA targets and the experimental validation of miRNA–mRNA interactions represent crucial steps in revealing the role of miRNAs toward cellular functions. In the two following sections, computational and experimental approaches for identifying miRNA targets are explained.

### **2.2.1 miRNA Target Prediction**

The prediction of miRNA targets using computational approaches is based mainly on the fact the miRNAs exhibit imperfect or perfect sequence complementarity to their target mRNAs. Many statistical and machine learning-based prediction approaches have been developed for the identification of miRNA–target interactions.

In animal genomes, identifying miRNA targets is rather challenging. MiRNAs preferentially bind to the 3' UTR of mRNAs but target sites may also occur in the ORF or the 5' UTR. Moreover, less complementary bases between miRNAs and their targets are required for functional miRNA–target interactions, which suggests that hundreds of potential mRNA targets are post-transcriptionally controlled per miRNA. Indeed, it was shown that miRNAs induce “widespread changes in protein synthesis” [12], but in reality the number of targets per miRNA ranges between only a handful and several hundred. Therefore, the main challenge for prediction

algorithms is to reliably identify functional miRNA–target pairs. Moreover, it is desired to filter the most efficiently regulated targets for a given miRNA or the most efficient regulator for a given target.

The first generation of target prediction algorithms (rule-based) has been developed based on prior biological knowledge and experimental observations [53].

Most of these algorithms are based on similar principles, which include: (a) sequence complementarity between a miRNA and its target (with focus on the seed region), (b) thermodynamic stability of the miRNA–mRNA duplex (characterized by the hybridization energy), (c) evolutionary conservation of target sites, (d) compositional and sequence features of the target site, and (e) homologous target sites in related species.

Some of these algorithms use secondary structure prediction based on a dynamic programming approach to determine possible miRNA–mRNA duplex structures and predict duplex hybridization energies. The underlying assumption is that stable duplexes are more likely to be functional.

Machine learning-based approaches, as the next generation of algorithms, make predictions based on statistical models. These models are trained with sequence and structural patterns of validated miRNA–target pairs and/or target expression profiles from miRNA transfection/knockout experiments. Furthermore, based on selected features, miRNA target prediction algorithms apply different scoring schemes that are used to mirror the likelihood or confidence on a predicted miRNA–target pair. Although, data-driven algorithms came later to tackle the false-positive rate of rule-based methods, the problem is not sufficiently solved yet.

The reason is that many of them are biased due to a limited and homogenous set of validated miRNA–target interactions used to design or train an algorithm. Furthermore, there is no clear best performing algorithm because also the test sets used in benchmarks are often homogeneous and limited to those miRNA–target interactions that have been validated based on predictions from the early developed algorithms. Therefore, it is no surprise that, e.g. predictions from miRanda [53], one of the first algorithms proposed, largely overlap with those of more recent target prediction algorithms. In fact, miRanda predictions have the largest relative overlap with predictions from other algorithms [54]. Two reviews on miRNA–target prediction algorithms are provided by [55, 56].

### 2.2.2 Experimental Target Validation

Reliable identification of miRNA–target pairs requires support through experimental methods. One way is by transfecting cells with mature miRNAs or adenoviral vectors that induce overexpression of a certain miRNA. Subsequent mRNA expression experiments (e.g. microarray, RT-PCR, or RNAseq) may give rise to miRNA-induced target regulation, especially in case the expression

of the predicted target gene is downregulated. This approach, among others, led to the validation of many miRNA–target interactions which can now be found in databases of experimentally validated targets, e.g. TarBase [57] and miRTarBase [58].

Another widely used approach to support computationally predicted targets for specific cells or tissues is done by performing transcriptomics experiments for both miRNAs and mRNAs. Inversely correlated miRNA–target pairs may reflect miRNA-induced target repression [59–61]. Expression correlation is considered as feature in some of the recently developed data-driven target classification algorithms, as for example GenMir3, MirTarget2, and T-REX [62–64]. However, transcriptomics-based evidence does not account for any direct interaction. Reporter assays have been employed widely to determine a direct link whereby expression of a chimeric construct of luciferase reporter-3' UTR will be altered through manipulation of a regulatory miRNA. However, this technique fails to illustrate the exact miRNA binding site. This problem can be addressed by site-directed mutagenesis experiments in which miRNA binding sequences are mutated and analyzed in reporter gene assays [65]. A mutated target site prevents a miRNA from hybridizing with its target and therefore prevents target repression.

Since miRNAs can also repress translation without detectable differences in target mRNA levels [66], proteomic approaches are often employed to confirm miRNA-induced target repression at the post-translational level.

Selbach and colleagues [12] recently developed a high-throughput proteomics approach called pSILAC (pulsed stable isotope labeling with amino acids in cell culture) that directly measures genome-wide changes in protein synthesis in response to changes in miRNA expression. Another recently developed approach provides strong evidence for RISC hybridization, with a target mRNA. The HITS-Clip (high-throughput sequencing of RNAs isolated by cross-linking immunoprecipitation) method developed by [67] can be used to identify Argonaute-mRNA as well as Argonaute-miRNA hybrids through cross-linking immunoprecipitation. The sequence of the binding site and the identity of the miRNA are determined through RNAseq. This method allows the exact localization of miRNA binding sites in the target sequence.

### 2.2.3 Prediction of Target Repression Efficiency

Several approaches for the prediction of the target repression efficiency, i.e. the reduction of target protein concentration upon miRNA regulation, have been proposed. However, none have tested the performance in an unbiased way. Certain features of a predicted miRNA–target hybrid are considered to impact upon the target repression efficiency, e.g. the target site accessibility, the target site location, the number of other sites in the same target and in other targets, and the thermodynamic stability (binding energy)

of the duplex. As indicated before, one cannot say with certainty to which degree any particular feature influences target repression efficiency [68].

In our own work we have combined algorithms from bioinformatics, computational biology, 3D structural simulations, and mathematical modeling to simulate target repression efficiency as a consequence of post-transcriptional regulation by single or multiple miRNAs [69, 70].

#### 2.2.4 Functional Characterization of miRNAs

One approach to further understand gene regulation is the integration of miRNA expression profiles with mRNA profiles. Integrated analysis of miRNA and mRNA expression profiling in one condition compared to the other may give rise to miRNA function by identifying inversely correlated mRNAs and predicted targets of dysregulated miRNAs [71]. Details on miRNA–target prediction approaches are described in Subheading 2.2.2 of this chapter.

Next, functional enrichment analysis can be performed on this set of predicted and inversely correlated target genes to look for co-occurrences in, e.g. biological processes or biochemical pathways. This is done, for example, by determining associations of the predicted and/or validated miRNA targets to the gene product descriptions or ontologies (e.g. in the form of gene ontology terms). Biologically relevant molecular networks and pathways affected by targets of differentially expressed miRNAs (DEMs) can be identified by using comprehensive web resources, such as Ingenuity Pathways Analysis (IPA, [www.ingenuity.com](http://www.ingenuity.com)), Kyoto Encyclopedia of Genes and Genomes (KEGG, [www.kegg.jp](http://www.kegg.jp)), Panther ([www.pantherdb.org](http://www.pantherdb.org)), KeyMolnet ([www.immd.co.jp](http://www.immd.co.jp)), and Reactome ([www.reactome.org](http://www.reactome.org)).

Statistical methods are used in this context to test if the number of co-occurrences in a functional class exceeds the number that could be observed in a reference background [37, 72]. The Database for Annotation, Visualization and Integrated Discovery (DAVID) is an example of a widely used tool for the functional analysis of large gene lists [73].

Finally, the functionally inverse relationship between miRNAome and targetome is validated by loss-of-function or gain-of-function experiments in an in vitro and/or in vivo model [74].

### 2.3 RNA Systems Biology Web Resources

Data integration is one of the key tasks in systems medicine research to take full advantage of, e.g. omics data from related studies for further understanding the operation of complex biological systems and ultimately to develop predictive models of human diseases and optimized medical treatment of individual patients. Many cancer-related web-accessible databases are already established. Some of them provide gene expression profiles or information on somatic mutations for different cancer entities, e.g. COSMIC—Catalogue Of Somatic Mutations In Cancer ([cancer.sanger.ac.uk](http://cancer.sanger.ac.uk); [75]),

OncoPrint® ([www.oncoPrint.org](http://www.oncoPrint.org)), or TCGA—The Cancer Genome Atlas (<http://cancergenome.nih.gov>). However, in recent years a number of resources were designed that provide information on the involvement of functional RNAs (miRNAs, lncRNAs) in cancer. Most of these public databases contain carefully selected information from the scientific literature. For example, the miR2Disease database, a manually curated database that is a resource for miRNA deregulation in human pathologies, contains in its current release 349 miRNAs and 163 diseases (release April 2008; [76]) including most cancer types. miRCancer is another resource for miRNA–cancer associations, which was established using a text mining algorithm to scan the literature for relevant information in this context. The current version of miRCancer contains 878 relationships between 236 miRNAs and 79 human cancers by processing more than 26,000 published articles [77]. HMDD (the Human microRNA Disease Database) is a curated experiment-supported evidence database for human miRNA-associated diseases that collected 10,368 entries, containing 572 miRNA genes, and 378 diseases, from 3511 publications. Another resource, miRandola, is a manually curated online database that gathers all the available data regarding human circulating miRNAs and comprises as of now 132 entries, with 581 unique mature miRNAs and 21 types of samples [78]. PhenomiR is a manually generated database which provides systematic and comprehensive access to the 365 articles showing the association of dysregulated miRNAs and diseases. The database contains 632 entries with 675 unique miRNAs and 145 diseases [79].

The SomamiR DB is a database that identifies miRNA-related mutations (within miRNAs or their target sites) in cancer. From three major sources of cancer somatic mutation data repositories—Catalogue of Somatic Mutations in Cancer (COSMIC), Pediatric Cancer Genome Project (PCGP) and International Cancer Genome Consortium (ICGC)—they have collected 43 datasets and together they include 15,783 cancer somatic mutations in miRNA–target sites [80].

Recently, Chen et al. [81] presented a curated database for lncRNA-associated diseases (lncRNADisease) which contains 321 lncRNAs associated with 221 diseases from ~500 publications.

---

### 3 RNAs in Cancer Diagnosis and Prognosis

High-throughput expression profiling across different pathological conditions allows the identification of genome-wide expression changes which can be used for predicting the presence and type of tumors, and estimating its progression and the effectiveness of therapeutic treatment [24].



However, sophisticated computational and statistical methods need to be applied to analyze massive high-throughput data in order to find biomarkers for cancer progression, metastasis, or drug resistance. Another aim in this context is the computational, unsupervised reconstruction of large cancer relevant networks from gene expression data. This can be done purely data-driven or in combination with additional knowledge from online databases of biomolecule interactions [82].

### **3.1 RNA Expression Profiling**

The computational analysis of high-throughput transcriptomic data involves steps of data pre-processing (including quality control and data normalization), statistical analysis (including multiple testing correction), pattern recognition (supervised/unsupervised clustering), and the functional interpretation of the results. For details on the analysis of high-throughput gene expression experiments and biomarker identification, see the chapter on “Network-assisted disease classification and biomarker discovery” in this book.

Gene expression profiling is one of the most powerful technologies in cancer research to stratify cancer patients for guiding clinical management. However, recent studies show that miRNA expression data are more informative than mRNA expression profiles in classifying tumors. For example Lu et al. [9] using a bead-based hybridization technology demonstrated that expression data of 217 miRNAs stratified tumors better than expression pattern of 16,000 mRNAs. The reason might be that single miRNAs can regulate the expression of hundreds of mRNAs and consequently induce significant effects on gene expression networks. The small size of miRNAs makes them more stable than mRNAs; this allows profiling their expression using a variety of tissue sources, including frozen, fixed, and paraffin-embedded tissues, as well as blood and other body fluids [83].

In many diseases, miRNAs are found to be differentially expressed compared to healthy controls. Sometimes one can even observe distinct miRNA expression profiles through the course of disease progression (e.g. in [41, 84]). Notably, miRNA profiles at different stages of cancer cannot only give insight into potential biomarkers associated with tumor progression but also separate groups of patients into clinically relevant classes. The importance of further scientific investigation into miRNA expression and function can therefore not be understated.

However, the profiling of miRNA expression because of their small size, low abundance, and the developmental stage and disease state specificities is a challenging issue and not as easy as mRNA expression profiling. Plus, miRNA profiling must take into consideration the difference between precursor miRNAs and their matured form, and should also be able to discriminate between miRNAs with accuracy of as little as a single nucleotide in difference [85].

Several techniques have been adapted to identify miRNA expression profiling from a variety of biological materials, including patient samples. For instance, the bead-based flow cytometric technique is able to profile only limited subsets of miRNAs (<100 miRNAs per profile) due to the limitations of the color-decoding system on beads. The quantitative PCR (qPCR), as a popular technology for profiling miRNA levels, is a highly sensitive methodology, and excellent for microarray data validation. However, qPCR is a low-throughput technique, and can be limited by scalability if a large number of miRNAs need to be studied. Oligonucleotide miRNA microarray analysis is the most popular high-throughput technique, allowing for the genome-wide assessment of miRNA expression levels in a large number of samples [83]. MiRNA microarrays are being used to explore the biogenesis of miRNAs, differential miRNA expression between normal and abnormal states, disease characterization, and mechanism of carcinogenesis, discovery of novel disease biomarkers as well as identification of therapeutic targets [86].

In addition to microarray technologies, cancer genomics and transcriptomics have recently experienced huge help by the advent of the new massively parallel sequencing technologies, commonly referred to as NGS (Next-Generation Sequencing). Emerging next-generation sequencing (NGS) platforms and associated bioinformatics tools have also revolutionized the field of genomic research by enabling researchers to detect specific mutations through sequencing the DNA [87]. Further areas of application are transcriptomics (gene expression analysis), non-coding RNA discovery, and the identification of their target genes. However, the assembly of short-length reads generated by NGS requires sophisticated bioinformatics approaches includes the following steps: (1) pre-processing (filtering), (2) graph construction and simplification, and (3) post-processing [88]. A number of tools for the assembly of NGS reads have been developed.

### **3.2 Prognostic and Predictive Biomarkers**

The aim of achieving a personalized medicine approach, where the selection of treatment for each patient is becoming individualized, has to take advantage of a fine grained patient stratification procedure based on biomarkers. Prognostic and predictive markers also permit a molecular characterization of cancer signatures and thus deliver valuable information for a more personalized therapy, e.g. by on estimating the response to a particular treatment.

Recently, many studies have shown associations between miRNAs and various cancers and introduced miRNAs as novel class of biomarkers for cancer. Biomarkers in the form of miRNA sets which can classify samples are derived from supervised machine learning approaches such as random forests, support vector machines, Bayes classifiers, k nearest neighbors, or combinations thereof [89]. For example, Keutgen et al. [90] analyzed miRNA

expression profiles in 29 ex vivo indeterminate fine needle aspiration (FNA) samples and determined the prognostic effects of miRNA expression on final pathological diagnosis. They developed a predictive model using a SVM-RBF approach and found that the expression of 4 miRNAs (miR-222, miR-328, miR-197, and miR-21) is able to accurately differentiate malignant from benign indeterminate FNA thyroid lesions with 100 % sensitivity and 86 % specificity.

In another study, Kuo et al. [91] benefited from an inverse correlation between the expression of miRNAs and their direct mRNA targets and suggested miRNA-29a/c as potential biomarkers to predict early recurrence of CRC by combining the computational approaches and the empirical experiment. This group inferred miRNA expression profiles from mRNA expression data by bioinformatics approaches and performed computational analysis to identify miRNAs associated with CRC recurrence. Their analysis and further meta-analysis of the six mRNA expression datasets demonstrated that two miRNAs, miR-29a/c are extremely significant based on the Fisher  $p$ -value combination. The IMRE method (Imputed microRNA regulation based on weighted ranked expression and putative microRNA targets) has been applied to calculate a score representing miRNA expression level based on six mRNA expression datasets of CRC patients from the GEO and MicroCosm database (miRNA target database).

### **3.3 Prognostic Models of Cancer Progression and Patient Survival**

Over the recent years, researchers have started to develop prognostic models in order to predict clinical outcomes in people with a given disease. Prognostic models use a combination of multiple prognostic factors to calculate the individual patients' risk [92].

In a recent study, Fan et al. [93] developed a combined prognostic model of both clinical and genomic parameters using clinical-pathological variables and 323 gene expression "modules" in patients with lymph node-negative ER-positive breast cancer. They demonstrated that a combined clinical and genomics model is more accurate for outcome predictions for newly diagnosed patients with node-negative breast cancer than either clinical or genomic model alone [93].

Another study proposed a probabilistic graphical model (PGM) methodology based on factor graphs known as PARADIGM (Pathway Recognition Algorithm using Data Integration on Genomic Models), which can integrate multiple high-throughput data sets together to disclose disease subclasses and specific class expression signatures that may help to stratify patients for personalized diagnosis, prognosis, and therapy [43].

In another study, the authors proposed a multimarker prognostic model, an independent determinant of melanoma survival, generated by using a genetic algorithm on a subset of 38 candidate proteins and evaluated on a cohort of 192 primary melanomas.

The proposed prognostic assay can help clinicians to triage patients at increased risk of recurrence. This group could successfully validate the prognostic assay on an independent cohort of 246 primary melanomas [94]. Motzer et al. [95] used a stepwise modeling approach based on Cox proportional hazards regression to develop a multivariate model in order to characterize prognostic factors for survival in patients with metastatic renal-cell carcinoma (RCC). They assessed the performance of the predictive model using a two-step nonparametric bootstrapping process. They presented five prognostic factors for predicting survival and could categorize metastatic RCC patients into three risk groups that can be used in directing therapy and patient management.

---

## 4 Non-coding RNA-Based Therapy Design

The regulatory role of miRNAs in biological processes such as cell proliferation, differentiation, and apoptosis as well as cancer initiation and progression make them promising therapeutic targets for cancer management [96].

According to existing evidence, deregulated miRNAs in cancer may act as oncogenes or tumor suppressors. Based on their functions, two main therapeutic strategies can be applied to modulate miRNA activity [97]: (1) miRNA inhibition using miRNA sponges, antisense oligonucleotides, antagomirs or locked nucleic acid (LNA) constructs, which aim to inhibit miRNA functions by preventing stable binding to their targets; (2) miRNA replacement therapy using either synthetic miRNA mimics or viral expression constructs which aim to restore the levels of tumor suppressor miRNAs.

Oncogenic miRNAs (oncomiRs), usually overexpressed in cancer, downregulating tumor suppressor genes are potential therapeutic targets for inhibition. For example, transfection of anti-miR-155 oligonucleotides into pancreatic cancer could enhance apoptosis by de-repressing the synthesis of the tumor suppressor protein 53-induced nuclear protein (*TP53INP1*) [98]. By using the mentioned approaches for miRNA inhibition silencing of a single miRNA is possible, which however might not be effective enough for harnessing cancer. Recently, Lu and colleagues [99] developed a multiple-target anti-miRNA antisense oligodeoxy ribonucleotide (MTg-AMO) that can simultaneously inhibit several miRNAs.

On the other hand tumor suppressor miRNAs, which are down-regulated in cancer, are mimicked in the form of a replacement therapy. For example, systemic delivery of a miRNA mimic for miR-26a in a murine model of hepatocellular carcinoma could significantly reduce the tumor size [100]. The first synthetic miRNA mimic to enter clinical testing was MRX34, a miR-34a replacement [101]. MiR-34 is a well-known tumor suppressive miRNA which is involved in the p53 and wnt/ $\beta$ -catenin pathways.

Another area of application for miRNAs is in the alteration of the chemotherapeutic response during cancer treatment. As miRNAs can act as potent regulators of multiple key signaling pathways, they also have the ability to mediate drug resistance. This property makes miRNA replacement or silencing-based therapy a promising approach to improve the effects of, e.g. chemotherapy [83].

There is increasing evidence that miRNAs play a master role in the response of tumor cells to different cancer treatments such as chemotherapy, radiotherapy, and monotherapy. Thus, the characterization of miRNA expression alterations may help oncologists to pursue the most effective therapeutic approach for cancer treatment and improve patient classification for disease surveillance and clinical benefit.

Since for cancer patients, losing time with improper treatment can lead to an irrecoverable disease state, miRNAs that can predict responses to specific treatments offer a promising opportunity for clinicians to be able to design the most effective treatment for a patient. For example, breast cancer cells with overexpressed miR-221/222 show resistance to tamoxifen chemotherapy. Zhao et al. [102] demonstrated that knockdown of miR-221 and miR-222 in MDA-MB-468 cells restores ER $\alpha$  expression partially and sensitizes them to tamoxifen-induced cell growth arrest and apoptosis.

Some other studies used miRNA replacement therapy to reintroduce depleted miRNAs in cancer cells to reactivate cellular pathways that drive a therapeutic response. For example, using a mimic of miR-34 in pancreatic cancer cells could significantly inhibit clonogenic cell growth and invasion, induce apoptosis, and sensitize cells to chemotherapy and radiation [103]. Another study demonstrated that by applying an antagomir against miR-21 colon cancer cells are sensitized to 5-Fluorouracil (5-FU) which is a chemotherapeutic agent against cancer [104]. In addition, some other recent studies have illustrated the importance of miRNAs in response to radiotherapy. Zhang et al. [105] showed that overexpression of miR-124 could radiosensitize CRC cells by downregulating PRRX1, which is associated with EMT and cancer stem cells.

Interestingly, some recent studies assigned a crucial role to long ncRNAs (lncRNAs) in the development of chemoresistance. A prominent example is the HOX antisense intergenic RNA (HOTAIR) whose expression is linked to metastasis, drug resistance, and poor clinical prognosis [106]. Another lncRNA, urothelial cancer-associated I (UCAI), increases the cisplatin resistance during chemotherapy for bladder cancer cells by increasing the expression of Wnt6, and thus represents a potential target to overcome chemoresistance in bladder cancer [107].

#### **4.1 Identification of Therapeutic Targets**

Using bioinformatics tools, mathematical and computational methods, and integrative approaches is indispensable for understanding highly interconnected regulatory networks, and help

directing and analyzing experimental studies for the purpose of identifying putative therapeutic targets. Different bioinformatics and systems biology approaches have been used to understand the complexity of cancer-related pathways and likely mechanisms underlying disease, and to identify suitable therapeutic targets. The types of approaches used rely on the availability of pre-existing data and the type of biological questions to be answered. However, the main goal of these approaches is to identify probable drug targets *in silico* that can be used for future *in vitro* or *in vivo* analysis.

For example, Chowdhury et al. [108] used databases and the literature to reconstruct the complete hedgehog pathway which is an important target in cancer therapy to perform computational analyses of the pathway and to identify probable drug targets. They performed a structural analysis by means of a graph theoretical approach and a logical analysis using a Boolean model to derive structural and topological properties of the network as well as to identify drug target candidates. Furthermore, they established a Boolean model describing all interactions between the proteins and created instances for different cancer scenarios like glioma, colon and pancreatic cancer based on gene expression signatures to comprehend the regulations of the molecules. The model-based simulations lead to the identification of important proteins that can be used as probable drug targets for these cancers. Their results are supported by experimental findings described in the literature. Thereafter, they performed perturbation analysis to ascertain the main and minimal set of proteins that can be used as drug targets for these three types of cancers. They introduced few novel combinations of target proteins for cancer therapy and proposed GLII, which forms a “hub” node in the network, as the most important protein. *In silico* perturbations (i.e. mutation, malfunction, high or low expression, etc.) of this protein in the Hedgehog signaling network demonstrated the effect on normal network function and suggest causes for several types of cancers [79]. In another study, Rateitschak et al. [109] showed how parameter identifiability and sensitivity analysis are able to provide a basis for the prediction of potential therapeutic targets. In a case study they investigated interferon-gamma ( $\text{IFN}\gamma$ ) which has anti-proliferative effects mediated by STAT1 signaling in two cell types, stellate and cancer cells, which have a critical role in the development of pancreatic cancer. They used calibrated mathematical models to focus on the common situation in which variations between profiles of experimental time series, from different cell types, are observed. In order to find out the role of biochemical reactions in observed variations, a parameter identifiability analysis has been done. They could discover reactions which are different in pancreatic stellate and cancer cells by comparing confidence intervals of parameter value estimates and the variability of model trajectories. Afterwards, the consequences of parameter uncertainties have been considered for

the prediction of common therapeutic targets in both cell types. To this end they performed sensitivity analysis to identify those parameters, which sensitively influence protein concentrations during the time of observation [80].

An attractive therapeutic target for cancer management are miRNAs which are able to target many genes simultaneously and thereby affect different signaling pathways which are associated with a specific disease. To date, several computational studies have been used to explain the association of miRNAs with cancer and present them as promising targets for cancer treatment. However, the focus of most of them is on finding DEMs and in more advanced cases integrates DEMs with gene expression data interactions to identify regulatory modules in which the miRNA expression is found to be negatively correlated with putative targets in the cancer samples.

Recently, Wu and Chan [110] presented a novel approach to predict therapeutic miRNA targets based on their function in the human metabolic network which affects cancer growth. As metabolism plays a crucial role in cell growth, genes regulating metabolism can be used as drug targets in the treatment of cancer. To this end they simulated a condition-specific metabolic system for human hepatocellular carcinoma (HCC) wherein overexpression of each miRNA was simulated to predict their role on reducing cancer cell growth. By integrating metabolic modeling, context-specific gene expression data and miRNA–target prediction, they could predict not only the altered miRNAs in cancer but also potential miRNA targets that could change the cancer metabolic phenotype if perturbed [82].

#### **4.2 Predicting Therapy Efficiency**

Since the introduction of different treatments for cancer, huge efforts have been carried out to maximize therapy efficiency and minimize drug resistance. In fact, because of the biological heterogeneity, only patients with a specific type and grade of cancer benefit from a particular treatment [111]. However, the mechanism and the condition under which one can get an effective response to therapy still remains unclear. Recently, systems biology-based studies came to improve our understanding of complex biological systems involved in drug resistance.

For example, Zeng et al. [112] developed a new approach named as module network rewiring-analysis (MNR) to systematically study dynamical drug sensitivity and resistance to interferon therapy based on gene expression profiling of patients with hepatitis C virus infection (HCV). They demonstrated that MNR is able to predict dynamical drug sensitivities and resistances by investigating module rewiring on the module network or functional reorganization of the complex biological system during treatment.

In a retrospective study, Kanagavel et al. [113] assessed the prognostic significance of clinical factors in 126 patients with



metastatic gastric cancer, who were treated with second-line chemotherapy and could develop a prognostic model to benefit oncologists in selecting the subset of patients who would give better response to second-line chemotherapy. Their results demonstrated that patients with good performance status (PS), higher hemoglobin (Hb) level along with higher time-to-progression (TTP) under first-line therapy would show a better response to second-line chemotherapy [113]. In another study, Vera et al. [40] presented a kinetic model to study the emergence of drug resistance in tumor cells. Using the model the effect of periodic injections of a conventional genotoxic drug treatment on an *in silico* tumor cell population was simulated considering heterogeneous genetic signatures. They found that a signature with high E2F1 and low miR-205 expression levels is responsible for the resistance to genotoxic drugs and that an imbalance in the p73/DNp73 ratio mediates downregulation of miR-205. Their results suggested that the E2F1-p73/DNp73-miR-205 pathway which is tied up in regulating pro- and anti-apoptotic genes renders drug resistance [40].

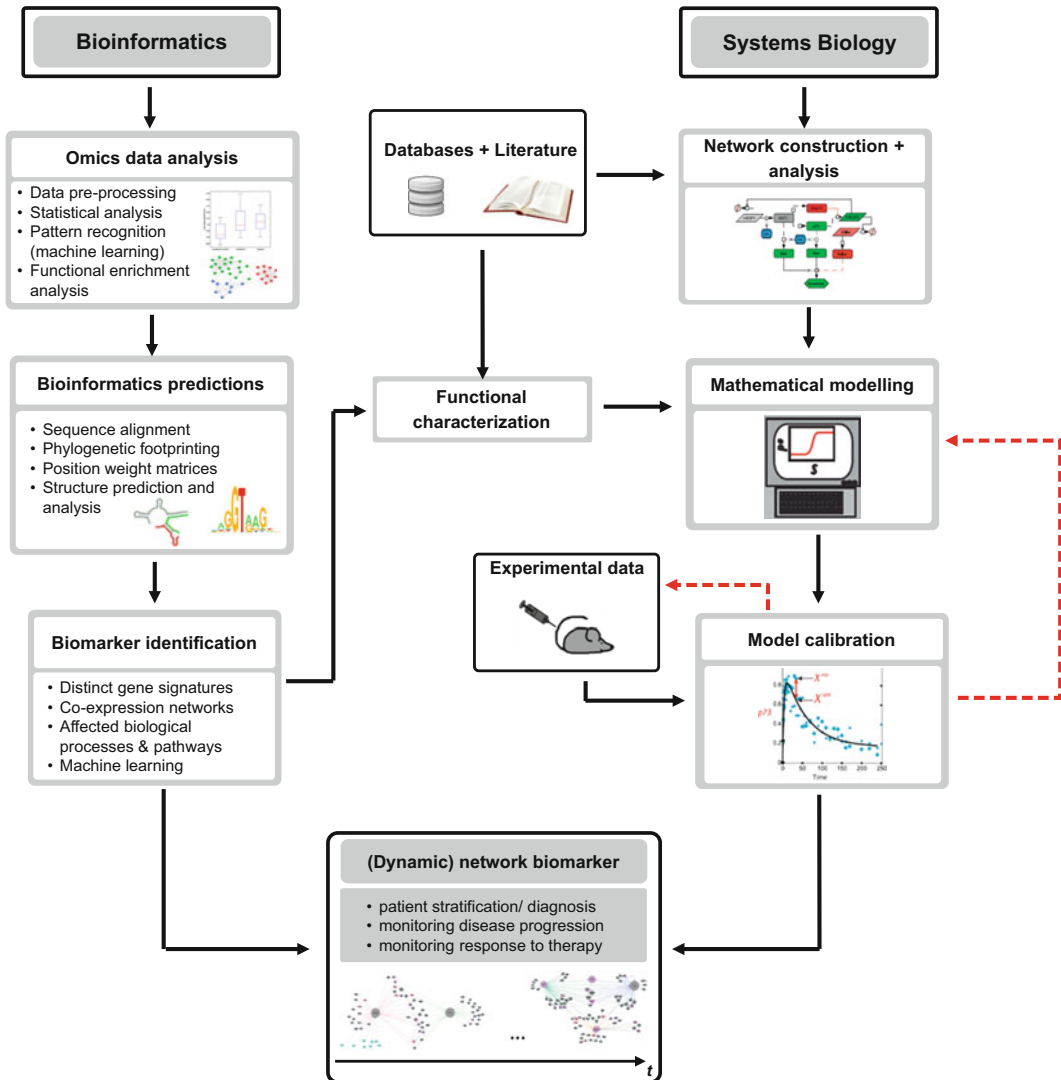
---

## 5 RNA Systems Medicine Workflows

In this chapter we discussed about how to use bioinformatics tools to analyze omics data and system biology to model and simulate biological complex networks in order to comprehensively address the existing challenges in cancer transcriptomics. Consequently, a major goal for a successful systems medicine approach relies on the integration of bioinformatics and system biology into comprehensive workflows (see an exemplary workflow in Fig. 4).

Systems medicine is a multidisciplinary approach which combines quantitative experimental data, mathematical modeling and bioinformatics tools as well as computational biology to improve our understanding of the structure and the dynamics of biological systems.

In this approach an important component is the development of mathematical models that describe disease progression mechanisms and can be used to simulate the effect of therapeutic interventions. In recent years, this approach got more and more attention by biomedical researchers because it enabled them (1) to analyze and integrate diverse types of biomedical high-throughput data obtained from *in vitro* or *in vivo* experiments and patient samples, (2) to reconstruct the large multilevel regulatory networks associated with critical disease phenotypes, (3) to analyze the dynamic behavior and regulation emerging from disease-relevant biochemical networks, and (4) to simulate signaling pathways which are influenced by, e.g. ncRNA regulation [37]. Therefore, it provides a promising way to personalize conventional therapies and find new drug targets and novel robust biomarkers.



**Fig. 4** Sketch of a possible RNA systems medicine workflow. The workflow integrates approaches from bioinformatics and systems biology to identify novel gene and network biomarkers to classify patients for diagnosis and prognosis as well as to design personalized therapies. Bioinformatics approaches are a fundamental necessity for the analysis of omics data and the prediction of functional ribonucleic acid and protein structures; while system biology provides tools to reconstruct huge regulatory networks from databases and publications and to translate them into mathematical models. These models, after parameterization, can be used to make predictive simulations, e.g., for studying cancer-regulatory mechanisms or the effects of therapeutic interventions on the system behavior

In summary, the RNA systems medicine approach integrates algorithms, methodologies, software tools, and web resources into analytical workflows in order to (1) explore and understand the molecular mechanisms underlying complex diseases, (2) enhance our understanding about biological systems through the interactions of its components, and (3) overcome shortcomings in the analysis of complex regulatory networks involving ncRNA regulation [37].

## References

1. Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 5:140
2. Gilbert W (1986) Origin of life: the RNA world. *Nature* 319(6055):618
3. Jeffares DC, Poole AM, Penny D (1998) Relics from the RNA world. *J Mol Evol* 46(1):18–36
4. Poole AM, Jeffares DC, Penny D (1998) The path from the RNA world. *J Mol Evol* 46(1):1–17
5. Lau NC et al (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294(5543):858–862
6. Soifer HS, Rossi JJ, Sætrom P (2007) MicroRNAs in disease and potential therapeutic applications. *Mol Ther* 15(12):2070–2079
7. Calin GA et al (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc Natl Acad Sci U S A* 101(9):2999–3004
8. Calin GA et al (2004) MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proc Natl Acad Sci U S A* 101(32):11755–11760
9. Amirkhah R, Schmitz U, Linnebacher M, Wolkenhauer O, Farazmand A (2015) MicroRNA-mRNA interactions in colorectal cancer and their role in tumor progression: miRNA targets in colorectal cancer. *Genes, Chromosome Canc* 54(3):129–141
10. Kozomara A, Griffiths-Jones S (2011) miR-Base: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 39(Database issue):D152–D157
11. Landgraf P et al (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 129(7):1401–1414
12. Selbach M et al (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature* 455(7209):58–63
13. Garzon R, Calin GA, Croce CM (2009) MicroRNAs in cancer. *Annu Rev Med* 60:167–179
14. Yuan JH, Yang F, Wang F, Ma JZ, Guo YJ, Tao QF, Liu F, Pan W, Wang TT, Zhou CC, Wang SB, Wang YZ, Yang Y, Yang N, Zhou WP, Yang GS et al (2014) A long noncoding RNA activated by TGF-beta promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer Cell* 25(5):666–681
15. McCann KL, Baserga SJ (2012) Long non-coding RNAs as sinks in Prader-Willi syndrome. *Mol Cell* 48(2):155–157
16. Hung T et al (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* 43(7):621–629
17. Feng J et al (2006) The Evt-2 noncoding RNA is transcribed from the Dlx-5/6 ultra-conserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev* 20(11):1470–1484
18. Poliseno L et al (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465(7301):1033–1038
19. Salmena L et al (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146(3):353–358
20. Leucci E et al (2013) microRNA-9 targets the long non-coding RNA MALAT1 for degradation in the nucleus. *Sci Rep* 3:2535
21. Li CH, Chen Y (2013) Targeting long non-coding RNAs in cancers: progress and prospects. *Int J Biochem Cell Biol* 45:1895–1910
22. Kemena C et al (2013) Using tertiary structure for the computation of highly accurate multiple RNA alignments with the SARA-Coffee package. *Bioinformatics* 29(9):1112–1119
23. Gibb EA, Brown CJ, Lam WL (2011) The functional role of long non-coding RNA in human carcinomas. *Mol Cancer* 10
24. Yang QQ, Deng YF (2014) Long non-coding RNAs as novel biomarkers and therapeutic targets in head and neck cancers. *Int J Clin Exp Pathol* 7(4):1286–1292
25. Shen Z et al (2014) Long non-coding RNA profiling in laryngeal squamous cell carcinoma and its clinical significance: potential biomarkers for LSCC. *PLoS One* 9(9):e108237
26. Zheng HT et al (2014) High expression of lncRNA MALAT1 suggests a biomarker of poor prognosis in colorectal cancer. *Int J Clin Exp Pathol* 7(6):3174–3181
27. Gupta RA et al (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464(7291):1071–1076
28. Li R, Zhang L, Jia L, Duan Y, Li Y, Bao L et al (2014) Long non-coding RNA BANCR promotes proliferation in malignant melanoma by regulating MAPK pathway activation. *PLoS*

- One 9(6):e100893. doi:10.1371/journal.pone.0100893
29. Tang L et al (2013) Long noncoding RNA HOTAIR is associated with motility, invasion, and metastatic potential of metastatic melanoma. *Biomed Res Int* 2013
  30. Khaitan D et al (2011) The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. *Cancer Res* 71(11):3852–3862
  31. Garzon R et al (2006) MicroRNA expression and function in cancer. *Trends Mol Med* 12(12):580–587
  32. Schmitz U, Wolkenhauer O, Vera J (2013) MicroRNA cancer regulation advanced concepts, bioinformatics and systems biology tools. Springer, Dordrecht
  33. Mendes ND, Freitas AT, Sagot MF (2009) Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res* 37(8):2419–2433
  34. Friedman Y, Balaga O, Linial M (2013) Working together: combinatorial regulation by microRNAs. In: Schmitz U, Wolkenhauer O, Vera J (eds) *MicroRNA cancer regulation*. Springer, Dordrecht, The Netherlands, pp 317–337
  35. Lai X, Schmitz U, Gupta SK et al (2012) Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs. *Nucleic Acids Res* 40(18):8818–8834
  36. Vera J et al (2013) MicroRNA-regulated networks: the perfect storm for classical molecular biology, the ideal scenario for systems biology. *Adv Exp Med Biol* 774:55–76
  37. Vera J, Wolkenhauer O, Schmitz U (2014) Current achievements in cancer systems biology. In: eLS. Wiley, Chichester. <http://www.els.net>
  38. Bhattacharya A et al (2012) Regulation of cell cycle checkpoint kinase WEE1 by miR-195 in malignant melanoma. *Oncogene* 32:3175–3183
  39. Alla V et al (2012) E2F1 confers anticancer drug resistance by targeting ABC transporter family members and Bcl-2 via the p73/DNp73-miR-205 circuitry. *Cell Cycle* 11(16):3067–3078
  40. Vera J et al (2013) Kinetic modeling-based detection of genetic signatures that provide chemoresistance via the E2F1-p73/DNp73-miR-205 network. *Cancer Res* 73(12):3511–3524
  41. Bhattacharya A et al (2015) miR-638 promotes melanoma metastasis and protects melanoma cells from apoptosis and autophagy. *Oncotarget* 6(5):2966–2980
  42. Knoll S et al (2014) E2F1 induces miR-224/452 expression to drive EMT through TXNIP downregulation. *EMBO Rep* 15(12):1315–1329
  43. Vaske CJ et al (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26(12):i237–i245
  44. Ruan K, Fang X, Ouyang G (2009) MicroRNAs: novel regulators in the hallmarks of human cancer. *Cancer Lett* 285(2):116–126
  45. Cascione L et al (2013) Elucidating the role of microRNAs in cancer through data mining techniques. *Adv Exp Med Biol* 774:291–315
  46. Liu DF et al (2012) MicroRNA expression profile analysis reveals diagnostic biomarker for human prostate cancer. *Asian Pac J Cancer Prev* 13(7):3313–3317
  47. Khanin R, Vinciotti V (2008) Computational modeling of post-transcriptional gene regulation by microRNAs. *J Comput Biol* 15(3):305–316
  48. Lee Y et al (2010) Network modeling identifies molecular functions targeted by miR-204 to suppress head and neck tumor metastasis. *PLoS Comput Biol* 6(4):e1000730
  49. Nikolov S et al (2011) A model-based strategy to investigate the role of microRNA regulation in cancer signalling networks. *Theory Biosci* 130(1):55–69
  50. Weber M et al (2013) Dynamic modelling of microRNA regulation during mesenchymal stem cell differentiation. *BMC Syst Biol* 7:124
  51. Röhr C et al (2013) High-throughput miRNA and mRNA sequencing of paired colorectal normal, tumor and metastasis tissues and bioinformatic modeling of miRNA-1 therapeutic applications. *PLoS One* 8(7):e67461
  52. Batagov AO et al (2013) Role of genomic architecture in the expression dynamics of long noncoding RNAs during differentiation of human neuroblastoma cells. *BMC Syst Biol* 7(Suppl 3):S11
  53. Enright AJ et al (2003) MicroRNA targets in *Drosophila*. *Genome Biol* 5(1):R1
  54. Ritchie W, Flamant S, Rasko JEJ (2009) Predicting microRNA targets and functions: traps for the unwary. *Nat Methods* 6(6):397–398
  55. Ritchie W, Rasko JEJ, Flamant S (2013) MicroRNA target prediction and validation.

- In: Schmitz U, Wolkenhauer O, Vera J (eds) *MicroRNA cancer regulation*. Springer, Dordrecht, The Netherlands, pp 39–53
56. Sethupathy P, Megraw M, Hatzigeorgiou AG (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods* 3(11):881–886
  57. Vergoulis T et al (2011) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res* 40(D1):D222–D229
  58. Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A et al (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* 42:D78–D85
  59. Lim LP et al (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433(7027):769–773
  60. Farh KK-H et al (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* 310(5755):1817–1821
  61. Huang JC et al (2007) Using expression profiling data to identify human microRNA targets. *Nat Methods* 4(12):1045–1049
  62. Wang X, El Naqa IM (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* 24(3):325–332
  63. Volinia S et al (2009) Identification of microRNA activity by Targets' Reverse EXpression. *Bioinformatics* 26(1):91–97
  64. Huang JC, Frey BJ, Morris QD (2008) Comparing sequence and expression for predicting microRNA targets using GenMiR3. *Pac Symp Biocomput* 2008:52–63
  65. Bhattacharya A, Kunz M (2013) Target identification, microRNA. In: Dubitzky W et al (eds) *Encyclopedia of systems biology*. Springer, New York, pp 2138–2142
  66. Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* 9(2):102–114
  67. Licatalosi DD et al (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456(7221):464–469
  68. Schmitz U (2013) MicroRNA target regulation. In: Dubitzky W et al (eds) *Encyclopedia of systems biology*. Springer, New York, pp 1346–1350
  69. Lai X et al (2013) A systems' biology approach to study microRNA-mediated gene regulatory networks. *Biomed Res Int* 2013
  70. Schmitz U et al (2014) Cooperative gene regulation by microRNA pairs and their identification using a computational workflow. *Nucleic Acids Res* 42(12):7539–7552
  71. Pritchard CC, Cheng HH, Tewari M (2012) MicroRNA profiling: approaches and considerations. *Nat Rev Genet* 13(5):358–369
  72. Gupta SK, Schmitz U (2011) Bioinformatics analysis of high-throughput experiments. In: Singh MP, Agrawal A, Sharma B (eds) *Recent trends in biotechnology*. Nova Science Publishers, Inc., New York, NY, pp 129–156
  73. Dennis G Jr et al (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 4(5):P3
  74. Satoh J-I (2012) Molecular network analysis of human microRNA targetome: from cancers to Alzheimer's disease. *BioData Min* 5(1):17
  75. Forbes SA et al (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43(D1):D805–D811
  76. Jiang Q et al (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 37(Database issue):D98–D104
  77. Xie B et al (2013) miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* 29(5):638–644
  78. Russo F et al (2012) miRandola: extracellular circulating microRNAs database. *PLoS One* 7(10):e47786
  79. Ruepp A, Kowarsch A, Theis F (2012) PhenomiR: microRNAs in human diseases and biological processes. *Methods Mol Biol* 822:249–260
  80. Bhattacharya A, Ziebarth JD, Cui Y (2013) SomamiR: a database for somatic mutations impacting microRNA function in cancer. *Nucleic Acids Res* 41(Database issue):D977–D982
  81. Chen G et al (2012) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* 41(D1):D983–D986
  82. Cun Y, Frohlich H (2013) Network and data integration for biomarker signature discovery via network smoothed T-statistics. *PLoS One* 8(9):e73074

83. Stahlhut C, Slack FJ (2013) MicroRNAs and the cancer phenotype: profiling, signatures and clinical implications. *Genome Med* 5(12):111
84. Ueda T et al (2010) Relation between microRNA expression and progression and prognosis of gastric cancer: a microRNA expression analysis. *Lancet Oncol* 11(2):136–146
85. Davison TS, Johnson CD, Andruss BF (2006) Analyzing micro-RNA expression using microarrays. *Methods Enzymol* 411:14–34
86. Yin JQ, Zhao RC, Morris KV (2008) Profiling microRNA expression with microarrays. *Trends Biotechnol* 26(2):70–76
87. Thomas RK et al (2007) High-throughput oncogene mutation profiling in human cancer. *Nat Genet* 39(3):347–351
88. El-Metwally S et al (2013) Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput Biol* 9(12):e1003345
89. Ren X et al (2012) A unified computational model for revealing and predicting subtle subtypes of cancers. *BMC Bioinformatics* 13:70
90. Keutgen XM et al (2012) A panel of four miRNAs accurately differentiates malignant from benign indeterminate thyroid lesions on fine needle aspiration. *Clin Cancer Res* 18(7):2032–2038
91. Kuo TY et al (2012) Computational analysis of mRNA expression profiles identifies microRNA-29a/c as predictor of colorectal cancer early recurrence. *PLoS One* 7(2):e31587
92. Steyerberg EW et al (2013) Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 10(2):e1001381
93. Fan C et al (2011) Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Med Genomics* 4:3
94. Gould Rothberg BE et al (2009) Melanoma prognostic model using tissue microarrays and genetic algorithms. *J Clin Oncol* 27(34):5772–5780
95. Motzer RJ et al (1999) Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma. *J Clin Oncol* 17(8):2530–2540
96. Shah MY, Calin GA (2014) MicroRNAs as therapeutic targets in human cancers. *Wiley Interdiscip Rev RNA* 5(4):537–548
97. Wu W (2010) MicroRNA: potential targets for the development of novel drugs? *Drugs R D* 10(1):1–8
98. Gironella M et al (2007) Tumor protein 53-induced nuclear protein 1 expression is repressed by miR-155, and its restoration inhibits pancreatic tumor development. *Proc Natl Acad Sci U S A* 104(41):16170–16175
99. Lu Y et al (2009) A single anti-microRNA antisense oligodeoxynucleotide (AMO) targeting multiple microRNAs offers an improved approach for microRNA interference. *Nucleic Acids Res* 37(3):e24
100. Kota J et al (2009) Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. *Cell* 137(6):1005–1017
101. Misso G et al (2014) Mir-34: a new weapon against cancer? *Mol Ther Nucleic Acids* 3:e194
102. Zhao JJ et al (2008) MicroRNA-221/222 negatively regulates estrogen receptor alpha and is associated with tamoxifen resistance in breast cancer. *J Biol Chem* 283(45):31079–31086
103. Ji Q et al (2009) MicroRNA miR-34 inhibits human pancreatic cancer tumor-initiating cells. *PLoS One* 4(8):e6816
104. Deng J et al (2014) Targeting miR-21 enhances the sensitivity of human colon cancer HT-29 cells to chemoradiotherapy in vitro. *Biochem Biophys Res Commun* 443(3):789–795
105. Zhang Y et al (2014) MiR-124 radiosensitizes human colorectal cancer cells by targeting PRRX1. *PLoS One* 9(4):e93917
106. Malek E, Jagannathan S, Driscoll JJ (2014) Correlation of long non-coding RNA expression with metastasis, drug resistance and clinical outcome in cancer. *Oncotarget* 5(18):8027–8038
107. Fan Y et al (2014) Long non-coding RNA UCA1 increases chemoresistance of bladder cancer cells by regulating Wnt signaling. *FEBS J* 281(7):1750–1758
108. Chowdhury S, Pradhan RN, Sarkar RR (2013) Structural and logical analysis of a comprehensive hedgehog signaling pathway to identify alternative drug targets for glioma, colon and pancreatic cancer. *PLoS One* 8(7):e69132
109. Rateitschak K et al (2012) Parameter identifiability and sensitivity analysis predict targets for enhancement of STAT1 activity in pancreatic cancer and stellate cells. *PLoS Comput Biol* 8(12):e1002815
110. Wu M, Chan C (2014) Prediction of therapeutic microRNA based on the human metabolic network. *Bioinformatics*

111. Wu M, Chan C (2014) Prediction of therapeutic microRNA based on the human metabolic network. *Bioinformatics*. doi:10.1093/bioinformatics/btt751
112. Zeng T et al (2014) Prediction of dynamical drug sensitivity and resistance by module network rewiring-analysis based on transcriptional profiling. *Drug Resist Updat* 17(3):64–76
113. Kanagavel D et al (2010) A prognostic model in patients treated for metastatic gastric cancer with second-line chemotherapy. *Ann Oncol* 21(9):1779–1785



## Mathematical Models of Pluripotent Stem Cells: At the Dawn of Predictive Regenerative Medicine

Pinar Pir and Nicolas Le Novère

### Abstract

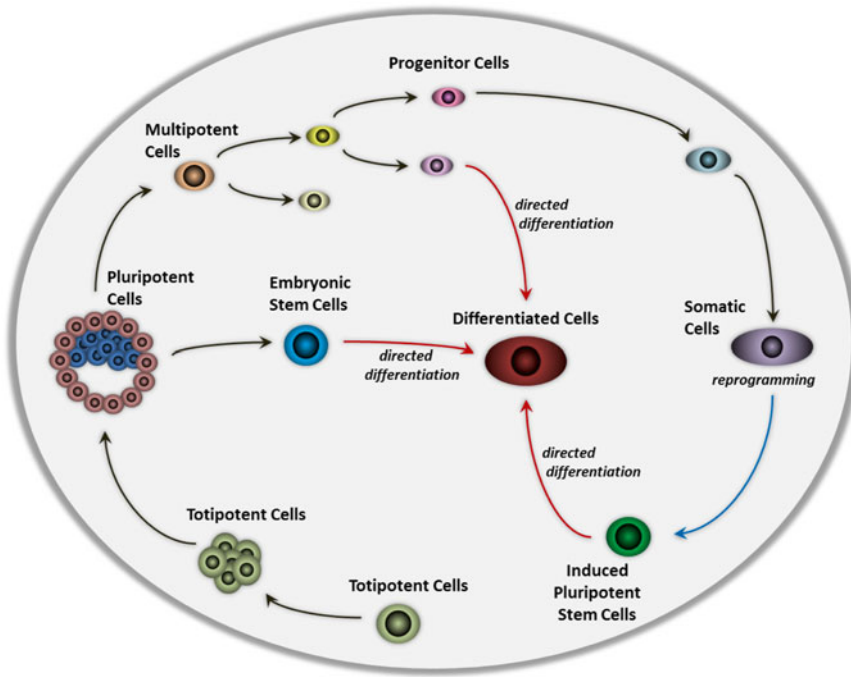
Regenerative medicine, ranging from stem cell therapy to organ regeneration, is promising to revolutionize treatments of diseases and aging. These approaches require a perfect understanding of cell reprogramming and differentiation. Predictive modeling of cellular systems has the potential to provide insights about the dynamics of cellular processes, and guide their control. Moreover in many cases, it provides alternative to experimental tests, difficult to perform for practical or ethical reasons. The variety and accuracy of biological processes represented in mathematical models grew in-line with the discovery of underlying molecular mechanisms. High-throughput data generation led to the development of models based on data analysis, as an alternative to more established modeling based on prior mechanistic knowledge. In this chapter, we give an overview of existing mathematical models of pluripotency and cell fate, to illustrate the variety of methods and questions. We conclude that current approaches are yet to overcome a number of limitations: Most of the computational models have so far focused solely on understanding the regulation of pluripotency, and the differentiation of selected cell lineages. In addition, models generally interrogate only a few biological processes. However, a better understanding of the reprogramming process leading to ESCs and iPSCs is required to improve stem-cell therapies. One also needs to understand the links between signaling, metabolism, regulation of gene expression, and the epigenetics machinery.

**Key words** Regenerative medicine, Systems biology, Mathematical modeling, Predictive models, Stem cells, Pluripotency

---

### 1 Introduction

Regenerative medicine aims to repair or regenerate tissues or organs with impaired functions, as an alternative to organ transplantation from donors. Approaches based on the use of the patient's own cells have the potential to overcome obstacles such as ethical concerns, limited donor availability, or transplant rejection. Such approaches often require derivation, generation, or manipulation of stem cells. Establishment of techniques for generation of stem cells to be used in regenerative medicine is a very active field of research [1].



**Fig. 1** Differentiation and reprogramming of cells. Embryonic stem cells can be derived from the inner cell mass (illustrated as *blue* cells) at the blastula stage of embryos, and can be differentiated into various cell types under laboratory conditions. Somatic cells can also be re-programmed into induced pluripotent stem cells, which can then be differentiated into desired cell types. See the main text for definition of potency of illustrated cell types

Stem cells have the potential to differentiate into more specialized cells in the multi-cellular organisms (Fig. 1). They also have the ability of self-renewal or proliferation, i.e., to generate cells identical to themselves via mitosis. The differentiation of stem cells is a progressive process, specialization increasing at each step. The zygote and the daughter cells generated via the first couple of cell divisions which can give rise to all cell types in the embryo and the placenta are called *totipotent*. More specialized embryonic stem cells which can give rise to all cell types in the embryo, but cannot contribute to the placenta are called *pluripotent* [2]. The pluripotent cells in pre-implantation embryo and post-implantation have distinct gene expression profiles, developmental and functional characteristics, and are called *naive* and *primed*, respectively [3]. As development of the embryo progresses, more specialized cells are generated, such as *multipotent* stem cells that can give rise to several tissues, *oligopotent* (or *progenitor*) cells that can only generate a set of closely related tissues, and terminally differentiated somatic cells. With the exception of the germinal lineage, mammals are mostly constituted of somatic cells. However, a limited number of adult stem cells keep regenerating tissues throughout the lifetime

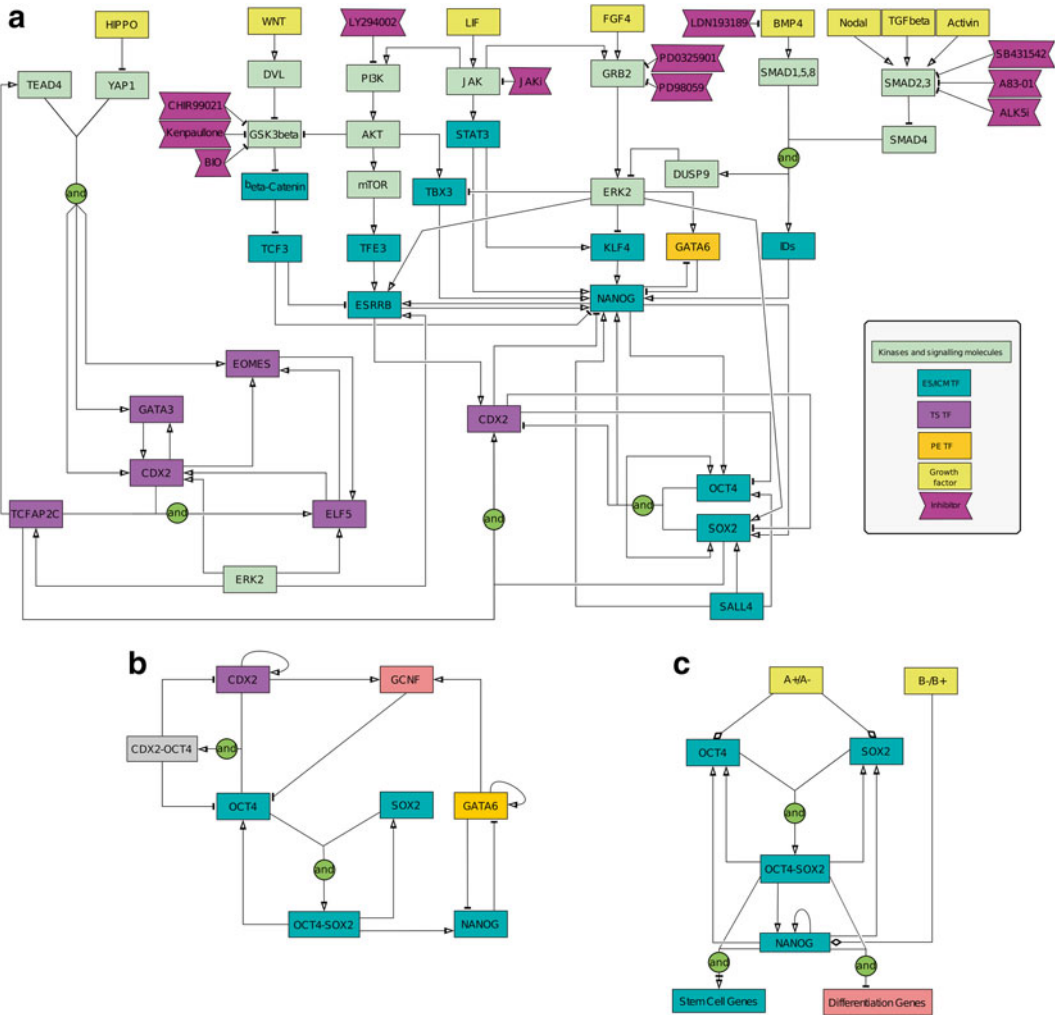
of the organism, such as hematopoietic stem cells giving rise to blood cells, muscle satellite cells, or neural stem cells [4].

Stem cells can be derived from a developing embryo or an adult, and the derived cell lines can be kept proliferating indefinitely or be differentiated into somatic cells in controlled niches. It is also possible to reprogram differentiated somatic cells into pluripotent cells via forced expression of proteins or continued exposure to external cues. Such reprogrammed cells are called *induced pluripotent stem cells* (iPSC) and can subsequently be re-differentiated into tissues [5]. Therefore, stem cells have the potential to be an infinite source of “spare-parts” for regenerative medicine as well as providing a model for drug and disease pathway discovery [1, 6, 7]. Stem cell science recently achieved two milestones, both of which came as a result of decade-long efforts: (1) The generation of human naive iPS cells from somatic cells [8, 9] and (2) the first clinical trial of a stem cell therapy via directed differentiation of human iPS cells into retinal pigment epithelium cells [10]. These milestones increased the confidence in stem cell science and raised the expectations from regenerative medicine.

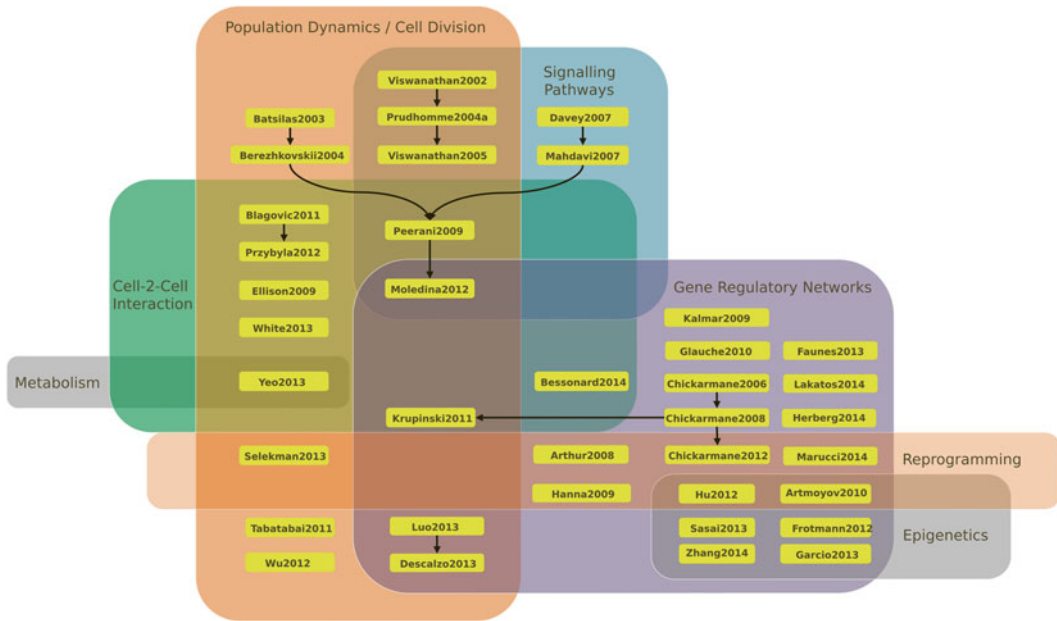
Computational biology is central to large-scale studies on stem cells, aiding extraction of meaningful information from large data sets, abstraction of biological systems for systematic analysis of their properties and ultimately predicting previously unknown relationships to guide future studies. Mathematical models and simulations have been used for a long time to understand the regulation of stem cell pluripotency and their differentiation and recent developments saw the rise of complex experimentally based models [11, 12]. It is likely that as for any other areas of bioengineering, models will provide major contributions to regenerative medicine research in the near future.

Reprogramming and differentiation of stem cells are regulated by the activity of gene regulatory networks controlled by several signaling pathways and the epigenetics machinery. Signaling pathways are triggered by the signaling molecules that bind their ligands on the cell surface or by the physical forces experienced by the cells; leading to a cascade of intracellular protein–protein interactions which in turn modulate the activity of the transcription factors in the downstream gene regulatory networks. Resulting gene expression profiles establish the composition of the cellular protein pool and ultimately, the phenotype.

The signaling pathways and gene regulatory networks that are actively modulated in pluripotent cell lines and pre-implantation embryos are shown in Fig. 2a. In addition, epigenetic factors may have a role in regulating the gene expression. Factors such as RNA interference, DNA methylation, nucleosome occupancy, and modifications on histone tails correlate with gene expression levels; these epigenetic factors are widely accepted as the barrier between transformation of cells into other cell types [13].



**Fig. 2** Pluripotency gene regulatory networks: activity flow map of the pathways regulating pluripotency (using the SBGN Activity Flow notation [110]). The nodes represent activities (of proteins or small chemicals) whereas the edges represent the effect of activities on each other. Proteins typically expressed in embryonic stem cells are shown in *turquoise*, proteins typically expressed in trophoblast stem cells are shown in *violet*, proteins typically expressed in primitive endoderm are shown in *orange* in all three panels. **(a)** Signaling networks and GRN in mouse pluripotent stem cells (modified from [90]) *Yellow boxes* indicate growth factors (or cytokines), whereas *light green boxes* indicate components of signaling pathways downstream to growth factors leading to up- or downregulation of transcription factors. Inhibitors are shown in *fuschia*. **(b)** Core pluripotency network in pre-implantation mouse embryo (modified from [44]) OCT4 forms protein complexes either with CDX2 or SOX2. OCT4–SOX2 protein complex upregulates NANOG expression, whereas CDX2–OCT4 complex leads to reduction of CDX2 and OCT4 pools. CDX2 and GATA6 upregulate the nuclear receptor GCNF, which downregulates OCT4 expression. **(c)** Stem cell box GRN (modified from [46]) OCT4 and SOX2 expression can be regulated by NANOG and external signals such as growth factors. OCT4–SOX2 protein complex takes part in autoregulation of its components and also regulation of NANOG. NANOG also regulates its own expression. The three proteins activate expression of stem cell-related proteins, whereas their absence leads to expression of differentiation-related proteins



**Fig. 3** Classification of the knowledge-based mathematical models of pluripotency. The classification is based on the relevant biological processes represented by transparent *boxes* with different background colors. The models are labeled with the first author's surname and the year of the publication. Models located on overlapping regions represent multiple cellular processes. The *arrows* show models derived from each other

Models of stem cells have been developed for a long time, mostly based on prior knowledge on the signaling pathways, gene regulatory networks, and epigenetic factors (Fig. 3). Recent improvements in omics data, in particular at a single cell level, coupled to ever more accurate computational inference methods, now permit to build and parameterize models directly from experimental results. In this chapter, we provide a critical overview of two general types of mathematical models, namely knowledge-based and data-based models. We focused on models describing the cellular mechanisms that maintain stem cell pluripotency in mammals and the priority was given to models describing well documented *in vivo* and *in vitro* systems. Theoretical analysis of cell fate (see for example [14]) can provide interesting insights in system behavior, however models of hypothetical systems were excluded in this chapter due to space limitations. The sections are organized to follow a coarse chronological order, in accordance with the experimental discoveries in the field.

## 2 Knowledge-Based Models

Most of the mechanistic models of stem cells are built following a bottom-up or knowledge-based approach. Information on the components to include and their relationships is obtained from

scientific literature or public databases, that contain previously generated models or information to incorporate in building blocks. For a general presentation of methods to build models of genes and molecular networks, see Le Novère [15].

### **2.1 Predicting the Behavior of Stem Cell Populations**

The bone marrow contains hematopoietic stem cells (HSC) which are multipotent cells and has self-renewal ability as well as differentiating into blood cells [16–18]. Transplantation of bone marrow has been in clinical use since 1950s as the first successfully established stem cell therapy [19]. In line with clinical applications, early mathematical models of stem cell proliferation and differentiation focused on HSC. These models investigated the cell population dynamics in response to external stimuli such as stress [20] and hypoxia [21]. They aimed at relating the stimuli to the propagation of HSC, and to improve efficiency of the clinical applications via optimization of the parameters. An overview of early predictive mathematical models of hematopoietic cell populations can be found in [22–24].

More recently, population behavior has often been simulated in combination with detailed models of molecular interactions in the stem cells (see sections below); however, minimal models of embryonic stem cell populations omitting the underlying molecular mechanisms were also shown to reproduce the population dynamics [25–28].

### **2.2 Predicting Cell Fate Control by Signaling Pathways**

Pluripotency is defined as the ability of the cells to generate all cell types in an embryo via multiple differentiation steps. Following the derivation of pluripotent stem cells from mouse embryos in 1981 [29], embryonic cells culture has been a fundamental tool in stem cell research and more largely in molecular biology. However, maintenance of the pluripotency in defined growth media has been a challenge until the design of the LIF+2i medium [30]. This medium contains the cytokine leukemia inhibitory factor (LIF) and inhibitors of ERK and GSK3 pathways (Fig. 2a), underlining how important is the prediction of niche-dependent factors' effect on self-renewal of stem cells, a major aim of modeling efforts since.

The first mathematical model describing the embryonic stem cell renewal as a function of cytokine concentration appeared in the literature in 2002 [31], soon after the discovery of the role of the LIF/JAK/STAT3 pathway in the maintenance of pluripotency [32]. A deterministic model describing the ligand–receptor dynamics on the cell surface was used to determine the cytokine thresholds in cell fate decisions. The simulation results demonstrated that LIF has stronger influence on the maintenance of pluripotency in comparison to a fusion protein derivative of IL-6 (HIL-6) and the differences between the potency of the two cytokines could be explained by receptor binding properties and the stoichiometry of binding.

A growth-rate-based deterministic model of differentiating and self-renewing stem cell populations was developed to predict the response to the cytokines LIF and FGF4 in addition to the extracellular matrix components laminin and fibronectin [33]. Each factor was hypothesized to have a dose effect on cell growth rate and kinetic parameters were estimated based on measured growth-rates. The model was validated by comparing the simulation results to the fraction of cells expressing high levels of OCT4 in a 4-factor culture. Its applicability to predicting the kinase activity level of the JAK/STAT3 pathway was demonstrated [34]. The experimental data on the kinase activity and the 4-factor growth experiment of embryonic stem cells (ESCs) were further used to construct a Bayesian network model of the same system [35] and the network was used to identify causal interactions between the components of the JAK/STAT3, AKT, and MAPK signaling pathways. This study is unique among the ES pluripotency network inference models in the sense that components of the signaling pathways have been predicted rather than gene regulatory network (GRN, see sections below). Predictions of the model were in accordance with known interactions, however the data collected at steady-state conditions were limiting the models' potential to represent timescales of the responses. For instance, phosphorylation of Stat3 is expected to be much faster than cell fate determination.

The timescale of the responses was addressed by parameterizing a compartmental model of JAK/STAT3 signaling pathway [36] using time course data of protein phosphorylation and gene expression [37]. The model was shown to reproduce the response of the pathway to its inhibitors and to predict the stem cell fate decisions. The sensitivity analysis indicated that ESC self-renewal was controlled by the frequency of the LIF stimulation. Further, this model was merged with a stochastic spatial model [38] to describe the colonial behavior in response to autocrine and paracrine signals, which is LIF in this particular case. An optimal colony size was suggested based on the simulations to maintain ES cells in niche [39].

Reproducible, robust, and efficient expansion of stem cells in well-controlled growth niches is one of the essential stages in stem cell research. Coarse-grained models relating the external cues directly to the cellular phenotype found applicability in the design of such bioprocesses. The variability of the micro-environment in a microfluidics system was modeled dynamically to predict the expansion of ES cells in response to signals [28, 40, 41]. A growth rate kinetics-based model to predict the response of ES cells to toxic material accumulation was shown to represent the population dynamics in batch and continuous ES cultures [42]. The spatial JAK/STAT3 model mentioned above [39] was extended further to analyze cell growth in a microfluidics system, and impact of the



system parameters like flow rate, position in the flow field, and local cell organization was demonstrated [43].

Activation of signaling pathways by external cues is a major determinant of cell fate. However, gene transcription is the stage where fate decision is “executed,” by providing the cells with the specific set of proteins required to generate the cellular phenotype.

### **2.3 Predicting Cell Fate Control by Gene Regulatory Networks**

Transcription factors (TFs) are the major actors in gene expression regulation. TFs often bind to each others’ promoters, therefore construct a regulatory network with feedback loops. Genes targeted by gene regulatory networks (GRN) can be inferred from differential mRNA expression levels, as well as the detection of DNA–TF complexes. The differentiation of morula cells into trophectoderm (TS), inner cell mass (ICM) and further differentiation of ICM into primitive endoderm (PE) and epiblast (EPI) was used as an experimental model in the discovery of key pluripotency TFs. The identification of GRNs that regulate pluripotency and cell differentiation allowed to represent a new level of complexity in mathematical models of stem cells. The excellent review by Niwa ([44] and the references therein) gives an overview of the early discoveries in pluripotency GRN in the pre-implantation mouse embryo. The core GRN described by Niwa, including NANOG, OCT4, SOX2, CDX2, GATA6 and GCNF, has been the basis of most models since (Fig. 2b).

The first model of the core GRN in embryonic stem cells [45] focused on the “stem cell box” (Fig. 2c), the tight regulation loop between NANOG, OCT4, and SOX2 which maintains the pluripotency. The model was used to analyze the response of the network to external signals as a function of model parameters such as strength of the feedback loops. The model was extended [46] to represent a larger network including the TFs CDX2, GATA6, and GCNF (Fig. 2b), which regulates the differentiation of stem cells into trophectoderm or endoderm in the mouse embryo. The network was analyzed to identify the factors that mediate the reprogramming and concluded that NANOG overexpression is a more robust way of reprogramming as opposed to suppression of its repressor GATA6. This model was further extended [47] to take into account the cell–cell interactions and asymmetric cell division that play important roles in regulating the GRN in the development of mouse embryos [48]. The model was able to reproduce the early development of the embryo in 3D and capture the experimentally observed patterns in cell fate decisions together with the heterogeneity in the embryo as a function of spatial mechanical forces, signaling pathways, and GRN. The effect of the FGF4 signal, via the MAPK pathway, was investigated with a 2D 25-cell model [49] where cells can receive FGF4 from neighboring cells. The simulations indicated the existence of a tri-stable network in

pre-implantation mouse embryos, which corresponds to ICM, EPI, and PE cells.

Heterogeneity in NANOG expression has been shown to be a property of stem cell populations [50]. Understanding the origin of such heterogeneity is important since homogeneous cell populations are more desirable in cell therapies. LIF + 2i growth medium, designed for maintenance of stem cells in culture [30], activates NANOG expression via activation of the JAK/STAT3 pathway, inhibition of the MAPK pathway, and inhibition of the GSK3 $\beta$  phosphorylation. LIF + 2i was shown to reduce the heterogeneity in NANOG levels [51]. These experimental discoveries were complemented by mathematical models investigating the origin and consequences of NANOG heterogeneity in stem cells [52–54]. Simulation of the “stem cell box” together with the autocrine signaling of FGF4 demonstrated that the autocrine feedback loops are also a likely source of heterogeneity in NANOG levels [55].

Serum has often been used as a growth media component together with LIF before LIF + 2i was designed. Comparison of the two media has been of interest as serum is known to have BMP4, which is a signal that regulates the pluripotency GRN via SMAD signaling pathways (Fig. 2a). A model which represents the regulation of the stem cell box with all four external factors was used to analyze the three steady states of NANOG, effect of inhibitors, and noise in the cells growing in media with combinations of the external factors [56]. In another study, relative levels of the protein complexes formed between the components of the pluripotency network have been shown to regulate the OCT4 levels in LIF + 2i and LIF + serum [57], at the ground state and exit from pluripotency. The minimal model, representing the dynamics of the GRN and post-transcriptional regulation mediated by the protein complexes, was able to recapitulate the effect of gene deletions in the GRN. A recent model [58] has combined the mathematical modeling of the “stem cell box” with experimental validation via a downstream reporter, REX1 in LIF + 2i. The mechanism of stem cell box regulation by beta-catenin in serum + LIF and 2i was shown to be significant in differentiation compared to maintenance of pluripotency by modeling the Wnt/beta-catenin pathway together with the stem cell box [59]. The impact of beta-catenin on NANOG during reprogramming was also investigated [60].

#### **2.4 Predicting Cell Reprogramming Trajectories Controlled by Epigenetics**

Epigenetics has been long known to have an important role in the cell fate as conceptually described by Waddington [61–63], also see [64] for an excellent review of Waddington’s work in a philosophical context of systems biology and mathematical modeling. Currently, epigenetics is defined as relatively stable and potentially heritable changes in the transcriptional potential of the cells without any changes in the DNA sequence. Epigenetic factors are being defined as small molecules such as methyl- groups deposited on

DNA, the 3-D chromatin structure dictated by a set of DNA-binding proteins such as histones, together with the small molecules deposited on the DNA-binding proteins, and non-coding RNA with regulatory functions [65]. Identification of molecular details of epigenetic factors in embryonic stem cells followed shortly after the identification of the core pluripotency GRN. It has been shown that the enzymes and structural components of the epigenetic machinery work together with the signaling pathways and GRN on the gene regulation (see for example [66]), and further mediate transmitting the phenotype of the mother to its daughter cells via silencing the transcription of a set of genes.

Following the Yamanaka's and Gurdon's discoveries on reprogramming of somatic cells into stem cells [67], identification of the factors that increase the efficiency of reprogramming has been a new avenue for mathematical modeling of pluripotency. Understanding the underlying epigenetic profiles in stem cells became even more crucial with the recognition that epigenetic factors are the barrier in the reprogramming of somatic cells into pluripotent cells [13].

A GRN-based stochastic model for reprogramming of differentiated cells into pluripotent state was proposed [68]. The noise in gene expression levels was shown to be adequate for reprogramming if the level of the noise is large enough to overcome the silencing in stem cell box genes imposed by differentiation genes. The reprogramming efficiency has been described by a mathematical model as a function of the cell doubling time [69]. The results have suggested that all cells in a given population can be reprogrammed into induced pluripotent stem cells (iPS cells), whereas the number of cell divisions required for reprogramming differs between the cells. These two models explained the reprogramming trajectories without taking the epigenetic factors into account explicitly.

A first model explicitly representing the activity in GRNs in interaction with epigenetic profiles in pluripotent stem cells was constructed as a cell-cycle-based binary model [70] by simplifying the GRN and epigenetic networks into cell-type-specific modules. The simulations assumed that gene expression takes place only in the interphase of a two-phase cell cycle, whereas changes in epigenetic profiles take place only in the telophase. Active, silent, and poised states were represented in epigenetic modules where genes were represented either as being expressed or silenced. The model was able to reproduce the observed trajectories of differentiation and reprogramming. This model was modified as a Markov model with simplified rules in comparison to the original model and was shown to reproduce reprogramming and gene expression profiles [71].

Another Boolean model representing both the pluripotency GRN and epigenetics in finer detail [72] was able to reproduce the

profiles observed in cell differentiation and reprogramming as response to external modifications such as gene silencing or inhibition of epigenetic regulation. Optimization of reprogramming speed and efficiency was proposed as a function of dynamics of DNA methylation and chromatin structure.

An ODE-based model of the stem cell box with most studied epigenetic marks H3K4me3 and H3K27me3 (tri-methylation on Lysine4 and Lysine27 residues on Histone3 proteins) was constructed [73] and simulated stochastically. The model was used to analyze the observed bistability, inducibility, stochasticity, and reprogramming profiles of the cells as response to external stimuli.

Recently, *Nanog* expression in pre-implantation mouse embryos and ES cells grown in LIF+2i was shown to be under allelic regulation via differential epigenetic silencing between the alleles [74]. The allelic regulation of *Nanog* was investigated in a mathematical model based on the core GRN of pre-implantation mouse embryos extended with epigenetic regulation of gene expression [75]. The model did not take into account any external signals, however was able to demonstrate the bistable behavior of *Nanog* expression as a function of slow epigenetic dynamics that leads to differentiation or self-renewal of ES cells. Impact of slow epigenetic kinetics on stochastic cell fate decisions was shown with another GRN model, an extended pluripotency network with KLF4 and PBX1 [76].

Reprogramming and directed differentiation of stem cells usually require step-wise protocols. Recently, few models were proposed to represent the discrete changes taking place in signaling, GRN, and epigenetics profiles of the cells depending on the stage of the process [77, 78].

Representation of epigenetic effects in the mathematical models of stem cells is still in its infancy, however it can be envisioned that the cell commitment and reprogramming models with epigenetics will be superior in terms of their accuracy in making predictions, given the fact that epigenetics is the barrier that has to be overcome by transforming cells.

---

### 3 Data-Based Models

The models we have mentioned so far were built on a priori information about the systems to be investigated, and aimed to predict the phenotype as a function of selected factors and parameters. Building knowledge-based models can be tedious as the relevant molecular interactions have to be curated from the literature. In an emerging field like stem cell research, there is also the possibility that the list of interactions derived from the literature will be incomplete. Further, available measurements and biochemical information may not be adequate to parameterize the models.

In the following section we will be giving examples of models built without a priori information, i.e., built using data-based methods to identify network components or network structures in the pluripotent stem cells. Data-based methods are becoming more and more attractive in mathematical model reconstruction as they do not require tedious literature surveys, but indeed rely on analysis of high-throughput data which is accumulating with a tremendous speed in databases.

A partial least squares regression model (PLS, a singular value decomposition-based linear regression method) [79] relates the external factors to signaling pathways and the cellular responses (i.e., growth rate kinetics determined in [33]). The predicted effect of the tested factors on the phospho-proteome and growth patterns agreed well with the literature, demonstrating that data-driven unsupervised models can be used to build plausible models. While this model is a reconstruction of signaling pathways based on phospho-protein measurements, GRNs can be reconstructed based on gene expression data. The reconstructed GRN models are built using the predictive power of gene expression levels on other relevant readouts [80–83]. Gene expression levels can be further combined with data providing evidence on gene regulation, such as DNA–protein binding data [84] or miRNA [85] for reconstruction of GRNs. On the other hand, gene expression levels can be used to validate models built using other types of high-throughput data such as histone modifications [86] or chromatin structure [87].

Data-based models can also improve our understanding of the barriers hindering the reprogramming efficiency, for example, gene expression-based GRNs supplemented by known interactions were analyzed to identify the response pathways that may hinder reprogramming efficiency [88, 89]. Differences between human and mouse ESC have been a matter of debate [90], the active signaling pathways in mouse and human ESC were compared using gene expression-based GRNs [84, 91] to answer the open questions in the field.

Hybrid models of pluripotency were proposed as an alternative to purely qualitative or quantitative models. For example, a meta-model derived from an ODE-based model of the human PI3K/AKT pathway was used for efficient parameter sensitivity analysis of the steady state of the pathway in hESC [92], whereas complementing a gene expression-based unsupervised GRN with literature-curated regulatory interactions was shown to lead to more accurate predictions on the active GRN of mESCs differentiating into the three germ lines, in parallel with the gastrulation process in the embryo [93].

The high-throughput data-driven nature of unsupervised methods has the advantage that their predictions are genome-wide and open to context-dependent interpretation. Further, the

algorithms underlying the models can be readily implemented to new systems or datasets. Therefore, taking the extra mile to provide a web application or software package facilitates the re-use of the models and their predictions by the stem cell community. Few such applications were specifically validated for pluripotent stem cells: CellNet is a collection of tools for construction and analysis of gene expression-based GRNs in stem cells [94]. RE:IN [11] is another web-based application for gene expression-based GRN reconstruction. RE:IN was used to propose a minimal pluripotency network in mES which was shown to correctly predict the phenotype of double knock-out mutants. ExprEssence [95] has been developed as a Cytoscape [96] plugin to build networks of gene expression data; the tool was verified with pluripotency GRN.

The recent developments in the technology allow generation of high-throughput data on a variety of cellular features relevant to maintenance of pluripotency, such as profiles of epigenetic marks and 3-D chromosome structure [97], in addition to proving more efficient and more accurate techniques over the established techniques of collecting transcriptome, proteome, metabolome, protein–protein, and DNA–protein interaction data. The potential of such features in unsupervised modeling of pluripotency has not been fully utilized yet, although few recent applications initiated such efforts: StemSight [98] integrates gene expression and DNA–protein binding data to provide a verified network of mouse pluripotent stem cells. StemSight has been further extended to include the human pluripotency network [99]. The ESCAPE database (former iScMiD, [100]) stores data from sources such as phosphoproteins, miRNA interactions, histone modifications, and gene expression. ESCAPE also provides a collection of data analysis and network construction tools for mouse and human pluripotent stem cells, verified for pluripotency network in mESCs [12]. Repositories such as multi-organism STRING [101] and mouse-specific MouseNET [102] integrate data from diverse sources including gene expression levels, protein–protein interaction, text mining, and functional relatedness to build interaction maps of known and predicted interactions between the genes.

---

## 4 Perspectives on Predictive Models in Regenerative Medicine

In this chapter, we tried to give an overview of computational modeling of pluripotent stem cells in line with the experimental discoveries in the field. The pioneering modeling efforts have started from minimal probabilistic population dynamics models, and progressed to represent most levels of regulation in cell fate decisions. The approach and granularity of the models have also progressed, a range of approaches from deterministic ODE-based models to unsupervised logical models were adopted, whereas

building models in finer granularity became feasible with increasing availability of quantitative data, evidence on molecular interactions, and more powerful computational resources.

#### **4.1 All Regulatory Levels Need to Be Represented**

It is becoming clear that the pluripotency is regulated at many levels of cellular activities, i.e., signaling, gene regulatory networks, epigenetics, and metabolism [103]. The existing computational models of pluripotency often focus on only one or two of these levels. For instance, metabolism has not yet been taken into consideration in modeling the pluripotency although metabolism dependency of embryos and induced pluripotent stem cells has been reported [104–106]. Dependency of the cell cycle on GRNs, post-transcriptional regulation by non-coding RNAs, and post-translational regulation of protein activities are other phenomena that have been overlooked by the current models. The success of next-generation predictive models of pluripotency will rely on the representation of cell fate regulation at *all* levels.

#### **4.2 Modularity Needed for Integrating Different Modeling Approaches**

The choice of the modeling approach depends on the nature of the cellular activities to be modeled. For example, gene expression is a noisy process; the molecular crowd in the nucleus and low copy number of the regulatory proteins introduce an intrinsic stochasticity that leads to stochastic cell fate decisions and heterogeneous cell populations. Cell division and unequal partition of low copy number species between the daughter cells is another source of heterogeneity in cell populations. However, metabolism and signaling are faster processes and take place in a relatively homogeneous environment, where assumption of continuity and steady state may hold in most cases. Therefore, a complete model representing all levels of cellular activities may require the use of different approaches for each level.

#### **4.3 Re-usability of the Models Needs to Be Promoted**

Re-usability of the models needs to be taken into consideration by the computational biologists. Distribution of the models in standard formats with adequate documentation is crucial for efficient use of resources. Only three of the models we have mentioned in this chapter are available at the BioModels database [107], and none could be found in the CellML [108] or JWS [109] model repositories, which demonstrates that re-use of the pluripotency models is currently not straightforward and potentially a time-consuming task.

The computational cost of parameterizing and simulating models increases with increasing granularity. Therefore, the molecular interactions represented in the model have to be chosen with care to provide enough details to answer the biological question being asked while omitting the details which are not relevant for the current question or observable with the current experimental



tools. Re-use of existing models to build larger and modular models can keep the computational cost at feasible levels.

Software tools and databases (particularly specialized on stem cells) are invaluable resources for computational biologists. However, often very useful tools and databases are not maintained or updated, therefore become obsolete quickly. Continued maintenance by dedicated teams could prevent the waste of resources and provide the community with reliable and up-to-date tools and databases.

#### **4.4 Validation of the Predictive Power Missing for Most Models**

Mathematical models aim to explain the biological phenomena observed and predict the outcome under the conditions not yet experimentally tested. Validation of models can rely on use of existing experimental observations, however accuracy of the predictions often are not tested. Lack of follow-up experiments often hinders the usefulness of the models. An iterative systems biology approach is needed to utilize the potential of predictive models, where model construction, prediction, and validation have to be designed a priori for the biological question being asked. Such iterative approaches require close collaboration between experimental and computational biologists.

We would like to conclude with the observation that predictive modeling in regenerative medicine is at its dawn. Availability of detailed knowledge and genome-wide data on cellular and organismal level will promote the construction of larger and modular models with higher predictive power. The predictions from the models will lead to targeted clinical applications with higher rates of success in regenerative medicine.

---

## **Acknowledgements**

We would like to thank editors for the critical reading of the manuscript and the constructive comments. PP and NL are funded by the BBSRC Institute Strategic Programme BBS/E/B/000C0419.

## **References**

1. Robinton DA, Daley GQ (2012) The promise of induced pluripotent stem cells in research and therapy. *Nature* 481:295–305. doi:[10.1038/nature10761](https://doi.org/10.1038/nature10761)
2. Kolios G, Moodley Y (2013) Introduction to stem cells and regenerative medicine. *Respiration* 85:3–10. doi:[10.1159/000345615](https://doi.org/10.1159/000345615)
3. Nichols J, Smith A (2009) Naive and primed pluripotent states. *Cell Stem Cell* 4:487–492. doi:[10.1016/j.stem.2009.05.015](https://doi.org/10.1016/j.stem.2009.05.015)
4. Rezza A, Sennett R, Rendl M (2014) Adult stem cell niches: cellular and molecular components. *Curr Top Dev Biol*. doi:[10.1016/B978-0-12-416022-4.00012-3](https://doi.org/10.1016/B978-0-12-416022-4.00012-3)

5. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126:663–676. doi:[10.1016/j.cell.2006.07.024](https://doi.org/10.1016/j.cell.2006.07.024)
6. Wu SM, Hochedlinger K (2011) Harnessing the potential of induced pluripotent stem cells for regenerative medicine. *Nat Cell Biol* 13:497–505. doi:[10.1038/ncb0511-497](https://doi.org/10.1038/ncb0511-497)
7. Nishikawa S, Goldstein RA, Nierras CR (2008) The promise of human induced pluripotent stem cells for research and therapy. *Nat Rev Mol Cell Biol* 9:725–729
8. Theunissen TW, Powell BE, Wang H et al (2014) Systematic identification of defined conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* 1–17. doi: [10.1016/j.stem.2014.07.002](https://doi.org/10.1016/j.stem.2014.07.002)
9. Takashima Y, Guo G, Loos R et al (2014) Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* 158:1254–1269. doi:[10.1016/j.cell.2014.08.029](https://doi.org/10.1016/j.cell.2014.08.029)
10. Sheridan C (2014) Stem cell therapy clears first hurdle in AMD. *Nat Biotechnol*. doi:[10.1016/S0140-6736\(14\)61376-3](https://doi.org/10.1016/S0140-6736(14)61376-3)
11. Dunn S-J, Martello G, Yordanov B et al (2014) Defining an essential transcription factor program for naïve pluripotency. *Science* 344:1156–1160. doi:[10.1126/science.1248882](https://doi.org/10.1126/science.1248882)
12. Xu H, Ang Y-S, Sevilla A et al (2014) Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS Comput Biol* 10:e1003777. doi:[10.1371/journal.pcbi.1003777](https://doi.org/10.1371/journal.pcbi.1003777)
13. Papp B, Plath K (2011) Reprogramming to pluripotency: stepwise resetting of the epigenetic landscape. *Cell Res* 21:486–501. doi:[10.1038/cr.2011.28](https://doi.org/10.1038/cr.2011.28)
14. Buzi G, Lander AD, Khammash M (2015) Cell lineage branching as a strategy for proliferative control. *BMC Biol*. doi:[10.1186/s12915-015-0122-8](https://doi.org/10.1186/s12915-015-0122-8)
15. Le Novère N (2015) Quantitative and logic modelling of molecular and gene networks. *Nat Publ Gr* 16:146–158. doi:[10.1038/nrg3885](https://doi.org/10.1038/nrg3885)
16. Mazo IB, Massberg S, von Andrian UH (2011) Hematopoietic stem and progenitor cell trafficking. *Trends Immunol* 32:493–503. doi:[10.1016/j.it.2011.06.011](https://doi.org/10.1016/j.it.2011.06.011)
17. Moignard V, Woodhouse S, Fisher J, Göttgens B (2013) Transcriptional hierarchies regulating early blood cell development. *Blood Cells Mol Dis* 51:239–247. doi:[10.1016/j.bcmd.2013.07.007](https://doi.org/10.1016/j.bcmd.2013.07.007)
18. Sive JI, Göttgens B (2014) Transcriptional network control of normal and leukaemic haematopoiesis. *Exp Cell Res* 329:255–264. doi:[10.1016/j.yexcr.2014.06.021](https://doi.org/10.1016/j.yexcr.2014.06.021)
19. Thomas ED, Lochte HL, Cannon JH et al (1959) Supralethal whole body irradiation and isologous marrow transplantation in man. *J Clin Invest* 38:1709–1716. doi:[10.1172/JCI103949](https://doi.org/10.1172/JCI103949)
20. Till JE, McCulloch EA, Siminovitch L (1963) A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *PNAS* 51:29–36
21. Loeffler M, Wichmann HE (1980) A comprehensive mathematical model of stem cell proliferation which reproduces most of the published experimental results. *Cell Tissue Kinet* 13:543–561
22. Viswanathan S, Zandstra PW (2003) Towards predictive models of stem cell fate. *Cytotechnology* 41(2–3):75–92. doi:[10.1023/A:1024866504538](https://doi.org/10.1023/A:1024866504538)
23. Foster SD, Oram SH, Wilson NK, Göttgens B (2009) From genes to cells to tissues – modelling the haematopoietic system. *Mol Biosyst* 5:1413–1420. doi:[10.1039/B907225j](https://doi.org/10.1039/B907225j)
24. Pisu M, Concas A, Cao G (2007) A novel simulation model for stem cells differentiation. *J Biotechnol* 130:171–182. doi:[10.1016/j.jbiotec.2007.02.028](https://doi.org/10.1016/j.jbiotec.2007.02.028)
25. Tabatabai MA, Bursac Z, Eby WM, Singh KP (2011) Mathematical modeling of stem cell proliferation. *Med Biol Eng Comput* 49:253–262. doi:[10.1007/s11517-010-0686-y](https://doi.org/10.1007/s11517-010-0686-y)
26. Wu J, Tzanakakis ES (2012) Contribution of stochastic partitioning at human embryonic stem cell division to NANOG heterogeneity. *PLoS One* 7:e50715. doi:[10.1371/journal.pone.0050715](https://doi.org/10.1371/journal.pone.0050715)
27. White DE, Kinney MA, McDevitt TC, Kemp ML (2013) Spatial pattern dynamics of 3D stem cell loss of pluripotency via rules-based computational modeling. *PLoS Comput Biol* 9:e1002952. doi:[10.1371/journal.pcbi.1002952](https://doi.org/10.1371/journal.pcbi.1002952)
28. Blagovic K, Kim LY, Voldman J (2011) Microfluidic perfusion for regulating diffusible signaling in stem cells. *PLoS One* 6:e22892. doi:[10.1371/journal.pone.0022892](https://doi.org/10.1371/journal.pone.0022892)
29. Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. *Nature* 292:154–156

30. Ying Q-L, Wray J, Nichols J et al (2008) The ground state of embryonic stem cell self-renewal. *Nature* 453:519–523. doi:[10.1038/nature06968](https://doi.org/10.1038/nature06968)
31. Viswanathan S, Benatar T, Zandstra PW et al (2002) Ligand/receptor signaling threshold (LIST) model accounts for gp130-mediated embryonic stem cell self-renewal responses to LIF and HIL-6. *Stem Cells* 20:119–138
32. Niwa H, Burdon T, Chambers I, Smith A (1998) Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev* 12:2048–2060. doi:[10.1101/gad.12.13.2048](https://doi.org/10.1101/gad.12.13.2048)
33. Prudhomme WA, Duggar KH, Lauffenburger DA (2004) Cell population dynamics model for deconvolution of murine embryonic stem cell self-renewal and differentiation responses to cytokines and extracellular matrix. *Biotechnol Bioeng* 88:264–272. doi:[10.1002/bit.20244](https://doi.org/10.1002/bit.20244)
34. Viswanathan S, Davey RE, Cheng D et al (2005) Clonal evolution of stem and differentiated cells can be predicted by integrating cell-intrinsic and -extrinsic parameters. *Biotechnol Appl Biochem* 42:119–131. doi:[10.1042/BA20040207](https://doi.org/10.1042/BA20040207)
35. Woolf PJ, Prudhomme W, Daheron L et al (2005) Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. *Bioinformatics* 21:741–753. doi:[10.1093/bioinformatics/bti056](https://doi.org/10.1093/bioinformatics/bti056)
36. Davey RE, Onishi K, Mahdavi A, Zandstra PW (2007) LIF-mediated control of embryonic stem cell self-renewal emerges due to an autoregulatory loop. *FASEB J* 21:2020–2032. doi:[10.1096/fj.06-7852com](https://doi.org/10.1096/fj.06-7852com)
37. Mahdavi A, Davey RE, Bhola P et al (2007) Sensitivity analysis of intracellular signaling pathway kinetics predicts targets for stem cell fate control. *PLoS Comput Biol* 3:e130. doi:[10.1371/journal.pcbi.0030130](https://doi.org/10.1371/journal.pcbi.0030130)
38. Batsilas L, Berezhkovskii AM, Shvartsman SY (2003) Stochastic model of autocrine and paracrine signals in cell culture assays. *Biophys J* 85:3659–3665. doi:[10.1016/S0006-3495\(03\)74783-3](https://doi.org/10.1016/S0006-3495(03)74783-3)
39. Peerani R, Onishi K, Mahdavi A et al (2009) Manipulation of signaling thresholds in “engineered stem cell niches” identifies design criteria for pluripotent stem cell screens. *PLoS One* 4:e6438. doi:[10.1371/journal.pone.0006438](https://doi.org/10.1371/journal.pone.0006438)
40. Ellison D, Munden A, Levchenko A (2009) Computational model and microfluidic platform for the investigation of paracrine and autocrine signaling in mouse embryonic stem cells. *Mol Biosyst* 5:1004–1012. doi:[10.1039/b905602e](https://doi.org/10.1039/b905602e)
41. Przybyla LM, Voldman J (2012) Attenuation of extrinsic signaling reveals the importance of matrix remodeling on maintenance of embryonic stem cell self-renewal. *Proc Natl Acad Sci U S A* 109:835–840. doi:[10.1073/pnas.1103100109](https://doi.org/10.1073/pnas.1103100109)
42. Yeo D, Kiparissides A, Cha JM et al (2013) Improving embryonic stem cell expansion through the combination of perfusion and Bioprocess model design. *PLoS One* 8:e81728. doi:[10.1371/journal.pone.0081728](https://doi.org/10.1371/journal.pone.0081728)
43. Moledina F, Clarke G, Oskooei A et al (2012) Predictive microfluidic control of regulatory ligand trajectories in individual pluripotent cells. *Proc Natl Acad Sci U S A* 109:3264–3269. doi:[10.1073/pnas.1111478109](https://doi.org/10.1073/pnas.1111478109)
44. Niwa H (2007) How is pluripotency determined and maintained? *Development* 134:635–646. doi:[10.1242/dev.02787](https://doi.org/10.1242/dev.02787)
45. Chickarmane V, Troein C, Nuber UA et al (2006) Transcriptional dynamics of the embryonic stem cell switch. *PLoS Comput Biol* 2:e123. doi:[10.1371/journal.pcbi.0020123](https://doi.org/10.1371/journal.pcbi.0020123)
46. Chickarmane V, Peterson C (2008) A computational model for understanding stem cell, trophoblast and endoderm lineage determination. *PLoS One* 3:e3478. doi:[10.1371/journal.pone.0003478](https://doi.org/10.1371/journal.pone.0003478)
47. Krupinski P, Chickarmane V, Peterson C (2011) Simulating the mammalian blastocyst – molecular and mechanical interactions pattern the embryo. *PLoS Comput Biol* 7:e1001128. doi:[10.1371/journal.pcbi.1001128](https://doi.org/10.1371/journal.pcbi.1001128)
48. Ralston A, Rossant J (2008) Cdx2 acts downstream of cell polarization to cell-autonomously promote trophoblast fate in the early mouse embryo. *Dev Biol* 313:614–629. doi:[10.1016/j.ydbio.2007.10.054](https://doi.org/10.1016/j.ydbio.2007.10.054)
49. Bessonard S, De Mot L, Gonze D et al (2014) Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development* 3637–3648. doi:[10.1242/dev.109678](https://doi.org/10.1242/dev.109678)
50. Singh AM, Hamazaki T, Hankowski KE, Terada N (2007) A heterogeneous expression pattern for Nanog in embryonic stem cells. *Stem Cells* 25:2534–2542. doi:[10.1634/stemcells.2007-0126](https://doi.org/10.1634/stemcells.2007-0126)
51. Canham MA, Sharov AA, Ko MSH, Brickman JM (2010) Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. *PLoS Biol* 8:e1000379. doi:[10.1371/journal.pbio.1000379](https://doi.org/10.1371/journal.pbio.1000379)

52. Kalmar T, Lim C, Hayward P et al (2009) Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol* 7:e1000149. doi:[10.1371/journal.pbio.1000149](https://doi.org/10.1371/journal.pbio.1000149)
53. Glauche I, Herberg M, Roeder I (2010) Nanog variability and pluripotency regulation of embryonic stem cells – insights from a mathematical model analysis. *PLoS One* 5:e11238. doi:[10.1371/journal.pone.0011238](https://doi.org/10.1371/journal.pone.0011238)
54. Chickarmane V, Olariu V, Peterson C (2012) Probing the role of stochasticity in a model of the embryonic stem cell: heterogeneous gene expression and reprogramming efficiency. *BMC Syst Biol* 6:98. doi:[10.1186/1752-0509-6-98](https://doi.org/10.1186/1752-0509-6-98)
55. Lakatos D, Travis ED, Pierson KE et al (2014) Autocrine FGF feedback can establish distinct states of Nanog expression in pluripotent stem cells: a computational analysis. *BMC Syst Biol* 8:112. doi:[10.1186/s12918-014-0112-4](https://doi.org/10.1186/s12918-014-0112-4)
56. Luo Y, Lim CL, Nichols J, Martinez-Arias A, Wernisch L (2012) Cell signalling regulates dynamics of Nanog distribution in embryonic stem cell populations. *J R Soc Interface*. [Epub ahead of print].
57. Muñoz Descalzo S, Rué P, Faunes F et al (2013) A competitive protein interaction network buffers Oct4-mediated differentiation to promote pluripotency in embryonic stem cells. *Mol Syst Biol* 9:694. doi:[10.1038/msb.2013.49](https://doi.org/10.1038/msb.2013.49)
58. Herberg M, Kalkan T, Glauche I et al (2014) A model-based analysis of culture-dependent phenotypes of mESCs. *PLoS One* 9:e92496. doi:[10.1371/journal.pone.0092496](https://doi.org/10.1371/journal.pone.0092496)
59. Faunes F, Hayward P, Descalzo SM et al (2013) A membrane-associated  $\beta$ -catenin/Oct4 complex correlates with ground-state pluripotency in mouse embryonic stem cells. *Development* 140:1171–1183. doi:[10.1242/dev.085654](https://doi.org/10.1242/dev.085654)
60. Marucci L, Pedone E, Di Vicino U et al (2014)  $\beta$ -catenin fluctuates in mouse ESCs and is essential for Nanog-mediated reprogramming of somatic cells to pluripotency. *Cell Rep* 8:1686–1696. doi:[10.1016/j.celrep.2014.08.011](https://doi.org/10.1016/j.celrep.2014.08.011)
61. Waddington CH (1956) *Principles of embryology*. G. Allen, London
62. Lee HJ, Hore TA, Reik W (2014) Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* 14:710–719. doi:[10.1016/j.stem.2014.05.008](https://doi.org/10.1016/j.stem.2014.05.008)
63. Tomizawa S, Shirakawa T, Ohbo K (2014) Stem cell epigenetics: insights from studies on embryonic, induced pluripotent, and germline stem cells. *Curr Pathobiol Rep* 2:1–9. doi:[10.1007/s40139-013-0038-3](https://doi.org/10.1007/s40139-013-0038-3)
64. Fagan MB (2011) Waddington redux: models and explanation in stem cell and systems biology. *Biol Philos* 27:179–213. doi:[10.1007/s10539-011-9294-y](https://doi.org/10.1007/s10539-011-9294-y)
65. Boland MJ, Nazor KL, Loring JF (2014) Epigenetic regulation of pluripotency and differentiation. *Circ Res* 115:311–324. doi:[10.1161/CIRCRESAHA.115.301517](https://doi.org/10.1161/CIRCRESAHA.115.301517)
66. Griffiths DS, Li J, Dawson MA et al (2011) LIF-independent JAK signalling to chromatin in embryonic stem cells uncovered from an adult stem cell disease. *Nat Cell Biol* 13:13–21. doi:[10.1038/ncb2135](https://doi.org/10.1038/ncb2135)
67. Gurdon JB, Melton DA (2008) Nuclear reprogramming in cells. *Science* 322:1811–1815
68. MacArthur BD, Please CP, Oreffo ROC (2008) Stochasticity and the molecular mechanisms of induced pluripotency. *PLoS One* 3:e3086. doi:[10.1371/journal.pone.0003086](https://doi.org/10.1371/journal.pone.0003086)
69. Hanna J, Saha K, Pando B et al (2009) Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* 462:595–601. doi:[10.1038/nature08592](https://doi.org/10.1038/nature08592)
70. Artyomov MN, Meissner A, Chakraborty AK (2010) A model for genetic and epigenetic regulatory networks identifies rare pathways for transcription factor induced pluripotency. *PLoS Comput Biol* 6:e1000785. doi:[10.1371/journal.pcbi.1000785](https://doi.org/10.1371/journal.pcbi.1000785)
71. Hu Z, Qian M, Zhang MQ (2011) Novel Markov model of induced pluripotency predicts gene expression changes in reprogramming. *BMC Syst Biol* 5(Suppl 2):S8. doi:[10.1186/1752-0509-5-S2-S8](https://doi.org/10.1186/1752-0509-5-S2-S8)
72. Flöttmann M, Scharp T, Klipp E (2012) A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front Physiol* 3:216. doi:[10.3389/fphys.2012.00216](https://doi.org/10.3389/fphys.2012.00216)
73. Grácio F, Cabral J, Tidor B (2013) Modeling stem cell induction processes. *PLoS One* 8:e60240. doi:[10.1371/journal.pone.0060240](https://doi.org/10.1371/journal.pone.0060240)
74. Miyanari Y, Torres-Padilla ME (2012) Control of ground-state pluripotency by allelic regulation of Nanog. *Nature* 483:470–473. doi:[10.1038/nature10807](https://doi.org/10.1038/nature10807)
75. Sasai M, Kawabata Y, Makishi K et al (2013) Time scales in epigenetic dynamics and phenotypic heterogeneity of embryonic stem cells. *PLoS Comput Biol* 9:e1003380. doi:[10.1371/journal.pcbi.1003380](https://doi.org/10.1371/journal.pcbi.1003380)
76. Zhang B, Wolyne PG (2014) Stem cell differentiation as a many-body problem. *Proc Natl Acad Sci U S A* 111:10185–10190. doi:[10.1073/pnas.1408561111](https://doi.org/10.1073/pnas.1408561111)

77. Muraro MJ, Kempe H, Verschure PJ (2013) Concise review: the dynamics of induced pluripotency and its behavior captured in gene network motifs. *Stem Cells* 31:838–848. doi:[10.1002/stem.1340](https://doi.org/10.1002/stem.1340)
78. Selekman JA, Das A, Grundl NJ, Palecek SP (2013) Improving efficiency of human pluripotent stem cell differentiation platforms using an integrated experimental and computational approach. *Biotechnol Bioeng* 110:3024–3037. doi:[10.1002/bit.24968](https://doi.org/10.1002/bit.24968)
79. Prudhomme W, Daley GQ, Zandstra P, Lauffenburger DA (2004) Multivariate proteomic analysis of murine embryonic stem cell self-renewal versus differentiation signaling. *Proc Natl Acad Sci U S A* 101:2900–2905. doi:[10.1073/pnas.0308768101](https://doi.org/10.1073/pnas.0308768101)
80. Sun Y, Li H, Liu Y et al (2008) Evolutionarily conserved transcriptional co-expression guiding embryonic stem cell differentiation. *PLoS One* 3:e3406. doi:[10.1371/journal.pone.0003406](https://doi.org/10.1371/journal.pone.0003406)
81. Chavez L, Bais AS, Vingron M et al (2009) In silico identification of a core regulatory network of OCT4 in human embryonic stem cells using an integrated approach. *BMC Genomics* 10:314. doi:[10.1186/1471-2164-10-314](https://doi.org/10.1186/1471-2164-10-314)
82. Trott J, Hayashi K, Surani A et al (2012) Dissecting ensemble networks in ES cell populations reveals micro-heterogeneity underlying pluripotency. *Mol Biosyst* 8:744–752. doi:[10.1039/c1mb05398a](https://doi.org/10.1039/c1mb05398a)
83. Tan MH, Au KF, Leong DE et al (2013) An Oct4-Sall4-Nanog network controls developmental progression in the pre-implantation mouse embryo. *Mol Syst Biol* 9:632. doi:[10.1038/msb.2012.65](https://doi.org/10.1038/msb.2012.65)
84. Walker E, Ohishi M, Davey RE et al (2007) Prediction and testing of novel transcriptional networks regulating embryonic stem cell self-renewal and commitment. *Cell Stem Cell* 1:71–86. doi:[10.1016/j.stem.2007.04.002](https://doi.org/10.1016/j.stem.2007.04.002)
85. Gu P, Reid JG, Gao X et al (2008) Novel microRNA candidates and miRNA-mRNA pairs in embryonic stem (ES) cells. *PLoS One* 3:e2548. doi:[10.1371/journal.pone.0002548](https://doi.org/10.1371/journal.pone.0002548)
86. Markowitz F, Mulder KW, Airoidi EM et al (2010) Mapping dynamic histone acetylation patterns to gene expression in nanog-depleted murine embryonic stem cells. *PLoS Comput Biol* 6:e1001034. doi:[10.1371/journal.pcbi.1001034](https://doi.org/10.1371/journal.pcbi.1001034)
87. Teif VB, Vainshtein Y, Caudron-Herger M et al (2012) Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol* 19:1185–1192. doi:[10.1038/nsmb.2419](https://doi.org/10.1038/nsmb.2419)
88. Mah N, Wang Y, Liao MC et al (2011) Molecular insights into reprogramming-initiation events mediated by the OSKM gene regulatory network. *PLoS One* 6:e24351. doi:[10.1371/journal.pone.0024351](https://doi.org/10.1371/journal.pone.0024351)
89. Qin H, Diaz A, Blouin L et al (2014) Systematic identification of barriers to human iPSC generation. *Cell* 158:449–461. doi:[10.1016/j.cell.2014.05.040](https://doi.org/10.1016/j.cell.2014.05.040)
90. Hassani SN, Totonchi M, Gourabi H et al (2014) Signaling roadmap modulating naive and primed pluripotency. *Stem Cells Dev* 23:193–208. doi:[10.1089/scd.2013.0368](https://doi.org/10.1089/scd.2013.0368)
91. Peterson H, Abu Dawud R, Garg A et al (2013) Qualitative modeling identifies IL-11 as a novel regulator in maintaining self-renewal in human pluripotent stem cells. *Front Physiol* 4:303. doi:[10.3389/fphys.2013.00303](https://doi.org/10.3389/fphys.2013.00303)
92. Mathew S, Sundararaj S, Mamiya H, Banerjee I (2014) Regulatory interactions maintaining self-renewal of human embryonic stem cells as revealed through a systems analysis of PI3K/AKT pathway. *Bioinformatics* 30:2334–2342. doi:[10.1093/bioinformatics/btu209](https://doi.org/10.1093/bioinformatics/btu209)
93. Lutter D, Bruns P, Theis FJ (2012) An ensemble approach for inferring semi-quantitative regulatory dynamics for the differentiation of mouse embryonic stem cells using prior knowledge. *Adv Exp Med Biol* 736:247–260. doi:[10.1007/978-1-4419-7210-1\\_14](https://doi.org/10.1007/978-1-4419-7210-1_14)
94. Cahan P, Li H, Morris SA et al (2014) Cell net: network biology applied to stem cell engineering. *Cell* 158:903–915. doi:[10.1016/j.cell.2014.07.020](https://doi.org/10.1016/j.cell.2014.07.020)
95. Warsow G, Greber B, Falk SSI et al (2010) ExprEssence – revealing the essence of differential experimental data in the context of an interaction/regulation network. *BMC Syst Biol* 4:164. doi:[10.1186/1752-0509-4-164](https://doi.org/10.1186/1752-0509-4-164)
96. Cline MS, Smoot M, Cerami E et al (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2:2366–2382. doi:[10.1038/nprot.2007.324](https://doi.org/10.1038/nprot.2007.324)
97. Sarda S, Hannenhalli S (2014) Next-generation sequencing and epigenomics research: a hammer in search of nails. *Genomics Inform* 12:2–11. doi:[10.5808/GI.2014.12.1.2](https://doi.org/10.5808/GI.2014.12.1.2)
98. Dowell KG, Simons AK, Wang ZZ et al (2013) Cell-type-specific predictive network yields novel insights into mouse embryonic stem cell self-renewal and cell fate. *PLoS One* 8:e56810. doi:[10.1371/journal.pone.0056810](https://doi.org/10.1371/journal.pone.0056810)
99. Dowell KG, Simons AK, Bai H et al (2014) Novel insights into embryonic stem cell self-renewal revealed through comparative human



- and mouse systems biology networks. *Stem Cells* 32:1161–1172
100. Xu H, Baroukh C, Dannenfels R et al (2013) ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database* (Oxford) 2013:bat045. doi:[10.1093/database/bat045](https://doi.org/10.1093/database/bat045)
  101. Franceschini A, Szklarczyk D, Frankild S et al (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41:D808–D815. doi:[10.1093/nar/gks1094](https://doi.org/10.1093/nar/gks1094)
  102. Guan Y, Myers CL, Lu R et al (2008) A genomewide functional network for the laboratory mouse. *PLoS Comput Biol* 4:e1000165. doi:[10.1371/journal.pcbi.1000165](https://doi.org/10.1371/journal.pcbi.1000165)
  103. Hackett JA, Surani MA (2014) Regulatory principles of pluripotency: from the ground state up. *Cell Stem Cell* 15:416–430. doi:[10.1016/j.stem.2014.09.015](https://doi.org/10.1016/j.stem.2014.09.015)
  104. Kondoh H, Leonart ME, Nakashima Y et al (2007) A high glycolytic flux supports the proliferative potential of murine embryonic stem cells. *Antioxid Redox Signal* 9:293–299. doi:[10.1089/ars.2006.1467](https://doi.org/10.1089/ars.2006.1467)
  105. Xu X, Duan S, Yi F et al (2013) Mitochondrial regulation in pluripotent stem cells. *Cell Metab* 18:325–332. doi:[10.1016/j.cmet.2013.06.005](https://doi.org/10.1016/j.cmet.2013.06.005)
  106. Varum S, Rodrigues AS, Moura MB et al (2011) Energy metabolism in human pluripotent stem cells and their differentiated counterparts. *PLoS One* 6:e20914. doi:[10.1371/journal.pone.0020914](https://doi.org/10.1371/journal.pone.0020914)
  107. Li C, Donizelli M, Rodriguez N et al (2010) BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol* 4:92. doi:[10.1186/1752-0509-4-92](https://doi.org/10.1186/1752-0509-4-92)
  108. Lloyd CM, Halstead MDB, Nielsen PF (2004) CellML: its future, present and past. *Prog Biophys Mol Biol* 85:433–450. doi:[10.1016/j.pbiomolbio.2004.01.004](https://doi.org/10.1016/j.pbiomolbio.2004.01.004)
  109. Snoep JL, Olivier BG (2002) Java Web Simulation (JWS); a web based database of kinetic models. *Mol Biol Rep* 29:259–263
  110. Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villéger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H (2009) The systems biology graphical notation. *Nat Biotechnol* 27(8):735–741. doi:[10.1038/nbt.1558](https://doi.org/10.1038/nbt.1558)

# Part IV

## Tools and Methodologies



## Network-Assisted Disease Classification and Biomarker Discovery

Sonja Strunz, Olaf Wolkenhauer, and Alberto de la Fuente

### Abstract

Developing improved approaches for diagnosis, treatment, and prevention of diseases is a major goal of biomedical research. Therefore, the discovery of biomarker signatures from high-throughput “omics” data is an active research topic in the field of bioinformatics and systems medicine. A major issue is the low reproducibility and the limited biological interpretability of candidate biomarker signatures identified from high-throughput data. This impedes the use of discovered biomarker signatures into clinical applications. Currently, much focus is placed on developing strategies to improve reproducibility and interpretability. Researchers have fruitfully started to incorporate prior knowledge derived from pathways and molecular networks into the process of biomarker identification. In this chapter, after giving a general introduction to the problem of disease classification and biomarker discovery, we will review two types of network-assisted approaches: (1) approaches inferring activity scores for specific pathways which are subsequently used for classification and (2) approaches identifying subnetworks or modules of molecular networks by differential network analysis which can serve as biomarker signatures.

**Key words** Biomarker discovery, Classification, Feature selection, Pathways, Molecular networks

---

## 1 Introduction

A major goal of biomedical research is the development of better approaches for diagnosis, treatment, and prevention of diseases. In this context, researchers often seek biological parameters, so-called biomarkers, which are indicative for specific health or disease characteristics [1]. As defined by the Biomarkers Definitions Working Group, a biomarker is a “characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention” [2]. The applications of such indicators are manifold: they can be utilized (a) to recognize an overt disease, (b) to screen a subclinical disease, (c) to predict the course of a disease including the patients’ response to therapeutic interventions, (d) to estimate the risk of developing a disease, or (d) to categorize a disease severity [1].

The discovery of prognostic or diagnostic biomarkers is thus a major step toward personalized medicine.

Whereas many of the already well-established, clinically relevant biomarkers are single molecular species or physiological characteristics, modern molecular biology provides data sources for predictive signatures combining multiple molecular constituents. The discovery of molecular signatures from high-throughput “omics” data is an active research topic in the field of bioinformatics [3]. A typical workflow for high-throughput data analysis and biomarker discovery, at which biologists and bioinformaticians are equally involved, is illustrated in Fig. 1. Ideally, the experimentally validated results obtained within this process enhance the understanding of the biological system under study and motivate the design of new experiments. The depicted procedure is basically applicable to many types of “omics” data, even though single steps may vary in their implemented approaches. In what follows, each step is discussed in more detail, whereby the description is mainly dedicated to the analysis of gene expression data.

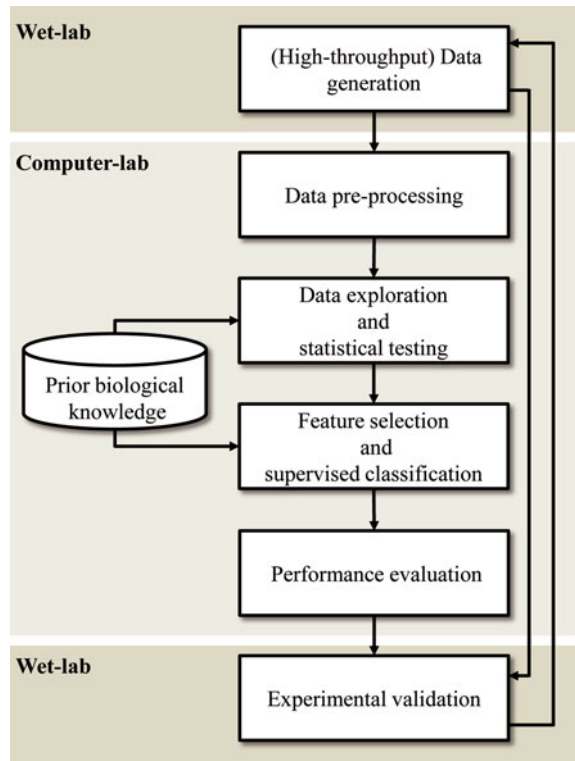
---

## 2 Computational Concepts

After high-throughput data generation, there are four main steps for computational data analysis and biomarker discovery which have to be passed: (1) data pre-processing, (2) data exploration and statistical testing, (3) feature selection and supervised classification, and (4) performance evaluation (*see* Fig. 1).

### 2.1 Data Pre-processing

First, the experimental raw data have to be processed. Typical issues of data pre-processing include quality assessment, removal of systematic sources of variation, scaling of raw data, and the detection of outliers. For microarray experiments, the pre-processing normally consists of (a) background subtraction, which is based on the assumption that the measured signal intensity is composed of the fluorescence of the spot and some background noise, (b) scaling of signal intensities by log-transformation, and (c) data normalization. Whereas the need for background correction has been controversially discussed [4], scaling and normalization are inherent parts of each analysis. Microarray experiments are subject to multiple sources of technical variations, including differences in mRNA preparation, cDNA labeling or hybridization efficiency, which can considerably limit the biological interpretability of the data [5]. Normalization is a means to adjust for such effects of technical errors within and between different arrays. A variety of different techniques for normalization has been developed, but the choice for the most appropriate method is dependent on both the context of the performed study and the array technology used for gene expression profiling [6].



**Fig. 1** Generic workflow for high-throughput data analysis and biomarker discovery

## 2.2 Data Exploration and Statistical Testing

After pre-processing the raw data, the first step of high-throughput data analysis is often to investigate the underlying structure of the given data in an explorative way by cluster analysis. The goal of cluster analysis is to divide the measured data into groups (i.e. clusters) of similar data points, whereby the data points within one group are more similar to each other than data points of distinct groups. For a microarray experiment, in which genes are measured under different conditions, either genes with similar expression patterns across various conditions or samples with similar expression patterns across the measured transcriptome are identified. A review of existing clustering methods for gene expression data is given by Jain [7].

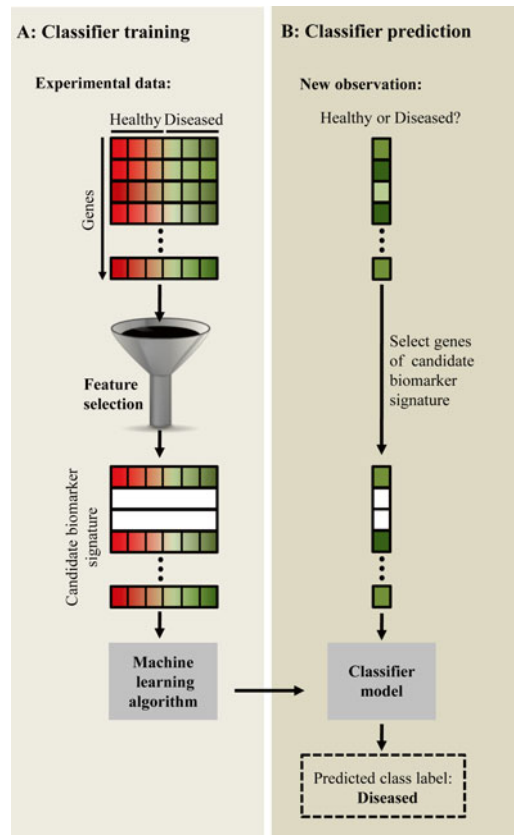
The identification of differentially expressed genes, e.g. genes which show significant changes in gene expression between different conditions, is one of the basic goals of microarray experiments. To this end, statistical testing has to be carried out and often the main difficulty lies in the choice of the appropriate test statistic, which is most of all dependent on the experimental design (e.g. number of conditions and influencing factors, repeated measurements) and the nature of the underlying data.

A detailed discussion of this issue is beyond the scope of this introduction and interested readers are referred to the review by Cui and Churchill [8].

When testing for differentially expressed genes, two types of errors can occur: either a gene is declared as differentially expressed when it is not (i.e. Type I error) or a truly differentially expressed gene is not identified as such (i.e. Type II error). Since each statistical test has a specified Type I error probability, the chance of committing some Type I errors increases with the number of hypothesis tested [9]. For this reason, in the context of microarray analysis, a considerable number of genes may be identified as differentially expressed simply by chance. For a scenario of a large number of simultaneously tested hypothesis, the Type I error rate can be controlled by applying multiple testing procedures. Dudoit et al. [9] discussed different approaches for multiple hypothesis testing in the context of microarray experiments and compared the procedures on microarray and simulated data sets. The discovery of genes with similar expression patterns across various conditions by clustering, or the identification of differentially expressed genes by statistical testing, is just an intermediate step toward the in-depth understanding of biological systems under study. The functional interpretation of the results is indispensable. The systematized knowledge about gene and protein function provided by the Gene Ontology (GO) consortium [10] or the information of biological pathway databases, like Kyoto Encyclopedia of Genes and Genomes (KEGG) [11] are just two of a myriad of available resources to derive biologically meaningful results from high-throughput data experiments.

### **2.3 Feature Selection and Supervised Classification**

Given a high-dimensional data set representing two or more biological conditions (such as healthy and diseased), the goal of biomarker discovery is to find a subset of all measured features, with which the conditions under study, also referred to as classes or class labels, can be accurately predicted. Dependent on the underlying data, the features can be, for example, proteins, metabolites or, if gene expression profiling was performed in the first step, genes. If more than one feature is selected for class prediction, one often speaks of a molecular signature functioning as a biomarker. For biomarker discovery, feature selection methods in combination with supervised classification algorithms are commonly used. The basic principle, as illustrated in Fig. 2, is that with a feature selection method a subset of features is extracted from data for which the classes are known. The measured values of these features (e.g. gene expression levels) together with their class labels (i.e. the condition under which they were measured) serve as an input for the supervised classification algorithm in order to train a classifier. Classifier training is the process of learning a set of rules or a mathematical model, which can then be used to



**Fig. 2** Scheme of a supervised classification procedure of using biomarker signatures for outcome prediction. **(a)** A subset of features is selected from the given experimental data by applying a feature selection method. This subset represents the candidate biomarker signature. In the here presented example, a gene expression signature is selected from microarray samples of patients which are either healthy or diseased. The measured values of the signature (e.g. gene expression levels) together with their class labels (i.e. the condition under which they were measured, like healthy and diseased) serve as an input for the supervised classification algorithm in order to train a classifier. **(b)** The trained classifier model can be utilized to predict the class labels of new observations. This process is called classifier prediction. Here, the new observation is a microarray measurement of a patient with unknown health status. From this all features of the candidate biomarker signature are selected. The classifier model predicts based on the measurements of the selected features the class label of the new observation (e.g. *diseased*)

predict the class labels of new observations. The prediction of class labels of new observations, by which the features but not the biological conditions under which they were measured are known, is called classifier prediction.

The extraction of feature subsets for classification has several advantages: it reduces the dimensionality of high-throughput data,

limits the risk of overfitting during classification (see next section for explanation), supports a computationally faster classifier training, and may help to gain a deeper insight into the underlying processes that generated the data [12]. The existing feature selection methods can be categorized into three main approaches, namely the filter, wrapper, and embedded methods [13]. Filter techniques score the relevance of individual features or feature subsets without incorporating any classification scheme. Typical scoring procedures include statistical testing or the analysis of relationships (e.g. correlations) between feature measurements and class labels. The features are ranked according to their relevance scores and a certain set of best-ranked features are selected for subsequent classifier training. Wrapper techniques, on the other hand, traverse the space of possible feature subsets and evaluate their prediction performance by applying a predefined supervised machine learning algorithm for classifier training. The space of possible feature subsets is often pruned by heuristic search algorithms and the subset associated with the highest performance is then selected. The third category of methods, the embedded techniques, directly integrates the search and selection of feature subsets into the process of classifier training, which means that feature selection and classifier training cannot be separated from each other. For all categories many approaches have been proposed, differing in their complexity and thereby in their interpretability. As will be discussed in more detail later in the chapter, additional biological information can be utilized to support the process of feature selection. The integration of prior biological knowledge from external repositories, such as pathway information, aims at a reduction of the dimensionality by only selecting features which are known to be relevant in the given biological context. Once a subset of features is identified, their measurements and class labels serve as an input for classifier training by applying supervised classification algorithms. Supervised classification algorithms can be statistical methods like discriminant analysis or machine learning techniques. For the latter, numerous approaches have been proposed in the last decades, which range from simple approaches like the instance-based learning methods to more complex approaches like support vector machines (SVMs) and random forests.

For a more detailed description of other machine learning algorithms, like support vector machines, decision trees, neuronal networks or Bayesian methods, including a discussion on their individual advantages and shortcomings, the reader is referred to the reviews by Kotsiantis et al. [14] and Larrañaga et al. [15]. The choice of the most appropriate learning algorithm is a critical step and highly dependent on the underlying data and context. Systematic evaluation studies comparing different supervised learning approaches applied to different data sets may support this decision [16, 17].

## 2.4 Performance Evaluation

The fourth step of the here described biomarker discovery pipeline is, as illustrated in Fig. 1, the evaluation of the classifier performance. The performance is a measure to describe how well a classifier discriminates the classes, i.e. predicts the class labels of new observations. A whole bundle of different evaluation measures has been proposed, all of them assessing different characteristics of a classification. The interested reader is referred to the work of Sokolova et al. [18] and Sokolova and Lapalme [19] in which the commonly accepted performance measures are compared and discussed in more detail.

A key concept of the performance evaluation of supervised classifications is that the performance has to be evaluated with data which were not used for feature selection and classifier training before. Only when using a disjoint data set a reliable conclusion can be drawn that the learned characteristics of the data in the training phase can be generalized to new observations. The use of distinct data sets for classifier training and performance evaluation (i.e. classifier testing) is often a concern in practical applications because the amount of experimental data is usually limited. Especially in high-throughput experiments, the number of measured samples, i.e. conditions, is far smaller than the number of measured features. Such data sets are prone to a phenomena called overfitting, meaning that a classifier is highly specialized for its training examples together with its peculiarities which do not represent the structure of the total set of data. Overfitted classifiers appear to have a good performance on the training data but show only limited capabilities of generalization and are therefore unable to accurately predict new observations. A common strategy to deal with the limited amount of data and with the problem of overfitting is to perform cross-validation procedures [20]. Here, a classifier is trained on a subset of data (called training set) and tested on the remainder (called test set). Repeating this systematically by using different partitioning of the data, cross-validation has the potential of employing the entire training set for testing, albeit not at once, and simultaneously creating the largest possible test set for a fixed training set [21]. By aggregating the classifier performances obtained for each pair of training and test sets, an overall performance for the classifier can be obtained. Depending on the way the data are partitioned into training and test sets, various types of cross-validation procedures can be distinguished. Some of the most common types are hold-out cross-validation,  $k$ -fold cross-validation or leave one-out cross-validation. The hold-out cross-validation partitions the data into exactly two subsets, one for training and one for testing the classifier. It is common practice to use two-thirds of the data for classifier training and the remaining one-third for classifier testing. For the  $k$ -fold cross-validation, the data are partitioned into  $k$  randomly selected subsets of approximately equal size. Each subset is utilized once as a test set, whereas the remaining ones function as the corresponding training



set. Thus classifier training and testing is repeated  $k$  times. In practice  $k$  is often set to 10. If the data are partitioned in a way that the class proportion of each subset (e.g. the proportion of healthy and diseased samples) is approximately the same as for the entire data set, one speaks of a stratified cross-validation. If the number of subsets  $k$  is equal to the number of observations, a leave one-out cross-validation is realized.

This splitting approach for performance evaluation is called internal validation. After a potential biomarker signature (i.e. feature subset) is extracted and its ability for class prediction was evaluated by an internal validation strategy, it is recommended to employ an additional external validation. An external validation, comprising of a classifier prediction with a distinct data set of new biological samples, supports drawing meaningful conclusions on the capability of generalization of the identified biomarker signature.

In the next step, the potential biomarker signature should be experimentally verified. For example, if a potential gene signature was identified based on microarray gene expression data, it is common practice to additionally confirm the gene expression alterations with quantitative real-time polymerase chain reaction (qRT-PCR) using an independent, newly generated data set.

---

### 3 Current Limitations and Challenges of Biomarker Discovery

A major obstacle, researchers are confronted with, is the low reproducibility and the limited biological interpretability of candidate biomarker signatures identified from high-throughput data. The lack of reproducibility of candidate biomarker signatures across different data sets often impedes their transfer into clinical applications [22]. It is therefore a fundamental challenge of current biomedical research to develop strategies to overcome this limitation. In this context, the concept of stable feature selection recently gained importance in the field of computational biomarker discovery. For high-dimensional data, an important step toward the identification of potential biomarkers is to rank features (e.g. genes) according to their relevance or importance. Based on such ranked lists, the final set of candidate biomarkers, like the top- $k$  ranked features, is often selected. The term *stability* refers to the similarity of ranked lists obtained either by applying the same feature selection method to slightly modified versions of the underlying data set (e.g. using different subsamples of the original data set) or by applying different feature selection methods on the same data set [23]. It has previously been shown that genes which are selected for outcome prediction are highly dependent on the training samples generated by a resampling strategy [24]. High stability of features with respect to sampling variations is a good indicator for biomarker reproducibility [25], since markers which are

tolerant against variations within data have a higher chance of having a high discriminatory power for experimental data from different studies generated in different laboratories.

In the study by He and Yu [25], three major causes of feature instability are mentioned:

1. The existence of multiple true markers, all allowing an equally accurate outcome prediction, is one main cause of instability.
2. The small number of observations compared to the extremely high number of measured features in high-dimensional data can lead to instability [25, 26]. As demonstrated by Kim [27], the overlap between independently developed gene signatures increases linearly with more observations. Based on a newly developed mathematical model, Ein-Dor et al. [26] concluded that as a minimum, thousands of observations are needed to achieve a typical overlap of 50 % between two predictive lists of genes obtained from breast cancer studies.
3. As the third cause of instability, He and Yu [25] name the application of algorithms that are primarily designed to select feature subsets providing the best prediction accuracy and do not explicitly attach importance to the stability.

Whereas researchers have only partial influence on the first two causes, the possibilities to improve the algorithmic design to support biomarker stability are manifold. One promising idea to improve not only the reproducibility of biomarkers but also their functional interpretation is to incorporate prior biological knowledge into the process of feature selection [28, 29]. It has been shown that the integration of secondary data sources, such as pathway and molecular network information, has great potential to overcome the above-mentioned limitations of current disease-related biomarker research [30]. In the remainder of this chapter, we will discuss methods for pathway and network-assisted disease classification and biomarker discovery.

---

## 4 Pathways and Networks

Three types of prior network information are relevant to the here discussed methods: (1) pathways, (2) protein interaction networks, and (3) gene co-expression networks. This section will provide some details on these resources.

### 4.1 Pathway Resources

“Pathways” is used as a general term typically referring to metabolic pathways and signal transduction pathways. Pathways are sets of sequential biochemical reactions, such as substrate to product conversion, in the case of metabolic pathways, or post-translational modifications, such as phosphorylation reactions, in the case of

signal transduction pathways. In the context of the present discussion, pathways are essentially lists of genes involved in a common set of such biochemical reactions.

In recent years a lot of effort has been made to catalog our knowledge about functional relationships into pathways, and this knowledge is made available through online resources. Frequently used pathway sources are KEGG [11]—<http://www.genome.jp/kegg/>, Reactome [31]—<http://www.reactome.org/>, the MSigDB (Molecular Signature Database) [32]—<http://www.broadinstitute.org/msigdb>, Pathguide [33]—<http://www.pathguide.org>, Biocarta—<http://www.biocarta.com/>, PID [34]—<http://pid.nci.nih.gov/>, and the GO annotation of gene products [10]. Many other pathway resources are available [35–38], each having their own representation and conventions. BioPAX (Biological Pathway Exchange—<http://www.biopax.org>) is an attempt to integrate these resources by defining an open file format specification for the exchange of biological pathway data [39].

## **4.2 Protein Interaction Networks**

Protein interaction networks are networks in which the nodes correspond to proteins and undirected edges represent pair-wise binding interactions. Such networks are typically established with high-throughput techniques, such as the yeast two-hybrid system [40, 41] and tandem affinity purification tagging [42–44], through literature mining [45, 46], and computational predictions based on, e.g. sequence and genomic information [47]. Several online resources could be consulted for these networks, such as STRING [48]—<http://string-db.org/>, BioGRID (The Biological General Repository for Interaction Data sets) [49]—<http://www.thebiogrid.org/>, DIP (Database of Interacting Proteins) [50]—<http://dip.doe-mbi.ucla.edu/>, HPRD (Human Protein Reference Database) [51]—<http://www.hprd.org/>, MINT (the Molecular INteraction database) [52]—<http://mint.bio.uniroma2.it/mint/>, IntAct [53]—<http://www.ebi.ac.uk/intact/>, and UniProt [54]—<http://www.uniprot.org/>.

## **4.3 Gene Co-expression Networks**

Gene co-expression networks are networks in which the nodes correspond to gene expression levels and the undirected edges represent pair-wise associations between them. Such networks are typically inferred from gene expression data. Association measures, such as the Pearson or Spearman correlation [55], and mutual information [56], can be employed to derive a gene co-expression network. Several techniques to prune edges from co-expression networks have been proposed, in order to distinguish between “direct” and “indirect” edges [57–60]. The indirect edges in co-expression networks appear due to transient causal effects (e.g. a path  $A \rightarrow B \rightarrow C$  will give rise to correlations between A and C) or confounding effects (e.g. a common regulator B,  $A \leftarrow B \rightarrow C$ , will give rise to correlations between A and C). In the

context of the here discussed work it is relevant to consider changes in the co-expression structure between disease states, as these reflect changes in regulatory networks possibly relevant to the disease phenotypes [61].

---

## 5 Methods for Pathway and Network-Based Biomarker Discovery

High-throughput data are a veritable treasure trove for biomarker discovery. Often thousands of transcripts, gene products, or metabolites are simultaneously measured under multiple conditions and time points. When researchers started to utilize these data for the identification of putative biomarker signatures, they often selected the candidate features independent from each other and thus neglected that they may interact with each other. By ignoring potential dependencies between the features, the selected signature may contain redundant information which impedes a synergistic improvement of the discriminatory power [62]. For this reason, researchers started to incorporate prior knowledge derived from pathways and molecular networks into the process of biomarker identification.

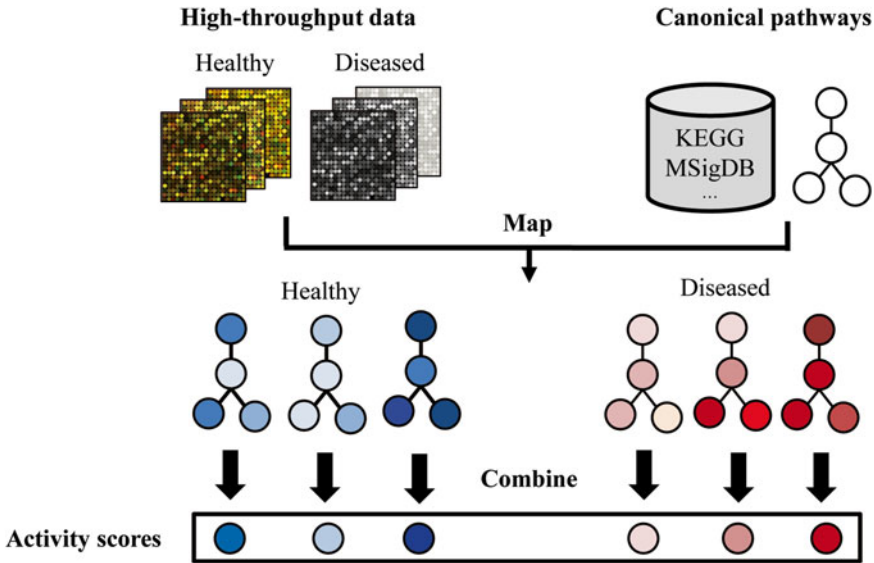
Cun and Fröhlich proposed to categorize current methods for pathway and network-assisted biomarker discovery into network-centric and data-centric approaches [30]. In contrast to the data-centric approaches, network-centric approaches do not directly integrate the knowledge-driven feature selection into the machine learning procedure. Thus, feature selection, incorporating prior information about pathways and molecular networks, and the supervised or unsupervised classification are uncoupled processes [30].

In the following, we will introduce two types of network-centric approaches: (1) approaches inferring activity scores for specific pathways which are subsequently used for classification and (2) approaches identifying subnetworks or modules of a (reconstructed) molecular network by differential network analysis which can serve as biomarker signatures.

Numerous other methods, as for example the identification of dynamical network biomarkers, have been published in the last years. Interested readers will find extensive information on those methods and classification schemes in the valuable reviews by Cun and Fröhlich [30], Zhao and Guimin [63], and Zeng et al. [64].

### 5.1 Pathway Activity Analysis

Methods of this category use two types of input: First, the experimental data, which are most frequently gene expression measurements, and second, the information about curated pathways. They combine the measurements, mapped onto the members of a pathway, to a pathway activity score. By combining the expression levels of numerous pathway genes into only one pathway activity score, a dimension reduction is achieved. Instead of using individual gene



**Fig. 3** Schematic representation of pathway activity-based methods. First, the methods map experimental data, such as microarray gene expression data, onto the members of a pathway of interest. The different colors of the pathway members illustrate that they are mapped to gene expression measurements ranging from low (indicated as *blue*) to high (indicated as *red*) levels. Second, the expression levels of all pathway members are combined (e.g. averaged) to one pathway activity score. The pattern of activity scores, also called composite features, obtained from different samples measured under different conditions and/or obtained from different pathways can be used for disease classification

activities as features for classification, the pathway activity scores are used. The latter can be referred to as composite-feature classification [65]. A schematic representation of pathway activity analysis is given in Fig. 3. The principle difference between the methods of this category is within the way of calculating the pathway activity. In addition, different supervised and unsupervised classification schemes as well as different pathway resources are incorporated (see Table 1).

One of the first studies, in which a pathway activity score is calculated from GO functional modules was published by Guo et al. [66]. To discriminate between different tumor types, the authors defined the pathway activity score as the mean or median expression value of genes which are part of GO modules enriched with differentially expressed genes. By applying this method to four publicly available microarray gene expression data sets, the authors were able to discriminate between different tumor types. However, averaging the gene expression values is obviously a very simplistic way of summarizing coherent pathway expression patterns and valuable information for prediction may be lost [62, 67].

Another way to infer a pathway activity score from a predefined set of pathway genes is to employ principal component analysis (PCA). Here, the pathway activity score is defined as the weighted

**Table 1**  
**Methods for biomarker discovery using pathway activity analysis**

Method	Pathway activity measure	Pathway resource	Classification method
Guo et al. [66]	Mean, median	GO	Decision tree
Tomfohr et al. [68]	PCA	KEGG, Biocarta	–
Bild et al. [70]	PCA	–	Binary probit regression model
Liu et al. [69]	PCA	MSigDB	SVM
Lee et al. [71]; Yang et al. [72]	Z-score of condition-responsive genes	MsigDB	Logistic regression, SVM
Su et al. [62]	Log-likelihood ratios	MsigDB	Logistic regression, linear discriminant analysis
Vaske et al. [73]	Estimated parameters of factor graph models	PID	Unsupervised clustering
Kim et al. [77]	GSEA (leading edge genes)	MsigDB, KEGG, manually curated data sets	Hierarchical SVM
Pyatnitskiy et al. [78]	SNEA	ResNet	Unsupervised clustering

sum of standardized expression levels of all pathway genes, whereas the weights are given by the first principal component [68]. It has been shown that PCA-derived activity scores can be used to discriminate between cancer and non-cancer samples [69] as well as between specific cancer and tumor subtypes [70]. By constructing pathway interaction networks, where each node represents a pathway and each edge represents an inter-pathway interaction, the PCA-based method by Liu et al. also considers cross-talks between pathways [69].

Summarizing the expression values of all genes belonging to the same pathway might be problematic, since not all genes of a specific pathway need to be associated with the disease under study and even if a pathway gene is disease-associated, the mRNA expression level is not necessarily altered. For this reason, methods were proposed in which only a subset of all pathway genes are used to infer pathway activity scores [71, 72]. Using a heuristic search method, Lee et al. identified for each pathway a subset of all member genes, whose combined expression delivers optimal discriminative power for the disease phenotype [71]. For each sample, the expression levels of the subset genes, which are called condition-responsive genes, are summarized as a combined Z-score. An alternative approach is the probabilistic inference of pathway activity scores. Su et al. developed a probabilistic framework for the inference of pathway activities by focusing on the difference in

distribution of gene expression levels under different conditions [62]. By summing up the log-likelihood ratios between two disease phenotypes obtained for each gene, a probabilistic indicator which phenotype is more likely based on the pathway expression levels is calculated. Another probabilistic approach was proposed by Vaske et al. [73]. Using factor graphs the authors converted pathways into distinct probabilistic models and predicted the degree to which a pathway's activity is altered between different disease phenotypes by incorporating known pathway interactions.

Several approaches for pathway-based biomarker discovery utilize the output of gene set enrichment algorithms, such as GSEA [74], Signaling Pathway Impact Analysis (SPIA) [75], Subnetwork Enrichment Analysis algorithm (SNEA), or Differential Expression Analysis for Pathways (DEAP) [76]. Kim et al., for example, proposed a method which utilizes GSEA to automatically infer pathway activity scores [77]. Specifically, a weighted linear combination of gene expression levels of leading edge genes as identified by GSEA was used to train hierarchical SVMs for disease prediction. In a more recent study, SNEA, a method for constructing de novo user-defined subnetworks from a global literature-extracted protein-protein regulation network, was used to identify significant regulators of subnetworks. Afterwards, pathway activity scores for each cluster of regulators were calculated by summing up the median expression values of all downstream targets multiplied by the number of targets for each regulator within the cluster [78].

The integration of pathway information into the process of feature selection has become a valuable tool for biomarker discovery. However, most of the above-mentioned approaches share two major limitations:

First, our knowledge about pathways, their members and the interactions between them, is constantly increasing but far from being complete; so far, pathways cover only a small fraction of all known human proteins. Since this type of methods is not designed to identify novel relevant gene sets, the limited coverage of pathways can lead to the fact that a true or optimal biomarker cannot be identified. To employ protein-protein interaction networks instead of curated pathway information for biomarker identification is one way to at least partially address this challenge [79]. Chuang et al., for example, integrated protein-protein interaction networks with gene expression data to identify subnetworks with which a metastatic and non-metastatic disease status can be classified [79]. For this purpose, an activity score was calculated for each subnetwork based on the expression values of the member genes. Using a greedy search algorithm, discriminative subnetworks were identified to infer a subnetwork activity matrix for classification.

Second, many (but not all) methods neglect the topology of the pathways, which means they only consider expression changes in a set of pathway genes, but do not explicitly incorporate



information about the network structure, such as correlations between the pathway genes. Recently, DINA, a network-based algorithm, was developed to determine whether genes in a known pathway are significantly co-regulated in specific conditions, but not in others [80]. Thus DINA incorporates topological information by identifying differentially co-regulated pathways, but since it is based on known pathway information interesting de novo sub-networks cannot be identified by this method.

The concept of differential network analysis for biomarker discovery tackles the above-mentioned problems. The methods focus on the identification of changes in the topology of condition-specific molecular networks. For network analysis, they usually integrate transcriptome data with the constantly increasing information about protein–protein interactions.

## **5.2 Differential Network Analysis**

Physiological and disease phenotypes are the consequence of a complex interplay of genes. To study genes in the context of regulatory systems, modeled as molecular networks, has become a valuable research tool to elucidate the mechanisms that orchestrate the activities of gene products [61]. It has been shown that molecular interactions can change due to environmental or pharmacological stresses or disease states [81]. The dynamic nature of molecular networks allows us to compare networks across different conditions in order to identify “rewired” genes or subnetworks (i.e. modules) [82]. With the aim to identify genes or gene sets which exhibit such condition-specific interactions, differential network analysis goes beyond the conventional differential expression analysis in which significant expression changes but no differential co-regulations are detected. Differential network analysis is thus a first step toward the identification of dysfunctional regulatory systems correlated to a disease and holds great promise for biomarker discovery [61, 82].

The concept of differential network analysis for biomarker discovery and disease prediction has been employed in various studies. Liu et al. implemented a differential interaction-based approach with which they successfully identified 34 gastric cancer genes serving as network biomarkers [83]. This method takes two types of inputs: First, a microarray gene expression data set, which can be divided into different disease phases. At each phase the data were separated into control and disease samples. Second, a human protein–protein interaction network downloaded from the HPRD database. By mapping data-derived correlations to the protein–protein interaction network the method infers for each disease phase two correlation networks, one for the control and one for the disease samples. The exclusion of interactions which are present in both correlation networks results in a differential protein–protein interaction network. Genes which contribute to a differential interaction and which are expressed in both networks across all phases were used as potential biomarkers.

With the aim to develop a network-based biomarker for lung cancer prognosis, Wang and Chen also integrated gene expression and protein–protein interaction data [84]. The authors constructed two networks, a cancer and a non-cancer protein association network. In these networks, an association between two proteins quantifies the expression relation between the interacting proteins. They further calculated a carcinogenesis relevance value, indicating to which extent the protein associations differentiate between the two networks. With the differential network-based carcinogenesis relevance value, the authors identified 40 proteins significantly related to lung carcinogenesis.

Another strategy to model and detect statistically significant topological changes in molecular networks is to infer differential dependency networks [85–87]. Dependency networks are probabilistic graphical models, in which each node is assigned to a conditional probability distribution given its parents in the network [88, 89]. Recently, Tian et al. developed a novel approach in which differential dependency networks were used to infer biological networks with significant rewiring across different conditions by integrating experimental data and biological knowledge [86].

An interesting approach, which combines differential network analysis with differential expression analysis, was recently proposed by Sun et al. [90]. With the aim to investigate diabetes from a systems perspective, the authors developed a new type of molecular network, which they refer to as differential expression network. After integrating protein–protein interactions and gene expression data to a co-expression network, they identified differential and non-differential interactions. The latter are defined as interactions, which are not “rewired” between case and control samples, but whose interacting proteins exhibit a significant differential expression. Based on the two types of interactions, specific disease genes and interactions were identified which could serve as network biomarkers for disease prognosis in future.

Molecular networks are known to be structured in a modular manner, whereas modules are often defined as groups of interacting molecules driving a common cellular function [81]. Recently, the study of Islam et al. has shown that normal and cancer protein interaction networks exhibit a differential modular nature [91]. In the study it is demonstrated that cancer protein interaction networks show a higher level of clustering and a lower level modular overlapping compared to normal protein interaction networks. In line with these results, Taylor et al. proposed that altered network modularity of the human–protein interaction network can be used as an indicator for breast cancer prognosis [92]. By identifying hub genes that exhibit significant difference in the co-expression (i.e. correlation) with their interacting partners between patients who survived versus those who died from disease, the authors developed a prognostic signature to predict a good or poor disease prognosis.

---

## 6 Final Remarks

In the last decades, a lot of effort has been made in identifying predictive and prognostic molecular markers for human diseases. Especially in the light of personalized diagnostics, biomarker discovery has become an active field of research which still faces several challenges waiting to be addressed in future. One of the main problems of candidate biomarkers identified from high-throughput data is their lack of reproducibility. The poor reproducibility is undoubtedly one of the reasons why—as stated by the National Biomarker Development Alliance (NBDA)—less than 100 biomarkers are routinely used in the clinic, even though more than 150,000 were reported as “discovered” by 2012 alone (NBDA Fact Sheet). The causes for non-reproducibility are certainly manifold and some, as for example the “small  $n$  large  $p$  problem” of high-dimensional data, are difficult or impossible to avoid. From the computational point of view, it is therefore inevitable to develop new algorithms and approaches which explicitly support the identification of stable biomarker candidates. One promising approach to decrease the variability of prognostic signatures is the integration of prior biological knowledge into the process of feature selection and classification.

In this book chapter, we reviewed methods for a pathway and network-assisted discovery of biomarkers. All these methods aim at (1) an increased stability of candidate biomarkers to improve reproducibility, (2) an increased prediction accuracy, and (3) a better biological interpretability of the discovered signatures in order to gain a deeper insight into the underlying molecular mechanisms. However, it is still a matter of debate of whether pathway or network-assisted biomarker discovery results in higher accuracy and stability of the signatures identified. A comparative analysis of 14 methods used to predict early versus late relapse of breast cancer patients revealed that the eight network-assisted methods do not necessarily lead to a significantly increased prediction accuracy compared to classical prediction algorithms. However, the network-based SVM approach proposed by Zhu et al. [93] identified the most stable gene signatures, whereas the Reweighted Recursive Feature Elimination (RRFE) [94] and the computation of the average pathway activity resulted in an improved biological interpretability of the signatures identified. Two recent evaluation studies by Staiger et al., in which several network and pathway-based methods were used to predict the outcome of breast cancer, revealed that none of these methods outperform classical methods with respect to the prediction accuracy and feature stability [65, 67].

One inherent problem of methods for a pathway and network-assisted discovery of biomarkers is that gene expression is often interpreted in terms of protein levels. The integration of

experimental data with prior knowledge from different molecular levels involves the risk of obtaining erroneous results, since a gene's mRNA level is not necessarily correlated with its protein level [96, 97]. In addition, our currently incomplete knowledge about pathways and protein–protein interactions [95] may hamper a network-assisted discovery of predictive or prognostic signatures.

Finally, we want to stress the importance of a proper evaluation of the proposed methods using standardized benchmark data sets and independent evaluation protocols. The fact that each study is currently using different experimental and secondary data as well as different evaluation procedures impedes an objective assessment of classification performances and biomarker stability [65, 67]. In order to enable an objective comparison of classification and feature-selection methods, Staiger et al. recently developed the Amsterdam Classification Evaluation Suite (ACES), a python coded, publically available evaluation framework [65].

An effective strategy to verify research methods as well as to explore and analyze a specific research question is crowdsourcing. Crowdsourcing, as an engagement of an interested community to collaboratively solve a problem, gained importance in many research areas including biomedical research [98]. A prominent example of a successful academic community-wide effort to assess systems biology methods is the DREAM initiative (Dialogue on Reverse Engineering Assessments and Methods). By organizing challenges, which aim to profit from the collective intelligence, DREAM put effort, among others, to benchmark methods in cancer genomics to identify responders and non-responders to certain drug treatments or to facilitate disease prognosis [99]. Another crowdsourcing scheme, which focuses on the verification of systems biology methods and concepts in an industrial context, is sbv IMPROVER (Systems Biology Verification combined with Industrial Methodology for Process Verification in Research). These community-wide efforts and the provided benchmark data sets allow researchers to learn about the strengths and weaknesses of their methods [98] and simultaneously tackle the above-mentioned challenge of an objective and independent evaluation of methods.

In future, it will be important to investigate how the quality of prior biological knowledge influences pathway and network-assisted discovery of biomarkers. Performing extensive evaluation studies using also simulated data, as for example provided by the DREAM initiative, may help us to understand the impact of utilizing noisy, incomplete or even erroneous prior knowledge on the performance and stability of these biomarkers. This will be a prerequisite to develop methods which are capable of handling uncertainties in prior biological knowledge and thus improve the performance and reproducibility of candidate biomarkers.

## References

1. Vasan RS (2006) Biomarkers of cardiovascular disease: molecular basis and practical considerations. *Circulation* 113:2335–2362
2. Atkinson AJ, Colburn WA, DeGruttola VG et al (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 69:89–95
3. McDermott JE, Wang J, Mitchell H et al (2013) Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert Opin Med Diagn* 7:37–51
4. Zahurak M, Parmigiani G, Yu W et al (2007) Pre-processing {A}gilent microarray data. *BMC Bioinformatics* 8:142
5. Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet* 32(Suppl):496–501. doi:[10.1038/ng1032](https://doi.org/10.1038/ng1032)
6. Smyth GK, Speed T (2003) Normalization of c{DNA} microarray data. *Methods* 31:265–273
7. Jain AK (2010) Data clustering: 50 years beyond {K}-means. *Pattern Recognit Lett* 31:651–666
8. Cui X, Churchill GA (2003) Statistical tests for differential expression in c{DNA} microarray experiments. *Genome Biol* 4:210
9. Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. *Stat Sci* 18:71–103
10. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25–29. doi:[10.1038/75556](https://doi.org/10.1038/75556)
11. Kanehisa M, Goto S, Sato Y et al (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40:D109–D114. doi:[10.1093/nar/gkr988](https://doi.org/10.1093/nar/gkr988)
12. Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23:2507–2517. doi:[10.1093/bioinformatics/btm344](https://doi.org/10.1093/bioinformatics/btm344)
13. John GH, Kohavi R, Pfleger K (1994) Irrelevant features and the subset selection problem. *Proc Elev Int Conf Mach Learn* 129:121–129
14. Kotsiantis SB, Zaharakis ID, Pintelas PE (2007) Supervised machine learning: a review of classification techniques. *Front Artif Intell Appl* 160:3
15. Larrañaga P, Calvo B, Santana R et al (2006) Machine learning in bioinformatics. *Brief Bioinform* 7:86–112
16. Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 97:77–87
17. Lee JW, Lee JB, Park M, Song SH (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput Stat Data Anal* 48:869–885
18. Sokolova M, Japkowicz N, Szpakowicz S (2006) Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. *Adv Artif Intell (Lect Notes Comput Sci)* 1015–1021
19. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 45:427–437. doi:[10.1016/j.ipm.2009.03.002](https://doi.org/10.1016/j.ipm.2009.03.002)
20. Kohavi R (1995) A study of cross-validation and bootstrap for estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence*. Kaufman, Montreal, pp 1137–1143
21. Fung G, Rao RB, Rosales R (2008) On the dangers of cross-validation. An experimental evaluation (SIAM). In: Apte C, Park H, Wang K, Zaki MJ (eds) *Proceedings of the 2008 SIAM international conference on data mining*. doi:[10.1137/1.9781611972788.54](https://doi.org/10.1137/1.9781611972788.54), pp 588–596
22. Cun Y, Fröhlich H (2013) Network and data integration for biomarker signature discovery via network smoothed T-statistics. *PLoS One* 8:e73074. doi:[10.1371/journal.pone.0073074](https://doi.org/10.1371/journal.pone.0073074)
23. Boulesteix A-L, Slawski M (2009) Stability and aggregation of ranked gene lists. *Brief Bioinform* 10:556–568. doi:[10.1093/bib/bbp034](https://doi.org/10.1093/bib/bbp034)
24. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365:488–492
25. He Z, Yu W (2010) Stable feature selection for biomarker discovery. *Comput Biol Chem* 34:215–225. doi:[10.1016/j.compbiolchem.2010.07.002](https://doi.org/10.1016/j.compbiolchem.2010.07.002)
26. Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* 103:5923–5928. doi:[10.1073/pnas.0601231103](https://doi.org/10.1073/pnas.0601231103)
27. Kim S-Y (2009) Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics* 10:147. doi:[10.1186/1471-2105-10-147](https://doi.org/10.1186/1471-2105-10-147)
28. Haury A-C, Jacob L, Vert J-P (2010) Increasing stability and interpretability of gene expression signatures. *arXiv Prepr. arXiv1001.3109*
29. Sanavia T, Aiolfi F, Da San Martino G et al (2012) Improving biomarker list stability by

- integration of biological knowledge in the learning process. *BMC Bioinformatics* 13(Suppl 4): S22. doi:[10.1186/1471-2105-13-S4-S22](https://doi.org/10.1186/1471-2105-13-S4-S22)
30. Cun Y, Fröhlich H (2012) Biomarker gene signature discovery integrating network knowledge. *Biology (Basel)* 1:5–17. doi:[10.3390/biology1010005](https://doi.org/10.3390/biology1010005)
  31. Croft D, Mundo AF, Haw R et al (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.* doi:[10.1093/nar/gkt1102](https://doi.org/10.1093/nar/gkt1102)
  32. Liberzon A, Subramanian A, Pinchback R et al (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27:1739–1740. doi:[10.1093/bioinformatics/btr260](https://doi.org/10.1093/bioinformatics/btr260)
  33. Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. *Nucleic Acids Res* 34:D504–D506. doi:[10.1093/nar/gkj126](https://doi.org/10.1093/nar/gkj126)
  34. Schaefer CF, Anthony K, Krupa S et al (2009) PID: the pathway interaction database. *Nucleic Acids Res.* doi:[10.1093/nar/gkn653](https://doi.org/10.1093/nar/gkn653)
  35. Soh D, Dong D, Guo Y, Wong L (2010) Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics* 11:449. doi:[10.1186/1471-2105-11-449](https://doi.org/10.1186/1471-2105-11-449)
  36. Stobbe MD, Jansen GA, Moerland PD, van Kampen AHC (2014) Knowledge representation in metabolic pathway databases. *Brief Bioinform* 15:455–470. doi:[10.1093/bib/bbs060](https://doi.org/10.1093/bib/bbs060)
  37. Bauer-Mehren A, Furlong LI, Sanz F (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol Syst Biol* 5:290. doi:[10.1038/msb.2009.47](https://doi.org/10.1038/msb.2009.47)
  38. Wittig U, De Beuckelaer A (2001) Analysis and comparison of metabolic pathway databases. *Brief Bioinform* 2:126–142. doi:[10.1093/bib/2.2.126](https://doi.org/10.1093/bib/2.2.126)
  39. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I et al (2010) The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 28(9):935–942. doi:[10.1038/nbt.1666](https://doi.org/10.1038/nbt.1666), Epub 2010 Sep 9
  40. Walhout AJ, Vidal M (2001) High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* 24:297–306. doi:[10.1006/meth.2001.1190](https://doi.org/10.1006/meth.2001.1190)
  41. Ito T, Chiba T, Ozawa R et al (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98:4569–4574. doi:[10.1073/pnas.061034498](https://doi.org/10.1073/pnas.061034498)
  42. Krogan NJ, Cagney G, Yu H et al (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440:637–643. doi:[10.1038/nature04670](https://doi.org/10.1038/nature04670)
  43. Gavin A-C, Bösch M, Krause R et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141–147. doi:[10.1038/415141a](https://doi.org/10.1038/415141a)
  44. Pieroni E, De La Fuente Van Bentem S, Mancosu G et al (2008) Protein networking: insights into global functional organization of proteomes. *Proteomics* 8:799–816. doi:[10.1002/pmic.200700767](https://doi.org/10.1002/pmic.200700767)
  45. Hoffmann R, Valencia A (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics.* doi:[10.1093/bioinformatics/bti1142](https://doi.org/10.1093/bioinformatics/bti1142)
  46. Chen H, Sharp BM (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 5:147. doi:[10.1186/1471-2105-5-147](https://doi.org/10.1186/1471-2105-5-147)
  47. Valencia A, Pazos F (2002) Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol* 12:368–373. doi:[10.1016/S0959-440X\(02\)00333-0](https://doi.org/10.1016/S0959-440X(02)00333-0)
  48. Szklarczyk D, Franceschini A, Kuhn M et al (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* doi:[10.1093/nar/gkq973](https://doi.org/10.1093/nar/gkq973)
  49. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S et al (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* doi:[10.1093/nar/gks1158](https://doi.org/10.1093/nar/gks1158)
  50. Xenarios I, Fernandez E, Salwinski L et al (2001) DIP: the database of interacting proteins: 2001 update. *Nucleic Acids Res* 29:239–241. doi:[10.1093/nar/28.1.289](https://doi.org/10.1093/nar/28.1.289)
  51. Keshava Prasad TS, Goel R, Kandasamy K et al (2009) Human protein reference database – 2009 update. *Nucleic Acids Res* 37:D767–D772. doi:[10.1093/nar/gkn892](https://doi.org/10.1093/nar/gkn892)
  52. Licata L, Briganti L, Peluso D et al (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* doi:[10.1093/nar/gkr930](https://doi.org/10.1093/nar/gkr930)
  53. Kerrien S, Aranda B, Breuza L et al (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* doi:[10.1093/nar/gkr1088](https://doi.org/10.1093/nar/gkr1088)
  54. Apweiler R, Bairoch A, Wu CH et al (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 32:D115–D119. doi:[10.1093/nar/gkh131](https://doi.org/10.1093/nar/gkh131)
  55. D’haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16:707–726. doi:[10.1093/bioinformatics/16.8.707](https://doi.org/10.1093/bioinformatics/16.8.707)
  56. Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput* 418–429. doi:[10.1142/9789814447331\\_0040](https://doi.org/10.1142/9789814447331_0040)



57. Faith JJ, Hayete B, Thaden JT et al (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 5:e8. doi:[10.1371/journal.pbio.0050008](https://doi.org/10.1371/journal.pbio.0050008)
58. Margolin A, Wang K, Lim WK et al (2006) Reverse engineering cellular networks. *Nat Protoc* 1:662–671, doi: citeulike-article-id: 1224968
59. Schäfer J, Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21:754–764. doi:[10.1093/bioinformatics/bti062](https://doi.org/10.1093/bioinformatics/bti062)
60. De la Fuente A, Bing N, Hoeschele I, Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20:3565–3574. doi:[10.1093/bioinformatics/bth445](https://doi.org/10.1093/bioinformatics/bth445)
61. De la Fuente A (2010) From “differential expression” to “differential networking” – identification of dysfunctional regulatory networks in diseases. *Trends Genet* 26:326–333. doi:[10.1016/j.tig.2010.05.001](https://doi.org/10.1016/j.tig.2010.05.001)
62. Su J, Yoon BJ, Dougherty ER (2009) Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS One* 4:e8161. doi:[10.1371/journal.pone.0008161](https://doi.org/10.1371/journal.pone.0008161)
63. Zhao X-M, Guimin Q (2013) Identifying biomarkers with differential analysis. In: Shen B (ed) *Bioinformatics for diagnosis, prognosis and treatment of complex diseases*. Springer, Dordrecht, The Netherlands, pp 17–31
64. Zeng T, Sun S-Y, Wang Y et al (2013) Network biomarkers reveal dysfunctional gene regulations during disease progression. *FEBS J* 280:5682–5695. doi:[10.1111/febs.12536](https://doi.org/10.1111/febs.12536)
65. Staiger C, Cadot S, Györfy B et al (2013) Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front Genet* 4:289. doi:[10.3389/fgene.2013.00289](https://doi.org/10.3389/fgene.2013.00289)
66. Guo Z, Zhang T, Li X et al (2005) Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics* 6:58. doi:[10.1186/1471-2105-6-58](https://doi.org/10.1186/1471-2105-6-58)
67. Staiger C, Cadot S, Kooter R et al (2012) A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. *PLoS One*. doi:[10.1371/journal.pone.0034796](https://doi.org/10.1371/journal.pone.0034796)
68. Tomfohr J, Lu J, Kepler TB (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 6:225. doi:[10.1186/1471-2105-6-225](https://doi.org/10.1186/1471-2105-6-225)
69. Liu K-Q, Liu Z-P, Hao J-K et al (2012) Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics* 13:126. doi:[10.1186/1471-2105-13-126](https://doi.org/10.1186/1471-2105-13-126)
70. Bild AH, Yao G, Chang JT et al (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439:353–357. doi:[10.1038/nature04296](https://doi.org/10.1038/nature04296)
71. Lee E, Chuang H-Y, Kim J-W et al (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput Biol* 4:e1000217. doi:[10.1371/journal.pcbi.1000217](https://doi.org/10.1371/journal.pcbi.1000217)
72. Yang R, Daigle BJ, Petzold LR, Doyle FJ (2012) Core module biomarker identification with network exploration for breast cancer metastasis. *BMC Bioinformatics* 13:12. doi:[10.1186/1471-2105-13-12](https://doi.org/10.1186/1471-2105-13-12)
73. Vaske CJ, Benz SC, Sanborn JZ et al (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. doi:[10.1093/bioinformatics/btq182](https://doi.org/10.1093/bioinformatics/btq182)
74. Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102:15545–15550. doi:[10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102)
75. Tarca AL, Draghici S, Khatri P et al (2009) A novel signaling pathway impact analysis. *Bioinformatics* 25:75–82. doi:[10.1093/bioinformatics/btn577](https://doi.org/10.1093/bioinformatics/btn577)
76. Haynes WA, Higdon R, Stanberry L et al (2013) Correction: differential expression analysis for pathways. *PLoS Comput Biol*. doi:[10.1371/annotation/58cf4d21-f9b0-4292-94dd-3177f393a284](https://doi.org/10.1371/annotation/58cf4d21-f9b0-4292-94dd-3177f393a284)
77. Kim S, Kon M, DeLisi C (2012) Pathway-based classification of cancer subtypes. *Biol Direct* 7:21. doi:[10.1186/1745-6150-7-21](https://doi.org/10.1186/1745-6150-7-21)
78. Pyatnitskiy M, Mazo I, Shkrob M et al (2014) Clustering gene expression regulators: new approach to disease subtyping. *PLoS One*. doi:[10.1371/journal.pone.0084955](https://doi.org/10.1371/journal.pone.0084955)
79. Chuang H-Y, Lee E, Liu Y-T et al (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3:140. doi:[10.1038/msb4100180](https://doi.org/10.1038/msb4100180)
80. Gambardella G, Moretti M, de Cegli R et al (2013) Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics* 29:1776–1785, doi: citeulike-article-id:12415017, doi: [10.1093/bioinformatics/btt290](https://doi.org/10.1093/bioinformatics/btt290)
81. Mitra K, Carvunis A-R, Ramesh SK, Ideker T (2013) Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet* 14:719–732. doi:[10.1038/nrg3552](https://doi.org/10.1038/nrg3552)



82. Ideker T, Krogan NJ (2012) Differential network biology. *Mol Syst Biol*. doi:[10.1038/msb.2011.99](https://doi.org/10.1038/msb.2011.99)
83. Liu X, Liu Z-P, Zhao X-M, Chen L (2012) Identifying disease genes and module biomarkers by differential interactions. *J Am Med Inform Assoc* 19:241–248. doi:[10.1136/amiajnl-2011-000658](https://doi.org/10.1136/amiajnl-2011-000658)
84. Wang Y-C, Chen B-S (2011) A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. *BMC Med Genomics* 4:2
85. Zhang B, Li H, Riggins RB et al (2009) Differential dependency network analysis to identify condition-specific topological changes in biological networks. *Bioinformatics* 25:526–532. doi:[10.1093/bioinformatics/btn660](https://doi.org/10.1093/bioinformatics/btn660)
86. Tian Y, Zhang B, Hoffman EP et al (2014) Knowledge-fused differential dependency network models for detecting significant rewiring in biological networks. *BMC Syst Biol* 8:87. doi:[10.1186/s12918-014-0087-1](https://doi.org/10.1186/s12918-014-0087-1)
87. Zhang B, Wang Y (2012) Learning structural changes of Gaussian graphical models in controlled experiments. *Proceedings of the twenty-first conference on uncertainty in artificial intelligence*
88. Heckerman D, Chickering DM, Meek C et al (2000) Dependency networks for inference, collaborative filtering, and data visualization. *J Mach Learn Res* 1:49–75. doi:[10.1162/153244301753344614](https://doi.org/10.1162/153244301753344614)
89. Gámez J, Mateo J, Puerta J (2006) Dependency networks based classifiers: learning models by using independence tests. *Proceedings of the 3rd European workshop on probabilistic graphical models*. pp 115–122
90. Sun S-Y, Liu Z-P, Zeng T et al (2013) Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks. *Sci Rep* 3:2268. doi:[10.1038/srep02268](https://doi.org/10.1038/srep02268)
91. Islam MF, Hoque MM, Banik RS et al (2013) Comparative analysis of differential network modularity in tissue specific normal and cancer protein interaction networks. *J Clin Bioinforma* 3:19. doi:[10.1186/2043-9113-3-19](https://doi.org/10.1186/2043-9113-3-19)
92. Taylor IW, Linding R, Warde-Farley D et al (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* 27:199–204. doi:[10.1038/nbt.1522](https://doi.org/10.1038/nbt.1522)
93. Zhu Y, Shen X, Pan W (2009) Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics* 10(Suppl 1):S21. doi:[10.1186/1471-2105-10-S1-S21](https://doi.org/10.1186/1471-2105-10-S1-S21)
94. Johannes M, Brase JC, Fröhlich H et al (2010) Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics* 26:2136–2144. doi:[10.1093/bioinformatics/btq345](https://doi.org/10.1093/bioinformatics/btq345)
95. Stumpf MPH, Thorne T, de Silva E et al (2008) Estimating the size of the human interactome. *Proc Natl Acad Sci U S A* 105:6959–6964. doi:[10.1073/pnas.0708078105](https://doi.org/10.1073/pnas.0708078105)
96. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 19:1720–1730
97. Greenbaum D, Colangelo C, Williams K, Gerstein M (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 4:117. doi:[10.1186/gb-2003-4-9-117](https://doi.org/10.1186/gb-2003-4-9-117)
98. Meyer P, Alexopoulos LG, Bonk T et al (2011) Verification of systems biology research in the age of collaborative competition. *Nat Biotechnol* 29:811–815. doi:[10.1038/nbt.1968](https://doi.org/10.1038/nbt.1968)
99. Jarchum I, Jones S (2015) DREAMing of benchmarks. *Nat Biotechnol* 33:49–50. doi:[10.1038/nbt.3115](https://doi.org/10.1038/nbt.3115)

# Chapter 17

## Anatomy and Physiology of Multiscale Modeling and Simulation in Systems Medicine

Alexandru Mizeranschi, Derek Groen, Joris Borgdorff,  
Alfons G. Hoekstra, Bastien Chopard, and Werner Dubitzky

### Abstract

Systems medicine is the application of systems biology concepts, methods, and tools to medical research and practice. It aims to integrate data and knowledge from different disciplines into biomedical models and simulations for the understanding, prevention, cure, and management of complex diseases. Complex diseases arise from the interactions among disease-influencing factors across multiple levels of biological organization from the environment to molecules. To tackle the enormous challenges posed by complex diseases, we need a modeling and simulation framework capable of capturing and integrating information originating from multiple spatiotemporal and organizational scales. Multiscale modeling and simulation in systems medicine is an emerging methodology and discipline that has already demonstrated its potential in becoming this framework. The aim of this chapter is to present some of the main concepts, requirements, and challenges of multiscale modeling and simulation in systems medicine.

**Key words** Systems medicine, Complex disease, Modeling and simulation, Multiscale modeling and simulation

---

### 1 Introduction

A complex disease is a disease whose onset and progress are characterized by multiple contributing factors from the gene to the environment level [1]. Many chronic diseases are complex diseases with long duration and slow progression. Examples include cardiovascular diseases, cancer, chronic respiratory diseases, diabetes, rheumatologic diseases, Alzheimer's disease, scleroderma, Parkinson's disease, multiple sclerosis, osteoporosis, connective tissue diseases, kidney diseases, and autoimmune diseases. The burden (financial cost, morbidity, mortality) of chronic diseases is extremely high and continues to grow due to our aging population. To reduce this burden, a holistic multimodal integrated care, and multi-scale, multi-level systems approach has been proposed [2];

this approach is sometimes referred to as *systems medicine*. The medical and clinical drivers of systems medicine include [2–5]:

- Predictive, preventive, personalized, and participatory medicine (P4 medicine)
- Pharmaceutical drug and vaccine development, production, delivery, and safety combined with patient stratification approaches
- Discovery of effective diagnostics biomarkers and their multi-dimensional combination
- Methods to predict treatment response and disease progression in a given patient
- Rational design of combinatorial therapies (e.g., dosing)
- Reduction of healthcare costs through systems-based prevention, patient stratification, tailored therapies, disease management, and telemedicine (e-health) approaches

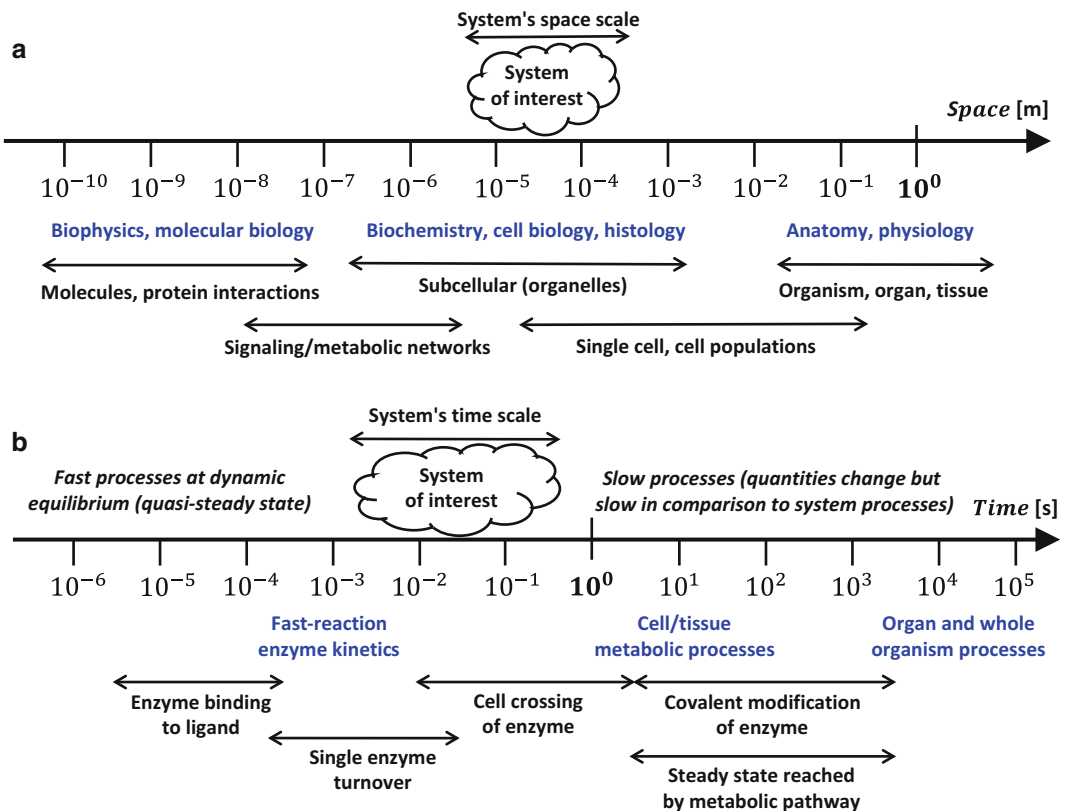
Systems medicine is the application of systems biology concepts, methods, and tools to medical research and practice. It aims to integrate data and knowledge from different disciplines into biomedical models and simulations. Systems medicine is an emerging multidisciplinary field that employs both hypothesis-driven and discovery-driven approaches to understand health and disease from a *systems* perspective [6]. The conceptualization of a complex disease as a system abandons reductionism which investigates contributing factors separately, under the assumption that they are independent of each other. In a systems view of complex disease, the relationships of system elements (disease-contributing factors) to one another make the elements *system-dependent* properties on the one hand, *and* give the system its global *emergent properties* (symptoms, abnormalities, dysfunction, anatomical and physiological changes), on the other hand.

Understanding a complex disease from a systems perspective involves the discovery of contributing factors and characterizing their contribution to the disease. One of the most important challenges of systems medicine is to understand the etiology and progression of complex diseases [2]. This is a difficult undertaking, because complex diseases are caused by a *combination* of genetic, environmental, and lifestyle factors, most of which are yet to be identified. Furthermore, genes, environment, and lifestyle are intangible variables, as they do not lend themselves to easy, accurate, and objective measurement. Another important source of complexity is the interactions among disease-influencing factors across multiple levels of biological organization from the environment to molecules.

Across the organizational levels of complex disease, we find an inherently hierarchical, modular arrangement of biological structure, function, and behavior. At each level, the organizational modules

could themselves be viewed as systems or subsystems in their own right [7]. Typically, each subsystem is defined by its own organizational anatomy and physiology, where structure and process co-produce function. The outer limits of the spatiotemporal expansion of a system or subsystem are often characterized by its typical size and life-span. For example, intestinal epithelial cells have an average life-span of about 5 days and an average size of approximately 40  $\mu\text{m}$ . In practice, the space and time scales of interest could lie well below these extremes. When we conceptualize a complex disease as a system of systems spanning multiple scales, the scale at which each system operates is typically described in terms of the characteristic organizational constitution (structure, process, function) of the system and/or the characteristic or relevant spatiotemporal operating range. Figure 1 illustrates some of the common space and time scales and organizational scales relevant to systems medicine [9, 10].

Mathematical modeling and simulation approaches aim to integrate knowledge and quantitative data about a complex system. A mathematical system model (short: *model*) refers to a mathematical description that represents the various components and elements of a physical or conceptual entity called *system*. A system simulation



**Fig. 1** Conceptualization of a biomedical system (cloud shape) and its associated spatiotemporal (part (a) shows space and (b) time scales) and organizational scales [8]

(short: *simulation*) represents the dynamic behavior/operation of a system over time. A modern approach to modeling and simulation implements a mathematical model in a computer program, and then executes the program to simulate the system. In this approach, a computer program *code* (passive instructions) *implements* the mathematical system model, and a computer process (executing code instance) running on a computer implements the system simulation. In this discourse, we use the term *dynamical model* when we refer to a computer implementation of a mathematical model capable of simulating the behavior of the system represented by the model.

By integrating the knowledge and quantitative data of a complex system into a dynamical model, we hope to improve our *understanding* of the system and *predict* the responses of the system to new inputs. Dynamical models facilitate understanding by providing a framework for interpreting and integrating data. New data can be interpreted (i.e., converted to information) by their implication obtained from simulating the system's response to these data. For example, a dynamical model may be used to convert ultrasound images of displacement (data) into estimates of tissue stiffness (information). We may interrogate a model in ways that may be difficult (time, cost) or impossible (ethics, regulatory issues) to do in the laboratory. For example, we may test the results of clinical procedures or drug administration before they are carried out. Dynamical models also provide a means to capture, share, generate, and communicate scientific knowledge.

A key challenge in developing a dynamical model for a complex system is to find a useful trade-off in terms of *model fidelity* (accuracy, validity) and *model complexity*. Model fidelity refers to how well a model mimics the salient aspects (features, structures, processes, functions, behaviors) of the system it represents. For example, a *valid* dynamical model of gene regulation is able to accurately predict the dynamic gene-regulatory response for a new condition. Model complexity refers to the number and type of elements (variables, parameters) and the number and type of relationships between the elements of a model. The more numerous and the more diverse its elements and relationships, the more difficult it becomes to understand and interpret the model and the results it produces. For example, if one were to incorporate all the currently available data on the heart (genes, molecules, pathways, reactions, cells, tissue, structural geometry), then the resulting model would be nearly as complex as the system it represents and therefore difficult to understand. As Norbert Wiener once put it: "The best material model of a cat is another, or preferably the same, cat." So the simpler a mathematical model is, the easier it is to understand and use. Simpler models tend to have fewer parameters to estimate and therefore are easier to fit to data. Furthermore, the chance of "overfitting" a model to the data with a small

number of parameters is lower than for models with many parameters (“With four parameters I can fit an elephant, and with five I can make him wiggle his trunk,” John von Neumann). Models that overfit the data often contribute little to the understanding of the system under study and their predictive performance on unseen data is usually very poor. While added model complexity usually improves the correspondence between the dynamical model and the biological reality that is being modeled (with the potential drawbacks of reduced interpretability and overfitting), it also increases the *simulation complexity*. Simulation complexity refers to the computational resources needed to run the simulations (computer processors, memory, and network bandwidth) and to other computational problems, such as numerical errors and instability. As we shall see, simulation complexity becomes an important issue when we go from conventional single-scale to multiscale modeling and simulation.

When we develop a conventional single-scale model, one of the first steps in the modeling process is to delineate the spatiotemporal and organizational scales at which the biological system of interest is assumed (conceptualized) to be operating. This is depicted by the two diagrams of Fig. 1. The degree to which the cloud shapes in the figure extend along the corresponding space and time scales is determined by the data available and the problems, questions, and hypotheses at hand. Crucially, by making particular scale assumptions, we are able to make certain model simplifications. For example, quantities that change much faster or slower than the quantities of interest can be either treated as constants or ignored altogether. Analogously, processes happening at much smaller or larger space scales than the scale of interest may be represented as constants or simply ignored. Making simplifications based on certain scale assumptions is effectively a form of methodological reductionism.

Understanding complex disease and other complex biomedical systems requires us to explicitly consider structures, processes, and functions across multiple spatiotemporal and organizational scales. For such problems, the traditional approach to model reduction and simplification is no longer adequate [7]. So how can we model such complex biomedical phenomena and systems and still retain human interpretability and computational tractability? The answer is *multiscale modeling and simulation*. Compared with the traditional approach of focusing on one scale, looking at a complex disease simultaneously from several different but interacting scales and different levels of detail is a much more sophisticated approach to modeling and simulation. This approach represents a fundamental shift in the way we view modeling and simulation [11].

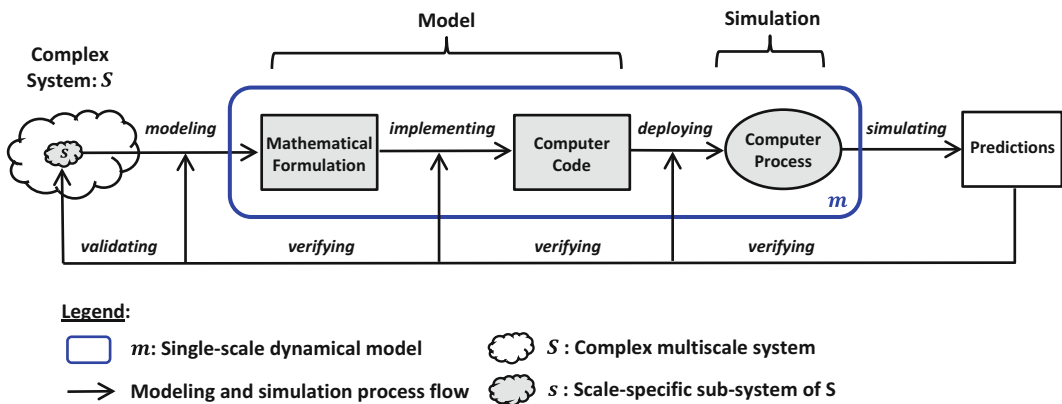
In the following sections, we briefly recapitulate the traditional approach to modeling and simulation, as this is crucial to understand the progression to multiscale modeling and simulation.

We then discuss some of the basic concepts of multiscale modeling and simulation. We emphasize the concepts model and simulation coupling, as these form the key “glue” in converting multiple single-scale models into an integrated multiscale dynamical model. Finally, we discuss some of the future requirements and challenges of multiscale modeling and simulation in systems medicine.

## 2 Traditional Single-Scale Modeling and Simulation

Modeling and simulation have a long history in biomedical science. Figure 2 depicts the typical elements and workflow for modeling and simulating a scale-specific, single-scale biomedical system. The final output of the process is a validated *dynamical model* (depicted by blue-outline box with rounded corners in Fig. 2). As we can see, a dynamical model consists of three components (grey boxes and oval): a *mathematical formulation*, *computer code*, and *computer process*. Each component is a realization of the same model but in a different form. Notice the way it is depicted in Fig. 2; the mathematical model description (box labeled Mathematical Formulation) refers to the *complete* model. This means that all parameters representing the system’s structure and properties are fully determined. In many modeling and simulation applications, the problem of determining the model parameters from experimental data is itself a considerable challenge. In this discourse, we do not cover the details of model parameter determination.

The mathematical formulation of the dynamical model describes a single-scale system using mathematical concepts and language. A wide variety of mathematical methods are used to express dynamical models of biomedical systems. These include



**Fig. 2** Components and processes of traditional single-scale modeling and simulation approaches. The processes modeling, implementing, deploying, and simulating involve conceptual, mathematical, and technical approximations of reality



methods describing continuum and discrete mechanics, classical/Newtonian mechanics (vectors) and analytical mechanics (scalars), statistical mechanics (probability), and network mechanics (graphs). Examples for specific mathematical methods and techniques include different types of differential equations, process algebras, constraint-based models, rule-based models, dynamic probabilistic networks, Petri networks, Potts models, dynamic logical networks, cellular automata, agent-based models, and molecular dynamics [12].

To simulate the dynamic behavior of a complex system, we need to implement the mathematical description of the model in a computer program or code (box labeled Computer Code in Fig. 2). Notice that, in this discourse, we use *computer code*, to mean either source code (in some programming language) or compiled object code, ready to be executed on compatible computer hardware. Once we have the model in the form of a computer code, an executing instance (oval labeled Computer Process in Fig. 2) of such a code represents a simulation of the system. The advantage of this approach is that we can explore different scenarios (varying inputs and conditions) simultaneously by running multiple instances of a model code in parallel. However, we need to be aware of some of the issues involved in terms of physical and virtual computer resources.

- (a) *Replicates, parameter sweep, sensitivity analysis.* To develop a deeper understanding of the system, we often want to simulate its response to a range of different inputs and conditions. This means that we have to run a simulation for each input, and thus need more computer resources.
- (b) *Numerical techniques.* Because most mathematical formulations of a dynamical model do not have a straightforward analytic solution, the computer codes implementing the mathematical model employ techniques that solve the model numerically (e.g., solving a system of ordinary differential equations). These techniques usually consume a large amount of the computer resources needed to run the simulations—the more accurate the numeric solutions need to be, the more resources are required.

Developing and deploying computer codes representing dynamical models is a complex task. A vast array of software tools have been developed to address this task. These include software libraries written in programming languages (e.g., Fortran, C/C++, Python, Java, R, Matlab) and comprehensive software tools. Some of these tools are specifically designed to address biomedical problems, e.g., the Complex Pathway Simulator [13], GNU Monte Carlo Simulations [14], Systems Biology Software Infrastructure [15], and JSim [16].

In addition to the main components of a dynamical model, Fig. 2 also depicts the process or workflow involved in modeling and simulating a single-scale dynamical model. The iterative process involves *model verification* and *model validation*. The value of a dynamical model representing a complex disease or another biomedical system or process is in part dependent on its ability to realistically mimic the targeted response variables [17]. Initially, parameterized models frequently fail to accurately fit observed data. To correct this situation, some parameters or functions are adjusted within reasonable limits (in accordance with global or contextual constraints such as physical laws). The new model's outputs are again compared with data from experimental and clinical data. This process is called model verification—it ensures that *one has built the model right*. In contrast, model validation is the comparison of the output from a verified model with independent data (data that was not used in the model construction process), followed by statistical analyses that test to what degree the model fits the data. Model validation ensures that *one has built the right model*.

One reason for verifying and validating a dynamical model is *error assessment*. We distinguish two main categories of errors: *modeling errors* and *numerical errors*. Modeling errors are due to the fundamental imperfections that arise when we make abstractions of reality in the form of a mathematical model. In Fig. 2, this is depicted by the arrow labeled *modeling* from the cloud shape representing the system of interest to the box representing the mathematical formulation of a dynamical model. A model, any model, is by definition an approximation of reality. The modeling error quantifies how good the approximation is. In Fig. 2, the arrow labeled *implementing* illustrates the process of mapping the mathematical formulation of a dynamical model to a form expressed in computer code. While the mathematical form essentially captures a *continuous* view of the physical quantities of the system (including space and time), the corresponding computer implementation is fundamentally discrete. This is known as the discretization error. Discretization errors, rounding errors, iteration errors, and similar errors are collectively referred to as *numerical errors*, which occur when we move from continuous mathematics to the discrete computer world (the processes labeled *implementing*, *deploying*, and *simulating* in Fig. 2). Ultimately, we need to assess whether the combination of modeling and numerical errors is still within acceptable bounds.

The general procedure for conventional modeling and simulation of single-scale dynamical models could be defined as follows (clearly, the entire procedure is informed by the nature and quality of the available data, the scientific problem at hand, and the concrete questions that the model is designed to answer):

1. **Identify scale:** Identify the spatiotemporal and organizational scale of interest.
2. **Construct model:** Iterate over the *modeling*, *implementing*, *deploying*, *simulating*, and *verifying* loop (Fig. 2), until you find a model that meets the verification criteria.
3. **Validate model:** Evaluate the verified model against the validation criteria.

---

### 3 Multiscale Modeling and Simulation

Traditional approaches to modeling and simulation focus on a single spatiotemporal or organizational scale. This approach more or less implicitly decomposes a complex multiscale problem into a “macroscopic” scale and a “microscopic” scale. If we are interested in system behavior on the macroscopic scale, we assume that nothing interesting is happening on the microscopic scale (e.g., we assume that the effects on the microscopic level are homogeneous, are constant or can be described by suitable constitutive equations). For example, when we model the biomechanical behavior of a bone, we assume homogeneous microscopic structures and properties. If we are interested in the behavior of a subsystem at the microscopic level, we assume that there is nothing interesting happening at the larger scale (e.g., the process is homogeneous at macroscopic scale). For instance, when modeling an osteocyte of a bone, we ignore to a large extent other bone cell types and macroscopic structures and properties of the bone.

Modeling complex diseases and other complex biomedical phenomena means that we need to explicitly incorporate factors and elements from more than one scale, because information from multiple scales is deemed important to fully understand the problem. Multiscale biomedical modeling and simulation aim to explore such multiscale complex systems quantitatively by incorporating several different modeling and simulation techniques [7]. The multiscale modeling and simulation approach adopts a divide-and-conquer philosophy similar to approaches found in other areas (e.g., multi-agent systems, ensemble approaches in machine learning, data fusion). The logic of these and similar approaches is to solve a complex problem by decomposing it into simpler subproblems, solving these subproblems individually, and then synthesizing the sub-solutions into a global solution. For multiscale modeling and simulation, the divide-and-conquer approach looks roughly as follows—details of this approach are discussed by Chopard et al. [18]:

1. **System decomposition:** Decompose the overall multiscale complex system of interest into multiple, scale-specific subsystems. Conceptually, this is a crucial modeling decision; a major mistake

at this point potentially leads to a large modeling error. Interestingly, for many multiscale modeling projects, this step is not performed explicitly, because they base their multiscale models on already existing scale-specific models. So they start at **step 3** or **4** of this procedure.

2. **Construction of scale-specific models:** Develop a dynamical model for each scale-specific subsystem (**steps 2** and **3** of the basic procedure for developing scale-specific dynamical models). This step yields multiple *validated* single-scale dynamical models that form the basis for constructing a multiscale dynamical model.
3. **Combination of scale-specific models into a multiscale model:** This is the most crucial modeling step, as it links the various scale-specific models in a particular way to allow information exchange across the various scales. Various terms are used to refer to this step: *scale linking*, *scale bridging*, *model coupling*, *code coupling*, or *simulation coupling*. Note that such scale bridging methods could be viewed as models in their own right, and should therefore be treated (error assessment, verification, validation) as if they were a scale-specific model.
4. **Model verification: Step 3** generates a multiscale model. We still need to verify this model. In principle, this requires us to iterate over **steps 1–4** multiple times, until we end up with a verified model. Since we often start with a set of verified and validated single-scale models, multiscale model verification may iterate only over **steps 3** and **4**, without the need to go back to the process that constructs single-scale models.
5. **Model validation:** After we have created a verified multiscale model in **step 4**, we need to validate the model before we use it to explore the multiscale complex system of interest.

**Steps 2–5** of the multiscale modeling and simulation process are illustrated in Fig. 3 based on a multiscale model involving two scale-specific models. The large cloud shapes in Fig. 3 show the complex multiscale system,  $S$ , and two of its scale-specific subsystems,  $s_1$  and  $s_2$ . The two subsystems relate to biomedical phenomena on different scales. Let us assume  $s_1$  operates on a microscopic and  $s_2$  on a macroscopic level. The dotted line between the two subsystems illustrates scale-crossing biomedical interactions between  $s_1$  and  $s_2$ . For each of the subsystems, a scale-specific dynamical model (small box shapes with rounded corners labeled  $m_1$  and  $m_2$ ) is constructed using the conventional modeling and simulation approach (Fig. 3a). Notice that we have omitted various process flow arrows. The single-scale models are supposed to be validated and the *modeling*, *verifying*, and *validating* tasks for  $m_1$  and  $m_2$  are no longer needed.



abstract level, the dotted lines relate to the more concrete coupling components and processes of *scale bridging*, *code coupling*, and *simulation coupling*. The dotted line labeled *scale bridging* denotes the procedures, algorithms, and models that translate information from one scale to another. The dotted line labeled *code coupling* depicts software and hardware components that enable the inter-operation between the two computer codes. The dotted line labeled *simulation coupling* depicts the coupled simulations (coordinated execution of code instances exchanging information at run-time) of the subsystems  $s_1$  and  $s_2$ . It is the coordinated running of two computer processes that simulates the behavior of the multiscale system  $S$ .

The dotted lines in Fig. 3 depict the various aspects of model coupling and simulation coupling between the scale-specific models used to form a multiscale model. These lines, however, do not fully convey the challenges and efforts involved in generating a suitable coupling among underlying individual models. Main activities in the coupling of scale-specific models into a “super model” include the following:

1. **Scale bridging:** The development of suitable scale bridging methods that realize the required information transformation between the scales of the constituent scale-specific models. The resulting solutions could themselves be viewed as models in their own right that capture and represent essential characteristics of the complex system of interest.
2. **Code coupling:** Software that enables the scaling linking between the computer codes that implement the scale-specific models. Ideally, such components should not require a modification of the computer *source* codes that implement the scale-specific models.
3. **Deploying:** Before the computer codes making up a multiscale model (scale-specific models and model coupling codes) can be executed, they need to be deployed, together with all dependencies such as libraries and data sources, on one or more machines. Whenever a code is changed, it needs to be re-deployed.
4. **Simulation coupling:** Once the computer code instances of a multiscale model have been deployed on the target computing platforms, they need to be executed in a coordinated fashion to simulate the behavior of the complex system of interest. For large-scale, multi-platform simulations, this may be a nontrivial task as computer resources need to be allocated, code instance execution needs to be monitored, input data needs to be staged, and simulation results (output data) need to be collected, processed, and analyzed [19, 20].

The code coupling, deployment, and simulation coupling may need to be repeated before a final verified model is obtained. Finally, the verified multiscale dynamical model should be validated against the validation data and criteria.

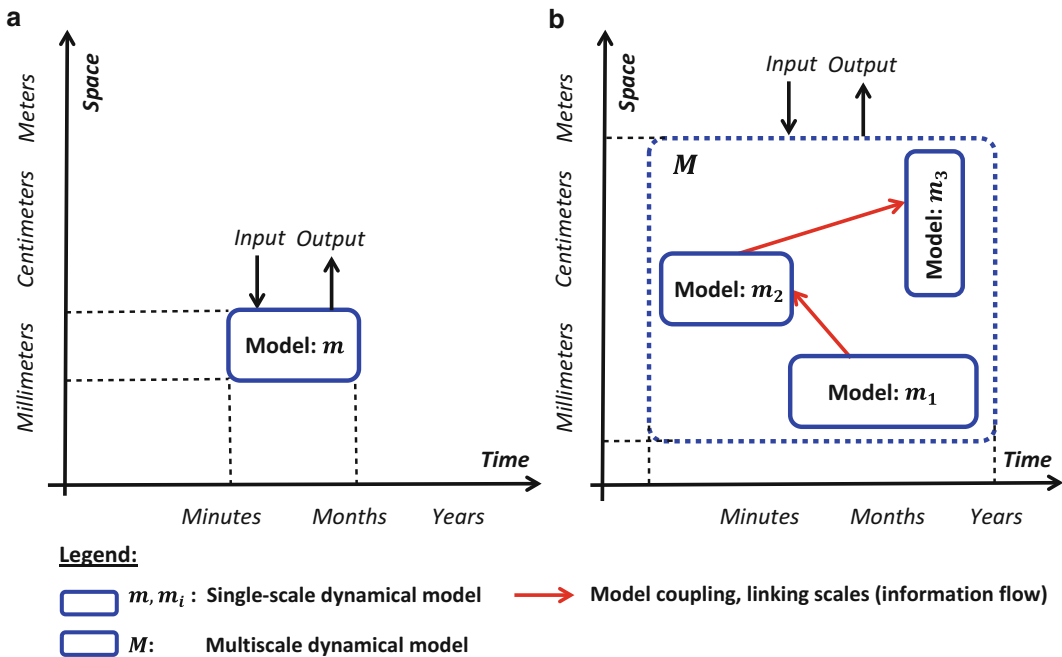
## 4 Combining Structure and Process

### 4.1 Model and Simulation Coupling

A *scale-separation map (SSM)* is a diagrammatic tool to depict the spatiotemporal organization of the single-scale constituents of a multiscale dynamical model and their interactions [18]. Figure 4a shows a single scale-specific model with its approximate scale expansion along the space and time coordinates on the SSM. The diagram in Fig. 4b shows the SSM of a multiscale model composed of three single-scale models.

The central idea in multiscale modeling and simulation is the *coupling* of multiple scale-specific models (codes) and their simulations (coordinated execution of code instances).

We say that two scale-specific models of a multiscale model are *coupled*, because the output of one model is used as an input by another. *Model coupling* (coupled models) defines the input–output structure of the model’s scale-specific sub-models, as well as scale-linking procedures that translate information from one scale



**Fig. 4** Scale-separation map. (a) Single scale-specific dynamical model occupying a defined spatiotemporal scale. (b) Multiscale dynamical model covering a spatiotemporal scale that includes the scales of the constituent scale-specific models



to another. The model coupling structures *represent* the scale-crossing interaction “skeleton” of the complex system of interest. It is possible for a model coupling structure to change dynamically as a result of external inputs or information produced by system simulations. The SSM in Fig. 4b shows the model coupling structure (arrows inside dotted box) of a multiscale model consisting of three scale-specific models. The direction of the arrows indicates the direction of the information flow between the sub-models.

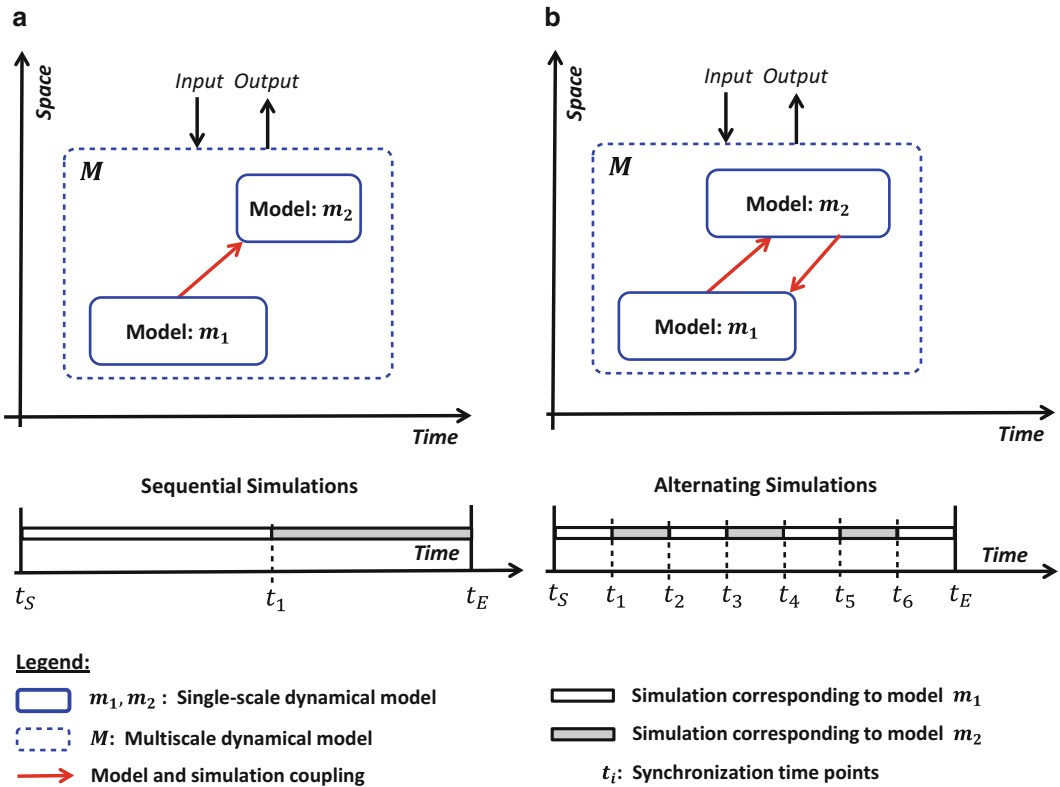
The coordinated execution of the coupled model code instances is referred to as *coupled simulations* (multiple interacting simulations forming the overall multiscale simulation representing the behavior of the modeled system). Coupled multiscale simulations form an abstraction adopted to simplify the complex dynamic characteristics of the system under study. In addition to accurately mimicking the behavior of the complex system of interest, a main consideration in practice is to make the computations involved in multiscale simulations tractable.

The coupling structure and simulation dynamics of a multiscale model are, of course, interrelated. We distinguish two fundamental modes of model and simulation coupling: *acyclic coupling* and *cyclic coupling*.

Two scale-specific models of a multiscale model are said to be acyclically coupled, if information exchange between them is only in one direction (structure aspect), and if the corresponding simulations are performed sequentially (timing aspect). The SSM and timing diagram in Fig. 5a illustrate the acyclic coupling between two scale-specific dynamical models  $m_1$  and  $m_2$ . Model  $m_2$  waits until  $m_1$  has performed its computations, and then takes the output of  $m_1$  as input for its own processing. In this acyclic coupling mode, the exchange of information between two coupled models is in one direction only and occurs once during the course of the overall multiscale simulation (it requires only a single synchronization point,  $t_1$ ).

Cyclic coupling occurs when two scale-specific dynamical models update *each other* during the course of a simulation; the output of one model is used as input for the other, and *vice versa*. This is illustrated in Fig. 5b. Model  $m_1$  performs part of its computations, passes the outputs to model  $m_2$ , and then waits for the output of model  $m_2$ , before it continues. In cyclically coupled multiscale simulations, the simulations of the corresponding sub-models are performed in an alternating fashion. This typically requires multiple synchronization points as illustrated in the timing diagram of Fig. 5b. Executing cyclically coupled model codes is challenging, because they involve finding reliable techniques for coordinating their execution, data exchange, memory allocation, and execution scheduling.

Defining the coupling logic (acyclic, cyclic) of a multiscale model determines the model’s basic information exchange structure.



**Fig. 5** Basic multiscale coupling topologies and the time diagrams of the corresponding simulations. Multiscale model consisting of (a) two acyclically coupled, and (b) cyclically coupled scale-specific models

The way that the information exchange works in detail is defined by the *scale linking* (or scale bridging) and *coupling synchronization* components and procedures of the model. Scale linking defines how the scale-specific information computed as the output of one sub-model is converted to the input of another sub-model operating on a different scale. Scale linking structures and procedures are specified either in the codes of the scale-specific models or by the tool that facilitates the coupling.

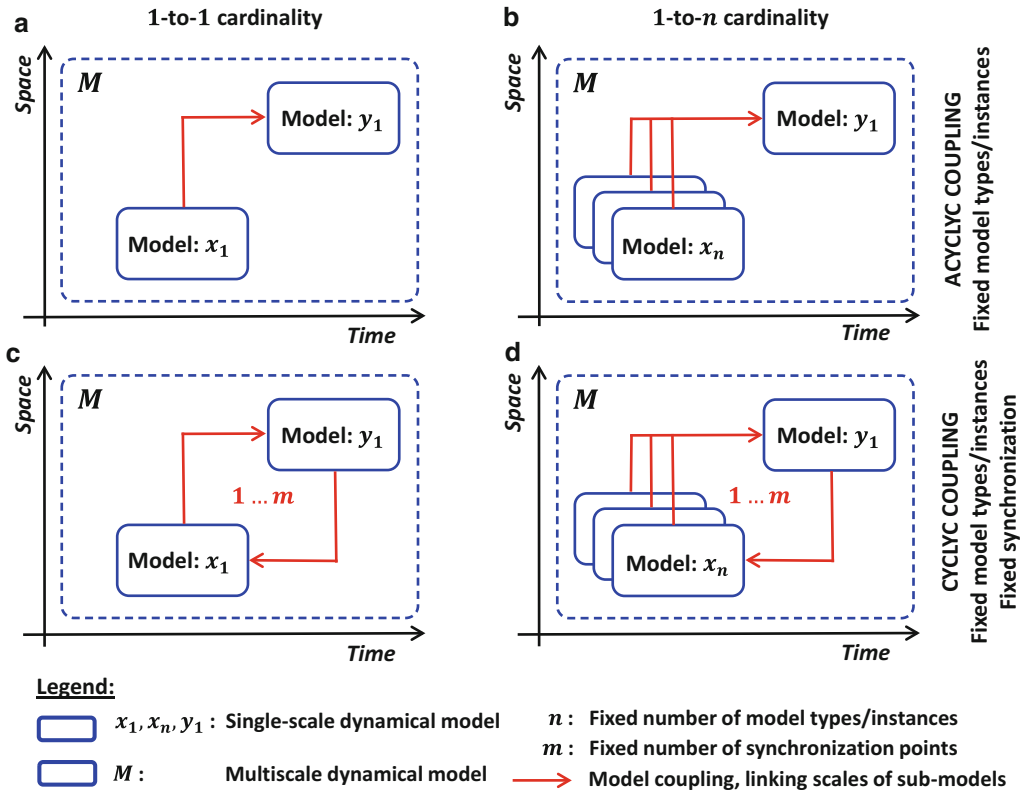
Coupling synchronization is required to orchestrate the inputs and outputs among coupled simulations at *run-time*. When two processes representing the simulation of two scale-specific systems exchange information, synchronization is a way to make this exchange meaningful. To do this, synchronization requires the two interacting processes to “join up” or “handshake” (i.e., synchronize) at a certain point, in order to reach an agreement or commit to a certain sequence of action. That “certain point” may be statically defined prior to simulation (e.g., at regular time or iteration points), or dynamically determined at run-time (e.g., based on defined patterns in the exchanged information). The synchronization points allow the two processes to adapt their future course

based on the information exchanged at the synchronization points. Because the information transfer in acyclic model coupling schemes is one-way (Fig. 5a), there is no need for sophisticated synchronization. However, even in acyclically coupled models, we require some form of orderly information “handover,” making sure that a simulation does not commence before it receives all the data computed by other simulations. In a synchronized communication between two cyclically coupled processes, both simulations unfold *dependent* on the other, and *vice versa*. We say the processes are interdependent (Fig. 5b). For cyclically coupled simulations, synchronization may also be defined statically, based on a fixed number of predefined synchronization points using time or iteration intervals, or dynamically, based on conditions in simulation outputs or external signals.

The examples in Fig. 5 show the coupling between two scale-specific models. Multiscale models representing more complex multiscale systems may be composed of more than two scale-specific models. Figure 4b illustrates a multiscale model consisting of three constituent single-scale models. In general, a multiscale model consists of at least two scale-specific dynamical models. Each of the sub-models of a multiscale model may be of a different type occupying a different scale. However, it is also possible that there are multiple instances of a particular single-scale model. While all sub-models of a multiscale model may be known and fully specified prior to simulation, it is also possible that the precise number of sub-models may be determined dynamically (e.g., based on information generated by the unfolding multiscale simulation). When more than two sub-models or sub-model instances are used to form a multiscale model, it is possible to define coupling cardinalities other than 1-to-1. Here, we consider only the 1-to- $n$  scenario in which the output of  $n$  sub-models forms the input of another sub-model (Fig. 6b, d), and ignore the  $n$ -to-1 scenario.

Based on coupling synchronization and coupling cardinality (1-to-1 and 1-to- $n$ ), we can distinguish a total of nine common multiscale coupling schemes [19]. Figure 6 shows the basic four of the nine schemes, assuming a fixed number of model types/instances and a fixed number of synchronization points. The remaining five schemes would be derived by allowing model types/instances and synchronization points to be created dynamically during the course of a simulation [18].

Three fundamental approaches are commonly distinguished both in multiscale modeling and simulation: “top-down,” “middle-out,” and “bottom-up.” *Top-down* approaches start with observed features on a macroscopic scale of a biomedical system and then attempt to deduce what mechanisms at a more fundamental microscopic scale could account for those observations. Starting from an initial hypothesis, top-down approaches allow to gradually increase the level of detail of the hypothesis with the starting level directly



**Fig. 6** Four basic coupling topologies. (a) One instance per model, acyclic, a single synchronization point. (b) Fixed number of multiple instances per model, acyclic, a single synchronization point. (c) One instance per model, cyclic, fixed number of multiple synchronization points. (d) Fixed number of multiple instances, cyclic, fixed number of multiple synchronization points

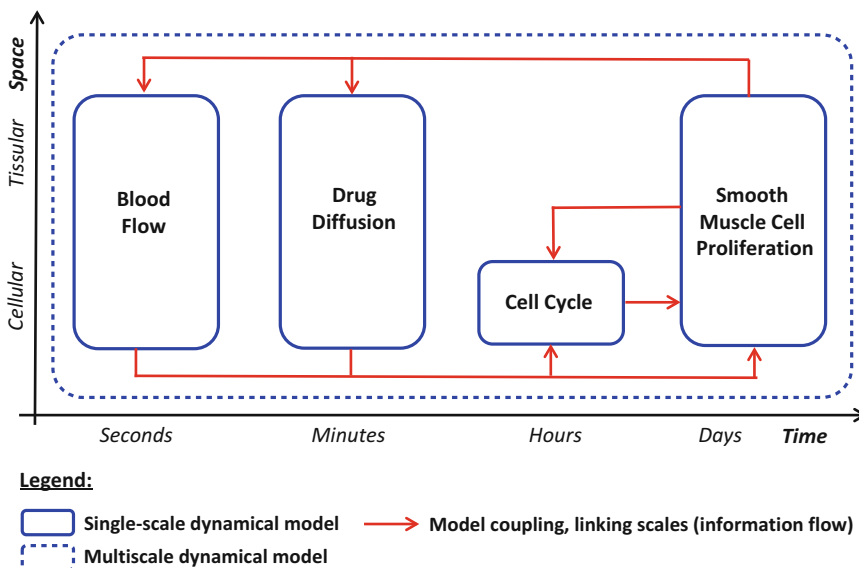
backed up by the data. However, because macroscopic properties often have multiple different potential underlying explanations on microscopic scales, adjacent scales of modeling do not unambiguously emerge from one another in the direction of increasing details. In contrast, *bottom-up* multiscale models aim to derive system behavior on macroscopic spatial, temporal, or organizational scales from the dynamics and interactions of system elements at microscopic scales. The coarse graining that connects the different scales involves identifying which types of collective behavior on fundamental microscopic scales give rise to a coherent behavior and function at a higher macroscopic scale. The major drawback of the bottom-up approach is the effort and time needed to construct the sub-models at the microscopic scales. However, constructing a model bottom-up has the advantage of unveiling gaps in our knowledge and pointing out new directions for experimental studies that without the modeling effort would be less apparent. The *middle-out* approach starts at any of the scales at which sufficient quantitative data is available [21, 22]. Once a sufficient understanding at

the chosen scale has been established, one can reach out to other scales. Eventually, one may reach down to the scale of genes and up to the scale of the entire organism, or even a population of organisms. The linking of the scales is important if one intends to interpret the genome in terms of its physiological function. Ultimately, reaching out and linking the various scales of biological organization is part of what systems medicine is all about. In complex diseases with interactions at many scales, the middle-out approach may be the most obvious option. Agent-based modeling is a prominent example of middle-out modeling and simulation where the initial, primary scale is the level of a cell or an individual in a population of individuals [23].

#### 4.2 A Multiscale Model of In-Stent Restenosis

Let us consider an example—a multiscale model of restenosis [24]. A stenosis is an abnormal narrowing in a blood vessel. Interventions treating stenotic arteries include balloon angioplasty followed by deployment of a metallic tubular mesh (a stent) into the walls of a blood vessel. In-stent restenosis describes an undesired vascular healing process leading to tissue growth after the insertion of a stent into the artery wall. This process could be viewed as a complex multiscale system, involving mechanisms spanning multiple biological phenomena from the cell to tissue levels and the minute-to-day scales. A simplified version of the multiscale model of in-stent restenosis [24] is depicted by the SSM in Fig. 7 [25].

The multiscale model depicted by Fig. 7 consists of four scale-specific models: blood flow, drug diffusion, cell cycle, and smooth muscle cell (SMC) proliferation. The SMC proliferation sub-model simulates the growth and proliferation of smooth muscle cells using



**Fig. 7** Multiscale dynamical model of in-stent restenosis; adapted and highly simplified from Evans et al. [24]

an agent-based approach (each agent representing an individual cell). The regulation of cell migration depends on the cells' cell cycle. This is represented by the rule-based model labeled cell cycle in Fig. 7. The blood flow sub-model uses the lattice-Boltzmann technique to simulate the flow of blood through the stent-supported artery. Modern stents are coated with anti-proliferative drugs to prevent tissue regrowth in and around the area where the stent has been implanted. The drug diffusion sub-model simulates the drug elution process based on a cellular automata approach. This multiscale model has been qualitatively validated against in vivo porcine data, and is used to formulate and test a number of hypotheses related to deeper understanding of this pathophysiology [26–29].

---

## 5 Challenges and Requirements of Multiscale Modeling and Simulation in Systems Medicine

While multiscale approaches are commonplace in some areas [30], multiscale modeling and simulation in the life sciences is a relatively recent development [7, 31]. As the need for a systems approach to human health and disease is increasing, so does the need for multiscale solutions. The mathematical and computing challenges of future multiscale systems medicine are enormous. The mathematical and modeling challenges include the development of novel solutions addressing various aspects relating to one or more of the main modeling and simulation steps (scale decomposition, construction of scale-specific models, model coupling, model verification, and validation) and fundamental modeling approaches (top-down, middle-out, bottom-up). Ultimately, multiscale dynamical models need to be executed on computer systems to simulate the underlying multiscale biomedical phenomena, processes, and systems. The challenges in this area range from programming models and languages, model/simulation coupling libraries, and system architectures, on one hand, and access, allocation, orchestration, and management of large-scale computer resources (physical: processors, storage, networks; virtual: files, memory slots, software), on the other hand. For a discussion on future challenges in the multiscale modeling and simulation, we refer to work by Hoekstra et al. [32]. Here, we focus specifically on challenges in relation to system medicine.

### 5.1 *Mathematical and Modeling Requirements and Challenges*

#### 5.1.1 *Scale Representation*

The majority of state-of-the-art methods and algorithms in systems medicine are designed to solve scale-specific problems. As a result, scale information in these methods is typically “implicit” and no particular scale operations are required. In the future, modeling complex scale-separated biomedical systems will require explicit representations of time, space, and organization structures. Ultimately, a scale-aware ontology for systems medicine is required

to support the development, sharing, and use of multiscale modeling and simulation applications in systems medicine. Such ontologies should be guided by the typical characteristic dimensions and structures (time, space, organizational mechanisms, entities, structures, hierarchies, processes, and systems) relevant to biomedicine. Some general attempts have been reported in the literature [33], but ontologies tailored to systems medicine need to be developed.

### 5.1.2 *Scale Decomposition*

Viewing a complex biomedical problem from a multiscale systems perspective requires us to decompose the problem into scale-separated subproblems. Such a decomposition is a key modeling step as it imposes a particular abstraction on the overall multiscale model. Errors made in scale decomposition are likely to have serious consequences for the quality of the final models and their results. For many concrete problems, a scale decomposition may be obvious, either because scale-specific models already exist or because the available data forces us to focus on a particular set of scales. Ultimately, a good decomposition of the problems into scale-separated subproblems should provide a considerable benefit to the overall modeling and simulation exercise. Many contemporary multiscale modeling and simulation projects start from a set of pre-existing single-scale models. This means that the scale decomposition for a particular multiscale model has already been done. Future and emerging multiscale modeling approaches may design their overall model from scratch, and thus are able to decompose the multiscale problem into scale-separated subproblems. Hence, effective methods and tools to support scale decomposition will become increasingly important. Indeed, it is conceivable that scale decomposition is performed in an automated fashion by algorithms that construct the single-scale models and a suitable scale decomposition simultaneously. Since the scale decomposition structure *is* part of the overall model, this task could be a target of automated or semiautomated model construction algorithms. For instance, as the complexity of a studied gene-regulatory network grows, its underlying structure is likely to have a time-scale-separated modular organization [34]. In such a scenario, the overall network is made up of subnetworks, each operating on a different time scale. If the scale decomposition is known, we could develop the overall multiscale gene-regulation model by constructing a scale-specific model for each distinct scale, and then define the model-coupling structure and scale-linking procedures for the multiscale model. If the scale decomposition is unknown, we may develop an algorithm that automatically constructs a multiscale model, including its mono-scale sub-models and associated model coupling structure. The latter approach effectively discovers an optimal scale separation for the given problem.



### 5.1.3 Scale Bridging

In general, scale-specific sub-models used to construct a multiscale model have been developed for a specific purpose and scale *independent* from the targeted multiscale model. A key requirement and challenge in multiscale modeling is the translation of information from the specific scale of one sub-model to that of another. For example, the properties of a biological tissue need to be computed from a population of cell models, where the output of each cell model is likely to be limited to a smaller spatiotemporal scale than the properties of the tissue-level model. Typically, when we link or couple a microscopic model to a macroscopic one, we need to compute macroscopic information from sub-models on the microscopic level, and we require suitable initial and boundary conditions in the microscopic sub-models. General scale bridging techniques include sampling, projection, homogenization and coarse graining, and constitutive models [32]. How such sub-models behave when integrated into a super-model is still not fully understood, and will need to be studied for each problem separately. Clearly, scale bridging approaches are strongly domain and problem dependent. Human biology complex but has many clearly defined organizational scales and a vast body of knowledge on how processes interact across these scales. This knowledge could be used to develop “standard” scale bridging techniques and tools for systems medicine.

In addition to the difference in scale, the sub-models intended for a multiscale model are often of different basic type (see below). Currently, the application of these methods in large multiscale models is not fully understood.

### 5.1.4 Model Construction

Many contemporary approaches to multiscale modeling and simulation start from a pre-existing set of scale-specific models. However, in the future, multiscale approaches may start from scratch, by first constructing all or some of the sub-models and then coupling these into a multiscale model. While the construction of sub-models creates some overhead, it also provides some advantages over approaches that start from a set of existing scale-specific models. First, it allows breaking down the overall problem into a set of sub-problems that provides an optimal balance in terms of scale decomposition, scale linking, model coupling, and the model error characteristics of the multiscale model. Second, it provides an opportunity to optimize the design, implementation, deployment, and execution performance (hardware/software resources) of the multiscale model. Starting from pre-constructed sub-models potentially forces one to compromise on all of these aspects, and may result in poor solutions (accuracy, efficiency).

### 5.1.5 Model Types

The sub-models of a multiscale model may come in all shapes and sizes. On a high level of abstraction, model types may be distinguished along the following dimensions: mathematical versus

algorithmic, continuous versus discrete, deterministic versus stochastic, mechanistic versus phenomenological, static versus dynamic, etc. More research is needed to understand the implications and challenges involved in combining sub-models of different types into multiscale dynamical models. There is a vast range of mathematical (quantitative) modeling techniques ranging from differential equations to cellular automata. Each method comes with its own advantages, limitations, and requirements. More research is needed to adapt these techniques to the requirements of multiscale modeling and simulation (e.g., incorporating explicit scale information), to realize scale linking between pairs of such techniques, and to couple models and simulations based on scale-specific models defined by these methods.

#### 5.1.6 *Model and Simulation Coupling*

Arguably, the definition of how sub-models interact to realize the composite multiscale model and the orchestration of the execution of the coupled model codes form the key elements in multiscale modeling and simulation. Various coupling and execution frameworks and technologies exist. The following is a brief list of coupling/execution technologies that have been employed in the life sciences. The Multiscale Coupling Library and Environment (MUSCLE 2) is a generic, portable framework supporting multiscale modeling and simulation applications on distributed computing resources [20, 35]. The run-time environment of MUSCLE 2 solves common distributed computing problems and couples sub-models of a multiscale model across various high-performance computer systems. GridSpace is a virtual laboratory framework enabling researchers to conduct virtual experiments on large-scale computing environments [36]. It facilitates a script-based development of experiments using languages such as Ruby, Python, and Perl. The Open Projet d'Assimilation par Logiciel Multi méthode (Open-PALM) facilitates the concurrent execution and the intercommunication of programs, as well as dynamic and parallel (component- and task-based) coupling [37]. Open-PALM supports various programming interfaces and provides a graphical user interface. In terms of agent-based coupling/modeling tools, MSI and Repast are commonly used. The Multiscale Systems Immunology (MSI) simulation framework provides a flexible model of the immune system [38]. Its modular, object-oriented design (implemented in C++ and Python) allows the execution of coupled multiscale simulations. Repast HPC addresses the computational issues in agent-based multiscale simulations [39]. In particular, Repast HPC enables the execution of large-scale multi-process simulations involving either a large number of relatively simple agents or a small number of very complex ones. Repast HPC is designed to support multiple model simulations that are typically used in agent-based models to account for stochastic variation in model outputs as well as to explore the possible range of outcomes.

We need a general description framework that allows us to define all aspects of model and simulation coupling, including the details of scale linking, coupling synchronization, and time delays. The XML-based Multiscale Modeling Language (MML) is a description language for specifying the coupling architecture of a multiscale simulation [19, 40]. Another issue in relation to model/simulation coupling is the development of general templates or patterns of model/simulation coupling that go beyond the basic nine scenarios that are implied by Fig. 6 [19]. Such templates or patterns could be defined on various levels that are of interest in practical multiscale problems, for example, based on scales and scale-linking patterns, based on model types, or based on technological aspects such as software and hardware architecture. In particular in systems medicine, with its numerous more or less clearly defined scales and levels of organization, and the paramount role of the cell scale, it should be possible to identify common coupling patterns. It is also conceivable to develop a workflow technology specifically dedicated to model and simulation coupling. An important technology aspect related to model/simulation coupling is the coupling of models and simulations in a *distributed* computing environment (see below).

#### 5.1.7 Errors

Understanding the systematic errors (bias) and random errors (variance) that we make in modeling and simulation is an area that is often dealt with in a superficial manner. Two main categories of error in multiscale modeling and simulation are *modeling errors* and *numerical errors*. The key modeling abstraction in multiscale modeling is the *coupling* of sub-models into an integrated multiscale model. Usually, the constituent sub-models are pre-existing and have not been developed with the view of integrating them into a large, multiscale model. Coupling such sub-models is a drastic modeling step and requires a good understanding of the modeling error we make. More research is needed to fully understand and characterize the modeling error in multiscale approaches. Furthermore, as we couple multiple sub-models to form a composition multiscale model, a big challenge is to estimate how numerical errors propagate from the sub-models to the overall model. More research is needed to better understand, describe, and estimate such errors.

#### 5.1.8 Model Verification and Validation

Related to the topic of errors in multiscale modeling and simulation are model verification and validation [17, 41]. Model verification is concerned with the correctness of the model construction process in terms of the conceptual and technical specifications and assumptions. In model verification, the implementation of the model is tested and errors are corrected. Model validation refers to the accuracy of a model in terms of its representation of a real-world system or phenomenon. Critical is that accuracy is

determined in relation to the model's intended purpose [17]. Because of the composite nature of multiscale models and simulations, model verification and validation need to be performed both on the level of sub-models and on the level of the coupled multiscale model. There seems to be a lack of guidelines and procedures on how to perform model verification and validation for multiscale models.

*5.1.9 Multiscale  
Parameter and Variable  
Exploration*

Once a conventional single-scale dynamical model has been validated, we would like to use it to study the system's response to new stimuli and conditions. To do this, we change the values of model parameters and/or variables in a systematic way, and run a new simulation for each parameter/variable configuration. In a multiscale dynamical model, we can perform parameter and variable exploration studies both on the "isolated" single-scale models individually, and on the fully coupled multiscale model. These approaches are compute intensive but could potentially provide a novel way to understand complex systems. Multiscale parameter and variable exploration may also be used to assess errors and error propagation in multiscale models.

*5.1.10 Multiscale  
Parameter Estimation  
(System Identification)*

Constructing dynamical models requires the estimation of concrete and optimal values of the model's parameters. Optimal model parameter values are those for which the model that performs well on unseen data. Automated parameter estimation studies are computationally very expensive because a simulation has to be performed for each of a potentially very large number of parameter value combinations. For a concrete multiscale dynamical model, we would typically start with a set of validated sub-models, for which good parameters have already been determined. However, we could in principle envisage an experiment that tries to re-estimate all model parameters in a coupled setup. While this is likely to be a highly compute-intensive task, it could potentially lead to new insights.

**5.2 Requirements  
and Challenges  
Relating  
to Methodologies,  
Software Tools,  
and Computer  
Systems  
and Infrastructures**

What is currently lacking are multiscale modeling and simulation frameworks and guidelines for systems medicine. These would present key concepts, tools, high-level methodologies, and guidelines on how to conceptualize, analyze complex multiscale biomedical problems, and how to develop, deploy, share, use, and maintain multiscale modeling and simulation solutions. Because human biology is organized into characteristic spatio-temporal and organizational scales, such frameworks and guidelines could be based on these characteristic biological scales. An example of a recently developed generic multiscale framework is MAPPER [35].

*5.2.1 Guidelines  
and Frameworks*

### 5.2.2 *Multiscale Model Management*

In order to facilitate the explicit representation of a multiscale model by multiple sub-models, multiscale modeling and simulation tools need to provide a framework and tools for annotating, managing, using, and sharing multiscale models. This includes description languages as well as repositories for storing and managing both the sub-models of a composite multiscale model and the specifications (scale bridging, coupling) that define the multiscale nature of the model. Because the components of multiscale models may be geographically distributed, model management solutions need to provide a suitable description framework.

### 5.2.3 *Multiscale Data and Information Management*

In addition to the model components themselves, multiscale modeling and simulation studies involve data (experimental, predicted) and information describing the studied systems and phenomena at different scales. Furthermore, once the results of multiscale simulations are reported, scientific, legal, and regulatory requirements may require the provenance information of the components and procedures that generate the results to be kept and communicated. Hence, good solutions are required to represent, record, and manage the results and associated provenance information of multiscale simulations [42]. Again, because of the multiscale and perhaps distributed nature of multiscale computing solutions, building such systems is a considerable challenge. The development of information management systems and services would benefit from an ontology for multiscale modeling and simulation in systems medicine.

### 5.2.4 *Large-Scale Computing Environments and Infrastructures*

As multiscale modeling and simulation approaches are growing in complexity, so are their requirements for large-scale resources and distribution of systems and system components. Below we list some of the key requirements in terms of large-scale computing environments and infrastructures for future and emerging multiscale approaches in systems medicine.

### 5.2.5 *Access to HPC Resources*

The need for large processing, storage, and network bandwidth capabilities in multiscale solutions arises from different aspects of multiscale modeling and simulation: automated reverse engineering of models (optimization, parameter estimation); exploration of different parameter and variable spaces (parameter sweep, sensitivity analysis; probabilistic projection of system evolution trajectories); replication of experimental conditions to estimate variability; repetition of computations as part of model verification and validation (Fig. 3a, b); and sharing and integration and management of “big data” (omics, clinical, lifestyle, environment) feeding into the modeling and simulation process [43]. Many organizations do not have local large-scale computing resources and access to external resources may be too expensive, technically too complicated, or hampered by access policies and regulations. These issues need to be addressed.

### 5.2.6 Resource Management

Large-scale and distributed computing environments employ resource management systems [44] to co-ordinate the availability, provision, and use of physical and virtual computer resources, including data storage devices, processing units, computer networks, software applications, data resources, instruments, files, network connections, memory areas, and even people. Multiscale modeling and simulation solutions that use substantial computer resources require sophisticated resource management solutions. Particular requirements for distributed multiscale computing applications are *resource co-allocation* [44] and *advance reservation* [45] technologies. Resource co-allocation and management technologies facilitate the allocation of resources and associated deployment, orchestration, and execution of computing jobs (programs) on more than one computing system and possibly across administrative domains.

### 5.2.7 Concurrent and Parallel Computing

In acyclically coupled multiscale simulations, it is possible that multiple instances of a particular type of sub-model, or multiple sub-model instances of different type, provide the input to another sub-model (Fig. 6b). In this case, it might be possible to execute the input models in parallel. In cyclically coupled models, a typical interaction or communication pattern is characterized by alternating processing—model  $x_1$  computes a few iterations while model  $x_2$  waits, then model  $x_2$  takes the output from model  $x_1$ , computes a few iterations while model  $x_1$  waits, and so on. Supporting this form of alternating processing requires suitable concurrent computing solutions.

### 5.2.8 Distributed Multiscale Computing

Multiscale modeling and simulation problems are complex and entail a highly heterogeneous research and computing environment in terms of scientific topics, concepts, disciplines, models, types of models, types of technology, resource requirements, researchers, organizations, etc. As a result, it is likely that many emerging and future multiscale systems medicine solutions will need to be developed and deployed across geographically dispersed research and computing environments. Such scenarios need to be supported by flexible distributed computing solutions that take the specific requirements of multiscale modeling and simulation and systems medicine into account. The European FP7 project MAPPER is an effort that has been designed to develop such distributed multiscale computing solutions [19, 35]. Systems medicine encompasses a potentially huge variety of problems, highly diverse communities operating in geographically dispersed locations. Developing multiscale modeling and simulation solutions for complex biomedical systems is likely to be based on distributed e-science infrastructures. Therefore, we anticipate that *distributed multiscale computing* will be a major R&D component of future and emerging systems medicine approaches.

### 5.3 Requirements and Challenges Beyond Dynamical Models

So far this discourse has focused on multiscale *dynamical* models. Such models allow quantitative simulation of the response of a complex biomedical system to system perturbations or stimuli. Developing such models usually requires considerable amounts of experimental time course data of the system(s) of interest. While the generation of such data in systems medicine is likely to become more affordable and more commonplace in the future, for certain problems the available data may still be too limited in the foreseeable future for developing detailed quantitative, multiscale dynamical models. However, there may be sufficient data about the relevant scales of the complex multiscale system to develop a useful multiscale model that is not dynamical. For example, it might be possible to generate a static network or graph structure for each scale-specific system, and then link the networks to form a *multiscale network* structure. Such a structure could then be subjected to analysis based on graph theory to reveal biomedically relevant interaction patterns, motifs, paths, etc. [46].

Networks form a natural way for linking information describing aspects of the same complex phenomenon or system at different spatiotemporal or organizational scales. A network is a logical abstraction (describable as a mathematical graph) and forms the main conceptual construct for representing the scale-crossing interactions of a multiscale system and model. Another form of network used to link data across different scales is *semantic ontologies* (an ontology is a formal specification of a shared conceptualization). Thus, ontologies could be used to analyze or mine multiscale data [47]. Central to such approaches would be physically or virtually integrated databases storing the multiscale data along with a multiscale ontology. Indeed, since multiscale data management, integration, and sharing are likely to become a prerequisite for all multiscale modeling and simulation projects in systems medicine, integrative analysis of such data may be the first step in any multiscale modeling exercise. Many existing models may not capture time and space in the way quantitative mechanistic models do, but capture knowledge in symbolic (artificial intelligence) or sub-symbolic (computational intelligence) structures. The former type includes formalisms such as logic-based rules, semantic networks, and decision trees, while the latter includes artificial neural networks, mathematical graphs, and cellular automata. This is a largely unexplored area in multiscale modeling and simulation, and it is likely to attract more attention in the future.

---

## 6 Conclusions

A major area of interest in systems medicine is complex diseases, which are characterized by multiple contributing factors originating from various levels of biological organization, from the environment



to genes. To understand, prevent, treat, and manage complex diseases will ultimately depend on our ability to integrate and use relevant knowledge and data. Multiscale modeling and simulation in systems medicine is currently emerging as the most promising methodology that could address the knowledge and data integration challenges in this field. This approach could be viewed as the ultimate consequence of a systems view of complex phenomena [6], because it conceptualizes such phenomena as *system of systems*, instead of system of “atomic” components [7]. Multiscale modeling and simulation rely on a combination of multiple scale-specific models to form a “super model” that represents the structure and behavior of complex biomedical systems across multiple organizational scales. While multiscale modeling and simulation have been successfully employed in nonmedical fields [30, 48], its adoption in areas related to systems medicine is a relatively recent development [7, 49–51].

Multiscale modeling and simulation is a comprehensive methodology that makes use of nontrivial information and communication technologies. Research in this area is ongoing and growing as the socioeconomic pressure to tackle large and complex problems in many areas continues to intensify [30]. Adopting the multiscale approach in biomedical research and practice involves a vast array of challenges, ranging from R&D on theoretical and technical aspects to issues of biomedical research and clinical practice. Some challenges and requirements of multiscale modeling and simulation in systems medicine are briefly mentioned in this chapter. Perhaps the biggest challenge of all lies in the sociocultural fabric of current scientific practice. The present scientific mind-set and culture in biomedicine are characterized by an interdisciplinary approach (where researchers and practitioners work jointly but still from a disciplinary-specific basis) which has evolved from multidisciplinary science (working in parallel or sequentially from a disciplinary-specific base) in the past century. The multiscale modeling and simulation framework outlined in this chapter could be viewed as a *transdisciplinary scientific paradigm*, in which researchers and practitioners work jointly using a *shared conceptual* (meaning of concepts) and *conceptual* (systems of explanation) framework and combined disciplinary-specific approaches to address complex R&D problems in systems medicine. Clearly, the paradigm of multiscale modeling and simulation in systems medicine will require considerable changes in currently prevailing scientific mind-sets and culture.

---

## Acknowledgements

A.G. Hoekstra acknowledges partial funding by Russian Scientific Foundation, grant # 14-11-00826.

## References

1. Craig J (2008) Complex diseases: research and applications. *Nat Educ* 1(1):184
2. Bousquet J, Anto JM, Sterk PJ, Adock IM, Chung KF, Roca J et al (2011) Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Med* 3(7):43
3. Hood L, Friend SH (2012) Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol* 8(3):184–187
4. Capobianco E (2012) Ten challenges for systems medicine. *Front Genet* 3(193):1–4
5. Calzolari D, Bruschi S, Coquin L, Schofield J, Feala JD, Reed JC, McCulloch AD, Paternostro G (2008) Search algorithms as a framework for the optimization of drug combinations. *PLoS Comput Biol* 4(12):e1000249
6. Von Bertalanffy L (1969) *General systems theory*. Braziller, New York
7. Sloot PMA, Hoekstra AG (2010) Multi-scale modelling in computational biomedicine. *Brief Bioinform* 11(1):142–152
8. Hunter PJ, Borg TK (2003) Integration from proteins to organs: the Physiome Project. *Nat Rev Mol Cell Biol* 4:237–243
9. Hunter P, Nielsen P (2005) A strategy for integrative computational physiology. *Physiology (Bethesda)* 20(5):316–325
10. Noble D (2002) Modeling the heart – from genes to cells to the whole organ. *Science* 295(5560):1678–1682
11. Dada JO, Mendes P (2011) Multi-scale modelling and simulation in systems biology. *Integr Biol* 2011(3):86–96
12. Machado D, Costa RS, Rocha M, Ferreira EC, Tidor B, Rocha I (2011) Modeling formalisms in systems biology. *AMB Express* 1:45
13. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U (2006) COPASI – a Complex Pathway Simulator. *Bioinformatics* 22(24):3067–3074
14. Bois FY (2009) GNU MCSim: Bayesian statistical inference for SBML-coded systems biology models. *Bioinformatics* 25(11):1453–1454
15. Adams R, Clark A, Yamaguchi A, Hanlon N, Tsorman B, Ali S, Lebedeva G, Goltsov A, Sorokin A, Akman OE, Troein C, Millar AJ, Goryanin I, Gilmore S (2013) SBSI: an extensible distributed software infrastructure for parameter estimation in systems biology. *Bioinformatics* 29(5):664–665
16. Butterworth E, Jardine BE, Raymond GM, Neal ML, Bassingthwaight JB (2014) JSim, an open-source modeling system for data analysis. *F1000Res* 2:288
17. Barlas Y (1994) Model validation in systems dynamics. *Int'l systems dynamics conference*, p 1–10
18. Chopard B, Borgdorff J, Hoekstra AG (2014) A framework for multi-scale modelling. *Phil Trans R Soc A* 372(2021):20130378
19. Borgdorff J, Falcone J-L, Eric Lorenz E, Bona-Casas C, Chopard B, Hoekstra AG (2013) Foundations of distributed multiscale computing: Formalization, specification, and analysis. *J Parallel Distrib Comput* 73:465–483
20. Borgdorff J, Mamonski M, Bosak B, Kurowski K, Ben Belgacem M, Chopard B, Groen D, Coveney PV, Hoekstra AG (2014) Distributed multiscale computing with MUSCLE 2, the multiscale coupling library and environment. *J Comput Sci* 2014(5):719–731
21. Walker D, Southgate JS, Hill G, Holcombe M, Hose D, Wood S, MacNeil S, Smallwood R (2009) The epitheliome: modelling the social behavior of cells. *Biosystems* 76:89–100
22. Noble D (2006) *The music of life: biology beyond the genome*. Oxford University Press, Oxford
23. An G, Mi Q, Dutta-Moscato J, Vodovotz Y (2009) Agent-based models in translational systems biology. *Wiley Interdiscip Rev Syst Biol Med* 1(2):159–171
24. Evans DJW, Lawford PV, Gunn J, Walker E, Hose DR, Smallwood RH, Chopard B, Krafczyk M, Bernsdorf J, Hoekstra A (2008) The application of multiscale modelling to the process of development and prevention of stenosis in a stented coronary artery. *Philos Trans A Math Phys Eng Sci* 366:3343–3360
25. Groen D, Borgdorff J, Bona-Casas C, Hetherington J, Nash RW, Zasada SJ, Saverchenko I, Mamonski M, Kurowski K, Bernabeu MO, Hoekstra AG, Coveney PV (2013) Flexible composition and execution of high performance, high fidelity multiscale biomedical simulations. *Interface Focus* 3(2):2013
26. Tahir H, Hoekstra AG, Lorenz E, Lawford PV, Hose DR, Gunn J, Evans DJW (2011) Multi-scale simulations of the dynamics of in-stent restenosis: impact of stent deployment and design. *Interface Focus* 1(3):365–373
27. Tahir H, Bona-Casas C, Hoekstra AG (2013) Modelling the effect of a functional endothelium on the development of in-stent restenosis. *PLoS One* 8(6):e66138
28. Amatruda CM, Casas CB, Keller BK, Tahir H, Dubini G, Hoekstra AG, Hose DR, Lawford P,

- Migliavacca F, Narracott AJ, Gunn J (2014) From histology and imaging data to models for in-stent restenosis. *Int J Artif Organs* 37(10):786–800
29. Tahir H, Bona-Casas C, Narracott AJ, Iqbal J, Gunn J, Lawford P, Hoekstra AG (2014) Endothelial repair process and its relevance to longitudinal neointimal tissue patterns: comparing histology with in silico modelling. *J R Soc Interface* 11(94):20140022
  30. Groen D, Zasada SJ, Coveney PV (2014) Survey of multiscale and multiphysics applications and communities. *Comput Sci Eng* 16(2):34–43
  31. Schnell S, Grima R, Maini PK (2007) Multiscale modeling in biology: new insights into cancer illustrate how mathematical tools are enhancing the understanding of life from the smallest scale to the grandest. *Am Sci* 95:134–142
  32. Hoekstra AG, Chopard B, Coveney P (2014) Multiscale modelling and simulation: a position paper. *Phil Trans R Soc A* 372(2021):20130377
  33. Yang A, Marquardt W (2009) An ontological conceptualization of multiscale models. *Comput Chem Eng* 2009(33):822–837
  34. Damle S, Davidson E (2012) Synthetic in vivo validation of gene network circuitry. *Proc Natl Acad Sci* 109(5):1548–1553
  35. Borgdorff J, Belgacem MB, Bona-Casas C, Fazendeiro L, Groen D, Hoenen O, Mizeranschi A, Suter JL, Coster D, Coveney PV, Dubitzky W, Hoekstra AG, Strand P, Chopard B (2014) Performance of distributed multiscale simulations. *Phil Trans A* 372(2021):20130407
  36. Ciepiela E, Wilk B, Harężlak D, Kasztelnik M, Pawlik M, Bubak M (2014) Towards provisioning of reproducible, reviewable and reusable in-silico experiments with the GridSpace2 Platform. In: Bubak M, Kitowski J, Wiatr K (eds) *eScience on distributed computing infrastructure*, LNCS, vol 8500. Springer, Switzerland, p 118–129. [http://link.springer.com/chapter/10.1007%2F978-3-319-10894-0\\_9](http://link.springer.com/chapter/10.1007%2F978-3-319-10894-0_9)
  37. Piacentini A, Morel T, Thevenin A, Duchaine F (2011) Open-PALM: an open source dynamic parallel coupler. *Proceedings of the 4th International conference on computational methods for coupled problems in science and engineering*, Kos, Greece, p 20–22
  38. Mitha F, Lucas TA, Feng F, Kepler TB, Chan C (2008) The Multiscale Systems Immunology project: software for cell-based immunological simulation. *Source Code Biol Med* 3:6
  39. Collier N, North M (2012) Repast HPC: a platform for large-scale agent-based modeling. In: Dubitzky W, Kurowski K, Schott B (eds) *Large-scale computing techniques for complex system simulations*. John Wiley, and Sons, Inc., Hoboken, NJ, pp 81–110
  40. Falcone J-L, Chopard B, Hoekstra A (2012) MML: towards a Multiscale Modeling Language. *Procedia Comput Sci* 1(2012):819–826
  41. Carson JS (2002) Model verification and validation. In: Yücesan E, Chen CH, Snowdon J, Charnes J (eds) *The 2002 winter simulation conference*, p 52–58
  42. Davison AP (2010) Challenges and solutions in replicability and provenance tracking for simulation projects. *BMC Neurosci* 11(Suppl. 1):P76
  43. Mark V (2013) Biology: the big challenges of big data. *Nature* 498:255–260
  44. Krauter K, Buyya R, Maheswaran M (2002) A taxonomy and survey of grid resource management systems for distributed computing. *Softw Pract Exp* 32:135–164
  45. Castillo C, Rouskas G, Harfoush K (2011) Online algorithms for advance resource reservations. *J Parallel Distrib Comput* 71(7):963–973
  46. Barabási A-L, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12:56–68
  47. Martone ME, Zaslavsky I, Gupta A, Memon A, Tran J, Wong W, Fong L, Larson SD, Ellisman MH (2008) The Smart Atlas: spatial and semantic strategies for multiscale integration of brain data. In: Burger A, Davidson D, Baldock R (eds) *Anatomy ontologies for bioinformatics*, p 267–286
  48. Fish J (2009) *Multiscale methods: bridging the scales in science and engineering*. Oxford University Press, Oxford
  49. Hoekstra AG, Copard B, Lawford P (2013) Multiscale modelling. In: Coveney P, Díaz-Zuccarini V, Hunter P, Viceconti M (eds) *Computational biomedicine*. Oxford University Press, Oxford, pp 138–159
  50. Viceconti M (2012) *Multiscale modeling of the skeletal system*. Cambridge University Press, New York
  51. Hunter P, Chapman T, Coveney PV, de Bono B, Diaz V, Fenner J, Frangi AF, Harris P, Hose R, Kohl P, Lawford P, McCormack K, Mendes M, Omholt S, Quarteroni A, Shublaq N, Skår J, Stroetmann K, Tegner J, Thomas SR, Tollis I, Tsamardinos I, van Beek JHGM, Viceconti M (2013) A vision and strategy for the virtual physiological human: 2012 update. *Interface Focus* 3:20130004

# Chapter 18

## Mathematical and Statistical Techniques for Systems Medicine: The Wnt Signaling Pathway as a Case Study

Adam L. MacLean, Heather A. Harrington, Michael P.H. Stumpf, and Helen M. Byrne

### Abstract

The last decade has seen an explosion in models that describe phenomena in systems medicine. Such models are especially useful for studying signaling pathways, such as the Wnt pathway. In this chapter we use the Wnt pathway to showcase current mathematical and statistical techniques that enable modelers to gain insight into (models of) gene regulation and generate testable predictions. We introduce a range of modeling frameworks, but focus on ordinary differential equation (ODE) models since they remain the most widely used approach in systems biology and medicine and continue to offer great potential. We present methods for the analysis of a single model, comprising applications of standard dynamical systems approaches such as nondimensionalization, steady state, asymptotic and sensitivity analysis, and more recent statistical and algebraic approaches to compare models with data. We present parameter estimation and model comparison techniques, focusing on Bayesian analysis and coplanarity via algebraic geometry. Our intention is that this (non-exhaustive) review may serve as a useful starting point for the analysis of models in systems medicine.

**Key words** Wnt signaling, Model development, Nondimensionalization, Asymptotic analysis, Parameter inference, Algebraic methods, Model selection

---

### 1 Introduction

Despite the growing number of therapeutic options available to clinicians, gaps remain in our fundamental understanding of many biological processes. Acquiring this additional knowledge requires that we focus on the molecular players that operate in intercellular and intracellular environments. Revealing the complex networks and dynamics that control cellular, tissue- and host-level behavior may enable us to improve existing treatments and design new drug targets.

Many intercellular signals are initiated by signaling proteins such as cytokines and hormones. When cytokines bind to receptors of a target cell, they trigger a cellular response by signal transduction

pathways: multistep sequences of intracellular signaling events and communication between molecules. Most of these molecules are proteins. Enzymes such as kinases and phosphatases, for example, catalyze (respectively) the addition/removal of a phosphate group to/from a substrate, and thus perform a crucial role in relaying information [1]. Phosphorylation (the addition of a phosphate group) can be associated with protein activation, and information can be communicated downstream, engaging multiple signaling cascades by successive chemical reactions. While some reactions are linear, with the output proportional to the input [2], many are complex, involving feedback loops or pathway redundancies. Often the output of these pathways is activation or inhibition of regulatory proteins called transcription factors, which modify gene transcription and the cellular state.

To turn a gene on, an activated transcription factor translocates from the cytoplasm into the nucleus, binds to the enhancer or promoter region of DNA, and RNA polymerase transcribes the DNA template to synthesize RNA. Then messenger RNA (mRNA) leaves the nucleus and enters the cytoplasm where ribosomes translate mRNA into protein [1]. Conversely, transcription factors may turn a gene off by repressing the recruitment of RNA polymerase. These possible responses thus regulate protein synthesis. In addition to the subcellular processes that changes in protein synthesis stimulate, proteins may be released by the cell and act as signaling molecules in other pathways.

Gene regulatory pathways are crucial to the normal functioning of cells, with many diseases caused by dysfunction of one or more pathways. For example, signaling pathways such as NF- $\kappa$ B, MAP Kinase, and Wnt/ $\beta$ -catenin are involved in a host of cellular processes and functions, including cancer. Due to their complexity, a systems approach is needed to understand normal and aberrant pathway function. Only by building theoretical models that describe how cells signal and validating/updating them using experimental data can we develop new drug therapies that target specific diseases.

The remainder of the chapter is organized as follows. In Subheading 2, we review methods used to model signal transduction pathways, and introduce an exemplary enzyme kinetics model. We then describe the biology of Wnt signaling, with reference to relevant models, and introduce two models of the Wnt signaling pathway that we focus on throughout the chapter to demonstrate various techniques. In Subheading 3, we detail methods that can be used to analyze a particular model and discuss the insight that each approach can generate. In Subheading 4, we introduce techniques that can be used to compare models, including some new methods for systems medicine. We conclude in Subheading 5 with a discussion of the different techniques, and ideas for their further application in systems medicine.

---

## 2 Mathematical Modeling

Signaling pathways are complex and may be difficult to understand by linear logic alone. Theoretical models can be used to gain insight into the dynamics of multiple biochemical interactions. Constructing a mathematical model is a nontrivial task that requires sufficient understanding of the system to determine not only the type of model that should be used to address a particular question but also the limitations of the model. After reviewing some of the modeling approaches that are used to study signaling pathways, we focus on ordinary differential equation (ODE) models. We introduce basic principles that can be used to construct ODE models and illustrate them by reference to enzyme kinetics and two models of the Wnt pathway.

### **2.1 Modeling Approaches for Systems Medicine**

Many processes associated with systems medicine in general, and signaling pathways in particular, can be modeled. These include: gene/protein abundances; gene/protein interactions; abundances of cellular species; the effects of cytokines, chemicals, drugs, or other interventions on system or tissue-level phenomena. Modeling strategies for systems medicine can be classified as either deterministic or stochastic; we describe stochastic approaches briefly here, since the methods introduced in later sections are generally only applicable to deterministic systems.

Deterministic approaches describe systems for which, given full details of the model (parameter values and initial conditions), its time evolution can be determined exactly. This means that if a system is restarted multiple times from the same initial state it will always return to the same future states. Ordinary and partial differential equations (PDEs) are two examples [3]. PDEs with two or more independent variables (e.g., space and time) are more flexible than ODEs, but their simulation and analysis can be computationally expensive. Deterministic methods provide accurate descriptions of population-level behavior if the population sizes are large enough that the effects of random fluctuations can be neglected.

Stochastic approaches describe systems whose temporal evolution has unpredictable elements due to randomness somewhere in the system. They are popular for modeling biological systems where randomness and heterogeneity abound, and should be used when population sizes are small enough that fluctuations cannot be ignored. In most cases, population averages will be recovered from a stochastic model when the abundances become large enough. One can, for example, construct stochastic models of protein dynamics with stochastic differential equations [4] (i.e., ODEs with noise terms—often Gaussian—added). Such models can be used to study the dynamics of species that fluctuate about a well-defined mean value.

Stochastic modeling can also be developed via agent-based approaches [5, 6]. Here, individual agents act according to a set of

rules. For example, within a given pathway, a protein could be phosphorylated or dephosphorylated with probabilities that depend on its environment. Such a framework treats protein species very differently to differential equation methods: each protein is viewed as an autonomous agent and population dynamics emerge in a “bottom up” manner. Whilst such methods may appeal to our intuition about protein heterogeneity, the approach is limited since analyses are often computationally expensive. As such, agent-based models should be used when population-averaged models fail to capture the behavior that the modeler seeks to describe.

Cellular automata are a subset of agent-based models that impose spatial structure on the system by constraining the agents to lie on a grid, in two or three dimensions [7, 8]. The agents are updated via rules which may be deterministic or stochastic. Each grid point may be occupied by a finite number of cells (typically only one) and the model can accommodate multiple cell types. Cellular automata can account for spatial relationships between different cell types and have the advantage of being easy to interpret biologically. A challenge associated with these models is that the update rules may not translate clearly into biological hypotheses. Additionally, as for other agent-based models, simulation of cellular automata can be computationally expensive. Fitting such models to data is at the limits of what is currently feasible since, despite significant advances in cellular imaging technology, obtaining cell data of sufficient resolution and quality to fit to a model is rare.

The above overview of modeling approaches is not exhaustive: in limited space, we make no mention of Boolean, semi-quantitative, hybrid, or branching processes. Instead, we continue by explaining how to develop ODE models for signaling pathways.

## **2.2 Formulating Mathematical Models of Signaling Pathways**

In this section our focus is on using ODEs to develop dynamic models of signaling pathways. Two basic principles are integral to the development of such models:

- *The Principle of Mass Balance* states that the rate of change of a species is equal to the difference between the rate at which the species is added to the system and the rate at which it is removed;
- *The Law of Mass Action* states that a reaction proceeds at a rate proportional to the product of its reactants.

If, for example, substrate  $A$  is irreversibly phosphorylated by enzyme  $B$ , to produce  $C$  then we write





where  $r_1$  is the rate at which phosphorylation occurs. We construct ODEs that describe the dynamics of  $A$ ,  $B$ , and  $C$  by appealing to the Principle of Mass Balance and the Law of Mass Action:

$$\frac{dA}{dt} = -r_1 AB, \quad \frac{dB}{dt} = -r_1 AB + r_1 AB \equiv 0, \quad \frac{dC}{dt} = r_1 AB. \quad (2)$$

By inspecting the above ODEs, it is straightforward to deduce that the following quantities are preserved:

$$A + C = A_0 + C_0, \quad \text{and} \quad B = B_0,$$

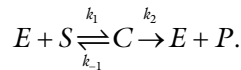
where  $A(t=0) = A_0$ ,  $B(t=0) = B_0$  and  $C(t=0) = C_0$  are prescribed as initial conditions. We can exploit these *Conservation Laws* to simplify the governing equations: in this case, we can eliminate both  $B$  and  $C$  and our model reduces to give

$$\frac{dA}{dt} = -r_1 B_0 A, \quad \text{with} \quad A(t=0) = A_0 \Rightarrow A(t) = A_0 e^{-r_1 B_0 t}.$$

Thus, substrate levels decay exponentially, at rate  $r_1 B_0$ .

### 2.2.1 Case Study I: The Enzyme Kinetics Model

We now consider a biochemical reaction that is catalyzed by an enzyme. In more detail, the enzyme  $E$  binds reversibly with the substrate  $S$  to form a complex  $C$ . While complexed with the substrate, the enzyme converts it into a product  $P$  and the enzyme is recovered. We represent these reactions as follows:



By applying the Law of Mass Action to this reaction scheme and appealing to the Principle of Mass Balance, we deduce that the following system of ODEs describes the time-evolution of  $S$ ,  $E$ ,  $C$ , and  $P$ :

$$\frac{dS}{dt} = -k_1 ES + k_{-1} C, \quad (3)$$

$$\frac{dE}{dt} = -k_1 ES + (k_{-1} + k_2) C, \quad (4)$$

$$\frac{dC}{dt} = k_1 ES - (k_{-1} + k_2) C, \quad (5)$$

$$\frac{dP}{dt} = k_2 C. \quad (6)$$

If we assume further that  $S(t=0) = S_0$ ,  $E(t=0) = E_0$ ,  $C(t=0) = 0$ , and  $P(t=0) = 0$ , and take suitable combinations of the governing ODEs, then we deduce

$$\frac{d}{dt}(E + C) = 0 \text{ and } \frac{d}{dt}(S + C + P) = 0, \Rightarrow E + C = E_0 \text{ and } S + C + P = S_0,$$

We can exploit these conservation laws to eliminate  $E$  and  $P$  and obtain the following reduced model:

$$\frac{dS}{dt} = -k_1 S(E_0 - C) + k_{-1} C, \quad (7)$$

$$\frac{dC}{dt} = k_1 S(E_0 - C) - (k_{-1} + k_2) C, \quad (8)$$

$$\text{with } S(t = 0) = S_0 \text{ and } C(t = 0) = 0. \quad (9)$$

### 2.3 Modeling Wnt Signaling

Wnt signaling is implicated in many biological processes. The pathway is activated when Wnt ligands bind to specific receptors on the cell surface, resulting in the stabilization and nuclear accumulation of the transcriptional co-activator  $\beta$ -catenin. *Canonical* Wnt signaling encompasses cellular responses to external Wnt stimuli mediated by  $\beta$ -catenin. *Noncanonical* signaling describes cellular signaling and responses to Wnt not mediated by  $\beta$ -catenin. The canonical Wnt pathway plays a key role in essential cellular processes ranging from proliferation and cell specification during development to adult stem cell maintenance and wound repair [9]. Dysfunction of Wnt signaling is implicated in many pathological conditions, including degenerative diseases and cancer [10–12]. Despite further molecular advances [13–15], certain details of the dynamics of the pathway are still not well understood.

The basic steps that constitute canonical Wnt signaling are as follows (although these are not undisputed; discussed below): Wnt binds to cell-surface receptors Frizzled and LRP5/6 [11] that transduce a signal via a multistep process involving Dishevelled (Dsh) to the so-called destruction complex (DC). The DC contains forms of Axin, adenomatous polyposis coli (APC), glycogen synthase kinase 3 (GSK-3), and casein kinase 1 $\alpha$  (CK1 $\alpha$ ). In the absence of a Wnt signal, the DC actively degrades  $\beta$ -catenin—which is being continually synthesized in the cell—by binding and phosphorylating the protein and thus marking it for proteasomal degradation. Following Wnt stimulation, degradation of  $\beta$ -catenin is inhibited through phosphorylation of DC member proteins. This leads to accumulation in the cytoplasm of free  $\beta$ -catenin, which is able to translocate to the nucleus where it can form a complex with T-cell factor (TCF) and lymphoid-enhancing factor (LEF) proteins and, thereby, influence the transcription of target genes associated with processes such as self-renewal and proliferation [16, 17].

In addition to these core mechanisms, evidence for other important processes has been found, some of which may challenge the Wnt signaling paradigm. Spatial localization within the cell has been found to be important not only for  $\beta$ -catenin but also for Dsh and DC member proteins including Axin, APC, and GSK-3 [18–23]. There is also evidence of competitive binding of  $\beta$ -catenin to cell membrane proteins such as E-cadherin [24] and intricate cross-talk with the Hippo pathway, this being mediated by Yap and Taz which promote translocation of cytoplasmic  $\beta$ -catenin to the nucleus via phosphorylation and then compete with TCF for  $\beta$ -catenin in the nucleus [25]. This spatial organization of Wnt pathway members may be key to understanding the pathway, as some modeling suggests [26, 27]. Equally, an alternative description for the degradation of  $\beta$ -catenin exists: in this picture,  $\beta$ -catenin can be actively degraded while still bound to the DC, rather than being released marked for degradation [28]. Discriminating between competing hypotheses is needed in order to fully elucidate canonical Wnt signaling: mathematical modeling is a natural framework within which to achieve this.

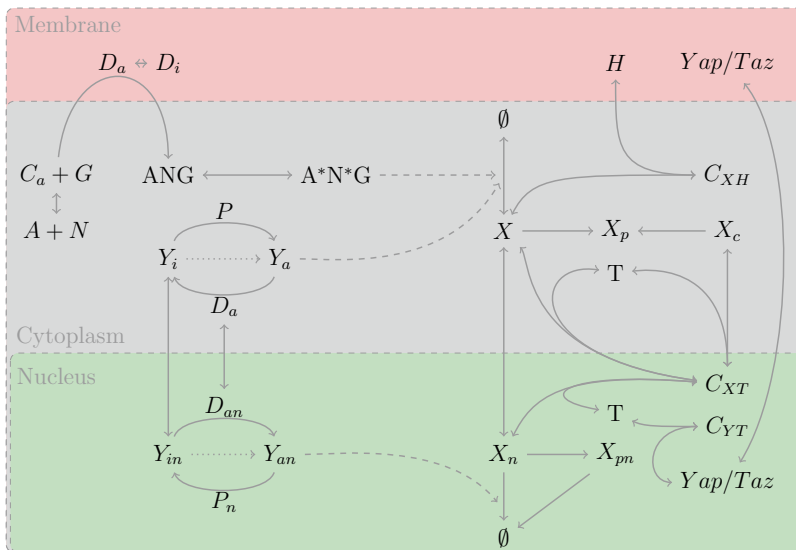
The first quantitative model of Wnt/ $\beta$ -catenin signaling was developed in 2003 [29], based on data from *Xenopus* extracts. Formulated as a system of ODEs, the model describes known interactions between core components of the canonical pathway, these being Wnt, Dishevelled, GSK3 $\beta$ , APC, Axin,  $\beta$ -catenin, and TCF. The DC is assumed to act only in the well-mixed cytoplasm and, hence, only cytoplasmic levels of pathway components are considered. Since its publication, the Lee model has been extended in many ways (for recent reviews of mathematical models of Wnt signaling, *see* [16, 30]). The effect of mutations in APC was investigated by Cho et al. [31], the action of Wnt inhibitors was studied by Kogan et al. [32], and the impact of Wnt-ERK cross-talk considered by Kim et al. [33]. The effect of competition for  $\beta$ -catenin with adhesion proteins was investigated by van Leeuwen et al. [34], while Schmitz showed how shuttling of core proteins between cytoplasm and nucleus could influence pathway dynamics [35, 36]. More recently, a new shuttling model was constructed that accounts not only for exchange of pathway proteins between the nucleus and cytoplasm, but also degradation of  $\beta$ -catenin while it is bound to active destruction complex (DC) and activation of the DC by dephosphorylation of its components [27]. Table 1 summarizes the key features of some of these models and Fig. 1 illustrates the localization and known interactions between key proteins involved in Wnt signaling.

We now present the Lee model [29] and the Schmitz model [36], using the notation presented in Table 2. These models, together with the enzyme kinetics model introduced above, will be

**Table 1**  
**Comparison of features across different models of Wnt signaling**

Biological feature	Lee	van Leeuwen	Schmitz	Shuttle
$\beta$ -Catenin production	✓	✓	✓	✓
$\beta$ -Catenin degradation (independent of DC)	✓	✓	×	✓
$\beta$ -Catenin degradation (dependent on DC)	✓	✓	✓	✓
$\beta$ -Catenin sequestration by DC	✓	✓	✓	✓
$\beta$ -Catenin sequestration by APC	×	×	×	×
Shuttling of species between cytoplasmic and nuclear compartments	×	×	✓	✓
Activation/inactivation of DC	✓	✓	✓	✓
Interaction with adhesion molecules	×	✓	×	×
Two $\beta$ -catenin forms: transcription only and transcription or adhesion	×	✓	×	×
DC is represented by its constituent parts	✓	×	×	×
$\beta$ -Catenin binds individual parts APC and Axin as well as DC	✓	×	×	×
$\beta$ -Catenin binds to TCF to promote transcription of target genes	✓	✓	✓	✓

For further details *see* [27, 29, 34, 36]



**Fig. 1** Reaction scheme that incorporates many different Wnt signaling models and additional molecular players (e.g., Yap/Taz). *Solid arrows* denote direct reactions; *long-dashed arrows* denote species that act as catalysts in degradation reactions; and *dotted arrows* denote alternative paths for the direct activation of  $Y$ . Note that active/inactive forms of  $Y$  are equivalent to active/inactive forms of  $ANG$ . *Species names are defined in Table 2*

revisited throughout the chapter to illustrate how the techniques discussed in Subheadings 3 and 4 are applied to specific models.

### 2.3.1 Case Study II: The Lee Model

In its original form, the Lee model comprises 15 time-dependent ODEs for protein species and complexes that participate in the canonical Wnt pathway, the reaction rates being based on mass action kinetics [29]. The model targets the assembly of the destruction complex from the constituent parts of APC, Axin, and GSK3 $\beta$ . It does not distinguish between nuclear and cytoplasmic compartments, instead assumes that all species are uniformly distributed throughout the cell. A schematic diagram of the reactions described in the Lee model is given in Fig. 2. Using the variable names defined in Table 2 and primes to denote differentiation with respect to time, the ODEs that specify this model are:

$$D'_i = -\alpha_1 D_i + \alpha_2 D_a, \quad (10)$$

$$D'_a = \alpha_1 D_i - \alpha_2 D_a, \quad (11)$$

$$\Upsilon'_a = \alpha_3 \Upsilon_i - \alpha_4 \Upsilon_a - \alpha_{10} X \Upsilon_a + \alpha_{11} C_{XY} + \alpha_{13} C_{XYp}, \quad (12)$$

$$\Upsilon'_i = \alpha_6 G C_{NA} - \alpha_5 D_a \Upsilon_i - \alpha_3 \Upsilon_i + \alpha_4 \Upsilon_a - \alpha_7 \Upsilon_i, \quad (13)$$

$$G' = \alpha_5 D_a \Upsilon_i - \alpha_6 G C_{NA} + \alpha_7 \Upsilon_i, \quad (14)$$

$$C'_{NA} = \alpha_5 D_a \Upsilon_i - \alpha_6 G C_{NA} + \alpha_7 \Upsilon_i + \alpha_8 NA - \alpha_9 C_{NA}, \quad (15)$$

$$A' = -\alpha_8 NA + \alpha_9 C_{NA} - \alpha_{21} XA + \alpha_{22} C_{XA}, \quad (16)$$

$$C'_{XY} = \alpha_{10} X \Upsilon_a - \alpha_{11} C_{XY} - \alpha_{12} C_{XY}, \quad (17)$$

$$C'_{XYp} = \alpha_{12} C_{XY} - \alpha_{13} C_{XYp}, \quad (18)$$

$$X'_p = \alpha_{13} C_{XYp} - \alpha_{14} X_p, \quad (19)$$

$$X' = -\alpha_{10} X \Upsilon_a + \alpha_{11} C_{XY} + \alpha_{15} - \alpha_{16} X - \alpha_{19} XT + \alpha_{20} C_{XT} - \alpha_{21} XA + \alpha_{22} C_{XA}, \quad (20)$$

$$N' = -\alpha_8 NA + \alpha_9 C_{NA} + \alpha_{17} - \alpha_{18} N, \quad (21)$$

$$T' = -\alpha_{19} XT + \alpha_{20} C_{XT}, \quad (22)$$

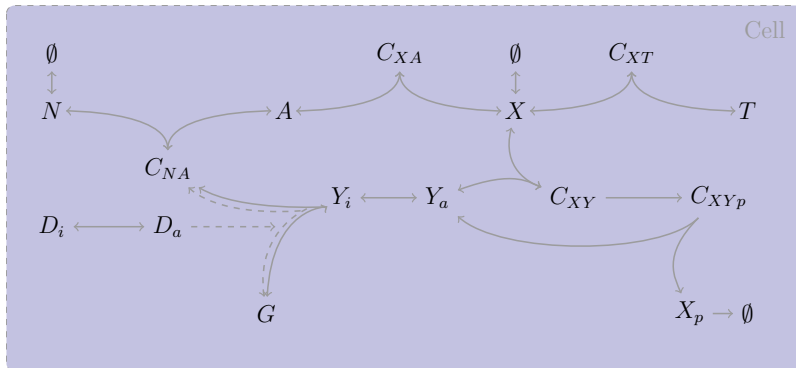
$$C'_{XT} = \alpha_{19} XT - \alpha_{20} C_{XT}, \quad (23)$$

$$C'_{XA} = \alpha_{21} XA - \alpha_{22} C_{XA}. \quad (24)$$

To facilitate comparison with the Schmitz model (see below), the nonnegative rate constants  $\alpha_k$ ,  $k \in (1, 2, \dots, 22)$  have been

**Table 2**  
**Definition of notation for the variables used by the Lee and Schmitz models**

Symbol	Species	Forms
$X$	$\beta$ -Catenin	$X_p$ —marked for proteasomal degradation
$Y$	Destruction complex (APC/Axin/GSK3 $\beta$ )	$Y_a$ —active $Y_i$ —inactive
$D$	Dishevelled	$D_a$ —active $D_i$ —inactive
$A$	APC	
$N$	Axin	
$G$	GSK3 $\beta$	
$T$	TCF	
$C$	Complex	$C_{XY}$ —complex of X and Y (etc.)



**Fig. 2** Schematic of the Lee model [29], which describes the activation of the destruction complex and its effect on  $\beta$ -catenin in a single cellular compartment (cytoplasm and nucleus combined). Notation of the model species is given in Table 2. *Solid arrows* represent reactions and *dashed arrows* represent catalytic processes

redefined from those used in [29]. Wnt dependence is incorporated via the parameter  $\alpha_1 = \alpha_1(W)$  that controls the activation of Dsh.

Inspection of Eqs. 10–24 reveals that there are four conservation laws:

$$\begin{aligned}
 D_0 &= D_i + D_a, \\
 G_0 &= G + Y_i + Y_a + C_{XY} + C_{XYp}, \\
 A_0 &= A + Y_i + Y_a + C_{XY} + C_{XYp} + C_{XA} + C_{NA}, \\
 T_0 &= T + C_{XT},
 \end{aligned}$$

the constants  $D_0, G_0, A_0$ , and  $T_0$  denote the (assumed constant) levels of Dishevelled, GSK3 $\beta$ , APC, and TCF initially present in the system. These conservation laws are consistent with experimental observations which suggest that levels of these proteins do not fluctuate during Wnt signaling (i.e., they are produced and degraded at the same rates). They can be used to eliminate four variables and, in so doing, to reduce the model from 15 to 11 ODEs. Further simplifications are achieved by assuming that all binding processes, except those for the binding of GSK3 $\beta$  to APC/Axin, reach equilibrium rapidly and that all species involving Axin are present at low levels. Under these assumptions, and after some algebra, the following expressions for  $D_0, G, A, T, X_p, C_{XT}, C_{XYp}$ , and  $C_{NA}$  are obtained:

$$D_i = D_0 - D_a, \quad G = G_0, \quad A = \frac{A_0}{1 + \frac{\alpha_{21}}{\alpha_{22}} X}, \quad T = \frac{T_0}{1 + \frac{\alpha_{19}}{\alpha_{20}} X}, \quad X_p = \frac{\alpha_{12}}{\alpha_{14}} C_{XY},$$

$$C_{XT} = \frac{X T_0}{1 + \frac{\alpha_{19}}{\alpha_{20}} X}, \quad C_{XA} = \frac{A_0 X}{1 + \frac{\alpha_{21}}{\alpha_{22}} X}, \quad C_{XYp} = \frac{\alpha_{12}}{\alpha_{13}} C_{XY}, \quad C_{NA} = \frac{\alpha_8}{\alpha_9} \frac{A_0 N}{1 + \frac{\alpha_{21}}{\alpha_{22}}}$$

and a reduced system of 7 ODEs for the remaining species is eventually recovered (equations not presented since they are rather involved and less instructive than Eqs. 10–24). In [29] and [37], this model reduction is performed in an ad hoc manner; it would be instructive to repeat it by first nondimensionalizing the governing equations (*see* Subheading 3.2) and using asymptotic analysis to perform the model reduction (*see* Subheading 3.3).

### 2.3.2 Case Study III: The Schmitz Model

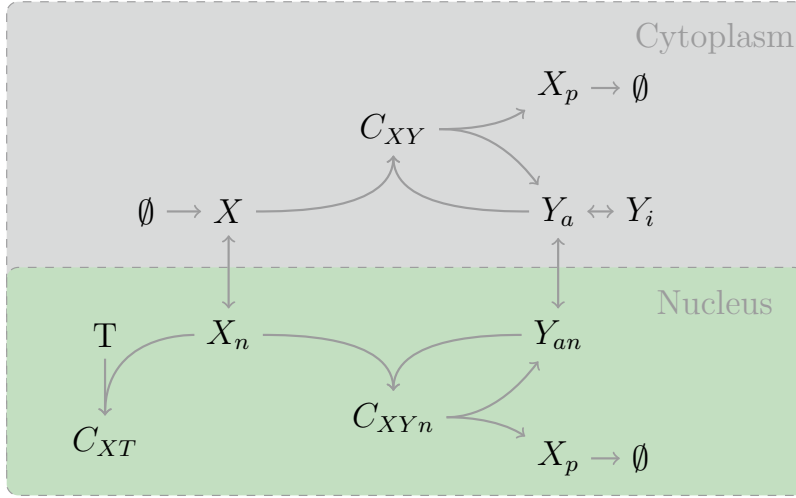
Like the Lee model, the Schmitz model [36] focuses on the canonical Wnt pathway. Key differences between the Lee and Schmitz models are that the latter distinguishes between the cytoplasm and nucleus and accounts for exchange of  $\beta$ -catenin and DC between these compartments (*see* Table 2 and Fig. 3 for further description). In each compartment, DC binding to  $\beta$ -catenin leads to its phosphorylation, and phosphorylated  $\beta$ -catenin is degraded. We use subscript  $n$  to denote species residing in the nucleus with the exception of TCF ( $T$ ) and the  $\beta$ -catenin-TCF complex ( $C_{XT}$ ); since these species are localized in the nucleus and to facilitate comparison with the Lee model, the subscript is omitted. Using notation that is modified from that used in [36], the ODEs that define the Schmitz model are:

$$X' = \delta_0 + (\delta_2 X_n - \delta_1 X) + (\delta_6 C_{XY} - \delta_5 X Y_n), \quad (25)$$

$$X'_n = (\delta_1 X - \delta_2 X_n) + (\delta_9 C_{XYn} - \delta_8 X_n Y_{an}) + (\delta_{12} C_{XT} - \delta_{11} X_n T), \quad (26)$$

$$X'_p = \delta_7 C_{XY} - \delta_{13} X_p, \quad (27)$$





**Fig. 3** Schematic of the Schmitz model [36], which describes the interaction between  $\beta$ -catenin and the destruction complex in two cellular compartments: cytoplasm and nucleus. Notation of the model species is given in Table 2

$$X'_p = \delta_{10} C_{XYn} - \delta_{14} X_p, \quad (28)$$

$$\Upsilon'_a = (\delta_4 \Upsilon_{an} - \delta_3 \Upsilon_a) + (\delta_6 C_{XY} - \delta_5 X \Upsilon_a) + \delta_7 C_{XY} + (\delta_{16} \Upsilon_i - \delta_{15} \Upsilon_a), \quad (29)$$

$$\Upsilon'_i = \delta_{15} \Upsilon_a - \delta_{16} \Upsilon_i, \quad (30)$$

$$\Upsilon'_{an} = (\delta_3 \Upsilon_a - \delta_4 \Upsilon_{an}) + (\delta_9 C_{XYn} - \delta_8 X_n \Upsilon_{an}) + \delta_{10} C_{XYn}, \quad (31)$$

$$C'_{XY} = (\delta_5 X \Upsilon_a - \delta_6 C_{XY}) + \delta_7 C_{XY}, \quad (32)$$

$$C'_{XYn} = (\delta_8 X_n \Upsilon_{an} - \delta_9 C_{XYn}) - \delta_{10} C_{XYn}, \quad (33)$$

$$T' = \delta_{12} C_{XT} - \delta_{11} X_n T, \quad (34)$$

$$C'_{XT} = \delta_{11} X_n T - \delta_{12} C_{XT}, \quad (35)$$

where  $\delta_k$  ( $k=1,2,\dots,17$ ) are nonnegative rate constants and  $\delta_{15} = \delta_{15}(W)$  so that Wnt acts to inactivate the destruction complex in the cytoplasm.

By taking appropriate combinations of Eqs. 25–35, it is straightforward to show that there are two conservation laws:

$$\Upsilon_i + \Upsilon_a + \Upsilon_{an} + C_{XY} + C_{XYn} = \Upsilon_{\text{TOT}} \quad \text{and} \quad T + C_{XT} = T_{\text{TOT}}, \quad (36)$$

the constants  $\Upsilon_{\text{TOT}}$  and  $T_{\text{TOT}}$  denoting, respectively, the total number of molecules of DC and TCF in the system, as determined from the initial conditions. These identities may be used to reduce

the order of the Schmitz model from 11 to 9. As explained below, further systematic simplifications may be possible following model nondimensionalization and parameter estimation.

### 3 Techniques for the Analysis of a Specific Model

Once model construction is complete, the modeler aims to extract from it new insight. This can be done in a number of ways: if no data are available, standard mathematical techniques can be used to increase understanding of the behavior of the model; however, if data are available, then it may be possible to estimate model parameters. In this section we describe a number of techniques, some standard and others less so, that can be used to analyze models. We demonstrate these methods by reference to the models of enzyme kinetics and Wnt signaling introduced in Subheading 2.

#### 3.1 Steady State Analysis

Broadly speaking, the behavior of an ODE model can be categorized as either transient or steady state. The latter describes the behavior at large timescales ( $t \rightarrow \infty$ ). For systems that reach single valued (i.e., not oscillating) steady states, we refer to the long time values that system variables take as the fixed points. Much theory exists for the analysis of fixed points, which can be helpful in characterizing model behavior and predicting the effects of perturbations [38]. We continue by calculating the steady states for the enzyme kinetics model and the Schmitz model (similar analysis can be performed for the Lee model but the resulting expressions are rather involved and therefore omitted).

##### 3.1.1 Case Study I: The Enzyme Kinetics Model (Steady State)

Setting  $\frac{d}{dt} = 0$  in Eqs. 3–6, we deduce that our model for enzyme kinetics evolves to the following unique, steady state solution:

$$S = 0, \quad E = E_0, \quad C = 0 \quad \text{and} \quad P = S_0.$$

Thus, as expected, the reaction proceeds until all of the substrate  $S$  has been converted to product  $P$ .

##### 3.1.2 Case Study III: The Schmitz Model (Steady State)

Setting  $\frac{d}{dt} = 0$  in Eqs. 25–35 and manipulating the resulting algebraic equations supplies the following expressions for  $\Upsilon_{an}, \Upsilon_i, X_p, X_{pn}, C_{XY}, C_{XTn}, T$ , and  $C_{XT}$  in terms of  $X, X_n$ , and  $\Upsilon_a$ :

$$\begin{aligned} \Upsilon_{an} &= \frac{\delta_3}{\delta_4} \Upsilon_a, \quad \Upsilon_i = \frac{\delta_{15}}{\delta_{16}} \Upsilon_a, \\ X_p &= \frac{\delta_7}{\delta_{13}} \frac{\delta_5}{\delta_6 + \delta_7} X \Upsilon_a, \quad X_{pn} = \frac{\delta_8}{\delta_{14}} \frac{\delta_{10}}{\delta_9 + \delta_{10}} X \Upsilon_a, \end{aligned}$$

$$C_{XY} = \frac{\delta_5}{\delta_6 + \delta_7} X \Upsilon_a, \quad C_{X\Upsilon_n} = \frac{\delta_3}{\delta_4} \frac{\delta_8}{\delta_9 + \delta_{10}} X_n \Upsilon_a,$$

$$T = \left( 1 + \frac{\delta_{11}}{\delta_{12}} X_n \right)^{-1} T_{\text{TOT}}, \quad C_{XT} = \frac{\delta_{11}}{\delta_{12}} \left( 1 + \frac{\delta_{11}}{\delta_{12}} X_n \right)^{-1} X_n T_{\text{TOT}},$$

wherein  $\Upsilon_a = \Upsilon_a(X, X_n)$  satisfies

$$\Upsilon_{\text{TOT}} = \Upsilon_a \left( 1 + \frac{\delta_3}{\delta_4} + \frac{\delta_{15}}{\delta_{16}} + \frac{\delta_5}{\delta_6 + \delta_7} X + \frac{\delta_3}{\delta_4} \frac{\delta_8}{\delta_9 + \delta_{10}} X_n \right),$$

while  $X_n$  depends linearly on  $X$  via

$$\begin{aligned} \left( 1 + \frac{\delta_3}{\delta_4} + \frac{\delta_{15}}{\delta_{16}} \right) &= \frac{\delta_5}{\delta_6 + \delta_7} \left( \frac{\delta_7}{\delta_0} \Upsilon_{\text{TOT}} - 1 \right) X \\ &+ \frac{\delta_8}{\delta_9 + \delta_{10}} \frac{\delta_3}{\delta_4} \left( \frac{\delta_{10}}{\delta_0} \Upsilon_{\text{TOT}} - 1 \right) X_n, \end{aligned} \quad (37)$$

and  $X$  solves a quadratic of the form

$$0 = AX^2 + BX + C \quad (38)$$

where the constant coefficients A, B, and C are functions of the model parameters. For physically realistic solutions, we require  $X, X_n > 0$ . Therefore, we conclude that this model has at most two steady states and at most one of them may be stable.

As models increase in complexity, the algebra usually prohibits the construction of analytical expressions for the steady state solutions. In the following sections we present other methods that can be used to generate insight in such situations.

### 3.2 Nondimensionalization

When a mathematical model is first developed, the independent and dependent variables typically represent physical quantities (e.g., protein levels) which are measured in dimensional units (e.g., protein levels may be measured as the number of molecules per unit volume or the number of molecules per cell). The model may also contain parameters which relate to physical processes (e.g., reaction rates, Michaelis–Menten constants) and are also dimensional (e.g., rates may be measured per second, per hour, or per day). Nondimensionalization involves recasting the model in terms of dimensionless (or unit-less) variables. This process is instructive for several reasons. First, the number of model parameters is typically reduced. Second, the resulting dimensionless parameter groupings can provide useful information about the system's behavior. Further, if estimates of these parameters can be obtained and then compared, it is possible to identify physical processes that

dominate on a particular timescale and, thereby, rationale to simplify the governing equations. We illustrate these concepts by nondimensionalizing the enzyme kinetics and Schmitz models.

3.2.1 Case Study I:  
The Enzyme Kinetics  
Model  
(Nondimensionalization)

We introduce the dimensionless variables  $\tau, s, \epsilon, c,$  and  $p$  where

$$t = T\tau, \quad S = S_0s, \quad E = E_0\epsilon, \quad C = E_0c, \quad P = S_0p.$$

and the timescale  $T$  is specified below. It is natural to scale the complex  $C$  with  $E_0$  since the amount of complex that forms is limited by the amount of enzyme present. If the enzyme is working effectively (i.e., serving as an efficient catalyst), then the amount of product created will be comparable to the amount of substrate. Therefore, we scale  $P$  with  $S_0$  rather than  $E_0$ .

There are several possible choices for the timescale  $T$ . Consider Eq. 3. Initially, when  $C=0$ , the maximum rate of uptake of  $S$  is  $k_1E_0$  and similarly the initial rate of uptake of  $E$  is  $k_1S_0$ . The associated timescales are  $T_1 = 1/(k_1E_0)$  and  $T_2 = 1/(k_1S_0)$ . Since enzyme levels are typically much smaller than substrate levels (i.e.,  $E_0/S_0 = \epsilon \ll 1$ ), it is clear that  $T_2/T_1 = E_0/S_0 \ll 1$ . We conclude that  $T_1$  represents a *long* timescale, associated with substrate depletion, while  $T_2$  represents a *short* timescale, associated with the initial rapid uptake of enzyme.

Rescaling on the longer timescale, so that  $t = \tau T_1 = \tau/(k_1E_0)$ , Eqs. 7–9 transform to give

$$\frac{ds}{d\tau} = -s(1-c) + \kappa_\epsilon c, \tag{39}$$

$$\epsilon \frac{dc}{d\tau} = s(1-c) - \kappa_m c, \tag{40}$$

$$s(\tau = 0) = 1, \quad c(\tau = 0) = 0, \tag{41}$$

$$\text{where } \epsilon = \frac{E_0}{S_0}, \quad \kappa_\epsilon = \frac{k_{-1}}{k_1S_0} \text{ and } \kappa_m = \frac{k_{-1} + k_2}{k_1S_0}. \tag{42}$$

Comparing Eqs. 7–9 and 39–42 we note that nondimensionalization has reduced the number of model parameters from five to three. We remark further that in Eq. 40, the initial conditions supply  $dc(0)/d\tau = 1/\epsilon$ . Thus, if  $\epsilon \ll 1$ , then  $c$  will initially increase very rapidly on the timescale  $\tau$ .

3.2.2 Case Study III:  
The Schmitz Model  
(Nondimensionalization)

The procedure for nondimensionalizing the Schmitz model is identical to that used for the enzyme kinetics model. As the dimension of the system increases, and more processes are included, the number of ways to rescale the independent and dependent variables

increases rapidly. In such situations, it is important to consider which variables are expected to vary and over what timescale: the answers to these questions should help to identify appropriate scalings.

When studying Wnt signaling, inactivation of the DC plays a key role in the system dynamics and therefore when we nondimensionalize the Schmitz model time is rescaled so that  $t = \tau/\delta_{15}$  ( $\delta_{15}^{-1}$  is the timescale for inactivation of the DC). Variables relating to free  $\beta$ -catenin (i.e.,  $X, X_n, X_p, X_{pn}$ ) are all rescaled with  $\tilde{B} = \delta_0/\delta_{15}$ , the amount of  $\beta$ -catenin produced during the typical timescale  $\tilde{t}$ . This scaling eliminates  $\delta_0$  from the dimensionless equations (see below). When choosing the scalings for variables involving DC and TCF, we aim to preserve conservation laws. Accordingly, guided by Eq. 36, we scale  $\Upsilon_a, \Upsilon_i, \Upsilon_{an}, C_{XY}$ , and  $C_{XYn}$  with  $\Upsilon_{\text{TOT}}$ , the total amount of DC in the system. Similarly, we scale  $T$  and  $C_{XT}$  with  $T_{\text{TOT}}$ , the total amount of TCF in the system. Summarizing, we have

$$\begin{aligned} (X, X_n, X_p, X_{pn}) &= \tilde{B}(x, x_n, x_p, x_{pn}), \\ (\Upsilon_a, \Upsilon_i, \Upsilon_{an}, C_{XY}, C_{XYn}) &= \Upsilon_{\text{TOT}}(y_a, y_i, y_{an}, c_{xy}, c_{xyn}), \\ (T, C_{XT}) &= T_0(\theta, c_{x\theta}), \end{aligned}$$

where  $x(\tau), x_n(\tau), \dots, c_{x\theta}(\tau)$  are dimensionless variables. Under these scalings, the Schmitz model gives the following nondimensional system:

$$x' = 1 + (\tilde{\delta}_2 x_n - \tilde{\delta}_1 x) + (\tilde{\delta}_6 c_{xy} - \tilde{\delta}_5 x y_a), \quad (43)$$

$$x'_n = (\tilde{\delta}_1 x - \tilde{\delta}_2 x_n) + (\tilde{\delta}_9 c_{xyn} - \tilde{\delta}_8 x_n y_{an}) + (\tilde{\delta}_{12} c_{x\theta} - \tilde{\delta}_{11} x_n \theta), \quad (44)$$

$$x'_p = \tilde{\delta}_7 c_{xy} - \tilde{\delta}_{13} x_p, \quad (45)$$

$$x'_{pn} = \tilde{\delta}_{10} C_{xyn} - \tilde{\delta}_{14} x_{pn}, \quad (46)$$

$$\frac{1}{\omega} y'_a = \frac{1}{\omega} (\tilde{\delta}_4 y_{an} - \tilde{\delta}_3 y_a) + (\tilde{\delta}_6 c_{xy} - \tilde{\delta}_5 x y_a) + \tilde{\delta}_7 c_{xy} + \frac{1}{\omega} (\tilde{\delta}_{16} y_i - y_a), \quad (47)$$

$$\frac{1}{\omega} y'_i = \frac{1}{\omega} (y_a - \tilde{\delta}_{16} y_i), \quad (48)$$

$$\frac{1}{\omega} y'_{an} = \frac{1}{\omega} (\tilde{\delta}_3 y_a - \tilde{\delta}_4 y_{an}) + (\tilde{\delta}_9 c_{xyn} - \tilde{\delta}_8 x_n y_{an}) + \tilde{\delta}_{10} c_{xyn}, \quad (49)$$

$$\frac{1}{\omega} c'_{xy} = (\tilde{\delta}_5 x y_a - \tilde{\delta}_6 c_{xy}) + \tilde{\delta}_7 c_{xy}, \quad (50)$$

$$\frac{1}{\omega} c'_{xyn} = (\tilde{\delta}_8 x_n y_{an} - \tilde{\delta}_9 c_{xyn}) - \tilde{\delta}_{10} c_{xyn}, \quad (51)$$

$$\frac{1}{v} \theta' = (\tilde{\delta}_{12} c_{xT} - \tilde{\delta}_{11} x_n \theta), \quad (52)$$

$$\frac{1}{v} c'_{x\theta} = (\tilde{\delta}_{11} x_n \theta - \tilde{\delta}_{12} c_{x\theta}), \quad (53)$$

where primes denote differentiation with respect to  $\tau$  and  $\tilde{\delta}_i$  ( $i = 1, 2, \dots, 16$ ) are the following dimensionless parameters:

$$\tilde{\delta}_1 = \frac{\delta_1}{\delta_{15}}, \quad \tilde{\delta}_2 = \frac{\delta_2}{\delta_{15}}, \quad \tilde{\delta}_3 = \frac{\delta_3}{\delta_{15}}, \quad \tilde{\delta}_4 = \frac{\delta_4}{\delta_{15}}, \quad \tilde{\delta}_5 = \frac{\delta_5 \mathcal{Y}_{\text{TOT}}}{\delta_{15}}, \quad (54)$$

$$\tilde{\delta}_6 = \frac{\delta_6 \mathcal{Y}_{\text{TOT}}}{\delta_0}, \quad \tilde{\delta}_7 = \frac{\delta_7 \mathcal{Y}_{\text{TOT}}}{\delta_0}, \quad \tilde{\delta}_8 = \frac{\delta_8 \mathcal{Y}_{\text{TOT}}}{\delta_{15}}, \quad \tilde{\delta}_9 = \frac{\delta_9 \mathcal{Y}_{\text{TOT}}}{\delta_0}, \quad \tilde{\delta}_{10} = \frac{\delta_{10} \mathcal{Y}_{\text{TOT}}}{\delta_0}, \quad (55)$$

$$\tilde{\delta}_{11} = \frac{\delta_{11} T_0}{\delta_{15}}, \quad \tilde{\delta}_{12} = \frac{\delta_{12} T_0}{\delta_0}, \quad \tilde{\delta}_{13} = \frac{\delta_{13}}{\delta_{15}}, \quad \tilde{\delta}_{14} = \frac{\delta_{14}}{\delta_{15}}, \quad \tilde{\delta}_{16} = \frac{\delta_{16}}{\delta_{15}}, \quad (56)$$

$$\omega = \frac{(\delta_0/\delta_{15})}{\mathcal{Y}_{\text{TOT}}} \quad \text{and} \quad v = \frac{(\delta_0/\delta_{15})}{T_0}. \quad (57)$$

### 3.3 Asymptotic Analysis

In applied mathematics, if the (dimensionless) governing equations contain a small parameter, it is common to assume that there is an asymptotic expansion for the solution, as a power series in the small parameter. As we demonstrate below, this technique can be used systematically to simplify a mathematical model and, in so doing, provide useful information about the dynamics of its components.

#### 3.3.1 Case Study I: The Enzyme Kinetics Model (Asymptotics)

A key assumption of the enzyme kinetics model is that initial enzyme levels are much smaller than substrate levels. This assumption is represented in the dimensionless model equations via the small parameter  $\varepsilon = E_0/S_0 \ll 1$ . We exploit this small parameter by seeking a solution to Eqs. 39–41 of the form

$$s(\tau) \sim s_0(\tau) + \varepsilon s_1(\tau), \quad c(\tau) \sim c_0(\tau) + \varepsilon c_1(\tau). \quad (58)$$

Substituting with Eq. 58 in the governing equations and equating to zero terms of  $O(\varepsilon^n)$ , we deduce that, at leading order,  $s_0$  and  $c_0$  satisfy

$$\frac{ds_0}{d\tau} = \kappa_e c_0 - s_0(1 - c_0), \tag{59}$$

$$0 = s_0(1 - c_0) - \kappa_m c_0, \tag{60}$$

$$s_0(0) = 1, \quad c_0(0) = 0. \tag{61}$$

Thus the ODE for  $c$  reduces to an algebraic relation, giving  $c_0$  in terms of  $s_0$ , and an ODE for  $s_0$ , with the implicit solution

$$\kappa_m \log s_0(\tau) + s_0(\tau) = A - \kappa\tau, \quad c_0 = \frac{s_0}{\kappa_m + s_0}, \tag{62}$$

where  $A$  is a constant of integration. A problem arises when we attempt to impose the initial conditions: it is not possible simultaneously to satisfy both initial conditions. This is because the leading order problem is of lower order than the original one.

In order to resolve this problem, we use *matched asymptotic expansions*. We recall that  $c$  varies rapidly near  $\tau=0$  and, hence, examine the system dynamics near  $\tau=0$  by switching to the short timescale  $T = \tau/\varepsilon$ . In terms of  $T$ , the model becomes

$$\frac{d\tilde{s}}{dT} = \varepsilon(\kappa_e \tilde{c} - \tilde{s}(1 - \tilde{c})), \tag{63}$$

$$\frac{d\tilde{c}}{dT} = \tilde{s}(1 - \tilde{c}) - \kappa_m \tilde{c}, \tag{64}$$

$$\tilde{s}(0) = 1, \quad \tilde{c}(0) = 0. \tag{65}$$

where  $\tilde{s}(T) = s(\tau)$ ,  $\tilde{c}(T) = c(\tau)$ . As before, we seek asymptotic expansions for  $\tilde{s}$  and  $\tilde{c}$  in terms of  $\varepsilon \ll 1$ , of the form specified at Eq. 58. In this way, we obtain the following leading order solutions for  $\tilde{s}_0(T)$  and  $\tilde{c}_0(T)$ :

$$\tilde{s}_0(T) = 1, \quad \tilde{c}_0(T) = \frac{1 - e^{-(1+\kappa_m)T}}{1 + \kappa_m}. \tag{66}$$

The above approximate solution is accurate near  $\tau=0$  but not for  $\tau = O(1)$ , whereas Eq. 62 is accurate for  $\tau = O(1)$  but not for  $\tau \ll 1$ . The method of matched asymptotics involves choosing the constant of integration  $A$  to match Eqs. 62 and 66 [39]. By imposing the matching conditions

$$\lim_{\tau \rightarrow 0}(s_0(\tau), c_0(\tau)) = \lim_{T \rightarrow \infty}(\tilde{s}_0(T), \tilde{c}_0(T)),$$

we deduce that  $A=1$ .



In practice, similar asymptotic analyses can be used to study ODE models of signaling pathways. As we have seen, such models may involve large numbers of variables and parameters, and estimates for many parameters may be lacking. In such cases, progress can be made by using *order of magnitude* estimates for certain processes. For example, in [29], the authors assume that all binding reactions are rapid, apart from the binding of GSK3 $\beta$  to APC/Axin. Under this *fast kinetics* assumption, the ODEs for the relevant species reduce to algebraic equations, in the same way that, for the enzyme kinetics model, on the longer timescale the ODE for the complex  $c$  reduces to an algebraic relation (see Eq. 60).

To the best of our knowledge, the Schmitz model has yet to be subject to such asymptotic analysis. Referring to Eqs. 43–53, and by analogy with the asymptotic analysis of the enzyme kinetics model presented above, we note that the dynamics of the system will be strongly influenced by the ratios  $\omega$  and  $\nu$ . For example, if typical levels of  $\beta$ -catenin are much greater than levels of TCF and DC, then we could construct approximate solutions to the Schmitz model in the limit for which  $\nu \ll 1 \ll \omega$ . Such an analysis of the Lee model was performed by Mirams et al. [40]. Since the details are rather involved, we summarize the key points below and refer the interested reader to [40] for further details.

3.3.2 Case Study II:  
The Lee Model  
(Asymptotics)

Numerical simulations of the Lee model generated using parameter estimates reported in [29] (see Fig. 5) suggest that the processes involved in the Wnt signaling pathway act over at least two different timescales. Lee et al.’s parameter estimates indicate that the basal rate at which  $\beta$ -catenin is degraded is much smaller than the rate at which the DC becomes inactive. This discrepancy is exploited to define a small parameter,  $\eta = \alpha_{16}/\alpha_{15}$ , which is the ratio of the rate at which  $\beta$ -catenin undergoes natural decay to the rate at which the DC becomes inactive. The dimensionless parameters are then rescaled by multiplying them by appropriate powers of  $\eta$  so that they are  $O(1)$ . By retaining terms of leading order, the following reduced model is obtained:

$$\frac{dD_a}{dt} = \bar{\alpha}_1 W(1 - D_a) - \bar{\alpha}_2 D_a, \tag{67}$$

$$\frac{dY_i}{dt} = -(\bar{\alpha}_5 D_a + \bar{\alpha}_3 + \bar{\alpha}_7) Y_i + Y_a + \frac{\bar{\alpha}_6 N}{1 + \eta \bar{K}_1 X}, \tag{68}$$

$$\eta \frac{dC_{XR}}{dt} = \bar{\alpha}_{10} X Y_a - \bar{\alpha}_{11} C_{XR}, \tag{69}$$

$$\frac{dN}{dt} = \left( (\bar{\alpha}_5 D_a + \bar{\alpha}_7) \Upsilon_i - \left( \frac{\bar{\alpha}_6}{(1 + \eta \bar{K}_1 X)} + \bar{\alpha}_{18} \right) N + 1 \right) \frac{1}{1 + \bar{K}_2}, \quad (70)$$

$$\frac{d\Upsilon_a}{dt} = \frac{\bar{\alpha}_3 \Upsilon_i - \Upsilon_a - \frac{dC_{XY}}{dt}}{1 + \bar{K}_3 X}, \quad (71)$$

$$\frac{1}{\eta} \frac{dX}{dt} = \bar{\alpha}_{15} - \bar{\alpha}_{10} X \Upsilon_a - \bar{\alpha}_{16} X. \quad (72)$$

We remark that Eq. 67 decouples and if a constant Wnt stimulus is applied ( $W(t) = W$ , constant), then

$$D_a \rightarrow \frac{\bar{\alpha}_1 W}{\bar{\alpha}_1 + \bar{\alpha}_2}.$$

We note further that the time derivatives in Eqs. 67–72 are premultiplied by three different powers of  $\eta$ . This suggests that model processes act on three distinct timescales, a prediction that is consistent with the rapid fluctuations and slow increases depicted in Fig. 5.

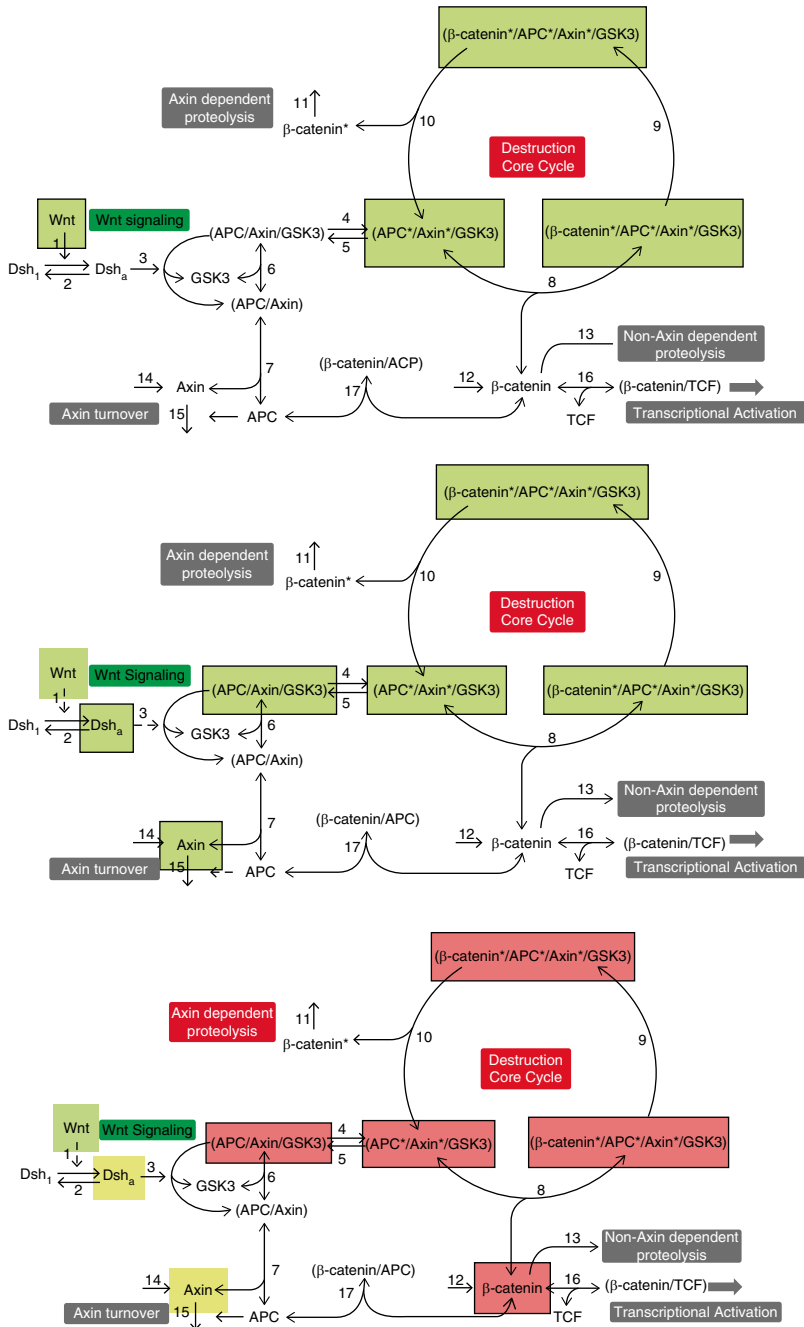
As we have already seen for the enzyme kinetics model (Eq. 58), it is possible to analyze the reduced Lee model on different timescales; here we have short, medium, and long timescales for which  $t = O(\eta)$ ,  $O(1)$ , and  $O(\eta^{-1})$ , respectively. In each case, asymptotic expansions in powers of the small parameter  $\eta$  are sought and used to simplify the governing equations. The results of this analysis can be summarized as follows (*see* [40] for details).

1. Short timescale ( $t = O(\eta)$ ): all model variables except  $\Upsilon_i$  and  $C_{XY}$  are constant, at leading order. The dominant reaction is phosphorylation of  $\beta$ -catenin by active destruction complex.
2. Intermediate timescale ( $t = O(1)$ ): the dominant reaction is found to involve inactivation of the destruction complex.
3. Long timescale ( $t = O(\eta^{-1})$ ): the dynamics are dominated by degradation of free  $\beta$ -catenin.

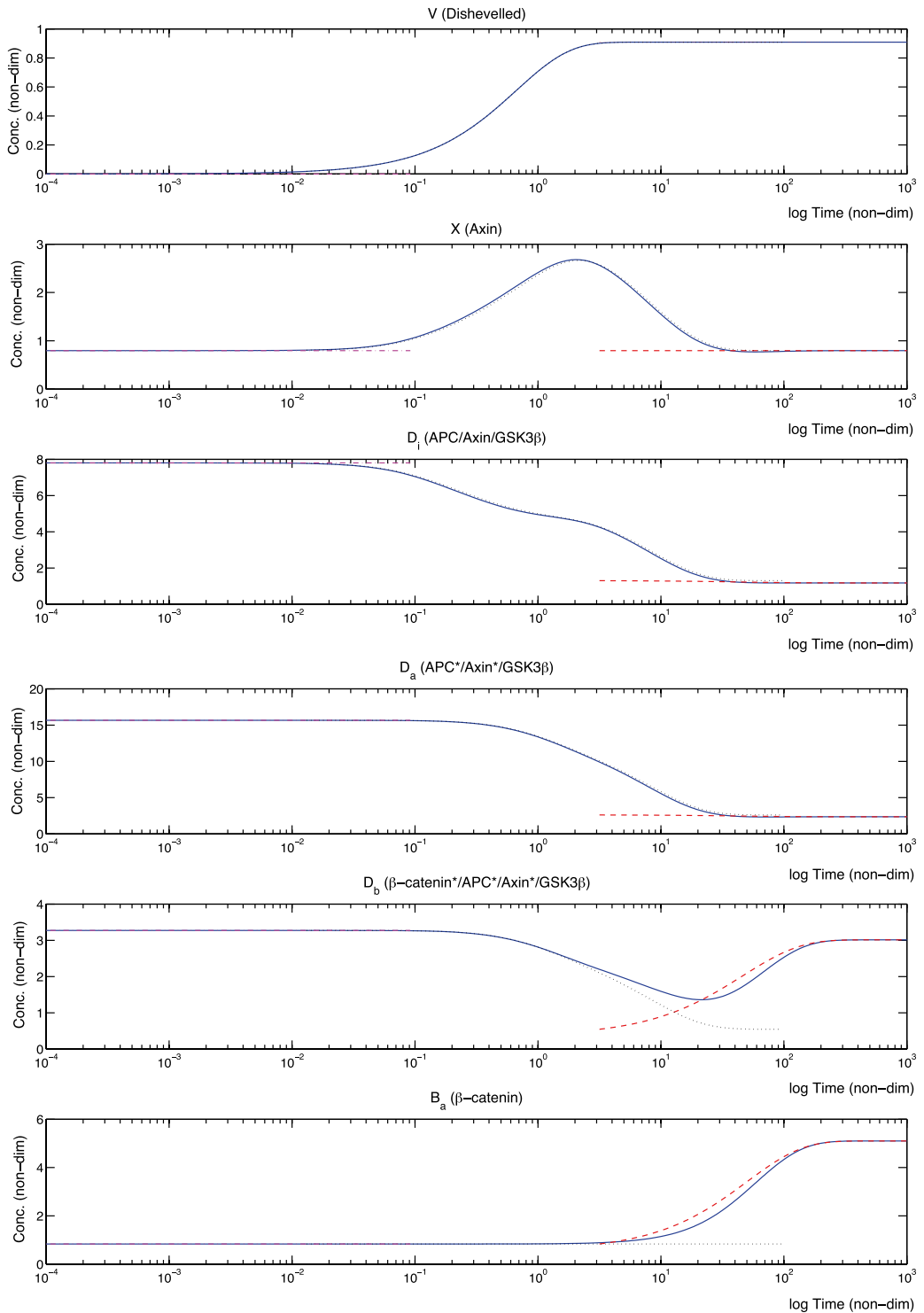
Pathway components acting on the short, intermediate, and long timescales are highlighted in Fig. 4, while Fig. 5 shows good agreement between the approximate solutions and those of the full model.

### 3.4 Parameter Analyses

The selection of model parameters, their physical meaning, and numerical values are especially important; parameter analysis examines the response of the system to changes in parameters. Many methods for estimating parameters depend on time course data. These data generally give a quantitative measure of the variable



**Fig. 4** Series of schematics showing which components of the Lee model of Wnt signaling are active on the short (*top*), medium (*middle*), and long timescales. The active components on each timescale are highlighted with *bold borders*. Figure reproduced from [40], with permission



**Fig. 5** Series of figures showing how the Lee model responds to a Wnt stimulus ( $W=1$ ) that is applied at  $t=0$  when the pathway is in equilibrium ( $W=0$ ) at  $t=0$ . Also shown is the asymptotic solution obtained by matching the short, medium, and long time approximations to the Lee model. There is good agreement between the approximate and numerical solutions at all timescales. Key: numerical simulations of the (dimensionless) Lee model, Eqs. 10–24 (solid line); short, medium, and long time approximations are represented by dash-dotted, dotted, and dashed lines, respectively. Figure reproduced from [40], with permission

level, such as mRNA or protein concentration level, at different time points. Testing a model against experimental data is a good way to *validate* or *invalidate* it; however, gathering experimental data is often too expensive to determine all parameter values and overfitting, i.e., describing noise instead of the relationship is a risk, as demonstrated for Wnt signaling later in this section. Following parameter estimation (using optimization) or parameter inference (using statistics), a good way to test a model is by performing parameter sensitivity analysis: this evaluates qualitative or quantitative relationships between parameters and their effect on the system outcome [41].

### 3.4.1 Parameter Estimation and Wnt Data

Ultimately, every model should be tested against data, a process that can either invalidate the model or provide evidence in its favor, if it provides a good fit under acceptable conditions. The aim is to estimate parameters that drive the model close to the data; this can be done using minimization techniques. Effectively, one calculates an objective function which is defined as the difference between the model simulated for particular value of parameters  $\kappa$  and the observations (data), and aims to minimize the error of the objective function, often performed iteratively [42–44].

Since the publication of the Lee model [29], where estimates of the parameters controlling Wnt signaling were based on data from *Xenopus* extracts, few studies have quantitatively studied the dynamics of the Wnt pathway. This knowledge gap means that currently it remains difficult to test the models that have arisen in recent years. This problem is not uncommon in systems medicine. We also remark that the *Xenopus* data gathered by Lee et al. may be markedly different from those for mammalian Wnt signaling. In [13], dynamic changes in  $\beta$ -catenin levels were investigated in *Xenopus* extracts. They demonstrated that absolute levels of  $\beta$ -catenin did not dictate the Wnt signaling outcome: rather the  $\beta$ -catenin fold-change was the crucial variable. They used the Lee model to test their experimental results and, via sensitivity analysis, identified that the model confirmed their experimental findings.

Quantification of Wnt signaling in mammalian cell lines was undertaken by Hernández et al. [14] and Tan et al. [15]. Discrepancies with data from *Xenopus* extracts (such as higher Axin levels and lower APC levels in mammalian cells) highlight the need for caution in data gathering and for further quantification of the pathway. Since these measurements were made at steady state, they do not yet permit elucidation of transient Wnt signaling. More recent measurements of cytoplasmic and nuclear  $\beta$ -catenin in response to a Wnt stimulus provide a valuable first look at the dynamics of the pathway [45].

The above studies provide preliminary insight into the Wnt pathway but much remains to be done. The data are not yet of

sufficient quality to discriminate between most models (which typically contain many molecular species). Caution must be taken when applying data. For example, where data generated from non-mammalian systems may be used in a model that addresses clinical outcomes. For systems medicine to have the greatest impact, modeling (with prediction) and experimentation (to test predictions) must proceed iteratively.

### 3.4.2 Parameter Inference

There are often cases where it is either infeasible or impossible experimentally to determine values for parameters that describe a given model. In such cases, we may be able to estimate (some of) the parameters using statistical inference. In general the aim is to identify the values of the parameters,  $\theta$  (ideally including corresponding confidence regions), for which a model best explains the data.

A reliable way of doing so is to focus on the likelihood  $L(\theta)$ , which is defined as the probability of observing the data ( $x$ ) given parameters ( $\theta$ ):

$$L(\theta) := P(x | \theta).$$

Varying  $\theta$  to identify the value for which this probability is maximized gives the maximum-likelihood estimate. There is a rich literature on this topic and how confidence of the estimates can be assessed [46].

Likelihood estimates center around the available data. In many circumstances we may have additional information, for example based on biophysical arguments, about which parameter values can be ruled out. Incorporating such *prior information* is hard in a pure likelihood framework, but lies at the heart of Bayesian inference [47]. Here inferences are based on the *posterior distribution* over model parameters. The posterior distribution can be described starting from Bayes rule:

$$P(\theta | x) \propto P(x | \theta)\pi(\theta). \quad (73)$$

$P(\theta | x)$ , the probability of  $\theta$  given  $x$ , is called the posterior probability,  $P(x | \theta)$  is the likelihood function, and  $\pi(\theta)$  is the prior probability (knowledge about parameters before we begin fitting to data) [48]. As well as the full (joint) posterior distribution, one may also analyze the marginal posterior distributions which are the individual distributions over each parameter.

In certain cases, such as for large, complex systems, computing the likelihood is impractical. In such cases approximate Bayesian computation (ABC) should be considered [49]. Instead of the likelihood, a distance function is used to compare the actual data with data simulated by a model, denoted  $x_m$ . If the underlying model is given by  $f = f(x_m | \theta)$ , then we express the ABC posterior function by

$$P_{\text{ABC}}(\theta | x) \propto \mathbb{I}(\Delta(x, x_m) \leq \varepsilon) f(x_m | \theta) \pi(\theta) \quad (74)$$

where  $\Delta(a, b)$  denotes a distance measure between  $a$  and  $b$ , and  $\varepsilon$  is the tolerance level that determines how well real and simulated data should agree.

By evaluating the posterior function, ABC allows the modeler to identify parameter regions that are of interest, and ignore those that are not. Furthermore, the posterior distribution gives information about joint distributions in parameter space and can reveal multivariate dependencies between parameters.

ABC for parameter inference has been implemented in the software package ABC-SysBio with support for parallelization [50]. For the examples given below, we used the CUDA implementation of ABC-SysBio with a Euclidean distance measure between model and data [51, 52]. Proceeding to analyze the Lee and Schmitz models, we do not try to infer all of the model parameters, since this is not possible with the data available, but instead study a 3D subset of parameter space. We choose free parameters that have direct (or strong) influence on the dynamics of  $\beta$ -catenin, since this is the species for which we have experimental measurements. The data used for fitting are published in [45]: they describe how the level of  $\beta$ -catenin changes over time in the cytoplasm and nucleus, following application of a Wnt stimulus to the system. These data, alongside the results of the parameter inference, are shown in Fig. 6.

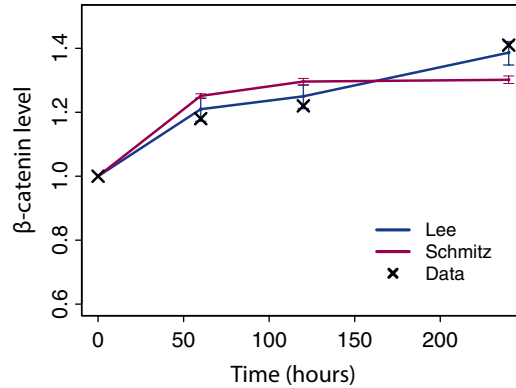
For the Lee model, we study the  $\beta$ -catenin-DC binding rate ( $\alpha_{10}$ ) that has a prior of  $[0, 100]$ , the  $\beta$ -catenin degradation rate that is independent of the DC ( $\alpha_{16}$ ), and the binding rate of  $\beta$ -catenin to TCF ( $\alpha_{19}$ ). The latter two parameters both have priors of  $[0, 1]$ . The marginal posterior distributions for these three parameters (Fig. 7) show that the  $\beta$ -catenin-DC binding parameter takes values over the lower half of its prior range, whereas the other two parameters can take any values spanning the prior range. This suggests that for this model the parameter that has the greatest impact on outcome is the  $\beta$ -catenin-DC binding rate; however, we note the larger prior range over this parameter.

For the Schmitz model, we study the  $\beta$ -catenin production rate ( $\delta_0$ ), the  $\beta$ -catenin shuttling rate ( $\delta_1$ ), and the binding rate of  $\beta$ -catenin to TCF ( $\delta_{11}$ ). The prior used for each parameter is  $[0, 1]$  and we see from Fig. 7 that the marginal posterior distributions are relatively stiff: each parameter is constrained to lie within a narrow range relative to its prior. In order to fit the data, the rates of  $\beta$ -catenin shuttling and binding to TCF must be low, while the rate of  $\beta$ -catenin production must be high.

### 3.4.3 Sensitivity Analysis

Sensitivity analysis investigates how a model responds to perturbations around a set of parameter values and characterizes its robustness: a *robust* system is one for which perturbations of the parameters





**Fig. 6** Data published in [45] were used to fit the Lee and Schmitz models using approximate Bayesian computation for parameter inference.  $\beta$ -catenin concentration units were normalized based on their initial values. From the inference, we can see that the Lee model provides a better fit to the data

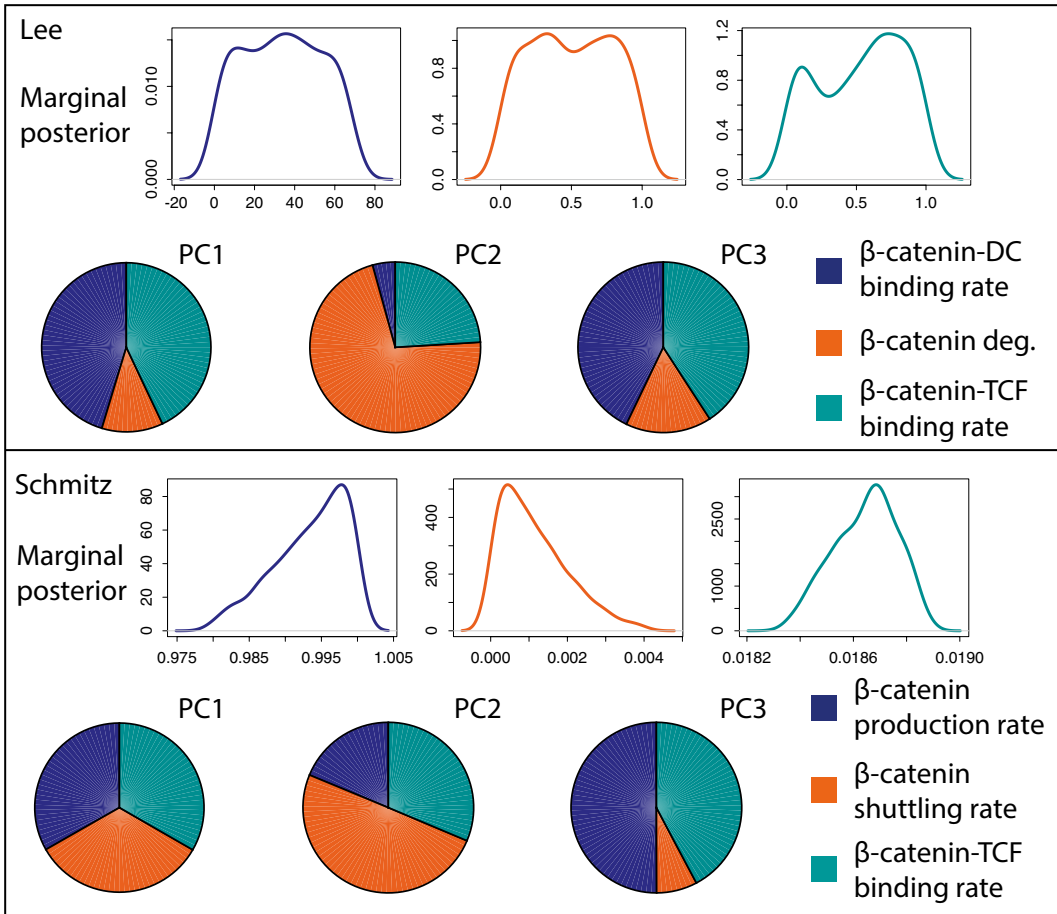
or initial conditions do not change the outcome. However, many trade-offs between sensitivity and robustness exist [53–55].

Local sensitivity analysis determines how parameter perturbations affect the output of a system. Estimated or inferred parameters can be used as a baseline for parameter sensitivity. If the output of  $dx/dt = f(x, \kappa)$  is approximated by a first-order Taylor series in a neighborhood of reference input values, then the local sensitivity coefficient  $s_{i,j}$  is the partial derivative of the  $i^{\text{th}}$  state to the  $j^{\text{th}}$  parameter:

$$s_{i,j}(t) = \frac{\partial x_i(t)}{\partial \kappa_j}, \quad (75)$$

The elements  $s_{i,j}$  define a sensitivity matrix  $S = \partial \mathbf{x} / \partial \kappa$ . This local method provides information about the sensitivity in a given parameter region but not the global sensitivity landscape. Local sensitivity analysis can reveal parameters that are sensitive or robust to perturbations in the region of interest.

Principal component analysis (PCA) offers another way to investigate system sensitivity. This technique can be readily applied to the posterior distribution obtained following Bayesian inference. The principal components are constructed by evaluating the eigenvalues and eigenvectors of the covariance matrix of the parameters: the first principal component (given by the largest eigenvalue) corresponds to the direction in which the posterior is most wide; the last principal component (given by the smallest eigenvalue) corresponds to the direction in which the posterior is most narrow [49, 56]. The last few principal components represent the most sensitive (or “stiff” parameters) [57].



**Fig. 7** Posterior distributions and sensitivity analysis for the Lee and Schmitz models. Histograms of marginal posteriors for each free parameter in the two models are shown. The marginal posterior is the probability distribution for a single parameter, given data describing  $\beta$ -catenin dynamics in cytoplasmic and nuclear compartments [45]. Principal component (PC) analysis allows us to assess the sensitivity of the parameters to small perturbations: the last PC (PC3), contains the most sensitive parameters. We see that for each model, two parameters dominate PC3 and, thus, are most sensitive in this system

In Fig. 7, sensitivity analysis via PCA for the Lee and Schmitz models is shown. The principal components (PC) are ordered 1–3, thus PC3 is the last component and contains the most sensitive parameter combinations. For both models, PC3 is dominated by two parameters: the rates of  $\beta$ -catenin binding to the destruction complex (DC) or to TCF for the Lee model ( $\alpha_{10}, \alpha_{19}$ ); and the rates of  $\beta$ -catenin production or binding to TCF for the Schmitz model ( $\delta_0, \delta_{11}$ ). These results suggest that the Lee model is more robust to changes in the  $\beta$ -catenin degradation rate ( $\alpha_{16}$ ), and that the Schmitz model is more robust to changes in the  $\beta$ -catenin shuttling rate ( $\delta_1$ ).

## 4 Techniques for the Comparison and Discrimination of Models

Given a set of models that describe similar biological phenomena, a challenge is to determine which model best describes the system, given the evidence available. In this section we describe two methods that enable comparison and discrimination between models. The first employs ABC, introduced above, and has already gained a strong foothold in systems medicine [50, 58–60]. The second is model discrimination with the use of algebraic matroids; as far as we know this is a recent addition to the modeler’s toolkit and holds great potential for advances in systems medicine.

### 4.1 Model Selection via ABC

Returning now to the Lee and Schmitz models, we consider how to choose between models using ABC model selection. We have already demonstrated how methods for parameter inference, such as ABC, can yield the posterior distributions over the parameters of a model (given data) and discussed briefly how this can be interpreted. For two or more models ( $M_i$ ,  $i = 1, \dots, n$ ) some measure of the evidence for each model is needed [61],

$$P(M_i | x) \propto P(x | M_i)\pi(M_i), \quad (76)$$

where (as previously)  $x$  represents the data and  $\pi$  the prior probability.

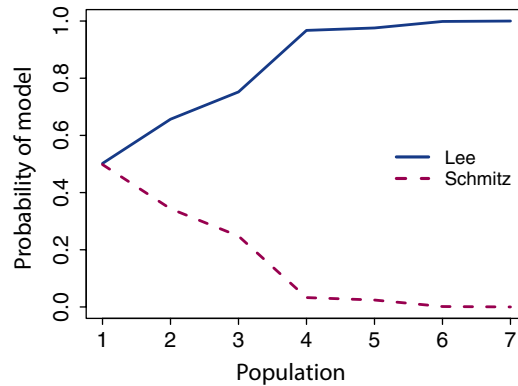
The ABC approach may be extended to parameter inference and model selection simultaneously using a joint space approach [49]. This may be performed for  $M$  models where  $M = [M_1, \dots, M_n]$ , by assigning to each model (and parameters therein) a prior distribution and perturbation kernel that designates weights for model transition. The algorithm accepts  $N$  particles at the  $\epsilon_F$  tolerance, which forms the joint posterior distribution  $P(\alpha, M | \mathbf{x})$  and upon marginalizing over parameters, the marginal posterior distribution  $P(M | \mathbf{x})$  is approximated, providing a measurement for model selection. Bayesian model selection, like other approaches including the likelihood ratio test or Akaike Information Criteria (AIC), also penalizes over-parameterization.

The AIC for model  $M_i$ , with  $i \in \{1, \dots, n\}$ , is defined as

$$\text{AIC}_i = -2 \log L(\theta_i^*; x, M_i) + 2k_i, \quad (77)$$

where  $L$  is the likelihood, and  $\theta_i^*$  and  $k_i$  are (respectively) the maximum likelihood parameter and number of parameters in model  $M_i$ . This criterion, probably the best known model selection tool, makes explicit the penalty for an increased number of parameters. However, as the amount of data increases, the AIC introduces bias and tends to favor models that are over-parameterized. Therefore the Bayesian information criterion (BIC),

$$\text{BIC}_i = -2 \log L(\theta_i^*; x, M_i) + k_i \log n, \quad (78)$$



**Fig. 8** Model selection via ABC for the Lee and Schmitz models. The results show that, over successive populations, evidence in favor of the Lee model grows until there is a high probability that this model will be selected, given the data published in [45]

may be preferred, as it remains unbiased for large samples,  $n$ . The BIC is effectively an approximation to the model probability (76); the penalty term, explicit in the AIC and BIC definitions, is implicit in (76), where it enters via the priors for each model.

Model selection chooses, from among a set of candidate models, the model that best explains observed data. Two things need to be kept in mind: (1) one model will always be chosen as the best but this does not mean that the model is necessarily a good one; ideally model selection should go hand-in-hand with model checking (and topological sensitivity analysis [62]). (2) Model selection depends on the data available for testing the different models; since different data may favor different models, careful experimental design should precede model selection. With these issues in mind we have the pragmatic choice about which statistical model selection framework to employ. Fully Bayesian, even in an ABC context, is more expensive than identifying the maximum likelihood parameter set and applying AIC or BIC.

Shown in Fig. 8 are the results of ABC model selection for the Lee and Schmitz models, with the probability of the model given for successive iterations (populations). We see that initially both models are equally probable, but subsequently the probability of selecting the Schmitz model drops to close to zero and we conclude that the Lee model is favorable given these data and parameter combinations.

#### 4.2 Model Discrimination Using Parameter-Free (Algebraic) Approaches

When parameter values are unknown or cannot be estimated from data, one may still be able to discriminate between competing models. We present two approaches, one that requires no data (rather qualitative insight into whether the system can have multiple responses) and another method which requires either highly resolved single cell data or multiple replicates of steady state measurements.

4.2.1 *Precluding/  
Asserting Behaviors  
via Chemical Reaction  
Network Theory*

Chemical reaction network theory (CRNT) studies the structure of a model (which can also be described as a network) constructed from chemical reactions without relying on specific parameter values. The aim here is to use such theory to preclude (and sometimes assert) possible qualitative behaviors in the positive orthant, i.e.,  $\mathbb{R}_{>0}$ . Cases where multiple positive states are stable (i.e., biologically accessible) are of particular biological importance for cellular decision making, for example, differentiation into one of two or more specialized cell lineages.

The field of CRNT initially focused on a structural property of a model called deficiency, which could preclude multiple steady states [63, 64]. Then theorems were proved for precluding/asserting multiple equilibria by studying the cycles in the graph of a network, or the sign of the determinant of the Jacobian; some of these approaches can provide conditions on the parameters for behaviors such as bistability and oscillations [65–70]. An excellent and comprehensive survey of techniques for multistationarity was written by Joshi and Shiu [71]. One main tool for precluding multistationarity of a model is testing whether it is injective (a model, including conservation relations, is *injective* if  $F(x, \kappa) = F(\tilde{x}, \kappa) \Rightarrow x = \tilde{x}$ ). Here we demonstrate the application of multistationarity tests (developed for chemical reaction networks) to Wnt signaling models.

We begin with the Lee model. First we test injectivity, noting that while injectivity precludes multistationarity, failure of injectivity does not imply multistationarity. We use the algorithms in the CRNT Toolbox to determine whether the system can ever admit multiple positive steady states—multistationarity [72]. The Lee model fails injectivity, but cannot admit multiple positive steady states for any values of the system parameters and/or total concentration amounts (algorithms within [72]). Conversely, the Schmitz model has the capacity for multiple steady states; however, as calculated earlier, only one can ever be stable. Therefore, in this example, since both models only can have one stable steady state, it is difficult to use only qualitative data to discriminate between them. Clearly, if data suggested two stable states could exist, and all of the data had the same initial conditions, then one could rule both models out.

4.2.2 *Model  
Discrimination Using  
Coplanarity via Algebraic  
Geometry*

When data from a model clearly supports a specific behavior—whether monostable, bistable, or oscillatory, qualitative approaches such as those mentioned above may be a good first step for classifying models, especially if the data are not sufficient to estimate parameters. However, if steady state data are available, then determining steady state invariants may be helpful for determining whether a model is compatible with given data using a statistical parameter-free model discrimination method.

Since often data are not available for all model species, variables must be eliminated. A systematic technique from algebraic geometry proceeds by computing the Gröbner Bases of the model variety (studying the model at steady state) and eliminating unobservable variables. The resulting steady state invariant enables us to focus on part of the system and to test whether the data suggests that the relationships between species still hold. Notions of dependence and independence between model variables can also be studied using algebraic matroids and were recently applied to steady state model discrimination [27].

For smaller models, the steady states can be determined explicitly. For example, for the Schmitz model, the steady state values can be expressed in terms of  $X$  and  $X_n$ : all other variables can be eliminated by exploiting conservation laws and using variable substitution (see Eqs.37–38). Either by hand, by computing the matroid, or by using Gröbner bases, the polynomial relationship/algebraic dependence between  $X$  and  $X_n$  in the Schmitz model gives the following invariant:

$$\mathcal{I} = \delta_0\delta_3\delta_4\delta_6(\delta_8 + \delta_9)X^2 + (\delta_0\delta_2\delta_7\delta_9(\delta_5 + \delta_6) - \delta_1\delta_3\delta_4\delta_6(\delta_8 + \delta_9))XX_n - \delta_1\delta_2\delta_7\delta_9(\delta_5 + \delta_6)X_n^2,$$

which vanishes at steady state (i.e.,  $\mathcal{I} = 0$ ). Effectively, we aim to test whether the data are coplanar with our model, via the steady state invariant transformation. Model compatibility is determined by computing the coplanarity error ( $\Delta$ ) via the singular value decomposition of the matrix

$$\begin{pmatrix} \widehat{X}^2 & \widehat{X}_n^2 & \widehat{X}\widehat{X}_n \end{pmatrix} \begin{pmatrix} \tilde{h}_1 \\ \tilde{h}_2 \\ \tilde{h}_3 \end{pmatrix} = 0,$$

where  $\widehat{X}$  denotes the observed value of species  $X$ . The null hypothesis (that the model is compatible with the data) can be rejected when the coplanarity error (normalized smallest singular value) is less than a statistical bound, which is determined by the Gaussian measurement noise in the data and the invariant structure [73]. This method was recently applied to  $\beta$ -catenin localization data (cytoplasmic,  $X$ ; and nuclear,  $X_n$ ) published in [27, 45]. The Schmitz model could be ruled out if data were perturbed less than  $10^{-5}$  by measurement error/noise; for higher levels of noise, the model is compatible.

---

## 5 Discussion

Paradoxically, technological advances sometimes create new challenges for clinicians. For example, as the number and variety of treatments for cancer increase, it can be difficult to identify the combination of treatments that will most benefit a given patient (if a unique, optimal treatment even exists). The situation is further complicated when we consider the different types of data that can be used as a basis for diagnosis and treatment planning; it is often impossible to integrate the available data by linear thinking alone. Systems medicine aims to address these challenges by developing mathematical and computational tools that integrate different types of information in order to generate objective decisions for patient treatment. In this chapter we have focused on ODE models, a class of models widely used in systems medicine, particularly to study signaling pathways. We have reviewed a variety of techniques that can be used to develop and analyze ODE models, using models of enzyme kinetics and the Wnt signaling pathway as test cases.

Many of the techniques that we have presented are already well established (such as model development, nondimensionalization, identification of steady state solutions, asymptotic analysis, and parameter sensitivity analysis); however, others are less well known (such as ABC, CRNT, and matroid-informed coplanarity). In addition to the benefit that these methods bring to the field, model development for systems medicine—in its increasing sophistication—is helping to stimulate further development and application of mathematical and statistical techniques.

Many of the challenges in systems medicine arise because most biological processes, including the actions of whole pathways, do not act in isolation. For example, at the subcellular level, pathway cross-talk can have a significant effect on cell function. In particular, there is growing evidence of cross-talk between Wnt and E-cadherin [74], Wnt and Erk [33]), and Wnt and the Hippo pathway [75]. Even simplistic models of such pathway cross-talk quickly become large and demand sophisticated techniques for their analysis. The situation becomes even more complex when we consider the impact of signaling pathways at the multicellular and tissue scales. The impact of Wnt signaling at the multicellular and tissue levels has been studied theoretically, most prominently in models of intestinal crypts [76–79]. These models (for example) introduce spatial dependence by imposing a graded Wnt distribution along the crypt axis [78] or provide comparison of a continuum model with a cell-based model that incorporates heterogeneity and noise [79]. In [74], a multiscale model of interactions between the pathways affecting  $\beta$ -catenin and E-cadherin is developed and used to study the role of epithelial–mesenchymal transitions in cancer growth and metastasis, whereas in [80] a simple rule-based



model for cross-talk between the Wnt and delta-notch pathways is embedded within discrete epithelial cell agents and used to study cell fate specification within the intestinal crypt. In addition to these theoretical studies (ever growing in complexity), more sophisticated data collection is urgently needed as a basis for hypothesis testing and model (in)validation.

We end by proposing two grand challenges, whose solutions will bear much fruit in systems medicine. The first is to incorporate multiple levels of information—from biochemical reactions within a single cell to tissue-level processes—into cohesive models. The second is to incorporate data which is resolved in space and time into a theoretical framework. There are, of course, many other important challenges, and work in these areas should provide many exciting opportunities for theoreticians in systems medicine for years to come.

---

## Acknowledgements

All authors acknowledge funding from King Abdullah University of Science and Technology (KAUST) KUK-C1-013-04 and the workshop funded by this grant on Model Identification (January 2014). HAH gratefully acknowledges funding from EPSRC Fellowship EP/K041096/1. All authors also thank Gary Mirams for his help with Figs. 4 and 5.

## References

1. Alberts B, Johnson A, Lewin J, Raff M, Roberts K, Walter P (2002) *Molecular biology of the cell*, 6th edn. Garland Science, New York
2. Stark J, Hardy K (2003) *Science* 301(5637):1192
3. Murray JD (2008) *An introduction to mathematical biology: Pt. 1*, 3rd edn. Springer, Berlin
4. Gardiner CW (2009) *Stochastic methods: a handbook for the natural and social sciences*, 4th edn. Springer, Berlin
5. Jost J (2005) *Dynamical systems: examples of complex behaviour*. Springer, Berlin
6. Gilbert N (2008) *Agent-based models: quantitative applications in the social sciences*. SAGE Publications, London
7. von Neumann J (1966) *Theory of self-reproducing automata*. University of Illinois Press, Urbana
8. Wolfram S (1983) *Rev Mod Phys* 55(3):601
9. Logan CY, Nusse R (2004) *Annu Rev Cell Dev Biol* 20:781
10. Polakis P (2000) *Genes Dev* 14(15):1837
11. Reya T, Clevers H (2005) *Nature* 434(7035):843
12. Vermeulen L, De Sousa E Melo F, van der Heijden M, Cameron K, de Jong JH, Borovski T, Tuynman JB, Todaro M, Merz C, Rodermond H, Sprick MR, Kemper K, Richel DJ, Stassi G, Medema JP (2010) *Nat Cell Biol* 12(5):468
13. Goentoro L, Kirschner MW (2009) *Mol Cell* 36(5):872
14. Hernández AR, Klein AM, Kirschner MW (2012) *Science* 338(6112):1337
15. Tan CW, Gardiner BS, Hirokawa Y, Layton MJ, Smith DW, Burgess AW (2012) *PLoS ONE* 7(2):e31882
16. Lloyd-Lewis B, Fletcher AG, Dale TC, Byrne HM (2013) *Wiley Interdiscip Rev: Syst Biol Med* 5(4):391
17. Clevers H, Nusse R (2012) *Cell* 149(6):1192
18. Franca-Koh J, Yeo M, Fraser E, Young N, Dale TC (2002) *J Biol Chem* 277(46):43844
19. Wiechens N, Heinle K, Englmeier L, Schohl A, Fagotto F (2004) *J Biol Chem* 279(7):5263

20. Cong F, Varmus H (2004) *Proc Natl Acad Sci USA* 101(9):2882
21. Henderson BR, Fagotto F (2002) *EMBO Rep* 3(9):834
22. Itoh K, Brott BK, Bae GU, Ratcliffe MJ, Sokol SY (2005) *J Biol* 4(1):3
23. Habas R, Dawid IB (2005) *J Biol* 4(1):2
24. Heuberger J, Birchmeier W (2010) *Cold Spring Harb Perspect Biol* 2(2):a002915
25. Barry ER, Camargo FD (2013) *Curr Opin Cell Biol* 25(2):247
26. Basan M, Idema T, Lenz M, Joanny JF, Risler T (2010) *Biophys J* 98(12):2770
27. MacLean AL, Rosen Z, Byrne HM, Harrington HA (2015) *Proc Natl Acad Sci USA* 112(9):2652
28. Li VSW, Ng SS, Boersema PJ, Low TY, Karthaus WR, Gerlach JP, Mohammed S, Heck AJR, Maurice MM, Mahmoudi T, Clevers H (2012) *Cell* 149(6):1245
29. Lee E, Salic A, Krüger R, Heinrich R, Kirschner MW (2003) *PLoS Biol* 1(1):e10
30. Kuhl M, Kracher B, Gross A, Kestler H (2014) In: Hoppler S, Moon R (eds) *Wnt signaling in development and disease: molecular mechanisms and biological functions*. Wiley, Hoboken, pp 153–160
31. Cho KH, Baek S, Sung MH (2006) *FEBS Lett* 580(15):3665
32. Kogan Y, Halevi Tobias KE, Hochman G, Baczmanska AK, Leyns L, Agur Z (2012) *Biochem J* 444(1):115
33. Kim D, Rath O, Kolch W, Cho KH (2007) *Oncogene* 26(31):4571
34. van Leeuwen IMM, Byrne HM, Jensen OE, King JR (2007) *J Theor Biol* 247(1):77
35. Schmitz Y, Wolkenhauer O, Rateitschak K (2011) *J Theor Biol* 279(1):132
36. Schmitz Y, Rateitschak K, Wolkenhauer O (2013) *Cell Signal* 25(11):2210
37. Krüger R, Heinrich R (2004) *Genome Inform* 15(1):138
38. Glendinning P (1994) *Stability, instability and chaos: an introduction to the theory of nonlinear differential equations*. Cambridge University Press, Cambridge
39. Kevorkian J, Kole JD (1981) *Perturbation methods in applied mathematics*. Applied mathematical sciences, 1st edn. Springer, Berlin
40. Mirams GR, Byrne HM, King JR (2010) *J Math Biol* 60(1):131
41. Saltelli A, Ratto M, Tarantola S, Campolongo F (2005) *Chem Rev* 105(7):2811
42. Brewer D, Barenco M, Callard R, Hubank M, Stark J (2008) *Philos Trans A Math Phys Eng Sci* 366(1865):519
43. Gershenfeld N (2011) *The nature of mathematical modeling*. Cambridge University Press, Cambridge
44. Beguerisse-Díaz M, Wang B, Desikan R, Barahona M (2012) *J R Soc Interface* 9(73):1925
45. Tan CW, Gardiner BS, Hirokawa Y, Smith DW, Burgess AW (2014) *BMC Syst Biol* 8(1):44
46. Cox D, Hinkley D (1979) *Theoretical statistics*. Chapman and Hall/CRC, London
47. Carlin B, Louis T (1996) *Bayes and empirical Bayes methods for data analysis*, 2nd edn. Chapman and Hall/CRC, Boca Raton
48. Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D (2014) *Bayesian data analysis*, 3rd edn. Chapman & Hall/CRC, Boca Raton
49. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH (2009) *J R Soc Interface* 6(31):187
50. Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MPH (2014) *Nat Protoc* 9(2):439
51. Liepe J, Barnes CP, Cule E, Erguler K, Kirk P, Toni T, Stumpf MPH (2010) *Bioinformatics* 26(14):1797
52. Zhou Y, Liepe J, Sheng X, Stumpf MPH, Barnes CP (2011) *Bioinformatics* 27(6):874
53. Blüthgen N, Herzog H (2003) *J Theor Biol* 225(3):293
54. Kitano H, Oda K (2006) *Mol Syst Biol* 2:2006.0022
55. Stelling J, Sauer U, Szallasi Z, Doyle FJ, Doyle J (2004) *Cell* 118(6):675
56. Secrier M, Toni T, Stumpf MPH (2009) *Mol Biosyst* 5(12):1925
57. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP (2007) *PLoS Comput Biol* 3(10):1871
58. MacLean AL, Filippi S, Stumpf MPH (2014) *Proc Natl Acad Sci USA* 111(10):3882
59. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C (2013) *PLoS Comput Biol* 9(1):e1002803
60. Ratmann O, Donker G, Meijer A, Fraser C, Koelle K (2012) *PLoS Comput Biol* 8(12):e1002835
61. Kirk P, Thorne T, Stumpf MPH (2013) *Curr Opin Biotechnol* 24(4):767
62. Babbie AC, Kirk P, Stumpf MPH (2014) *Proc Natl Acad Sci USA* 111(52):18507
63. Feinberg M (1987) *Chem Eng Sci* 42(10):2229
64. Feinberg M (1988) *Chem Eng Sci* 43(1):1
65. Craciun G, Feinberg M (2005) *SIAM J Appl Math* 65(5):1526
66. Craciun G, Feinberg M (2006) *IEE Proc Syst Biol* 153(4):179

67. Craciun G, Feinberg M (2006) *SIAM J Appl Math* 66(4):1321
68. Feliu E, Wiuf C (2011) arXiv <http://arxiv.org/abs/1109.5149v3>
69. Feliu E, Wiuf C (2013) *Bioinformatics* 29(18):2327
70. Craciun G, Tang Y, Feinberg M (2006) *Proc Natl Acad Sci USA* 103(23):8697
71. Joshi B, Shiu A (2014) arXiv <http://arxiv.org/abs/1412.5257>
72. Ellison P, Feinberg M, Ji H (2011). Available at <http://www.chbmeng.ohio-state.edu/~feinberg/crntwin/>
73. Harrington HA, Ho KL, Thorne T, Stumpf MPH (2012) *Proc Natl Acad Sci USA* 109(39):15746
74. Ramis-Conde I, Drasdo D, Anderson ARA, Chaplain MAJ (2008) *Biophys J* 95(1):155
75. Varelas X, Miller BW, Sopko R, Song S, Gregorieff A, Fellouse FA, Sakuma R, Pawson T, Hunziker W, McNeill H, Wrana JL, Attisano L (2010) *Dev Cell* 18(4):579
76. van Leeuwen IMM, Mirams GR, Walter A, Fletcher AG, Murray P, Osborne J, Varma S, Young SJ, Cooper J, Doyle B, Pitt-Francis J, Momtahan L, Pathmanathan P, Whiteley JP, Chapman SJ, Gavaghan DJ, Jensen OE, King JR, Maini PK, Waters SL, Byrne HM (2009) *Cell Prolif* 42(5):617
77. Fletcher AG, Breward CJW, Jonathan Chapman S (2012) *J Theor Biol* 300:118
78. Murray PJ, Kang JW, Mirams GR, Shin SY, Byrne HM, Maini PK, Cho KH (2010) *Biophys J* 99(3):716
79. Murray PJ, Walter A, Fletcher AG, Edwards CM, Tindall MJ, Maini PK (2011) *Phys Biol* 8(2):026011
80. Buske P, Galle J, Barker N, Aust G, Clevers H, Loeffler M (2011) *PLoS Comput Biol* 7(1):e1001045

## Modeling and Simulation Tools: From Systems Biology to Systems Medicine

Brett G. Olivier, Maciej J. Swat, and Martijn J. Moné

### Abstract

Modeling is an integral component of modern biology. In this chapter we look into the role of the model, as it pertains to Systems Medicine, and the software that is required to instantiate and run it. We do this by comparing the development, implementation, and characteristics of tools that have been developed to work with two divergent methodologies: Systems Biology and Pharmacometrics. From the Systems Biology perspective we consider the concept of “Software as a Medical Device” and what this may imply for the migration of research-oriented, simulation software into the domain of human health.

In our second perspective, we see how in practice hundreds of computational tools already accompany drug discovery and development at every stage of the process. Standardized exchange formats are required to streamline the model exchange between tools, which would minimize translation errors and reduce the required time. With the emergence, almost 15 years ago, of the SBML standard, a large part of the domain of interest is already covered and models can be shared and passed from software to software without recoding them. Until recently the last stage of the process, the pharmacometric analysis used in clinical studies carried out on subject populations, lacked such an exchange medium. We describe a new emerging exchange format in Pharmacometrics which covers the non-linear mixed effects models, the standard statistical model type used in this area. By interfacing these two formats the entire domain can be covered by complementary standards and subsequently the according tools.

**Key words** Systems biology, Software design, Standards development, SBML, Kinetic modeling, Constraint-based modeling, Quantitative and systems pharmacology, Physiology-based pharmacokinetics, Pharmacodynamics, Pharmacometrics

---

### 1 Introduction

For more than a decade the Systems Biology approach has led to an integration of theoretical, modeling, and experimental approaches directed toward the understanding of complex biological systems. In this process, modeling has become a key component of Systems Biology and is integral to both quantitative “bottom-up” and qualitative “top-down” approaches [1]. While the former includes detailed mechanistic approaches such as kinetic [2] and constraint-based modeling [3], the latter includes

qualitative methods that include Boolean and petri-nets modeling [4] and statistical inference [5]. However, supporting this modeling process is the underlying assumption that there exists software in which a mathematical model can be instantiated and interrogated by a user, thereby generating results ready for further analysis. This second process involves the development of simulation software, the implementation of new theory, and the use of standards for data model description and exchange.

While the Systems Biology community has an established record of tool development, in general, much of this software has been developed in academia as a tool designed primarily to answer a specific research question or as a vehicle to illustrate a newly developed analysis method or algorithm. In contrast, general-purpose simulation software provides an integrated package of modeling and simulation methods that can, generally, be applied to a specific class of model or modeling methodology. This often leads, unsurprisingly, to the situation where the insight gained from applying the software to a research problem is considered more important than the software development process itself.

In this chapter we will first look at the definition of “a model” as it applies to Systems Biology (SB), Quantitative and Systems Pharmacology (QSP), and finally to Pharmacometrics (PMX). We examine the various strategies used to encode them, software used to run them and investigate how these are relevant to the development of tools for Systems Medicine. We will discuss key issues, such as user interface, model description and instantiation, software architecture, and standards support, that should be considered when selecting or designing a modeling tool. An aspect that is specifically relevant for Pharmacometrics and which will be discussed in detail is that of an extended model definition as a result of the use of population datasets required in clinical context. Starting from two divergent perspectives we attempt to address a common question: “What is required of current and future software (SB, QSP, PMX) such that it is relevant for Systems Medicine?”

---

## 2 Models and Systems Medicine

Systems Medicine encompasses the “iterative and reciprocal feedback between clinical investigations and practice with computational, statistical and mathematical multi-scale analysis and modeling of pathogenic mechanisms, disease progression and remission, disease spread and cure, treatment responses and adverse events as well as disease prevention both at the epidemiological and individual patient level” [6]. As a result, going toward Systems Medicine means medicine will move away from reductionist concepts, and toward holistic understanding of health and disease. The

successful and assessable outcome would be a medical practice that revolves around systems-based approaches and that becomes more and more predictive. Therefore, Systems Medicine aims to implement Systems Biology approaches in medical concepts and research, as well as in medical practice. It is important to underline that in Systems Biology, models are key in explaining and predicting biological phenomena, and are as such indispensable for the process of resolving biological problems or research questions. The dynamic data integration that is required to account for the biological complexity cannot be done without computer-assisted analyses and simulations. As a consequence, the transition of Systems Biology to Systems Medicine in practice will, from the start, revolve around the use of models.

In biology, the term “model” can be used for any description that proposes to explain or represent a part of biological reality. This can then be anything, for example a cartoon depicting a suggested binding mechanism of some molecule to an enzyme. Although such models and other basic representations will remain useful in both biology and medicine, in Systems Biology and Systems Medicine, and hence in this chapter, when we refer to models we explicitly assume them to be mathematical, computable representations that reflect (a part of) biological reality.

In addition, there is a wide variety of different types of mathematical models and modeling approaches in Systems Biology, and many of these can be valuable for application in a medical context. Some of these are mechanistic, e.g. kinetic molecular models that integrate biochemical and biophysical processes, while others may be non-mechanistic, like for instance models based on statistical and machine learning theory for the analysis of large-scale data. In addition, models may operate at different scales, either in time or in space. Previous chapters have addressed case studies that illustrate where and how several different approaches can be applied in manners that are suitable for Systems Medicine. It is not within the remit of this chapter to provide an exhaustive overview of which modeling techniques are available in Systems Biology and whether or not they could be valuable for Systems Medicine. From the notion that mathematical models are an essential part for the successful implementation of Systems Medicine, follows that the software tools that are needed to instantiate and operate the models will constitute an indispensable part in Systems Medicine. Therefore, we focus on a subset of existing Systems Biology modeling software tools, which are used in mechanistic molecular Systems Biology, and make clear that they can well serve as a proxy for considerations on Systems Medicine requirements of mathematical modeling software tools in general.

Modeling, i.e. the verb as it is used in biological and biomedical research fields, is commonly understood to be the activity of

constructing some model by integrating all the relevant knowledge and data and running simulations on them, i.e. computing (desired) outcomes from the mathematically encoded model given a set of starting and/or boundary conditions. In order to consider how these activities essentially relate to the use of the underpinning software, and what the repercussions are for such software to be applicable in a medical setting, it is illuminating to break down the process from model encoding via model instantiation to dealing with the modeling outputs. This brings to light that considering modeling tools for medical use is not merely relying on the computational strengths and weaknesses of the tool per se, but more-over about requirements with respect to the software and tool structure itself.

As can be appreciated from previous chapters, the clinical and medical sectors are broad and heterogeneous. The requirements for modeling and modeling software will therefore not be unambiguous and depends to a large extent on the precise context within which it will be used. In pursuance of providing concrete insights into issues concerning the use of modeling software for Systems Medicine, it is useful to distinguish three major environments within the field. First, the academic research environment; this is where most of the Systems Biology is routinely performed, and where most of the current modeling tools are developed. Its research outcomes can have clinical and medical relevance, but mostly at the conceptual level as it does not have to deal with real-world clinical and medical issues. Next, there is the medical and clinical research environment, which focuses, for example, on drug target discovery and novel therapeutics, and has to deal with controlled clinical trials. And third there is daily medical practice, which includes for instance patient diagnosis, treatment decisions, and ultimately Personalized Medicine. In viewpoint discussions on Systems Medicine there has been much emphasis on Personalized Medicine, or P4 medicine [7]. Nonetheless, before personalizing for example drug treatment, the development of new drugs remains a prerequisite. There is a clear requirement for new drugs in treating, e.g. cancer, Alzheimer's, osteoporosis, to name but a few, but we also need novel antibiotic and antiviral drugs. It is here, in the medical and clinical research environment, where most of the modeling-driven research approaches are being put into clinical practice. On the one hand to aid the drug development pipeline, but also to predict population level drug effects using Pharmacometrics. Modeling concepts and approaches from Systems Biology are actively being carried out here, so in the last part of this chapter we will focus in more detail on the current state of modeling deployment in this field with a focus on the supporting software that is being utilized.

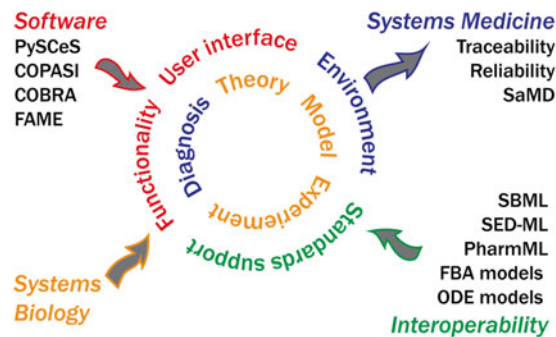


### 3 Selecting a Systems Biology Modeling Tool

In the previous section, we have seen that models and modeling are as central to Systems Medicine as they are to Systems Biology. As stated above, in this section we assume that “a model” is a quantitative, mechanistic description of a biological system that can be described as a set of coupled ordinary differential equations (ODEs). Furthermore, “a model” is assumed to be either a detailed kinetic description of such a system, usable for both time simulation and steady-state analysis [8–11], or a genome scale constraint-based model suitable for use in flux balance analysis (FBA) [12, 13]. Both these modeling formalisms are widely used in Systems Biology and supported by large, and active, software and standards development communities. In this section we consider the modeling process and look at the different strategies that can be used to design or select a modeling tool such that a model can be encoded, simulated in software and exchanged between tools. We then consider a hypothetical use case where a model is used as a diagnostic tool in Systems Medicine, a role that could be considered a medical device. Finally, we compare a selection of Systems Biology tools that highlight a variety of the aforementioned strategies. Figure 1 illustrates how software and standards can be combined to enable the transition from Systems Biology to Systems Medicine. To begin with let us consider the question: “How does one encode a model such that it can run on a computer?”

#### 3.1 Model Encoding

A widely used method for encoding a quantitative model is to write it as a human readable set of mathematical equations (e.g. as ODEs). If the model is to be simulated on a computer then this mathematical description still needs to be understood by a modeler and encoded in a programming language that can be interpreted by a computer. Though widely used, this approach has many points



**Fig. 1** Overview. The development of systems medicine will require new levels of software interoperability, expanded use of standards, relevant tool design, and a novel application of systems biology. Abbreviations used in this diagram are defined in the main text

of failure, for example, the accuracy of the model writer, the knowledge and skill of the person interpreting the mathematics and those responsible for writing the code that is then compiled and executed. A slightly higher-level approach is to encode the model directly in a computer language such as Fortran, C, C++, Java, or Python. While this has the advantage that, in principle, no human intervention is required to translate the model into a machine usable format, it makes understanding the model content and its validation (and debugging in the case of errors) very difficult for anybody not involved with the initial encoding effort. Another problem with this approach is that the model description starts to utilize features, or take on the characteristics, specific to the particular language that it is encoded in, e.g. usage of data structures and functionality provided by the programming language itself. In this way there is the potential that the model quickly becomes non-portable, i.e. not easily translatable into a format usable in any other software. A final consideration is that neither of these approaches scale well to the encoding of larger, more complex, problems, nor do they intrinsically support a portable method of annotating the individual model components.

Interactive model encoding, or the use of a custom model description language, are two approaches to model encoding often employed by standalone simulation software. Tools utilizing only a graphical user interface (GUI) generally allow users to create or edit models directly within the interface itself. A good example of this GUI-based approach is COPASI [14], while visual design tools such as JDesigner, described later in this section, offer an even higher-level interface. Alternatively, tools employing command line interfaces (CLIs) often make use of a text-based, human and machine readable, model description language. Examples of simulation tools that use this approach are JARNAC [15] and PySCeS [16], both of which use a format originally developed for use by SCAMP [17]. An analogous approach, often used in constraint-based modeling tools, e.g. CellNetAnalyser [18], is to describe the model as separate ASCII files containing lists of model components. While using a GUI is arguably the user-friendliest way of inputting a moderate-sized model typical of the average detailed kinetic model, it does not scale to the creation of large-scale models. In this case a text-based model description language may be more appropriate with the associated disadvantage that such approach can be error prone, especially for non-expert users. For very large-scale models such as genome scale reconstructions (GSRs), text or table-based model descriptions become necessary to deal with the encoding of hundreds or thousands of model components [19].

While the model encoding methods described above cover a range of use-cases, almost all of them are specific to a single tool or

family of tools. In today's Systems Biology software landscape, there are many different tools offering a wide range of functionality, the inability to exchange model descriptions between tools is a singular disadvantage. In a later section we will discuss interoperable model exchange format; for now we address the question "how do we instantiate a model in software?"

### **3.2 Model Instantiation**

At this stage "the model" is a document or structured dataset that can be found in an online database such as BioModels [20, 21]. At minimum the model can be interpreted as a structural representation of a biological network analogous to a DNA sequence stored in GenBank [22]. If the model components are sufficiently well annotated, it can be used as a structured dataset suitable for use in text mining and semantic analysis [23]. However, to fully realize a quantitative model's function it needs to be interpreted and instantiated in software on a computer. A variety of strategies can be employed for this purpose:

- Both model and analysis methods are directly implemented in a general programming language such as FORTRAN, C, C++, Java, or Python [24].
- Analysis methods are implemented in a mathematical environment, e.g. MATLAB™, R, Mathematica™, or SciPy (scipy.org), using the provided built-in tools.
- An existing mathematical environment is extended with additional functionality, for example, MathSBML for Mathematica™ [25] or the COBRA Toolbox for MATLAB™ [26].
- The model is instantiated in dedicated simulation software, developed as either standalone software like COPASI or as a web-based application like JWS Online [27].

All of these strategies have their own advantages and disadvantages. Directly encoding a model in a programming language or mathematical environment gives one complete control over its analysis, yet this process can be extremely time-consuming, can be error prone, and requires specialist knowledge of a programming language and numerical analysis. On the other hand, using dedicated simulation software may provide the required functionality at the "click-of-a-button," yet not provide the type of analysis required by the user and be difficult to extend with new functionality. Commercial mathematical environments such as MATLAB™ or Mathematica™ may provide support for their supplied algorithms but are not freely available outside of certain business sectors or academia (through academic licensing programs).

This interplay between high-performance commercial products and free open software also takes place at the algorithm implementation level. Consider a mixed-integer linear program (MILP)

solver that is becoming integral to software performing constraint-based modeling and analysis. While there are free solvers available, for example GLPK ([www.gnu.org](http://www.gnu.org)), for very large or complex models or models that require advanced analysis a commercial solver such as IBM CPLEX™ [28] or Gurobi™ [29] becomes necessary or even required. As licenses for these commercial solvers can be expensive, they are available in specialist commercial enterprises and academia (again through free licensing programs).

It is therefore of utmost importance to be aware that the strategy chosen to run a model *in silico* is both highly dependent on what analysis will be performed on the model, as well as the end-user environment where the software will be deployed. Once an appropriate strategy is chosen it can be used to assist in the selection of an existing simulation tool or in the design of a new one.

We previously highlighted that, in principle, every tool has its own encoding format, or way of inputting a model. Using one of the above strategies, “the model” is then instantiated in the software and can be used and interrogated. However, what if we want to exchange a model between different software tools?

### **3.3 Model Exchange**

The ability to exchange models has become an important feature of Systems Biology software that has independently developed across a diverse community of researchers [30, 31]. As quantitative models have increased in size and complexity and their use has become more widespread in the life sciences it has become critical to use them in ways not necessarily thought of by their original authors. For example, components or processes could be used in a different context or recombined with other independently developed models. In order to facilitate these processes it is vital that any model component’s identity can be unambiguously established and that it can be annotated with context-specific information.

In addition the number of analyses that can be performed using a model is rapidly expanding and no single modeling tool can incorporate all of them. Instead, multiple, individually developed, highly specialized tools will be required to work together to perform the “next generation” of simulation experiments. To do this will require the ability to seamlessly exchange models between different software, or, in other words, complete tool interoperability—independent of tool development language or operating environment. Standardized data formats can provide a platform that facilitates both of these processes, especially, when the foundations of this emerging tool interoperability are the development of open, community-driven standards and their implementation in both research and industrial software.

In the last 15 years or so, a set of standards for encoding and working with a variety of modeling methodologies have evolved in the Systems Biology community. While most of these standards have developed independent of one another, recent initiatives such

as the *Computational Modeling in BIology Network* (COMBINE) aim to enhance the interaction among existing standards and facilitate the development of new ones [32]. COMBINE (co.combine.org) incorporates a diverse range of standards and associated standardization efforts that include BIOPAX, a standard for the description of biological pathways, CellML for biological and physiological models [33] and the Synthetic Biology Open Language (SBOL) for exchanging genetic designs [34]. Many of the COMBINE standards incorporate aspects of two minimal guidelines relevant to a model or simulation. The first of these, the *Minimal Information Required In the Annotation of Models* (MIRIAM), is a set of guidelines that has been developed for the consistent and persistent annotation and curation of computational models in biology. MIRIAM can be implemented in any structured model format where individual components can be annotated [35]. The Minimal Information required for a Simulation Experiment (MIASE) provide the modeler or developer with a set of guidelines on the information necessary to reproduce a simulation experiment [36]. In the next section, we will look in more detail at two COMBINE standards that together implement both the MIRIAM and MIASE guidelines and are particularly relevant when selecting a kinetic or constraint-based modeling tool. The web addresses for some of the tools mentioned in this section are provided in Table 1.

*The Systems Biology Markup Language* (SBML) is a widely used standard for the interoperable encoding of Systems Biology models and is a machine-readable data format that can be used to encode biological processes and serialized, or written down, in the widely used eXtensible Markup Language (XML) [37]. SBML goes beyond pathway description and like CellML encodes the network reaction structure, the mathematical equations that describe the biological processes and the numerical values of model parameters. Furthermore, SBML provides an annotation mechanism that allows the model and each of its components to be fully annotated through a MIRIAM compliant annotation mechanism taking advantage of the resource and stable identifiers provided by the MIRIAM registry and the associated online resource, “[identifiers.org](http://identifiers.org)” [38, 39].

The SBML standard is itself evolving with major advancements referred to as “levels.” For kinetic simulation software Level 2 is, generally, more widely supported, with the equivalent Level 3 Core becoming more popular. One of Level 3’s advantages is that the language has been modularized so that the core model description can be extended by packages. Each package can be used alone or in combination with others to extend the core specification allowing new model types to be encoded. A good illustration of this is constraint-based models. While most existing SBML models are encoded in an SBML Level 2 dialect, the Level 3 *Flux Balance*

**Table 1**  
**Relevant URLs referred to in the text**

SBML Test Suite	<a href="http://sbml.org/Software/SBML_Test_Suite">sbml.org/Software/SBML_Test_Suite</a>
SBML Validator	<a href="http://sbml.org/Facilities/Validator">sbml.org/Facilities/Validator</a>
SBML Software matrix	<a href="http://sbml.org/SBML_Software_Guide">sbml.org/SBML_Software_Guide</a>
libSEDML	<a href="http://libsedml.sourceforge.net/libSedML/Welcome.html">libsedml.sourceforge.net/libSedML/Welcome.html</a>
JlibSEDML	<a href="http://sourceforge.net/projects/jlibsedml">sourceforge.net/projects/jlibsedml</a>
SED-ML Web Tools	<a href="http://sysbioapps.dyndns.org/SED-ML_Web_Tools">sysbioapps.dyndns.org/SED-ML_Web_Tools</a>
COMBINE archives	<a href="http://co.mbine.org/documents/archive">co.mbine.org/documents/archive</a>
ADAPT II	<a href="http://bmsr.usc.edu/software/adapt">bmsr.usc.edu/software/adapt</a>
Monolix	<a href="http://lixoft.com">lixoft.com</a>
NONMEM	<a href="http://www.iconplc.com/technology/products/nonmem">www.iconplc.com/technology/products/nonmem</a>
Phoenix NLME	<a href="http://www.certara.com/products/pkpd/phx-nlme">www.certara.com/products/pkpd/phx-nlme</a>
WinBUGS	<a href="http://www.mrc-bsu.cam.ac.uk/software/bugs">www.mrc-bsu.cam.ac.uk/software/bugs</a>
PharmML	<a href="http://ddmore.eu/pharmml">ddmore.eu/pharmml</a> and <a href="http://pharmml.org">pharmml.org</a>

*Constraints* (FBC) package provides all the necessary components required to encode a constraint-based or FBA model by extending SBML Level 3 Core with constructs describing, amongst other things, flux capacity constraints and objective functions [40]. Other packages developed include the *hierarchical composition* or combination of sub-models and the encoding of *qualitative models* such as petri-nets [41]. For tool developers, the SBML community makes free libraries available that provide bindings to multiple programming languages including C, Java, Python, MATLAB, and even Javascript. This allows software developers to easily access capabilities such as reading, writing, and modifying models using well maintained and supported libraries like libSBML [42] and JSBML [43]. In addition, SBML has a well documented and formally defined development process and management structure overseen by “editors” elected from the wider community. A comprehensive test suite that covers the interpretation of all aspects of the language, facilities for running the tests and a database of test results as well as a publicly available online validation tool are examples of the wide range of facilities provided to the modeling community.

*The Simulation Experiment Design Markup Language* (SED-ML) aims to implement the MIASE guidelines and is an important step along the path toward reproducible simulations. Whereas data formats such as SBML and CellML encode a model, SED-ML encodes a simulation experiment. At its core a simulation experiment consists of a quantitative model description, the set of parameter values which should be applied to the model, the algorithm that should be used to simulate the model and how the simulation output should be transformed in order to produce an expected numerical result [44]. In principle, SED-ML is independent of the format used to encode the model but in practice requires an XML-based model description. The initial SED-ML specification provided a comprehensive description of time course simulations, while recent versions have extended its capabilities to include steady-state parameter scans and now includes parameter optimization and, soon, constraint-based modeling. While there is a broad acceptance that SED-ML has an important role to play in the modeling process, up until now it has had a limited implementation even in kinetic modeling software. However, uptake should improve with the development of language libraries like libSedML and JlibSedML and community resources such as the SED-ML Web Tools (*see* Table 1).

### **3.4 Simulation Software as a Medical Device: A Thought Experiment**

Up until now the strategies and standards discussed have been focused primarily on research-oriented tools that have primarily been developed in an academic environment. However, the question now arises: “Are these tools suitable for use, as is, in Systems Medicine?” To answer this question we need to look at the ultimate usage of the model and its software instantiation in a simulation tool. In this thought experiment we make the hypothetical assumption that within the scope of Systems Medicine there will be the possibility of using a detailed quantitative model (as previously defined) that is simulated, in software, to perform some form of diagnostic or other human health-related function.

If we now consider the following definition from *The International Medical Device Regulators Forum* ([www.imdrf.org](http://www.imdrf.org)) of Software as a Medical Device:

The term “Software as a Medical Device” (SaMD) is defined as software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device [45].

Furthermore it notes (amongst other things) that:

- SaMD is a medical device and its definition includes in-vitro diagnostic medical devices.
- SaMD is capable of running on general-purpose computing hardware that has not necessarily been, specifically, designed for medical use.



- SaMD may be interfaced with other medical devices, including hardware medical devices and other SaMD, and other general-purpose software.

In our hypothetical example, the hardware could be a standard desktop computer, using a generic operating system, the input might be provided by a physician, the software could be a simulation program, the model some form of constraint-based or kinetic model, and the output some suggested diagnosis or treatment. As we are clearly using the model and its software instantiation for a medical purpose one could make the case that it easily falls within the scope of the SaMD definition.

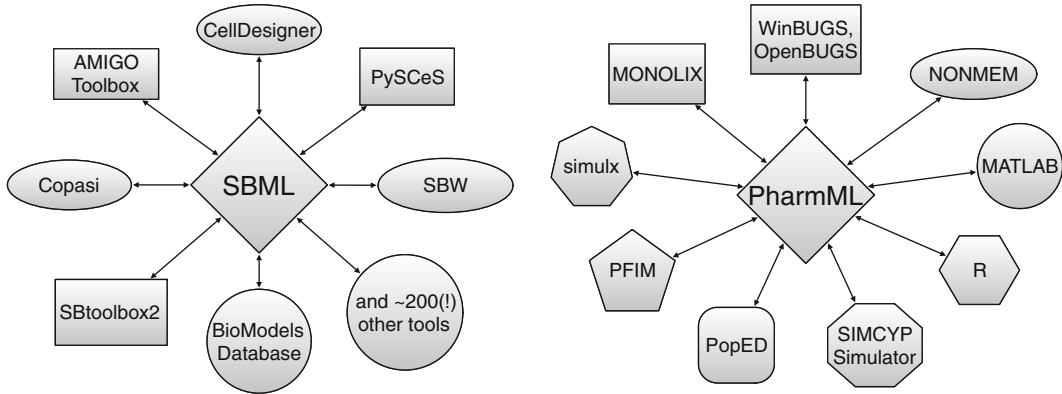
What differentiates our new software device from existing academic simulation tools is the fact that its output can have an effect on an individual's health and thereby introduces the concept of risk (to a patient). The question of risk management in medical devices is not new and is the subject of, amongst others, the IEC 62304 standard [46] that focuses on how the software development process should be structured and applied to manage the risk associated with developing a medical device. It does so by defining the majority of the software development and verification activities including development planning, requirement analysis, architectural design, software design, unit implementation and verification and release [47]. While an in-depth overview of IEC 62304 is beyond the scope of this chapter, one important aspect in the management of risk in the software development process is that of traceability [48]. Gotel et al. [49] define requirements traceability as:

[...] the ability to describe and follow the life of a requirement, in both a forwards and backwards direction (i.e., from its origins through its development and specification to its subsequent deployment and use, and through all periods of on-going refinement and iteration in any of these phases).

In practice traceability can be applied to every aspect of the software development process, from formulating requirements through software documentation and version control to testing the final implementation. Therefore, traceability, risk management, and its implementation in the entire software development process will be an important factor in the potential adaptation of existing Systems Biology tools into software medical devices. Of course this is a specific and hypothetical example, but it serves to highlight that changing the role of a piece of software from research to diagnostic, may not necessarily be as straightforward as changing the input data or interpreting the output in a different way.

### **3.5 An Overview of Some Systems Biology Tools**

Any web search for Systems Biology tools or software will return hundreds of potential matches and the user is referred to reviews of constraint-based modeling tools [50] and online resources such as the SBML software matrix for more extensive overviews of the



**Fig. 2** Comparison of interoperability in Systems Biology (*left*) and Pharmacometrics (*right*) achieved by the exchange formats SBML and PharmML, respectively

field (Table 1, Fig. 2). The following set of tools has been selected to highlight a variety of the strategies introduced earlier with particular reference to the following features: functionality, target audience, user interface, model definition and input, support for standards, and traceability.

*COPASI* ([www.copasi.org](http://www.copasi.org)) is an example of standalone software that has been developed for the simulation and analysis of dynamic biochemical reaction networks and aims to be usable by both beginners and advanced users through the use of a rich graphical user interface that runs on various operating systems. *COPASI* provides, amongst other things, stochastic and deterministic simulations as well as metabolic control analysis and parameter estimation. In addition, it also has a command line version (*CopasiSE*) and support for grid architectures [51]. Models can be created directly in the GUI or imported and exported in either SBML Level 1, 2, or 3, while support is also provided for MIRIAM-compliant model and component annotation. In terms of traceability, *COPASI* is Open Source Software that utilizes GitHub for version control. It also has extensive online documentation including version change logs, user forums, technical specifications, and video tutorials.

The *CO*nstraint-Based *Re*construction and *A*nalysis *T*oolbox (*COBRA*) is a widely used platform for constraint-based modeling [26] and the analysis of genome-scale stoichiometric models. It is aimed at an audience with intermediate to advanced technical skills, as it is a *MATLAB*<sup>TM</sup>-based tool that can be used interactively or scripted to perform various advanced analyses. *COBRA* ([open-cobra.sourceforge.net](http://open-cobra.sourceforge.net)) provides a wide range of optimization functionality that leverages either free solvers such as *GLPK* or commercial ones such as *MOSEK*<sup>TM</sup> via a *MATLAB*<sup>TM</sup> interface. It also includes reconstruction methods such as network gap filling and flux visualization and has recently been extended with a Python

version—COBRApy [52]. COBRA can load and save models in its own native formats and model building is either scripted or interactive and supports its own version of SBML Level 2 for model import and export. Due to its wide usage, COBRA-generated SBML is one of the most widely used formats for exchanging constraint-based models. COBRA provides a platform for the addition of user-defined modules and user support is provided through online API documentation and user forums while the source code is kept under version control on SourceForge and GitHub.

*The Systems Biology Workbench* (SBW) [53, 54] is a framework that allows independently developed software applications to interact and exchange data using a high-performance message passing system. The basic SBW package (sbw.sourceforge.net) contains a number of tools for model creation (JDesigner), simulation (RoadRunner), and analyses (e.g. Jarnac) that allow the user to perform, amongst other things, metabolic control analysis, structural analysis, network visualization, and FBA. SBW modules generally utilize GUIs, making them well suited for use by a wide range of users. Models can be visually created (by dragging and dropping components) in JDesigner, the Jarnac model description language, or Antimony [55]. Various levels of SBML are supported for model exchange as well as MIRIAM compliant model annotation. Programmable APIs are provided for advanced users and much of the functionality is made available as web-services. SBW is developed as Open Source Software and maintains its source code on SourceForge with documentation provided as individual help files.

The *PySCeS Constraint-Based Modeling platform* (CBMPy) and *Flux Analysis and Modeling Environment* (FAME) are independent tools that together provide both a standalone and web-based modeling solution. CBMPy (cbmpy.sourceforge.net) is a Python-based cross-platform framework for constraint-based modeling that provides a range of functionality using either commercial (CPLEX™) or free (GLPK) solvers. It is targeted toward more advanced modelers and algorithm developers but also provides terminal-based GUIs for simplifying tasks such as creating reactions. CBMPy supports importing and exporting models using the latest SBML Level 3 FBC standard as well as the older COBRA SBML Level 2 format. Models can be created using Python dictionaries, Excel spreadsheets, or interactively on the command line. CBMPy supports MIRIAM annotation of model components and the export of models as COMBINE archives. Designed using a flexible architecture, CBMPy exposes its functionality as SOAP-based web services using the PySCeS-Mariner extension. CBMPy and Mariner are both Open Source Software with the source code available from SourceForge. Documentation is limited and mostly consists of API references and installation instructions.

CBMPy's flexible design allows its functionality to be utilized by the web-based modeling environment FAME [56]. FAME (f-a-m-e.org) is targeted toward both beginner and intermediate users by way of a user-friendly graphical interface. It provides facilities for creating models based on KEGG pathways [57] and editing directly in the user interface, as well as importing and exporting SBML. FAME allows the results of optimizations to be visualized on either KEGG or user-supplied pathways. It is Open Source Software and both documentation and a tutorial are available. Interestingly, by coupling CBMPy and FAME using standard web-services, its combined functionality could be extended while still maintaining flexibility in each tool's separate development process.

This concludes the section on Systems Biology software. We now change perspective and look at how the modeling tools used in Systems Pharmacology can be relevant to Systems Medicine.

---

## 4 From Drug Discovery to Patient or from Systems Biology to Pharmacometrics

This section will focus on drug discovery and development and Pharmacometrics—the quantitative analysis of drug effects at the population level with special attention to the variability of drug responses, influences of covariates, and trial design [58]. After a short description of the typical clinical phases, we will list the corresponding areas of scientific computing associated with each of these phases [59] *see*, Fig. 3.

We start with the Exploratory/Discovery phase, which is when on one side the attempt is made to model basic biological processes, such as gene transcription and translation, cell cycle, signaling pathways, metabolic networks, etc., while on the other hand potential drug targets are analyzed, their influence on the whole biological process under consideration, and required receptor occupancy and response magnitudes are quantified. These phases are typically associated with Systems Biology and Quantitative and Systems Pharmacology (QSP). The latter considers the drug molecule as the major actor under investigation. It tries to validate targets and uncover mechanisms of action of existing therapeutics as well as discovering new ones [59].

**Early clinical** phases deal with candidate drug molecules and the whole organism, and are the bridging phase between preclinical discovery phases and large subject cohort studies of phase 3. It is here that drugs on predominantly healthy volunteers are tested, and where their bioavailability and basic PK parameters are assessed. At this stage Physiology-based Pharmacokinetics (PBPK) is the additional tool—besides QSP—that researchers have at their disposal [60–62]. The aspects of interest that PBPK can answer are summarized as ADME: absorption, distribution, metabolism, and

Clinical phase	<i>Preclinical</i>		<i>Early Clinical</i>		<i>Late Clinical</i>
Discipline	<i>Systems Biology</i>	<i>Systems Pharmacology</i>	<i>Translational PKPD</i>	<i>Pharmacometrics</i>	
Data type	<i>Frequently sampled single subject data</i>			<i>Sparse population data</i>	
Main objective	Drug - Target	Drug - Pathway/Tissue	Drug/PBPK - Organism	Drug - Disease/ Population	
Model exchange formats	 <b>SBML</b>				 <i>integrated with</i> <b>PharmML</b>

**Fig. 3** Clinical phases and corresponding model exchange formats. The *dotted SBML line* indicates that it covers only the structural models. PharmML provides the missing statistical layer required for the NLME models. See text for more details

excretion of drugs. Their understanding leads to readouts such as systemic exposure, concentration on the site of action, assessment of drug–drug interaction, etc., which are dependent on genetic constitution, disease, or sub-population ethnicity, amongst many other factors [62].

**Late clinical** phases look at the relationship between the disease and population by analyzing large subject cohorts. Main points of focus are safety and efficacy for the applied therapeutic dose. The approach of choice at this stage is Pharmacometrics. An essential aspect of PMX is that it is able to handle population data in contrast to SB and QSP, which require frequently sampled individual subject data. Population data in late clinical phases are often sparse, rendering SB and QSP unsuitable. PMX, on the other hand, is designed to utilize such data, constituting, in extreme cases, just one or two measurement records per subject. PMX applies statistical models that often revolve around the same deterministic prediction model as those used for SB or QSP [63, 64].

Before we turn our attention to the more technical aspects, it is worth looking at the role and effects that the application of computer models has had on the field to date. The attrition rates for new compounds vary strongly between development phases, but one can see that they also vary over the last years with a clear decreasing tendency [65]. Modeling and simulation influence is seen as the major reducing factor in PK-related attrition rates, falling between 1991 and 2001 from 40 % to as little as 10 % [66]. More modeling is required in pharmacodynamics (PD), especially as our understanding of the biology is improving and formulating this knowledge as mechanistic models finds its way to publicly available resources like the BioModels database [20] with ready-to-use, validated, and annotated mathematical models.

#### 4.1 Extension of the Model Definition

Definition of a model as used in SB was discussed in detail previously. Similar definitions are used in QSP and PBPK even though the complexity of the latter is often much higher, containing

frequently hundreds of ODEs and algebraic equations [61]. The model there is understood to define a deterministic prediction for simulations of a time course for a variable of interest.

PMX, on the other hand, utilizes statistical models (called non-linear mixed effect (NLME) models) in the majority of cases, and requires several extensions of this model definition [64]. The continuous data model, which consists of a deterministic prediction model as used in SB and QSP, now called *Structural Model*, is enclosed by a statistical layer with a number of NLME-specific model components described below. For discrete data, very common as endpoints in clinical trials, such deterministic prediction model is optional and only the suitable distribution needs to be defined in the *Observational Model*, e.g. Poisson distribution for count data, categorical distribution for categorical data and hazard or survival function for time-to-event data [63, 64]. The next element, the *Covariate Model*, describes the relationship between PK parameters and covariates. Covariates can be either demographic (e.g. age, body weight, sex), marker of organ functions (e.g. creatinine clearance, bilirubin), environmental indicators (degree of compliance, smoking status, concomitant medication), or other factors (pregnancy, disease progression state, genotypes, and phenotypes) [63]. The *Parameter Model* allows implementing the relationship between model parameters, fixed and random effects, and covariates. The *Variability Model* describes the parameter and residual error related, unexplained variability at a certain variability level, which is often related to the structure and features of the trial design.

Pharmacometrics distinguishes another characteristic model element, the *Trial Design Model*, which is an essential component for simulation and design optimization tasks [67]. In the latter case it is used to inform the drug development process about the optimal setup of the trial with respect to the number of patients, drug doses, and other factors. It is essential in the attempt to improve the treatment efficacy and to lower its costs.

## 4.2 Standards and Tools

SBML has proven very valuable in SB and QSP and became a *de facto* standard for model exchange across the field in both academia and pharma [68]. Looking from the perspective of Pharmacometrics, however, it lacks support for the model elements described above. The need for a comparable exchange format has been the stimulus for Drug Disease Model Resources (DDMoRe), an on-going Innovative Medicines Initiative (IMI) project that develops an interoperability platform for tools used in PMX. One of its crucial elements is the new exchange format called Pharmacometrics Markup Language [69], of which the second public version, PharmML 0.6, has been released in January 2015 (see PharmML related websites) and containing the missing model components discussed in the previous section (Table 1). Under development is

the interface for SBML-encoded models, which would provide the bridging element enabling the coverage of the entire drug discovery and development pipeline with two complementary model exchange formats. In other words, a mechanistic model developed for pre- or early-clinical use could then be taken forward and reused in the population context of late-clinical studies. This would have the advantage of seamless and error-free transition among hundreds of tools across the corresponding fields (Fig. 3). Next, we will briefly describe currently used tools for PBPK and PMX, and pointing out their most essential features (Fig. 2).

#### 4.2.1 PBPK Tools

**Simcyp Simulator** has a free academic license and is centered on sophisticated ADME (absorption, distribution, metabolism, excretion) and ADAM (advanced dissolution absorption and metabolism) modules and CYP-related metabolism in liver and other organs [70]. Its main merit lies not only in the ability to rescale in vitro data on drug metabolism and transport into in vivo data, but also in the incorporation of functional genetic polymorphism information resulting in differences in, for instance, hepatic clearance in a virtual patient population of interest [61]. It contains an extensive population library, including North European Caucasians, Japanese, healthy volunteers and several other datasets for specific populations and diseases with corresponding pharmacogenetic, physiological, and biochemical data. A pediatric module and pharmacodynamics, and drug–drug interaction modules complete the tool to a feature-rich, powerful simulation, and optimization platform. A recently added custom scripting facility further extends the capabilities of the Simcyp Simulator allowing encoding of sophisticated mechanistic disease models.

Physiology-based pharmacokinetics platform **PK-Sim/MoBi** has the ability to provide highly accurate PK profiles for over thirty organs, tissue and blood compartments in combination with user-defined mechanistic PD or disease models [71]. In addition to its anatomical and physiological features, it gives valuable insights into drug metabolism processes, e.g. in liver, as related to CYP enzymes, metabolic drug–drug interactions, or the occurrence of different metabolizers. The tool also supports scenarios where, for example, a drug's active metabolite is formed in the liver by an enzyme that has a known polymorphism, which alters the metabolism rate. This opens up new possibilities for simulations starting from gene expression up to whole organs and their pathology, in a multi-scale bottom-up approach. Similarly to the Simcyp Simulator, PK-Sim/MoBi comes with population libraries and pharmacodynamics, and pediatrics modules. Moreover, the extensibility of this platform has increased through interfacing with R and MATLAB along with an SBML interface. Comparable feature sets can be found in the commercial package **GastroPlus<sup>TM</sup>** (Simulations Plus Inc.).



#### 4.2.2 PMX Tools

Similarly to PBPK, PMX knows a much smaller tools selection than the SB community [72] (Fig. 2). This has multiple reasons. For once, the community is significantly smaller than that of computational biology around standards such as SBML and CellML. Missing exchange standards is another issue, but the major difficulty that tool designers face is the complexity of non-linear mixed effect (NLME) models and the demanding requirements with respect to the optimization algorithms vis-à-vis typical ODE-based models. A couple of software tools are available, such as Adapt II [73], Monolix [74], NONMEM [75], Phoenix NLME [72] or WinBUGS [76] and we will describe two of them that represent different modeling approaches that are relevant for discussion.

The best-known tool in Pharmacometrics, **NONMEM**, has been developed in the early 1980s and is still under active development [75] (Table 1). Based on FORTRAN, it allows users to encode almost any model scenario in the imperative language NMTRAN (in contrast to other tools using declarative languages). This comes with the setback that it requires non-trivial programming skills from the modeler in order to encode standard statistical models. It has a built-in library of PK models equipped with a number of ready-to-go routines that cover a few common compartmental models, while other models have to be encoded using ODEs. The supported dataset is of the event-driven type and provides observation and dosing records, covariates and—implicitly—comprehensive information about the underlying trial design with its phases, occasions, resetting events, etc. A number of third-party tools are at user disposal, enabling not only a more convenient working environment with this command-based tool but also complex analysis of the estimation results [77, 78].

**Monolix** is a relatively new software tool that is quickly gaining popularity, in part due to the introduction of a novel and powerful estimation method [79] (Table 1). One major advantage is the declarative modeling language MLXTRAN, which has a well-defined vocabulary and grammar, and clear boundaries. It allows defining models in a highly structured manner, is accessible to a wide audience, and is easy to learn for beginners, which becomes clear when dealing with discrete data models. Monolix handles any number of variability levels, mixture models, below limit of quantification (BLQ) data, etc. It is equipped with a powerful GUI enabling the user to interact with the tool by setting the dataset and structural model of interest, covariate model, parameter distributions, and residual error model type. It can be used to set initial values for fixed effects, standard deviations for random effects, residual error model parameters, and numerical algorithms settings. Three other tools accompany Monolix: *simulx*—a simulator for clinical trials; *mlxplore*—software for the exploration and visualization of complex PMX models; and *datxplore*—a tool for the exploration and visualization of data.

The majority of tools used in pharmacometrics are available only commercially. Some of them however are coming with a free academic license (e.g. Simcyp Simulator, PK-Sim, or Monolix). Compared to the mainly academically developed open source SB tools, their code is usually closed.

---

## 5 Conclusion: Standards and Tools for Systems Medicine

In this chapter, we have broadly approached the development of tools for use in Systems Medicine from two different perspectives: the Systems Biology and SB-QSP-PMX domains. In both cases we have highlighted different strategies, used in the development of a number of tools, and showed how these affect the choice of software for a particular task (Fig. 2). In the case of simulation tools, the issues of user interface, end-user environment, and especially support for open standards are highlighted as being critical not only for data interoperability and reproducibility, but also for traceability. When considering the substantial change in role that models, and their instantiation in simulation software, must go through when moving from simply being research tools to, for example, being “software as medical devices” one sees how this could, inevitably, have a large effect on how future tools are developed, maintained, and supported. However, these issues are not unique, a point which becomes clear when the question of tool development is approached from a Systems Pharmacology perspective.

There exist many hundreds of tools in the combined SB-QSP-PMX area and there arises the question of how to coordinate and streamline the process using this plethora of tools. Model exchange among tools requires often error prone manual re-coding and data conversions and should be avoided whenever possible. What is needed is, first, an interoperability platform based on a set of exchange formats assuring the required compatibility of available tools. Second, a generic data format is needed covering all features expected from single subject datasets as used in SB or QSP and the population data needed to feed the PMX tools. With SBML and PharmML we have, in fact, come very close to achieving this goal—we have a set of complementary formats covering major model types used at any stage of the drug discovery and development process. The above-discussed SBW framework can serve as a working example for such a platform driven by the SBML standard.

In the end, no matter the approach, the lessons learned in both these domains, especially with respect to standardization and interoperability, will be invaluable in the upcoming and challenging task of designing tools for Systems Medicine that have a real benefit to society.

## Acknowledgments

Brett Olivier is supported by a BE-Basic Foundation grant F08.005.001. Maciej Swat has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115156, resources of which are composed of financial contributions from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. The DDMoRe project is also supported by financial contribution from academic and SME partners.

## References

1. Bruggeman FJ, Westerhoff HV (2007) The nature of systems biology. *Trends Microbiol* 15:45–50
2. Garfinkel D, Garfinkel L, Pring M et al (1970) Computer applications to biochemical kinetics. *Annu Rev Biochem* 39:473–498
3. Savinell JM, Palsson BO (1992) Network analysis of intermediary metabolism using linear optimization. I Development of mathematical formalism. *J Theor Biol* 154:421–445
4. Hardy S, Robillard PN (2004) Modeling and simulation of molecular biology systems using petri nets: modeling goals of various approaches. *J Bioinform Comput Biol* 2:595–613
5. Liepe J, Kirk P, Filippi S et al (2014) A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat Protoc* 2:439–456
6. CASyM Consortium (2014) The CASyM roadmap: implementation of systems medicine, version 1.0. [https://www.casym.eu/lw\\_resource/datapool/\\_items/item\\_328/roadmap\\_1.0.pdf](https://www.casym.eu/lw_resource/datapool/_items/item_328/roadmap_1.0.pdf). Accessed 4 Dec 2004
7. Flores M, Glusman G, Brogaard K et al (2014) P4 medicine: how systems medicine will transform the healthcare sector and society. *Per Med* 10:565–576
8. Heinrich R, Rapoport SM, Rapoport TA (1977) Metabolic regulation and mathematical models. *Progr Biophys Mol Biol* 32:1–82
9. Wright BE, Kelly PJ (1981) Kinetic models of metabolism in intact cells, tissues and organisms. *Curr Top Cell Regul* 19:103–158
10. Massoud TF, Hademenos GJ, Young WL et al (1998) Principles and philosophy of modeling in biomedical research. *FASEB J* 12:275–285
11. Bakker BM et al (2010) Systems biology from micro-organisms to human metabolic diseases: the role of detailed kinetic models. *Biochem Soc Trans* 38:1294–1301
12. Orth JD et al (2010) What is flux balance analysis? *Nat Biotechnol* 28:245–248
13. Thiele I, Swainston N, Fleming RMT et al (2013) A community-driven global reconstruction of human metabolism. *Nat Biotechnol* 31:419–425
14. Hoops S, Sahle S, Gauges R et al (2006) COPASI: a COMplex PATHway SIMulator. *Bioinformatics* 22:3067–3074
15. Sauro HM, Hucka M, Finney A et al (2003) Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* 7:355–372
16. Olivier BG, Rohwer JM, Hofmeyr J-HS (2005) Modelling cellular systems with PySCeS. *Bioinformatics* 21:560–561
17. Sauro HM, Fell DA (1991) SCAMP: a metabolic simulator and control analysis program. *Mathl Comput Model* 15:15–28
18. Klamt S, Saez-Rodriguez J, Gilles ED (2007) Structural and functional analysis of cellular networks with Cell NetAnalyzer. *BMC Syst Biol* 1:2
19. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 5:320
20. Li C, Donizelli M, Rodriguez N et al (2010) BioModels Database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol* 4:92
21. Le Novère N, Bornstein B, Broicher A et al (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res* 34:D689–D691
22. Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2011) GenBank. *Nucleic Acids Res* 39:D32–D37

23. Wimalaratne SM, Grenon P, Hermjakob H et al (2014) BioModels linked dataset. *BMC Syst Biol* 8:91
24. Olivier BG, Rohwer JM, Hofmeyr J-HS (2002) Modelling cellular processes with Python and SciPy. *Mol Biol Rep* 29:249–254
25. Shapiro BE, Hucka M, Finney A et al (2004) MathSBML: a package for manipulating SBML-based biological models. *Bioinformatics* 20:2829–2831
26. Schellenberger J, Que R, Fleming RMT et al (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 6:1290–1307
27. Olivier BG, Snoep JL (2004) Web-based kinetic modelling using JWS Online. *Bioinformatics* 20:2143–2144
28. IBM Corporation (2014) IBM ILOG CPLEX optimizer. <http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer>. Accessed 29 Nov 2014
29. Gurobi Optimization, Inc. (2014) Gurobi optimizer reference manual. <http://www.gurobi.com>. Accessed 27 Nov 2014
30. Hucka M, Finney A (2005) Escalating model sizes and complexities call for standardized forms of representation. *Mol Syst Biol* 1:2005.0011
31. Keating SM, Le Novère N (2013) Supporting SBML as a model exchange format in software applications. *Methods Mol Biol* 1021:201–225
32. Waltemath D, Bergmann FT, Chaouiya C et al (2014) Meeting report from the fourth meeting of the Computational Modeling in Biology Network (COMBINE). *Stand Genomic Sci* 9:3
33. Miller AK, Marsh J, Reeve A et al (2010) An overview of the CellML API and its implementation. *BMC Bioinformatics* 11:178
34. Galdzicki M, Clancy KP, Oberortner E et al (2014) The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nat Biotechnol* 32:545–550
35. Le Novère N, Finney A, Hucka M et al (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* 23:1509–1515
36. Waltemath D, Adams R, Beard DA et al (2011) Minimum Information About a Simulation Experiment (MIASE). *PLoS Comput Biol* 7:e1001122
37. Hucka M, Finney A, Sauro HM et al (2003) The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* 9:524–531
38. Laibe C, Le Novère N (2007) MIRIAM Resources: tools to generate and resolve robust cross-references in systems biology. *BMC Syst Biol* 1:58
39. Juty N, Le Novère N, Laibe C (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res* 40:D580–D586
40. Olivier BG, Bergmann FT (2015) The Systems Biology Markup Language (SBML) Level 3 Package: Flux Balance Constraints. *Journal of Integrative Bioinformatics*, 12:269
41. Chaouiya C, Berenguier D, Keating SM et al (2013) SBML Qualitative Models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Syst Biol* 7:135
42. Bornstein BJ, Keating SM, Jouraku A et al (2008) LibSBML: an API library for SBML. *Bioinformatics* 24:880–881
43. Dräger A, Rodriguez N, Dumousseau M et al (2011) JSBML: a flexible Java library for working with SBML. *Bioinformatics* 27:2167–2168
44. Waltemath D, Adams R, Bergmann FT et al (2011) Reproducible computational biology experiments with SED-ML – The Simulation Experiment Description Markup Language. *BMC Syst Biol* 5:198
45. IMDRF SaMD Working Group (2013) Software as a Medical Device (SaMD): key definitions. <http://www.imdrf.org/documents/documents.asp>. Accessed 25 Nov 2014
46. Iec I (2006) 62304: 2006 Medical device software – software life cycle processes. International Electrotechnical Commission, Geneva
47. Buntz B (2011) Simplifying IEC 62304 compliance for developers. [http://www.emdt.co.uk/article/iec-62304-compliance?utm\\_source=emdt&utm\\_medium=articlebottom&utm\\_campaign=camilla](http://www.emdt.co.uk/article/iec-62304-compliance?utm_source=emdt&utm_medium=articlebottom&utm_campaign=camilla). Accessed 25 Nov 2014
48. Regan G, McCaffery F, McDaid K et al (2013) Medical device standards' requirements for traceability during the software development lifecycle and implementation of a traceability assessment model. *Comput Stand Int* 36:3–9
49. Gotel OCZ, Finkelstein CW (1994) An analysis of the requirements traceability problem. *Proceedings of the First International Conference on Requirements Engineering*. pp 94–101
50. Lakshmanan M, Koh G, Chung BKS et al (2014) Software applications for flux balance analysis. *Brief Bioinform* 15:108–122
51. Kent E, Hoops S, Mendes P (2012) Condor-COPASI: high-throughput computing for biochemical networks. *BMC Syst Biol* 6:91

52. Ebrahim A, Lerman JA, Palsson BO et al (2013) COBRApy: Constraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* 7:74
53. Bergmann FT, Vallabhajosyula RR, Sauro HM (2006) Computational tools for modeling protein networks. *Curr Proteomics* 3:181–197
54. Hucka M, Finney A, Sauro HM et al. (2002) The ERATO Systems Biology Workbench: enabling interaction and exchange between software tools for computational biology. *Pac Symp Biocomput* 450–461
55. Smith LP, Bergmann FT, Chandran D et al (2009) Antimony: a modular model definition language. *Bioinformatics* 25:2452–2454
56. Boele J, Olivier BG, Teusink B (2012) FAME, the Flux Analysis and Modeling Environment. *BMC Syst Biol* 6:8
57. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28:27–30
58. Byon W, Smith MK, Chan P et al (2013) Establishing best practices and guidance in population modeling: an experience with an internal population pharmacokinetic analysis guidance. *CPT Pharmacometrics Syst Pharmacol* 2:e51
59. Agoram BM, Demin O (2011) Integration not isolation: arguing the case for quantitative and systems pharmacology in drug discovery and development. *Drug Discovery Today* 16(23–24)
60. Mager D, Jusko W (2008) Development of translational pharmacokinetic-pharmacodynamic models. *Clin Pharmacol Ther* 83(6):909–912
61. Rostami-Hodjegan A, Tucker GT (2007) Simulation and prediction of in vivo drug metabolism in human populations from in vitro data. *Nat Rev Drug Discov* 6:140–148
62. Jones HM, Rowland-Yeo K (2013) Basic concepts in physiologically based pharmacokinetic modeling in drug discovery and development. *CPT Pharmacometrics Syst Pharmacol* 2:e63
63. Bonate P (2011) Pharmacokinetic-pharmacodynamic modeling and simulation, 2nd edn. Springer, New York
64. Lavielle M (2014) Mixed effects models for the population approach models, tasks, methods & tools, CRC biostatistics series. Chapman & Hall, Boca Raton, FL
65. Leil TA, Bertz R (2014) Quantitative systems pharmacology can reduce attrition and improve productivity in pharmaceutical research and development. *Front Pharmacol* 5:247
66. Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 3:711–716
67. Bazzoli C, Retout S, Mentré F (2010) Design evaluation and optimization in multiple response nonlinear mixed effect models: PFIM 3.0. *Comput Methods Prog Biomed* 98:55–65
68. Draeger A, Palsson BO (2014) Improving collaboration by standardization efforts in systems biology. *Front Bioeng Biotechnol*. doi:10.3389/fbioe.2014.00061
69. Swat MJ (2015) Pharmacometrics Markup Language (PharmML): opening new perspectives for model exchange in drug development. *CPT Pharmacometrics Syst Pharmacol* 4(6):316–319. doi:10.1002/psp4.57
70. Marciniak J (2009) The Simcyp population-based ADME simulator. *Expert Opin Drug Metab Toxicol* 5:211–223
71. Eissing T, Kuepfer L, Becker C et al (2011) A computational systems biology software platform for multiscale modeling and simulation: integrating whole-body physiology, disease biology, and molecular reaction networks. *Front Physiol* 2:4
72. Aarons L (1999) Software for population pharmacokinetics and pharmacodynamics. *Clin Pharmacokinet* 36:255–264
73. D’Argenio DZ, Schumitzky A, Wang X (2009) Adapt 5 user’s guide: Pharmacokinetic/pharmacodynamic systems analysis software. Tech. Rep., Biomedical Simulations Resource, Los Angeles
74. Lixoft. Monolix 4.3. <http://lixoft.com>
75. Beal SL, Sheiner LB, Boeckmann AJ et al. (2009) NONMEM User’s guides. Technical report. Icon Development Solutions, Ellicott City, MD, USA
76. Lunn DJ, Best N, Thomas A et al (2002) Bayesian analysis of population pk/pd models: general concepts and software. *J Pharmacokinet Pharmacodyn* 29:271–307
77. Lindbom L, Ribbing J, Jonsson EN (2004) Perl-speaks-NONMEM (PsN) – a Perl module for NONMEM related programming. *Comput Methods Programs Biomed* 75(2):85–94
78. Keizer R, Karlsson M, Hooker A (2013) Modeling and Simulation Workbench for NONMEM: tutorial on Pirana, PsN, and Xpose. *CPT Pharmacometrics Syst Pharmacol* 2:e50
79. Kuhn E, Lavielle M (2005) Maximum likelihood estimation in nonlinear mixed effects models. *Comput Stat Data Anal* 49:1020–1038

# INDEX

## A

- AAM. *See* Alternatively activated macrophages (AAM)
- ACES. *See* Amsterdam classification  
evaluation suite (ACES)
- AirPROM. *See* Airway Disease Predicting Outcomes  
through Patient Specific Computational  
Modeling Consortium (AirPROM)
- Airway Disease Predicting Outcomes through Patient  
Specific Computational Modeling Consortium  
(AirPROM).....127
- Airway smooth muscle (ASM) cells.....127
- Algebraic matroids ..... 432, 435
- ALK. *See* Anaplastic lymphoma kinase (ALK) gene
- AllergenPro database.....149
- Allergy epitopes  
databases.....146
- Alternatively activated macrophages  
(AAM).....126
- Amsterdam classification evaluation suite (ACES).....370
- Anaplastic lymphoma kinase (ALK) gene.....45
- Artificial intelligence (AI) .....210–215  
cancer-type hypothesis .....210  
computational systems.....210  
data/process integration.....210  
decision-support systems (DSS).....209–211  
modern approach  
AI winter .....212  
cognitive abilities .....211  
DSS .....210, 211  
experimental scientific methodology .....212  
hidden Markov models (HMMs).....212  
machine learning .....212, 213  
machine, collective components.....211  
medical domain .....214, 215  
MYCIN.....211  
realistic environments .....213  
systems medicine .....209
- ASM. *See* Airway smooth muscle (ASM) cells
- Asthma  
modeling.....127
- Asymptotic analysis  
enzyme kinetics model .....421–423  
Schmitz model.....423–424

## B

- Bacteroides thetaiotaomicron* .....274
- B cell epitopes.....140, 142–145  
BCR-binding site .....140  
databases.....146  
discontinuous  
3D structure-based methodologies .....144  
hybrid prediction methods.....145  
mimotope-based methodology .....144, 145  
linear  
amino acid propensity scale-based methods .....142  
machine learning-based methods .....143  
sequence-based methods.....140  
linear, prediction  
amino acid propensity scale-based methods.....142  
linear/discontinuous.....140, 142
- Big Bang model.....63
- Biocomplexity  
biological problems.....5  
computational models.....7–8  
modern systems theory .....5  
systems- oriented strategy .....5
- Bioinformatics  
biomedical research.....20, 21  
systems approaches .....34  
systems medicine .....21–22
- Biological networks .....29–33  
module-based approaches  
autism spectrum disorder (ASD).....31  
high-throughput omics.....31  
integrative modular technique .....32  
multi-omics datasets .....31  
network-based disease signatures.....33  
PARADIGM-based analysis.....32  
protein co-expression modules.....31  
species boundary.....33  
top-down network inference.....31  
WGCNA.....31  
omics data.....28  
top-down network reconstruction  
ARACNe.....29  
Bayesian networks .....29, 30  
clustering techniques .....29

Biological networks ( <i>cont.</i> )	
co-expression networks	29
correlation-based distance measures	29
eQTL	30
gene network inference algorithms	29
genotypic and gene expression data	30
information-theoretic approaches	29
mathematical and statistical techniques	29
pair-wise correlations	29
probabilistic techniques	30
reverse engineering	29
STRING database	30
Biomarker discovery	354, 361–363
challenges	360–361
computational concepts ( <i>see</i> Computational data analysis)	
limitations	360–361
molecular mechanisms	369
“omics” data	354
pathways and networks ( <i>see</i> Pathways and networks, biomarker discovery)	
prediction accuracy	369
reproducibility	369
Biomarker prediction	
IMRE	319
SVM-RBF approach	319
Biomedical applications	
COBRA	271–273
IEMs	273, 274
personalized metabolic models	274, 275
predictive medicine	276
whole-body modeling	275, 276
Brain-level networks	
anatomical, functional connections	235, 236
components and interactions	236, 237
disease-oriented approach	237
dMRI	238
EEG	238
fMRI	237, 239
longitudinal neuroimaging	238
neuroimaging approaches	236
radiolabeled tracers, neuroimaging	237
tractography approaches	238
<b>C</b>	
CAM. <i>See</i> Classically activated macrophages (CAM)	
CancerLinQ	214
Cancer metabolism	
aerobic conditions	271
aerobic glycolysis	272
constraint-based modeling	271
drug targets	271
fumarate hydratase and enzymes	271
gene expression data	272
generic cancer model	271
genome-scale model	271
human metabolic model	272
metabolic flux	272
physiological conditions	271
pyruvate carboxylase	271
renal cancer model	273
SOG pathway	272
solvent capacity constraints	271
Warburg metabolism	272
Cancer systems biology	183–188
databases and pathologies	
cancer genomics hub	184
cBioPortal for cancer genomics	184
CLARITY challenge	185
COSMIC	183
drug–gene interaction database	184
GDSC	184
ICGC data portal	184
modern decision- support systems	185
TCGA data portal	184
tumorportal	184
gate-keeper and care-taker genes	188
microRNAs	183
mutational processes	189
new cancer genes	183
next-generation sequencing (NGS)	182
passenger <i>vs.</i> driver mutations	
complex strategies	186
DNA replication	186
driver–hunter strategies	186
IntOGen-mutations	187
MutSig algorithm	186
NPV	188
ratiometric rule, Bert Vogelstein’s group	186
SNV	187
Cancer therapy	
antineoplastic agents	189
innovations	205
Canonical correlation analysis (CCA)	49
Caretaker gene	188
Catalogue of somatic mutations in cancer (COSMIC)	183
Causal biological networks (CBN)	35
CCA. <i>See</i> Canonical correlation analysis (CCA)	
CellDesigner software	51
Cellular interaction networks	232–234
astrocytes	229
Brainbow method	234
construction	
elements connection	233, 234
interactions, identification	233
DBS	232
disease-specific topology	
epilepsy	232
prion diseases	232
dopamine synthesis	232



elements and interaction types.....	231, 235
MEG.....	235
microglia.....	229, 231
molecular deregulation.....	232
neuronal interactions.....	229
oligodendrocytes.....	229, 231
plasticity.....	229
schematic overview, interactions.....	230
Chemical reaction network theory (CRNT).....	434
Chronic obstructive pulmonary disease (COPD).....	122, 288–289, 292
anxiety-depression.....	287
COPDKB ( <i>see</i> COPD knowledge base (COPDKB))	
description.....	287
enhanced stratification.....	295
synergy-COPD ( <i>see</i> Synergy-COPD)	
CIA. <i>See</i> Co-inertia analysis (CIA)	
Classically activated macrophages (CAM).....	126
Clinical decision-support systems (CDSS)	
chronic care management.....	296
design and implementation.....	296
families.....	294
Clinical medicine, transition.....	10
complex diseases.....	9
complex dynamic information.....	9
computational and mathematical tools.....	9
diagnostic findings.....	8
evidence-based medicine.....	8
metabolomics blood sample.....	8
modern diagnostics.....	8
P4 medicine.....	9
promise and challenges.....	9–11
education.....	10
multidisciplinarity.....	10
technological development.....	10
Clinical trial strategy	
driver-gatekeeper mutations.....	206
genetic/epigenetic characterization.....	206
MIM reconstructions.....	205
precision medicine, oncology.....	208
treatment strategy <i>vs.</i> genetic/epigenetic homogeneity.....	207
COBRA models	
catalytic function.....	263
cell type- and condition-specific models.....	269
cellular phenotypes.....	263
constraint-based modeling.....	268
enterocyte.....	268
high-throughput technologies.....	264
host-pathogen and host-gut microbe interactions.....	268
human disease conditions.....	268
human tissues and cell types.....	268
large-scale template.....	264
metabolic reconstructions.....	263
metabolomic datasets.....	269, 270
multi-tissue model.....	268
omics data.....	263, 264
Recon 1.....	264
single nucleotide polymorphisms.....	264
size and pathway topology.....	264
transcriptomic and proteomic data.....	268, 269
user-defined threshold.....	268
Co-inertia analysis (CIA).....	49
Colorectal cancer (CRC).....	64, 197, 199–201, 203, 204
biochemical interactions.....	196
CRC model.....	205
driver-gatekeeper mutations.....	196
dynamic model development	
fuzzy logic, pathway fragments.....	197, 199
multiple interconnected downstream pathways.....	197
ordinary differential equations (ODEs).....	197
virtual inhibitors.....	199
alterations, loss-of-function mutations.....	200
CRC cell lines.....	199
dominant mutations.....	200
HCT116 cancer line.....	199
HT29 cancer line.....	199
physiologic state.....	199
PTEN homozygous mutations.....	200
sponge effect.....	200
P-protein levels and mRNA regulation	
AKTP related pathways.....	203
MIM reconstruction.....	204
posteriori experimental verifications.....	201
semi-quantitative assessment.....	201
Spearman's rho test.....	203
pre-biological and biological evolution.....	196
protein neo-syntheses.....	197
signaling-network sub-region.....	196
targeted therapies.....	196
Combined targeted therapy.....	194–196
Onco-protein inhibitor combinationsm	
ALK and SRC inhibition.....	194
antineoplastic drugs.....	195
cell line.....	194
EGFR and FGFR inhibitors.....	194
gatekeeper mutation.....	194
genetic analysis.....	194
MEK inhibitor.....	194
NGS and omics-based strategies.....	194
pharmacologic platform strategy.....	195
resistant lung tumors.....	194
signaling pathway.....	195
tumor-resistant biopsies.....	194
signaling networks	
driver-gatekeeper.....	195
dynamic simulation approach.....	195

- Combined targeted therapy (*cont.*)  
  molecular pathology.....196  
  parameterization and pre-training .....196
- Comorbidity.....287, 288  
  clustering .....287, 288
- Computational data analysis.....354, 356–360  
  data exploration .....356  
  data pre-processing  
    log-transformation.....354  
    microarray experiments.....354  
  feature selection and supervised classification  
    advantages.....357  
    algorithms.....358  
    classes/class labels .....356  
    classifier prediction .....357  
    embedded techniques .....358  
    filter techniques .....358  
    high-dimensional data set.....356  
    SVMs.....358  
    wrapper techniques .....358  
  performance evaluation  
    disjoint data set .....359  
    distinct data sets.....359  
    k-fold cross-validation .....359  
    overfitted classifiers.....359  
    qRT-PCR.....360  
    splitting approach .....360  
    training /test set .....359  
  statistical testing .....356
- Computational modeling.....89  
  cellular scale  
    dynamic models .....89  
    interaction-based graphs.....89  
    stoichiometric models .....89  
  spatial-temporal modeling.....90, 91
- Computational Modeling in Biology Network*  
  (COMBINE) .....449
- Computer systems and infrastructures  
  concurrent and parallel computing .....400  
  HPC resources.....399  
  multiscale model management .....399
- Constrained-based modeling  
  biased and unbiased methods .....257, 258  
  biochemical network model.....254  
  biological systems .....254  
  biomass components.....257  
  CBMPy.....454  
  COBRA .....453  
  feasible solution space .....256  
  flux balance analysis (FBA).....258, 445  
  flux variability analysis.....258  
  gene-protein-reaction associations .....255  
  genome-scale network reconstruction.....254, 255  
  high-throughput techniques .....253, 254  
  human tissues/cells .....257  
  in vivo/in vitro systems .....256  
  metabolism .....257  
  MILP .....447  
  MIRIAM and MIASE guidelines.....449  
  objective functions .....257  
  sampling analysis .....258  
  SED-ML.....451  
  steady-state assumption .....256  
  stoichiometric matrix.....255  
  tools .....446  
  upper and lower bounds.....256
- Constraint-Based Reconstruction and Analysis Toolbox*  
  (COBRA) .....453
- Continuing medical education (CME)  
  European countries.....77  
  Medical Chamber of Slovenia .....82  
  requirements .....78
- COPD. *See* Chronic obstructive pulmonary disease (COPD)
- COPD knowledge base (COPDKB) .....292
- COSMIC. *See* Catalogue of somatic mutations in cancer  
  (COSMIC)
- CRC. *See* Colorectal cancer (CRC)
- CRNT. *See* Chemical reaction network theory (CRNT)
- Cross-scale network analysis .....240
- D**
- Data-based models, stem cells  
  CellNet .....343  
  chromosome structure .....343  
  data-driven partial least squares.....342  
  DNA-protein.....343  
  ESCAPE database.....343  
  gastrulation process, embryo .....342  
  gene expression .....342  
  GRNs .....342  
  growth rate kinetics .....342  
  high-throughput data .....343  
  human and mouse ESC.....342  
  hybrid models, pluripotency .....342  
  MouseNET .....343  
  network components .....342  
  phospho-proteome and growth patterns.....342  
  PI3K/AKT pathway .....342  
  pluripotent stem cells.....342  
  StemSight.....343
- Database for Annotation, Visualization and Integrated  
  Discovery (DAVID) .....315
- Data-driven modeling  
  genome-wide level .....96  
  GWAS.....96  
  network reconstruction algorithms .....97  
  signaling networks .....97
- Data integration .....315
- dbGAP. *See* Database of genotypes and phenotypes (dbGAP)
- DBS. *See* Deep brain stimulation (DBS)

DEAP. <i>See</i> Differential expression analysis for pathways (DEAP)	
Decision-support systems, oncology.....	208, 209
decision-making	
breast cancer (BC) lesions.....	208
chemotherapy (CT) regimens.....	209
co-morbidities.....	209
ductal and lobular breast cancer.....	208
HER2 tyrosine kinase membrane.....	208
multi-hit alteration.....	208
therapeutic strategies.....	208
therapy.....	208
Deep brain stimulation (DBS).....	232
Differential expression analysis for pathways (DEAP).....	366
Differential network analysis	
lung cancer prognosis.....	368
microarray gene expression data set.....	367
molecular networks.....	368
physiological and disease phenotypes.....	367
protein-protein interaction network.....	367
Diffusion MRI (dMRI).....	238
Digital Health Framework (DHF)	
building block strategy.....	292
Disease subtype discovery	
biomarkers.....	47
CellDesigner software.....	51
challenges.....	51
CIA.....	49
gene expression/DNA methylation.....	49
iCluster.....	49
integrative approach.....	48
MDI.....	49
multiple levels of granularity.....	52
multi-scale modeling.....	52
omics measurements.....	48
pathways and networks.....	50
PLS, CCA and sparse approaches.....	49
SBGN project.....	50
dMRI. <i>See</i> Diffusion MRI (dMRI)	
Doctoral training	
CASyM training tutorial.....	82
FEBS Advanced Lecture Course.....	83
Helmholtz Graduate School.....	82
Imperial College London.....	81
Royal College of Surgeons, Ireland.....	82
1st SyBSyM Como School.....	2014, 82
Trinity College Dublin.....	82
University College London.....	81
University of Ljubljana.....	81
<b>E</b>	
Education system	
basic and natural sciences.....	79
CME credits.....	77, 78
medicine, master's studies.....	81
multidisciplinary training.....	78
P4 medicine.....	78
professional dissemination.....	78
systems biology approaches.....	77
training programs.....	80
EEG. <i>See</i> Electroencephalography (EEG)	
Electroencephalography (EEG).....	238
Ellipro.....	143
Enzyme kinetics model.....	409–410
Epidemic modeling	
critical vaccination coverage.....	110
isolation/imposed travel restrictions.....	108
SEIR models.....	110
stochastic epidemic models.....	111
subpopulations of individuals.....	109
vaccination.....	109
eQTL. <i>See</i> Expression quantitative trait loci (eQTL)	
European Innovation Partnership on Active and Healthy Ageing (EIP-AHA).....	285
European Institute of Technology for Health (EIT-Health).....	285
European systems medicine community.....	12
Expression quantitative trait loci (eQTL).....	30
<b>F</b>	
FDA. <i>See</i> Food and Drug Administration (FDA)	
Federal Ministry of Education and Research (BMBF).....	13
FISH. <i>See</i> Fluorescence in situ hybridization (FISH)	
Fluorescence in situ hybridization (FISH).....	67
Flux balance analysis.....	258
Flux balance constraints (FBC).....	450
Flux variability analysis.....	258
fMRI. <i>See</i> Functional magnetic resonance imaging (fMRI)	
Food and Drug Administration (FDA).....	101
Functional magnetic resonance imaging (fMRI).....	237
<b>G</b>	
Gatekeeper genes.....	188
GDSC. <i>See</i> Genomics of drug sensitivity in cancer (GDSC)	
Gene co-expression networks	
direct and indirect edges.....	362
disease phenotypes.....	363
pair-wise associations.....	362
Gene expression profiling	
bead-based hybridization technology.....	317
NGS.....	318
qPCR.....	318
Gene Inactivation Moderated by Metabolism, Metabolomics, and Expression (GIM(3E).....	269
Gene regulatory networks (GRN).....	338
Generic cancer model.....	271
Genome scale reconstructions (GSRs).....	446
Genome-wide association studies (GWAS).....	26, 96

- Genomics of drug sensitivity in cancer  
(GDSC)..... 184, 185
- GIM(3)E. *See* Gene Inactivation Moderated by Metabolism,  
Metabolomics, and Expression (GIM(3)E)
- Global Initiative for Chronic Obstructive Pulmonary  
Disease (GOLD).....295
- Graphical user interface (GUI) .....446
- GRN. *See* Gene regulatory networks (GRN)
- GUI. *See* Graphical user interface (GUI)
- GWAS. *See* Genome-wide association studies (GWAS)
- ## H
- HCC. *See* Hepatocellular carcinoma (HCC)
- HD. *See* Huntington's disease (HD)
- Hematopoietic stem cells (HSC).....336
- Hepatocellular carcinoma (HCC) .....308
- Hereditary hemorrhagic telangiectasia  
(HHT).....274
- Heterogeneity
- Big Bang model.....63
  - clonal evolution theory .....62
  - CSCs .....63
  - data mining techniques.....69
  - digital pathology.....67
  - fluorophores and chromogens.....68
  - healthy tissue and host cells .....65
  - high-content analysis.....67
  - histopathology .....65
  - IF-based image analysis.....68
  - IHC and FISH.....67
  - image analysis .....67, 68
  - immunofluorescence, colorectal cancer .....67
  - inter- and intra-patient.....65, 66
  - LCM technique.....65
  - market-leading tissue imaging.....68
  - molecular genomic and proteomic profiling .....66
  - multi-parametric signature .....69
  - mutational/epigenetic aberrations.....65
  - nuclear morphometry .....68
  - RNA microarray chips and RNA sequencing  
technologies .....65
  - RPPA.....65
  - self-renewing cells .....63
  - software packages .....68
  - sophisticated data mining .....69
  - stromal and immune infiltrate .....63
  - stromal cancer-associated fibroblasts .....67
  - subpopulation segmentation and biomarker  
quantification.....69
  - transcriptomics .....67
  - tumorigenesis.....63
- HHT. *See* Hereditary hemorrhagic telangiectasia (HHT)
- High-throughput data (HTD).....151–157
- cell differentiation analysis
  - dendritic-cell function .....154
  - macrophage function and  
differentiation .....151–153
  - medical immune interventions
    - tumor-immunity interaction.....156, 157
    - vaccination, infectious diseases .....155, 156  - statistical analysis.....150
- Histopathology.....64, 65
- HIV modeling
- fundamental information.....112
  - in vivo experiments.....112
  - mathematical models.....111
  - systems-based interdisciplinary approaches.....113
- Homo Sapiens Recon 1.....259
- HOTAIR. *See* HOX antisense intergenic RNA (HOTAIR)
- HOX antisense intergenic RNA (HOTAIR).....308
- HTD. *See* High-throughput data (HTD)
- Human metabolic genome-scale reconstructions
- acylcarnitine/fatty acid oxidation module.....262
  - cellular compartments.....259
  - computational and experimental approach .....262
  - data mapping .....261
  - datasets .....261
  - dead-end metabolites and blocked reactions.....261
  - drug module capturing .....263
  - extracellular metabolite transporters.....263
  - host-pathogen and host-gut microbial  
interactions .....259
  - IEMs .....261
  - intestinal transport/absorption module.....262
  - Recon 1.....259, 260
  - Recon 2.....261
  - template function.....262
  - transport proteins and mechanisms .....263
- Huntington's disease (HD).....226
- ## I
- IBM Watson.....185, 214
- ICGC. *See* International Cancer Genome Consortium (ICGC)
- IEMs. *See* Inborn errors of metabolism (IEMs)
- IHC. *See* Immunohistochemistry (IHC)
- Immune interventions, HTD.....155, 156
- tumor-immunity interaction
    - data clustering techniques.....156
    - microarray analysis.....156  - vaccination, infectious diseases
    - ANOVA model.....155
    - genetic expression signatures .....155
    - malaria vaccination .....156
- Immunogenic epitopes .....143–149
- allergy epitopes .....149
  - B cell epitopes
    - discontinuous.....144, 145
    - linear.....143

computational approach, epitope-based vaccines	
designing.....	140, 141
T cell epitopes	
machine learning-based methods .....	146, 147
structure-based prediction methods.....	147–149
Immunogenicity .....	139–146, 149
antigen-mediated.....	169
immunogenic epitopes.....	146–149
allergy epitopes.....	149
B cell epitopes.....	140–146
T cell epitopes ( <i>see</i> T cell epitopes)	
immunoinformatics ( <i>see</i> Immunoinformatics)	
Immunohistochemistry (IHC).....	67
Immunoinformatics	
definition .....	139
Immunology	
advanced immunoinformatics.....	169
microbe–host interactions.....	158, 159
multiscale modeling.....	167
Imputed microRNA regulation (IMRE).....	319
IMRE. <i>See</i> Imputed microRNA regulation (IMRE)	
Inborn errors of metabolism (IEMs)	
AICA–ribosiduria.....	274
classification .....	273
comprehensive mathematical metabolic models .....	273
mouse metabolic network.....	274
newborn screening programs and systems	
biology.....	273
orotic aciduria.....	274
symptoms.....	273
Induced pluripotent stem cells (iPSC).....	333
Infectious diseases	
biological experiments and mathematical	
modeling.....	108
dynamic model development.....	108, 109
epidemic modeling .....	108–111
HIV modeling.....	111–113, ( <i>see also</i> TB–HIV
Coinfection) ( <i>see also</i> Tuberculosis modeling)	
Information management.....	23–26
clinical data	
biobanks.....	25
bioinformatics and medical informatics.....	25
downstream data analysis.....	24
GWAS.....	26
molecular data.....	24
OMIM database.....	26
phenotypic information .....	25, 26
PheWAS.....	25
sharing and integration.....	25
TCGA.....	25
omics data sets.....	20
phenotype databases .....	24
public databases	
gene ontology (GO) .....	24
GeneCards.....	23
integrating information.....	23
MalaCards .....	23
metabolic reconstruction.....	23
molecular biology database collection.....	23
omics data.....	23
In-stent restenosis	
blood vessel.....	392
cellular automata approach .....	393
complex multiscale system.....	392
lattice-Boltzmann technique .....	393
SMC.....	392
Integrated care	
barriers.....	285
ICT-supported .....	285
Integrative bioinformatics	
data integration.....	27
GWAS.....	26
molecular networks.....	26
single-omics disease signatures .....	26
SNPs.....	26, 27
systems medicine strategy.....	26
Integrative pathology.....	70
Interdisciplinary systems medicine	
approaches .....	84, 85
P4 medicine.....	84
PhD research training.....	82
stakeholders.....	84
iPSC. <i>See</i> Induced pluripotent stem cells (iPSC)	
<b>K</b>	
Kinetic modeling .....	446, 449,
451, 452	
in vivo BrdU (bromodeoxyuridine) labeling	
data .....	162 ( <i>see also</i> Systems biology)
Knowledge-based models, stem cells.....	336–341
cell fate control	
coarse-grained models .....	337
cytokine concentration.....	336
cytokines LIF and FGF4.....	337
embryonic stem cells (ESCs).....	337
gene regulatory networks.....	339
kinetics-based model .....	337
ligand-receptor dynamics.....	336
NANOG expression.....	339
pluripotency.....	336, 339
signaling pathways.....	337, 338
transcription factors (TFs).....	338
cell reprogramming trajectories, epigenetics	
barrier .....	341
Boolean model.....	340
core pluripotency GRN .....	340
cycle-based binary model.....	340
definition .....	339
DNA methylation and chromatin structure.....	341
epigenetic factors .....	340

- Knowledge-based models, stem cells (*cont.*)  
  gene silencing .....341  
  GRN-based stochastic model .....340  
  iPS cells .....340  
  Nanog expression .....341  
  ODE-based model .....341  
  phase cell cycle .....340  
  pre-implantation mouse embryos .....341  
  step-wise protocols .....341  
  stem cell populations, behavior .....336
- L**
- Laser capture microdissection (LCM) .....65  
LCM. *See* Laser capture microdissection (LCM)  
Leukemia inhibitory factor (LIF) .....336  
LIF. *See* Leukemia inhibitory factor (LIF)  
Lung diseases  
  AAM and CAM activation .....126  
  alveolar macrophages .....124  
  airways .....120  
  cancer .....122–123  
  cancer progression mechanisms .....128  
  clinical diagnosis and treatment .....123  
  gas exchange and inhaled pharmaceuticals .....129  
  IFN $\gamma$  and TNF $\alpha$  signaling .....126  
  inflammation and fibrosis .....128  
  neutrophils .....124  
  pneumococcal pneumonia .....124  
  pulmonary barrier failure, bacterial pneumonia .....125  
  pulmonary *Mycobacterium tuberculosis* (Mtb)  
    infection .....125, 126  
  sarcoid granulomas .....128  
  systems medicine consortia .....119, 120
- M**
- Machine learning  
  definition .....143  
  HMM .....143  
  SVM .....143  
Magnetoencephalography (MEG) .....235  
Mathematical and modeling requirements  
  errors .....397  
  model construction .....395  
  model types .....395–396  
  multiscale parameter estimation  
    (system identification) .....398  
  scale bridging .....395  
  scale decomposition .....394  
  simulation coupling .....396  
Mathematical modeling, immune-related  
  pathways .....160, 162–166  
  cytokine dynamics  
    CD4 $^+$  T cell differentiation .....163  
    in silico analysis .....163  
    single-cell quantification, IL-2 response .....163  
  host-pathogen interaction dynamics  
    Boolean model .....166  
    cell-mediated immune response .....165  
    in silico modeling .....165  
    *Mycobacterium tuberculosis* (Mtb) infection .....165  
  immune cell phenotypes .....159  
    B cells .....162  
    NK cells .....162, 163  
    T cells .....160  
  regulatory motifs .....159  
  tumor-immunity interaction and anticancer  
    immunotherapies  
      ODE-based models .....164  
      personalized mathematical model .....164  
      spatial-oriented approaches .....164  
Mathematical models  
  cell population dynamics .....336 (*see also* Data-based models)  
  embryonic stem cell .....336 (*see also* Knowledge-based models)  
  NANOG heterogeneity .....339  
  pluripotency .....340  
  pre-implantation mouse embryos .....341  
  stem cell regulation .....333  
MDI. *See* Multiple dataset integration (MDI)  
MEG. *See* Magnetoencephalography (MEG)  
Messenger RNAs (mRNA) .....306  
Metabolomic datasets  
  CCRF-CEM model .....270  
  GIM(3)E .....269  
  lymphoblastic leukemia cell lines .....270  
  Molt-4 model .....270  
  multi-omics approach .....270  
  network pruning .....269  
  noninvasive methods .....269  
  OAT1 transporter .....270  
  transcriptomic/proteomic data .....269  
Metastasis-associated lung adenocarcinoma transcript 1  
  (MALAT1) .....308  
MicroRNA (miRNA) .....312–314  
  functional characterization .....315  
  regulation models .....311  
  RNA systems medicine .....310  
  STRING .....311  
  target prediction  
    algorithms, principles .....313  
    computational approaches .....312  
    machine learning-based approaches .....313  
    validation .....313, 314  
Mimotopes  
  description .....144  
  mimotope-based prediction methodology .....144  
Mixed-integer linear program (MILP) .....447  
MNR. *See* Module network rewiring-analysis (MNR)  
Model development .....436

Modeling	
kinetic.....	441
qualitative methods.....	442
systems biology.....	441
Models and systems medicine	
biological and biomedical research fields.....	443
drug effects, pharmacometrics.....	444
iterative and reciprocal feedback.....	442
mathematical models and modelling approaches.....	443
medical and clinical research environment.....	444
modelling tools.....	444
P4 medicine.....	444
software.....	444
systems-based approaches.....	443
Model selection, Wnt signaling pathway.....	432–433
Module network rewiring-analysis (MNR).....	323
Molecular interaction networks.....	225–227
cellular networks, link.....	228
components.....	224
construction steps	
candidate molecules, identification.....	226
molecules connection.....	227
refinement and evaluation.....	227
disease-specific	
epilepsies.....	226
HD.....	226
neurodegenerative disorders.....	226
elements and interactions.....	224, 225
epilepsies, pathogenesis.....	229
interaction	
activation.....	225
catalysis.....	225
inhibition.....	225
Wnt signaling pathway.....	227
Molecular networks	
biomarker identification.....	363
biomarker signatures.....	363
rewired genes/subnetworks.....	367
secondary data sources.....	361
mRNA. <i>See</i> Messenger RNAs (mRNA)	
Multidisciplinarity.....	10
Multifactorial disease	
and complex diseases.....	74
genetic and environmental factors.....	74
noncommunicable diseases.....	74
Multiple dataset integration (MDI).....	49
Multiscale modeling and simulation.....	380–383, 401
biomedical systems.....	379
challenges and requirements	
dynamical models.....	401
chronic diseases.....	375
combination.....	384
complex disease.....	375, 376
construction.....	384
coupling components.....	386
dynamical model.....	378
in-stent restenosis.....	393
mathematical modeling.....	377
microscopic scale.....	383
model validation.....	384
model verification.....	384
scale-bridging.....	385, 386
single spatiotemporal/organizational scale.....	383
system decomposition.....	383
traditional single-scale modeling	
( <i>see</i> Single-scale modeling and simulation)	
Multiscale modeling, immunology.....	167
plasmacytoid-DC-mediated protection.....	166, 167
tuberculosis granulomas.....	167, 168
<b>N</b>	
National Biomarker Development Alliance (NBDA).....	369
NBDA. <i>See</i> National Biomarker Development Alliance (NBDA)	
Nervous system.....	222, 223
development	
connectome.....	222
neuroectoderm.....	222
neuronal and glial cell.....	222
Wnt pathway.....	223
Network contextualization	
genome-scale reconstructions.....	268
GIMME and iMAT.....	268
Network reconstruction.....	155
Neurodegenerative disorders.....	226
Neurological diseases.....	227, 229, 233, 239–241
brain pathophysiology.....	221
cellular networks ( <i>see</i> Cellular interaction networks)	
community-driven approaches.....	241
molecular and cellular processes.....	221
multi-scale, multimodal neuroimaging.....	242
network reconstruction	
brain.....	241
cellular.....	233
molecular.....	227
requirement, systems approach.....	223, 224
synthesis	
cross-scale network analysis.....	239, 240
networks representation.....	239, 240
systems biomedicine.....	241
Next-generation pathology	
biomarkers.....	62
CRC.....	64
datafication.....	64
histopathology ( <i>see</i> histopathology.....)	61, 64
immunohistochemistry.....	62
integrative pathology.....	70



- Next-generation pathology (*cont.*)  
  molecular signatures .....62  
  mRNA/DNA-based approaches .....61  
  multi-omics .....70, 71  
  prognosis.....64  
  systems pathology.....70–72  
  TNM staging.....64
- Next-generation sequencing (NGS).....318
- NGS. *See* Next-generation sequencing (NGS)
- Non-coding RNA-based therapy design .....320, 321  
  miRNAs  
    cancer treatment .....321  
    lncRNAs.....321  
    replacement therapy.....320  
  oncomiRs.....320
- Non-coding RNAs (ncRNAs)  
  amino acids, carrier .....306  
  classification scheme.....306, 307  
  functional RNAs.....306  
  mRNA.....306  
  ribosome, and transfer RNAs .....306
- Nondimensionalization, Wnt signaling pathway  
  enzyme kinetics model .....419  
  Schmitz model.....419–421
- O**
- ODE. *See* Ordinary differential equations (ODE)
- Omics data. *See* COBRA models
- Oncogenic miRNAs (oncomiRs) .....320
- Ordinary differential equations (ODE).....311, 407, 408
- P**
- Parameter analyses  
  ABC posterior function .....428–429  
  ABC-SysBio.....429  
  Bayes rule.....428  
   $\beta$ -catenin changes .....429, 430  
  estimation and Wnt data .....427–428  
  sensitivity analysis.....429–431  
  statistical inference.....428
- Partial least squares (PLS).....49
- Pathway activity analysis  
  curated pathways .....363  
  DEAP .....366  
  DINA .....367  
  GO functional modules .....364  
  heuristic search method .....365  
  microarray gene expression data .....364  
  protein–protein interaction.....366  
  SNEA.....366  
  SPIA.....366  
  subnetworks .....366
- Pathways and networks, biomarker discovery  
  differential network analysis .....368  
  gene co-expression networks .....363
- PBPK. *See* Physiology-based pharmacokinetics (PBPK)
- PCA. *See* Principal component analyses (PCA)
- PD. *See* Pharmacodynamics (PD)
- Personalized medicine  
  4P medicine.....297  
  convergent strategies.....287  
  exposome .....286  
  individual longitudinal health plan .....283
- Personalized metabolic models  
  adipocytes .....275  
  flux span ratio .....274  
  hepatocellular carcinoma .....275  
  HHT patient *vs.* controls .....274  
  mitochondrial pathways.....275  
  plasma androsterone levels.....275  
  potential drugs (antimetabolites) .....275
- PGM. *See* Probabilistic graphical model (PGM)
- Pharmaceutical R&D .....88–91, 98–100  
  animal models.....92  
  approaches .....102  
  biomarkers, identification .....88  
  computational modeling  
    biological organization.....89  
    experimental approaches.....89  
    holistic concept.....89  
  data-driven approaches (*see* Data-driven modeling)  
  dynamic models.....102  
  genotype data.....101  
  individualized therapeutic designs .....94  
  mechanistic modeling  
    benefit.....88  
    drug action.....88  
    multiscale representation, physiology.....89, 90  
  multiscale models.....95  
  PBPK (*see* Physiologically based pharmacokinetic (PBPK))  
  pediatric scaling .....95  
  PhysioSpace concepts  
    algorithm .....100  
    definition .....100  
    PCA.....99  
  SHM (*see* Structured hybrid modeling (SHM))
- Pharmacodynamics (PD) .....456, 458
- Pharmacometrics, drug discovery .....455, 456, 458–460  
  exploratory/discovery phase .....455  
  model definition .....457  
  pharmacodynamics (PD).....456  
  phases  
    early clinical.....455  
    late clinical.....456  
  physiology-based pharmacokinetics (PBPK).....455  
  quantitative and systems pharmacology (QSP) .....455  
  SB and QSP .....456  
  standards and tools  
    PBPK.....458  
    PMX.....459–460

Phenome-wide association study (PheWAS) .....	25
Physiologically based pharmacokinetic (PBPK)	
animal models .....	92
distribution model .....	92
drug concentration profiles .....	91
FDA .....	101
ibrutinib .....	101
individualized models .....	91
multiscale models .....	95
pediatric scaling .....	95
physiological properties .....	93
reference model .....	93
tissue-specific protein .....	92
uses .....	93
vertical model integration .....	95
workflow .....	93, 94
Physiology-based pharmacokinetics (PBPK) .....	455
PLS. <i>See</i> Partial least squares (PLS)	
Pluripotency	
cellular activities .....	344
definition .....	336
epigenetic factors .....	340
epigenetics and GRN .....	340
GRNs .....	338, 341
hybrid models .....	342
LIF/JAK/STAT3 pathway .....	336
network inference models .....	337
stem cell .....	333
time-consuming task .....	344
vivo and in vitro systems .....	335
4P medicine .....	299
Predictive models in regenerative medicine	
pluripotent stem cells .....	343
regulatory levels .....	344
re-usability .....	345
validation .....	345
Principal component analyses (PCA) .....	99
Probabilistic graphical model (PGM) .....	319
Protein interaction networks	
online resources .....	362
pair-wise binding interactions .....	362
Protrusion Index (PI) .....	143
P4 systems .....	44–47
<b>Q</b>	
qPCR. <i>See</i> Quantitative PCR (qPCR)	
qRT-PCR. <i>See</i> Quantitative real-time polymerase chain reaction (qRT-PCR)	
QSP. <i>See</i> Quantitative and systems pharmacology (QSP)	
Quantitative and systems pharmacology (QSP) .....	442
Quantitative PCR (qPCR) .....	318
Quantitative real-time polymerase chain reaction (qRT-PCR)	
Quantitative structure-activity relationship (QSAR) analysis .....	140

**R**

Reverse phase protein array (RPPA) .....	27, 65
RNA systems biology .....	305, 307, 308, 317, 318
bioinformatics tools .....	309
biomarkers ( <i>see</i> Biomarker prediction)	
central dogma, evolution .....	309, 310
expression profiling ( <i>see</i> Gene expression profiling)	
lncRNAs	
HCC .....	308
HOTAIR .....	308
medicine workflows .....	324, 325
miRNA expression profiling .....	318
miRNAs	
context-specific regulation .....	307
enzymatic process .....	307
metastamir .....	307
oncomirs .....	307
multimarker prognostic model .....	319
ncRNAs ( <i>see</i> Non-coding RNAs (ncRNAs))	
PGM .....	319
prognostic models .....	319, 320
role, miRNAs and lncRNAs .....	308, 309
therapeutic strategies .....	320
web resources .....	315, 316
RPPA. <i>See</i> Reverse phase protein array (RPPA)	

**S**

SBGN. <i>See</i> Systems biology graphical notation (SBGN) project	
SBML. <i>See</i> Systems biology markup language (SBML)	
Scale-separation map (SSM)	
multiscale dynamical model and interaction .....	387
timing diagram .....	388
SDAP. <i>See</i> Structural database of Allergenic Proteins (SDAP)	
Search tool for the retrieval of interacting genes (STRING) .....	311
SED-ML. <i>See</i> Simulation experiment design markup language (SED-ML)	
Signaling pathway impact analysis (SPIA) .....	366
Simulation coupling and model	
acyclic and cyclic coupling .....	388, 390
bottom-up approach .....	391
middle-out approach .....	391
SSM .....	387
top-down approaches .....	390
two scale-specific models .....	387
Simulation experiment design markup language (SED-ML) .....	451
Simulation software	
COPASI .....	447
medical device .....	452
research problem .....	442
Single-nucleotide variants (SNVs) .....	187

- Single-scale modeling and simulation .....380–382  
  components and processes .....380, 381  
  conventional modeling.....382  
  dynamical model.....380  
    computer code .....380, 381  
    computer process .....380, 381  
    mathematical formulation.....380, 381  
    modeling and numerical errors .....382  
  model verification and validation.....382  
  numerical techniques .....381  
  sensitivity analysis.....381
- SMC. *See* Smooth muscle cell (SMC)
- Smooth muscle cell (SMC) .....392
- SNEA. *See* Subnetwork enrichment analysis  
  algorithm (SNEA)
- Software as a medical device (SaMD).....451
- SPIA. *See* Signaling pathway impact analysis (SPIA)
- SSM. *See* Scale-separation map (SSM)
- Stem cells  
  computational biology .....333  
  embryo.....333  
  human iPSC .....333  
  iPSC .....333  
  knowledge-based and data-based models .....335  
  mathematical models and simulations .....333  
  multi-cellular organisms .....332  
  multipotent.....332  
  naive and primed .....332  
  pluripotent.....332  
  pre-implantation embryos .....333, 334  
  reprogramming and differentiation.....333  
  signaling pathways and gene regulatory  
    networks .....333  
  stem cell box .....334, 338  
  totipotent.....332
- STRING. *See* Search tool for the retrieval of interacting  
  genes (STRING )
- Structural database of Allergenic Proteins (SDAP).....149
- Structural model.....457
- Structured hybrid modeling  
  biomedical applications .....98  
  black box model.....98  
  hierarchical system.....99  
  iterative refinement.....98  
  pharmacodynamics models .....98
- Subnetwork enrichment analysis algorithm  
  (SNEA) .....366
- Support vector machines (SVMs) .....143, 358,  
  366, 369
- SVMs. *See* Support vector machines (SVMs)
- Synergy-COPD .....290, 292  
  biomedical challenges  
    lung cancer diagnosis .....290  
  ICT challenges  
    data harmonization, analytics and knowledge  
      generation .....292  
  Synthetic biology open language (SBOL).....449  
  System of systems (SoS).....136  
  Systems biology.....137  
    approaches .....443  
    biocomplexity .....6  
    biological phenomena.....443  
    clinical complexity .....7,  
      (*see* Constrained-based modeling)  
    description .....137  
    dynamic life processes.....6  
    ex vivo.....75  
    genetic and environmental factors .....75  
    in vivo .....75  
    mathematical modeling (*see* Mathematical modeling,  
      immune-related pathways)  
    mathematical models.....139  
    mathematical models and computer simulations .....6  
    mechanistic approaches .....441  
    mechanistic molecular .....443  
    modeling.....441, 443  
    noninvasive data collection .....76  
    omics data.....139  
    organisms and cell cultures .....75  
    PNPLA3 gene .....75  
    tool development .....442
- Systems biology approach  
  gene regulatory networks.....311  
  mathematical modeling .....312  
  ODE .....311  
  time delay model.....311
- Systems biology graphical notation (SBGN) project.....50
- Systems biology markup language (SBML) .....449
- Systems biology modeling tool.....445, 446, 448–451  
  COBRA .....453  
  constraint-based modeling platform  
    (CBMPy), 454, 455  
  COPASI.....453  
  flux analysis and modelling environment  
    (FAME), 454, 455  
  KEGG.....455  
  kinetic.....445  
  model encoding  
    ASCII files .....446  
    command line interfaces (CLIs) .....446  
    constraint-based modeling tools .....446  
    COPASI .....446  
    data structures.....446  
    GSRs .....446  
    GUI .....446  
    interactive model .....446  
    mathematical equations .....445  
    programming language .....445

model exchange	
COMBINE.....	449
FBC.....	450
language libraries.....	451
MIASE guidelines.....	451
MIRIAM.....	449
quantitative.....	448
SBML.....	449, 450
SBOL.....	449
SED-ML.....	451
standards.....	448
tool development language.....	448
model instantiation.....	448
ordinary differential equations (ODEs).....	445
simulation software, medical device.....	452
software and standards development.....	445
strategies.....	445
<i>systems biology workbench (SBW)</i> .....	454
Systems medicine.....	12, 13
chronic diseases.....	11
chronotherapy.....	11
computing methodologies.....	47
CRC.....	45
definition.....	3
Europe	
CASyM road map.....	12
ERA-NET.....	12
framework program.....	12
horizon 2020 work program.....	12
implementation strategy.....	12
large-scale community building process.....	12
national levels.....	13
sound strategic foundation.....	12
gene expression profiling.....	11
1000 Genomes Project.....	45
HDN and DGN network.....	44
heterogeneous and homogenous entity.....	44
integration.....	9
NSCLC.....	45
P4 medicine.....	11
P4 systems.....	44–47
practical feasibility.....	11
respiratory diseases.....	46
stakeholders.....	11
Synergy-COPD project.....	46
systems biology.....	43
systems-based approaches.....	11 ( <i>see also</i> Translational informatics)
Systems medicine education. <i>See</i> Education system	
Systems pathology.....	70–72
<b>T</b>	
Targeted therapy	
artificial intelligence approaches.....	193
biomarker.....	193
chronic myelogenous leukemia (CML).....	190
clinical reports.....	191
combination therapy.....	192
driver–gate–keeper mutations.....	190, 193
druggability.....	190
EGFR mutation.....	191
ethical and legal approaches.....	192
IBM Watson’s cognitive computing.....	193
kinase inhibitors.....	191
KRAS mutation.....	191
malignant transformation.....	192
MEK inhibitors.....	191
meta-analysis.....	192
metastatic CRC.....	191
oncogene addiction.....	190
predictive dynamic modeling.....	192
selective molecules.....	192
signaling-network sub-region.....	193
sorafenib.....	191
systems medicine perspectives.....	190
TSC1 mutations.....	191
vascular endothelial growth factor (VEGF).....	191
TB-HIV coinfection	
DOTS-based programs.....	115
epidemiology of.....	115
treatment plans improvement.....	116
T cell epitopes.....	147, 148
databases.....	146
hybrid methods.....	149
machine learning based methods	
ACS.....	147
ANN-based methods.....	147
BLOSUM encoding.....	147
TAP.....	147
structure based prediction methods	
QSAR.....	147
tumor immunology.....	148
TCGA. <i>See</i> The Cancer Genome Atlas (TCGA)	
TFs. <i>See</i> Transcription factors (TFs)	
The Cancer Genome Atlas (TCGA).....	25
Therapeutic target identification	
Boolean model.....	322
HCC.....	323
kinetic model.....	324
miRNAs.....	323
MNR.....	323
model-based simulations.....	322
Transcription factors (TFs).....	338
Transfer-associated protein (TAP).....	147
Translational informatics	
BioXM.....	55
components.....	53
dbGAP.....	54
GenomeSpace and Garuda frameworks.....	54

Translational informatics (*cont.*)

- high-dimensional phenotype datasets.....54
- Innovative Medicines Initiative .....55
- Matlab code and OMIC tools.....54
- T-MedFusion system.....56
- tranSMART platform .....55

Tuberculosis modeling

- “dormant” cells.....114
- antituberculosis drugs .....114
- in vitro and in vivo experiments.....113
- Mycobacterium tuberculosis* .....113
- profile likelihood method .....115

Tumor immunology.....148

**U**

U-BIOPRED. *See* Unbiased Biomarkers for the Prediction of Respiratory Disease Outcome Consortium (U-BIOPRED)

Unbiased Biomarkers for the Prediction of Respiratory Disease Outcome Consortium (U-BIOPRED).....127

**W**

Weighted gene co-expression network analysis (WGCNA).....31

WGCNA. *See* Weighted gene co-expression network analysis (WGCNA)

Wnt signaling pathway

- agent-based models .....408
- asymptotic analysis .....421–424
- cell-surface receptors .....410

- competitive binding .....411
- coplanarity via algebraic geometry.....434–435
- CRNT .....434
- cytokines and hormones .....405
- destruction complex (DC).....410, 411
- deterministic approaches .....407
- enzyme kinetics model .....409–410
- features .....411, 412
- gene regulatory pathways.....406
- Lee model
  - catalytic processes .....413, 414
  - conservations laws.....414
  - definition of notation.....413, 414
  - ODEs .....413
- lymphoid-enhancing factor (LEF) proteins.....410
- Mass balance.....409
- messenger RNA (mRNA) .....406
- model selection .....432–433
- noncanonical signaling .....410
- nondimensionalization .....418–421
- ODE models .....407, 408
- parameter analyses .....424–431
- phosphorylation.....406
- principles .....408
- quantitative model.....411
- Schmitz model.....415–417
- steady state analysis
  - enzyme kinetics model.....417
  - Schmitz model.....417–418
- stochastic approaches.....407
- systems medicine .....407
- T-cell factor (TCF).....410