Ming Chen
Ralf Hofestädt *Editors*

Integrative Bioinformatics

History and Future

Springer

Integrative Bioinformatics

Ming Chen • Ralf Hofestädt
Editors

Integrative Bioinformatics

History and Future

Springer

Editors
Ming Chen
College of Life Sciences
Zhejiang University
Hangzhou, China

Ralf Hofestädt
Faculty of Technology
Bielefeld University
Bielefeld, Nordrhein-Westfalen, Germany

ISBN 978-981-16-6794-7 ISBN 978-981-16-6795-4 (eBook)
https://doi.org/10.1007/978-981-16-6795-4

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
This work is subject to copyright. All rights are solely and exclusively licensed 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, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The unprecedented accumulation of high-throughput data from genomics, transcriptomics, proteomics, metabolomics, phenomics, etc., has resulted not only in new attempts to answer traditional biological questions and solve longstanding issues in biology but also in the formulation of novel hypotheses that arise precisely from this wealth of data. At present, with nearly 5000 biological data resources and information systems on the Internet, numerical bioinformatics tools, and exponential growths of omics data (big data), the storage, processing, description, transmission, connection, and integrative analysis of these data are still a great challenge for bioinformatics. Addressing this situation, we need information systems which realize the user-specific integration of data and analysis tools to help solve molecular questions. Therefore, the implementation of integrative information systems is the actual task, and the systems biology study by integrating different types of data at different levels is a key point in order to understand the mechanism of life.

With this idea in mind, the first book on Integrative Bioinformatics *Approaches in Integrative Bioinformatics—Towards the Virtual Cell* (ISBN: 978-3-642-41280-6) was published in 2014. It has been viewed/downloaded over 27 k times. Nevertheless, the past few years witnessed the rapid development of big omics data science and practical application of artificial intelligence in life sciences. New aspects and approaches have been emerged in the field of Integrative Bioinformatics. We felt encouraged to re-edit a new version of Integrative Bioinformatics to review the latest achievements and shed light on possible future development.

The initial idea of this book is based on a Symposium—Integrative Bioinformatics: History and Future, which took place in 2019 at Bielefeld University (Germany) and resulted in a special issue of the *Journal of Integrative Bioinformatics* (JIB) (https://www.degruyter.com/journal/key/JIB/16/3/html). This JIB special issue contains a unique compilation of invited and selected articles from JIB and annual meetings of International Symposium on Integrative Bioinformatics. Subjects covered include a summary of essential topics, basic introductions and latest developments, biological data integration and manipulation, modeling and simulation of networks, as well as a number of applications of Integrative Bioinformatics. It presents different views of Integrative Bioinformatics based on the

aspect of history and future, aiming to provide a basic introduction of biological information systems, and give guidance on the computational analysis of systems biology, covering a range of issues and methods that unveil the multitude of omic data integration performed and the relevance that Integrative Bioinformatics has today.

The book is divided into five parts:

Part I starts with a brief introduction of the history of Integrative Bioinformatics (Chap. 1). It is followed by a perspective on current developments in data management and data publication, particularly focusing on plant bioinformatics, from genotypes to phenotypes (Chap. 2).

In Part II, Chap. 3 introduces the data landscape that enables access to data resources for researching in the field of epidemiology. Chapter 4 describes major problems in database integration and presents an overview of important information systems. The information reconstruction and visualization process based on that integrated life science data are further discussed. In Chap. 5, a potential of the fully automated, graph-based data integration is described and a mapping tool BioDWH2 is explored. Chapter 6 introduces DaTo, a collection of published online biological databases and tools. It integrates a graphical interaction network browser to facilitate exploration of the relationship between different tools and databases with respect to their ontology-based semantic similarity. Chapter 7 describes bioinformatics approaches of using workflow-driven data integration and knowledge graphs for plant breeding. A customized instance of the open-source Galaxy computational platform and analyses of breeding data in a workflow-driven approach is presented.

Part III focuses on integrative analysis. Chapter 8 shows how to integrate and make sense of this wealth of data through digital applications that leverage knowledge graph models; as a significant use case, a genetic discovery platform, KnetMiner, leverages knowledge graphs built from molecular biology data sources. Chapter 9 explores plant transcription factor regulatory networks by integrating genome-wide datasets from ChIP-Hub database, to dissect the network structure to identify potentially important cross-regulatory loops in the control of developmental switches in plants. Chapter 10 introduces microbiome big data and databases and describes several microbiome applications to showcase the power of microbiome big data integration and mining for knowledge discovery and clinical applications. Chapter 11 presents potential applications of data integration for medical information systems towards e-healthcare.

Visualization, modeling, and simulation of complex biological networks are a major aspect of Integrative Bioinformatics and Systems Biology. In Part IV, Chap. 12 discusses the past, present, and future of the visualization of metabolic networks and pathways and provides links to several resources. One highlight shown is by an international consortium which started developing a standard for the graphical representation of cellular processes and biological networks including metabolism called the Systems Biology Graphical Notation (SBGN). Chapter 13 focuses on a Petri net formalism that covers discrete, continuous, as well as stochastic models among other features. VANESA and a Petri net library PNlib are introduced to model and simulate metabolic pathways. Chapter 14 discusses a few

promising immersive analytics and visualization-related approaches in the context of Integrative Bioinformatics.

Science has become more and more data-driven; data and analysis tools are available on the internet. In integrative data analysis, various tools, pipelines, and platforms have been developed. In Part V, Chap. 15 calls for action to develop the Internet of Science platform for scientific workflow management to facilitate a future focus on collaborative knowledge discovery. Chapter 16 describes a knowledge graph, AgroLD, to exploit the Semantic Web technology and some of the relevant standard domain ontologies, to integrate knowledge on plant crop species. Chapter 17 presents TBtools, an out-of-box solution to routine biological data analyses. It describes the design philosophy, development objectives, and main characteristics, and comprehensive introduction of TBtools. Chapter 18 introduces prominent integrated bioinformatics platforms, such as Galaxy and relevant framework applications.

Hangzhou, China — Ming Chen
Bielefeld, Germany — Ralf Hofestädt

Contents

About the Editors

Ming Chen is the Director of the Bioinformatics Laboratory at the College of Life Sciences, Zhejiang University (China). He received his PhD in Bioinformatics from Bielefeld University (Germany). His group research work covers bioinformatics, systems biology, noncoding RNA transcriptomics, and precision medicine. He published over 200 papers in peer-viewed journals. He is the President of the Bioinformatics Society of Zhejiang Province, China; Deputy director of Chinese Society for "Modeling and Simulation of Biological Systems," Committee executive member of Chinese Society for "Computational Systems Biology," and Committee member of Chinese Societies for "Functional Genomics & Systems Biology," "Biomedical Information Technology," and "Biophysics (Bioinformatics)."

Department of Bioinformatics, College of Life Sciences; The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China

Ralf Hofestädt studied Computer Science and Bioinformatics at the University of Bonn. He finished his PhD in Computer Science (University Bonn) and his Habilitation (Applied Computer Science and Bioinformatics) at the University of Koblenz. He was Professor for Applied Computer Science at the University of Magdeburg. Now he is Professor for Bioinformatics and Medical Informatics at the University of Bielefeld. He launched the *Journal of Integrative Bioinformatics* in the early 2000s. The research topics of the department concentrate on biomedical data management, modeling and simulation of metabolic processes, parallel computing, and multimedia implementation of virtual scenarios.

The Bioinformatics/Medical Informatics Department, Bielefeld University, Bielefeld, Germany

Part I
Introduction of Integrative Bioinformatics

Chapter 1
Integrative Bioinformatics: History and Perspective

Ming Chen, Ralf Hofestädt, and Jan Taubert

Abstract This chapter introduces the history of Integrative Bioinformatics. Particularly, it outlines major events in the field from Germany who took a leading role and from China who plays a rapid developing counterpart. The earliest bioinformatics database resources, projects and initiatives are mentioned. We are stepping into the biological big data era, which requires us to develop new methods, cutting-edge technologies to deal with the vast multi-scale and multi-dimensional data. Several directions that may lead to solve the bottleneck of Integrative Bioinformatics are discussed. As life-sciences become more data-driven, Integrative Bioinformatics aims to integrate various aspects together to comprehensively understand the mechanism of life, and make the outputs available for use in the industry.

Keywords Integrative bioinformatics · History · Future · Perspective · Industry

1.1 History

The "Human Genome Project" emphasized the significance of methods and concepts from applied computer science for genome analysis and biotechnology. This focus was the key argument of the German Ministry of Science (BMBF) to support Bioinformatics at the beginning of the 1990s (Schütt and Hofestädt, 1992). During the same time, the German Society of Computer Science (GI) founded a working

M. Chen (✉)
Department of Bioinformatics, College of Life Sciences, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
e-mail: mchen@zju.edu.cn

R. Hofestädt (✉)
The Bioinformatics/Medical Informatics Department, Bielefeld University, Bielefeld, Germany
e-mail: hofestae@techfak.uni-bielefeld.de

J. Taubert (✉)
KWS SAAT SE & Co. KGaA, Einbeck, Germany
e-mail: jan.taubert@kws.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_1

group (GI FG *Informatik in den Biowissenschaften*) to coordinate national activities (Nachrichten 1994; Hofestädt 2000). During that time, the first national conference on Bioinformatics was organized in Bonn 1993 (Hofestädt et al. 1993). At the same time, interdisciplinary activities started across the whole world. For example, the first ISMB conference was organized in 1993 Washington (Hofestädt 1993). In 1996, the German annual national conference was organized in Leipzig as an international conference—the so-called German Conference on Bioinformatics (GCB) (Hofestädt 1997; Hofestädt et al. 1997). In parallel, the GI working group specified the Bioinformatics curriculum and the German Research Foundation (DFG) offered special grants for German Universities to support faculties building up new studies for Bioinformatics. Furthermore, the German Ministry of Science (BMBF) offered a grant to support five Bioinformatics centers in Germany during the same time. Therefore, Bioinformatics was established in Germany and in many other countries, including China, at the end of the last century.

In China, Bioinformatics was initiated by several notable physicians and mathematicians who started bioinformatics research from the end of the 1980s (Wei and Yu 2008; Chen 2021). Due to the limitation of bioinformatics facilities and international collaboration at that time, Bioinformatics was not well developed until 1997–1998, two Bioinformatics centers were established in Peking University and Tianjin University successively (Luo 2021). In 1997, the first Xiangshan Science Conference on Bioinformatics was held in Beijing. Later, South Center (Shanghai) and North Center (Beijing) of China Human Genome Center (CHGC and CHGB) were established. Beijing Genomics Institute (BGI, currently known as the BGI group) was founded in 1999 to participate in the Human Genome Project (Waterman 2021; Dong et al. 2021). In the same year, the first Sino-German workshop of Bioinformatics was organized in Beijing. At the beginning of the new century, the first Chinese national conference on Bioinformatics was launched in Beijing in 2001. Since then, powered by returned scientists from overseas and the young talented generation, Bioinformatics in China continues to grow, promoting the accumulation of biological data, methods, tools, and contributing the rapid development of Bioinformatics. Two journals: *Chinese journal of Bioinformatics and Genomics*, Proteomics & Bioinformatics were launched in 2003. The Sino-German Integrative Bioinformatics cooperation started in 2009 when we founded the Sino-German Network supported by the German and Chinese Ministry of Science and Technology.[1]

From 1995 to 2004, the GI FG *Informatik in den Biowissenschaften* organized different international Dagstuhl seminars (Hofestädt et al. 1996; Collado-Vides et al. 1999; Collado-Vides and Hofestädt 2002; Hofestädt 2005), which discussed actual research topics of Bioinformatics. In 1995, the main topic was modeling and simulation based on molecular data and database systems. During that time, the internet revolution in combination with the new omics technology showed up. This was the starting point to develop and implement new information systems, which

[1] http://www.imbio.de/forschung2/

allow the systematic storage and analysis of molecular data. Besides the implementation of database systems (KEGG, TRANSFAC, PDB, etc.), the development and implementation of analysis tools became more and more important. Furthermore, the relevant molecular data and analysis tools became available via the internet during that time. The next step was to develop and implement software tools for the user-specific integration of data and analysis systems. Therefore, using computer science methods, new concepts and tools had to be developed for the integration and fusion of molecular data and analysis tools. At the beginning of this time, the concept of federated databases was common. The idea was/is to integrate worldwide running and supported data and database systems. Regarding aspects of data security and real-time access conditions, this approach failed. To eliminate these problems, the data warehouse concept was developed and implemented (Kormeier et al. 2011). Data warehouses can integrate heterogeneous and worldwide distributed user-specific data into one new and local organized database system (integrated user-specific in-house solution). Until now, data warehouse architectures are in use and still represent important and useful solutions. Overall, the main problem of this kind of integration is to overcome the heterogeneity problem of molecular data and database systems. One key problem is that a high percentage of these molecular information systems is represented by flat file systems until today. This is the main reason for the complexity of the integration process based on the implementation of specific adapter systems. One other important integrative aspect was and is to have access to literature information systems like for example PubMed. The access to all published papers (abstracts) could also be realized when the internet became available at the end of the last century. During recent years, methods of text mining and data mining got practical relevance. Today such tools are available and able to scan all PubMed abstracts (or papers). This mining and filtering process is useful to extend our knowledge based on annotated database and information systems. One more alternative integration concept was and is the specific definition and implementation of workflows, which integrate user-specific data and analysis tools directly.

The international journal of *Nucleic Acid Research* is trying to present the overview of all available molecular database systems at the beginning of each year. For integrative and analysis tools, we do not have this kind of service yet. The scientific relevance of these software techniques and their applications was the key motivation to organize the Integrative Bioinformatics Dagstuhl seminar 2004.[2] One result of this seminar was the foundation of the *Journal Integrative Bioinformatics (JIB)*, which is published by de Gruyter[3] since 2017. Furthermore, this Dagstuhl seminar in combination with the *Journal Integrative Bioinformatics* was also the beginning for the annual international conference of Integrative Bioinformatics (IB2022 will take place in Konstanz, Germany).

[2] https://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=04281

[3] http://www.degruyter.com/view/j/jib

1.2 Future Aspects

We are now increasingly in the big data era. Bioinformatics is facing much more heterogeneous biological data with huge volumes. Genome Projects like "1+ Million Genomes" Initiative are going on, leading to more and more individual sequences. It is not only for human but also for other species, as more and more species have been sequenced, e.g., "The Earth BioGenome Project" and "Million Microbial Genomes Project." Moreover, it does not simply measure whole tissue samples, but distinctly identify DNAs/RNAs or proteins at a cellular level. Single-cell sequencing and single-cell proteomics are generating millions of profiling datasets in a short time period. The multi-omics data brings us new challenges to develop appropriate integrative bioinformatics approaches to manipulate, integrate and model complex biological systems at spatial and temporal scales.

Since biological data are subjective and biased, often lacking standardization and reproducibility, and some databases are not well maintained, these resources are becoming more and more degraded. Although there are several bioinformatics methods developed to deal with a certain problem, often only one is widely used and highly cited, which encourages becoming a common/standard method. In many cases, we are not well aware of the original hypothesis of such methods, which may mislead the real problem. How to integrate the multi-omics data with different biological/technical conditions and bias? How to share/deposit data under an acceptable intelligence and ethic policy? Are our traditional data mining and machine learning methods suitable for big data? More powerful tools for multiple scale biological interactome modeling and simulation? How to uncover hidden patterns from such a huge and heterogeneous number of omics data and allowing the creation of predictive models for real-life applications? Nevertheless, advances in biological technologies and computational methodologies have provided a huge impetus. There are several directions which may lead to solve the bottleneck of Integrative Bioinformatics.

1. Integration of multiple biological data toward systems biology. Different omics data is reflecting different aspects of the biological problem. For instance, previously biological networks are regarded as gene regulatory network, protein–protein interaction network and metabolic networks. Now we know that non-coding RNAs, including microRNAs, siRNAs, lncRNAs, ceRNAs, cirRNAs, etc., play more important roles in regulations. Therefore, an integrative interactome model (e.g., a virtual cell) of known parts and non-coding RNAs needs to be built.
2. Integration of various bioinformatics methods and approaches. Often, to solve a problem, there are many different methods developed by many groups. These methods may perform differently, some good, some bad. However, individual results are often unreliable. In particular cases, the often-used methods may be unreliable or simply ineffective. It is suggested to depend on a variety of results by all methods. With various methods, we can integratively develop tailored bioinformatics pipelines to facilitate better understanding of biological problems.

3. To integrate multiple biological data and different methods/approaches, well-developed traditional data mining methods such as NN, SVM, HMM are available. However, they are not good enough to deal with high dimensional omics data and big data sets. So far, deep learning methods such as CNN, GNN and Transformer have been efficiently used. Combined with big data, and other approaches, artificial intelligence (AI) has been successfully applied in bioinformatics, especially in the field of biomedical image analysis.
4. Computing infrastructure development. Integrative Bioinformatics in the big data era requires a more advanced IT environment. To facilitate the related computing and visualization demands, both hardware (e.g., GPU, TPU) and software (e.g., TensorFlow, PyTorch) have been improved. Supercomputers are used. Cloud services are provided by more and more institutes and big companies.

1.3 Industry Aspects

During the turn of the century, the availability of the fully sequenced human genome and other model organisms sparked a boom of bioinformatics companies aiming to address the challenges in medicine, plant, and other life sciences using computational methods. Despite initial success like improving genome annotations or modeling of more complex protein structures, big promises like in silico drug discovery were not able to be kept and even huge players like Lion Biosciences diminished (McKelvey et al. 2004). Nevertheless, the enthusiasm and learnings of that time led to the establishment of dedicated bioinformatics functions within almost all of life sciences industries. These bioinformatics functions would be placed within the Research and Development functions of life science companies. As dedicated talent in bioinformatics was rare, biologists, computational scientists or even physicists and others strained in the new area of bioinformatics. The demand of industry for talent influenced the academic world and drove the creation of more and more bioinformatics or related curricula. Even though the large initial demand for bioinformaticians at the beginning of the century flattened, there is still a shortage of talent, especially when existing industry experience is required. It is almost certain that this trend will not change in the foreseeable future, as life science data continues to grow.

When such bioinformatics functions were embedded in the overall R&D ecosystem of a life science company, other surrounding data systems were and still are in-place concerning relevant data domains. These data systems can range from simple spreadsheets used by the scientists to Microsoft Access databases and relational database systems. Understanding the data stored in these systems and adding the contributions of bioinformatics tools and predictions to the R&D ecosystem heavily relies on integrative bioinformatics approaches. Breaking up data silos between functional units within the R&D ecosystem is a prerequisite to drive not only track and traceability of processes but also the discovery of new insights. Technologies like semantic web (Berners-Lee et al. 2001) or linked data provide the base

infrastructure of efficient bioinformatics functions. Ontologies (Smith et al. 2007) either reused from public repositories or customized together with R&D scientists establish a common language, which should also be machine interpretable. FAIR (Wilkinson et al. 2016) (findable, accessible, interoperable, reusable) principles of data management are increasingly being adopted in industry. This trend is supported by advancement of many public and proprietary bioinformatics tools implementing these principles.

Even though hesitant at first, industry is now steadily moving from on-premises data infrastructure to (private) cloud computing. Here the bioinformatics functions are beneath the early adopters of cloud computing (Sommer 2013) as they are commonly exposed to an ever faster changing portfolio of public and proprietary bioinformatics tools and services. As such they rely on the flexibility and power of cloud computing to evaluate new approaches or tools for use in life science industry. Such new approaches may also include artificial intelligence and machine learning techniques (Mak and Mallikarjuna 2019). Besides the current hype around these techniques, more and more use cases are being discovered by industry. Here the additional challenge arises to turn a proof of concept into a production ready system to be integrated into the R&D process. This requires not only a sound understanding of the data and algorithms, but also of the end users. Therefore, the classical role of business analysts in industry is supplemented by skills of user-centered design and user experience (Ziemski et al. 2019/20). The outcomes of this interaction are then driving either internal or external software development efforts or influence buying decisions. Still the adaption of free and open-source software (FOSS) in industry remains challenging due to complicated licensing and unclear legal terms (Vetter 2009).

To align pre-competitive industry efforts in common tasks of R&D data management, alliances like the Pistoia Alliance[4] have been formed. Here, life sciences industry, suppliers, academics, and start-ups discuss forthcoming challenges and evaluate common ground. In Europe, the ELIXIR Bioinformatics Industry Forum or the Innovation and SME Forums,[5] besides others, support engagement between ELIXIR member institutions and industry participants to exchange on services and the ecosystem of public data resources. One ELIXIR member, EMBL-EBI, since 1996 offers an industry program[6] to provide a forum for interaction and knowledge exchange for those employed at the forefront of industrial bioinformatics. These and many other initiatives help to disseminate cutting-edge technologies to industry at a time when life sciences becomes more data-driven.

[4] https://www.pistoiaalliance.org

[5] https://elixir-europe.org/industry

[6] https://www.ebi.ac.uk/industry

References

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284:34–43
Chen R (2021) Early bioinformatics research in China. Quant Biol 9(3):242–250
Collado-Vides J, Hofestädt R (eds) (2002) Gene regulation and metabolism—postgenomic computational approaches. MIT Press, Cambridge, MA. ISBN: 0-262-03297-X
Collado-Vides J, Hofestädt R, Mavrovouniotis M, Michal G (1999) Modeling and simulation of gene regulation and metabolic pathways. BioSystems 49:79–82
Dong W, Qiang B, Yang H (2021) The international Human Genome Project: a milestone for life sciences and humanity—the three stages and three major impacts of the HGP, and three contributions by China. Quant Biol 9(3):229–241
Hofestädt R (1993) Grammatical formalization of metabolic processes. In: ISMB-93 proceedings, pp 181–189
Hofestädt R (1997) Computer science and biology—the German Conference on Bioinformatics (GCB'96). BioSystems 43:69–71
Hofestädt R (ed) (2000) Bioinformatik 2000: Forschungsführer Informatik in den Biowissenschaften. BIOCOM Verlag. ISBN: 3-928383-11-6
Hofestädt R (2005) Dagstuhl seminar on integrative bioinformatics. In Silico Biol 5:81–82
Hofestädt R, Krückeberg F, Lengauer T (eds) (1993) Informatik in den Biowissenschaften. Informatik Aktuell. Springer, Heidelberg. ISBN-13: 978-3-540-56456-0
Hofestädt R, Mavrovouniotis M, Collado-Vides J, Loffler M (1996) Modeling and simulation of metabolic pathways, gene regulation and cell differentiation. BioEssays 18(4):333–335
Hofestädt R, Lengauer T, Löffler M, Schomburg D (eds) (1997) Bioinformatics. LNCS, 1278. Springer. ISBN: 3-540-63370-7
Kormeier B, Hippe K, Hofestädt R (2011) Data warehouses in bioinformatics: integration of molecular biological data. Inf Technol 53(5):241–249
Luo J (2021) Professor GU Xiaocheng and the Center for Bioinformatics at Peking University (in Chinese). SCIENTIA SINICA Vitae. https://doi.org/10.1360/SSV-2021-0332
Mak KK, Mallikarjuna RP (2019) Artificial intelligence in drug development: present status and future prospects. Drug Discov Today 24(3):773–780
McKelvey M, Annika R, Jens L-H (eds) (2004) The economic dynamics of modern biotechnology. Edward Elgar Publishing. ISBN: 978-1-84376-519-6
Nachrichten (1994) Nachrichten aus der Informations-technischen Gesellschaft und der Gesellschaft für Informatik. Informationstechnik und Technische Informtik 36(6):66–70
Schütt D, Hofestädt R (1992) Bioinformatik und Umweltinformatik—neue Aspekte und Aufgaben der Informatik. Informatik Forsch Entw 7:175–183
Smith B, Ashburner M, Rosse C et al (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25:1251–1255
Sommer T (2013) Cloud computing in emerging biotech and pharmaceutical companies. Commun IIMA 13(3):3
Vetter GR (2009) Commercial free and open source software: knowledge production, hybrid appropriability, and patents. Fordham Law Rev 77(5):2087
Waterman M (2021) The Human Genome Project: the beginning of the beginning. Quant Biol 9(1):4–7
Wei L, Yu J (2008) Bioinformatics in China: a personal perspective. PLoS Comput Biol 4(4):e1000020
Wilkinson MD, Dumontier M, Aalbersberg IJ et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3(1):1–9
Ziemski J, Fortenbacher S, Hoeksma J et al (2019) Evolution of user experience for life sciences. Drug Discov World Winter 20:59–66

Chapter 2
From Genotypes to Phenotypes: A Plant Perspective on Current Developments in Data Management and Data Publication

Daniel Arend, Sebastian Beier, Patrick König, Matthias Lange, Junaid A. Memon, Markus Oppermann, Uwe Scholz, and Stephan Weise

Abstract Integrative bioinformatics aims to combine information from various sources of different data domains in such a way that a cross-domain analysis becomes feasible. With this approach, insights may be gained, which would not be possible with an analysis restricted to a single domain. For example, relationships between genotypic characteristics (genotypes) and phenotypic characteristics (phenotypes) in their environmental context (environment) could be made visible. The efficient management of such data combined with the supply of corresponding machine-readable access possibilities are essential prerequisites to achieve the outlined goal. This awareness was the nucleus for the development of the concept of data life cycles. In such a cycle, the stages of planning, collecting, processing, analysing, preserving, sharing and reusing are represented. All these steps must be considered, mapped and carried out accordingly in data management.

This chapter will discuss this data life cycle. The description of the individual steps is always based on concrete applications of a modern plant research institution and is therefore allocated to the field of plant bioinformatics. The focus here is primarily on the three data categories "genotype", "phenotype" and "environment". The spectrum of activities ranges from local data management to making data available in public archives and thus includes project planning, metadata definition and collection, database storage solutions, data curation processes, data integration technologies, data access interfaces as well as data reusability. The ultimate goal is to make all research results available to the public according to the FAIR principles of Findable, Accessible, Interoperable and Reusable.

Keywords Plant genetic resources · Biodiversity · Data management · FAIR · Data life cycle · Plants · Genotype · Phenotype · Environment

D. Arend · S. Beier · P. König · M. Lange · J. A. Memon · M. Oppermann · U. Scholz (✉) · S. Weise
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany
e-mail: scholz@ipk-gatersleben.de

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_2

2.1 Introduction

This chapter is based on more than 20 years of data management experiences and activities at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK). The IPK is a leading international research institution in the field of crop plants and their wild relatives. Research focuses on the conservation of biodiversity and the performance of crop plants.

The Institute's distinguishing feature is the German Federal ex situ Genebank for Agricultural and Horticultural Crops. This is one of the world's largest genebanks and the largest of its kind in the European Union (EU27). The tasks of the genebank are the conservation of agrobiodiversity and the provision of plant genetic resources (PGR) for research and breeding. The IPK collection comprises about 151,000 samples, so-called accessions, which cover more than 3000 different species. The genebank represents a vault in which the biodiversity of cultivated plants is stored. To maintain this unique collection, regular multiplication trials have to be carried out. This involves recording a wide range of data, in particular phenotypic observations, but also environmental data (e.g. temperature, rainfall or UV radiation). As with all organisms, the phenotype of plants is influenced not only by the genotype but also by the environment. During cultivation, mainly phenotypic traits are recorded. In order to better understand the material, it also becomes useful to use genomic data, e.g. to explain the influence of genotypic variation on the phenotype.

While the data-side focus of the genebank has traditionally been on the passport data of the accessions and on phenotypic observation values, the extension of digital information services makes it possible to integrate data from other domains, e.g. genome or genotyping data, and thus successively develop the genebank into a bio-digital resource centre.

Concretely, in this chapter, we will discuss and include five data domains: plant genetic resources data (1), genomic data (2), genotyping data (3), phenotyping data (4) and environmental data (5). We will briefly explain how we define these terms in the following paragraphs.

1. **Plant genetic resources data**: On the one hand, this includes for each accession so-called ***passport data*** like country of origin, collection site, the genus/species, the full botanical name and recently also unique identifiers like a DOI (Digital Object Identifier). Furthermore, this includes ***characterization data***. These data describe the phenotype and are rather stable, e.g. the properties of organs such as the ear in cereals (e.g. two-row or six-row in barley). The third part is the ***evaluation data***. These are phenotypic characteristics that are collected during propagation cultivation. These include, for example, plant height, disease infestation or yield data such as the 1000 kernel weight.
2. **Genomic data**: These are, on the one hand, ***sequence data*** such as nucleotide sequences of entire chromosomes at pseudomolecule level, and the gene models (genes with their localization on the chromosomes, exons, introns, as well as the coding sequence and translated peptide sequence). Furthermore, it also includes

descriptive ***annotations*** of the structural regions of the genome like genes and their functions or information about non-coding regions such as repeats.
3. **Genotyping data**: This includes diversity information on how a specific genotype (e.g. one accession) or several genotypes (e.g. several accessions) differ from a reference genotype. The methods used to determine such differences are very diverse, e.g. SNP arrays or genotyping by sequencing (GBS). This also results in very heterogeneous data formats. One example is the so-called variant calling format (VCF). Here, the differences of several genotypes can be mapped to the reference, including qualitative assessments.
4. **Phenotyping data**: This includes all phenotypic traits that are collected outside the classical conservation cultivation in the genebank. This covers experiments both in the field and under controlled conditions, e.g. in the greenhouse.
5. **Environmental data**: These include weather data such as temperature, precipitation, humidity, wind speed or UV radiation. Furthermore, this includes data collected by environmental sensors in isolated environments, e.g. greenhouses. The data from environmental sensors complement the existing weather data and can therefore also be counted as part of it.

These characterized data domains are in the focus of the further described data management processes and systems.

2.2 Data Management Concepts in Plant Science

Data management plays an important role in achieving the goal to transform the IPK genebank into a bio-digital resource centre. In the beginning, data was managed analogously on paper or index cards. With the availability of computers, these were rapidly used for this purpose. In particular, database systems were identified as the more effective tool for this task. First databases were created in which different information could be stored and queried in a structured way. Often the results of scientific studies were imported and certain parts could be queried and extracted again. Unfortunately, a description and documentation of how the data acquisition was often missing. However, this is essential in order to be able to reuse the results and feed them as input into new studies.

In recent years, it has become clear that data management is a process that takes place over several stages and can be accompanied and supported by the use of databases. Ultimately, this process is transferable to all scientific fields and is currently a topic in the new scientific discipline of Data Science. Currently, this process is known as the Data Life Cycle (ELIXIR 2021) and is illustrated in Fig. 2.1. Each step of the Data Life Cycle is briefly described in the following paragraphs.

The process step **plan** defines a strategy for managing the data and documentation generated in the research projects. Consideration should be given in advance on how best to avoid problems associated with data management and to create the

Fig. 2.1 Data Life Cycle adapted from RDMkit (ELIXIR 2021)

conditions to ensure that all research data continue to have maximum impact in science beyond the end of the research project.

Data **collection** describes the process of gathering information for specific parameters either automatically, i.e., using instruments, as well as manually. During this process, data quality must be ensured regardless of the research field.

Data **processing** is the step in the cycle where data is converted into a format to prepare it for analysis. In addition to format conversion, this stage of the process includes quality checking and pre-processing according to standardized protocols. Furthermore, poor- or low-quality data is discarded in order to create a cleaned dataset that provides reliable results.

In the **analysis** step, the collected data is examined to identify the information contained in a dataset. These investigations can be performed multiple times in the process. Specifically, the data can be analysed directly or indirect analyses can be performed by using models, for example.

Data **preservation** includes all activities necessary to ensure the safety, integrity and accessibility of data for as long as it is required. Data preservation is more than storage and backup. It prevents data from becoming unavailable and unusable over time.

In the **sharing** phase, the data is made available to others. This can be sharing with collaborative partners or publishing the data to the whole research community. It is important to note that data sharing is not the same as making data open access. It is the decision of the data producer how the data will be shared. Thus, restricted access for different user groups is also possible, e.g. only for collaborative partners.

In the **reuse** phase, data is used for a new purpose for which it was not originally intended. This makes it possible to generate and also publish new results based on the same data. Reusability is an essential part of the FAIR principles.

In addition to the steps in the life cycle, the use of data standards as well as data concepts is, of course, essential basics in data management. Furthermore, it is crucial to manage and offer data according to the FAIR data principles. First formulated in 2016 Wilkinson et al. 2016), it is now established in more or less all data domains,

and several funding agencies have also made FAIR their central paradigm (Mons et al. 2017).

Behind this acronym is a guideline for handling research data in a sustainable way.

- **Findable:** Research data needs to be findable by humans via search engines, but also by machines using standardized harvesting formats.
- **Accessible:** A long-term stable access to research data is crucial for sharing research data within the research community and public users and to get credit for the data producers.
- **Interoperable:** Standardized metadata formats are essential for describing research data to integrate them and find possible interconnections.
- **Re-usable:** In order to exploit the full value of research data, it is necessary to provide a full technical description, which guarantees as far as possible a repeatability of the underlying process to create the data and allow users to use them for further investigations.

Therefore, FAIR has also been an important goal during the development of novel standards and updating of existing formats. Some popular examples are the MIAPPE recommendations for describing plant phenotypic experiments (Papoutsoglou et al. 2020) and the MCPD standard for describing plant genetic resources (Alercia et al. 2015). MIAPPE is a descriptive list of recommended minimal attributes, which are helpful to explain and document the experimental setup of phenotypic trials. It was originally described in 2015 and is still under active development. On the other hand, the MCPD (Multi-Crop Passport Descriptors) standard is relatively old and established across genebanks and plant genetic resource providers worldwide. It provides a comprehensive list of stable and well-defined attributes necessary to document genebank accessions.

But of course, meeting the FAIR recommendations requires not only an improvement of data formats and standards but also a re-design and update of existing infrastructures and databases. One obstacle is the homogenization of the vocabularies used in these resources. It is necessary for resources that offer data for exchange to access a standartised vocabulary established by the community. To this end, consortia have been formed with the mission of building such ontologies. For example, there are specialized ontologies that offer a controlled vocabulary for describing plant structures and growth stages (Jaiswal et al. 2005) or ontologies that describe more general concepts, like the Gene Ontology (Gene Ontology Consortium 2004). However, as both language and methodologies continue to evolve, this effort must be supported and sustained.

In all research areas, including the life sciences, the tasks of data management and publication are of essential importance. Only in this way can new findings be appropriately substantiated and are traceable. Initially, these tasks were performed exclusively in analogue form. With the broad emergence of computers, it became digital. Along with the triumph of the World Wide Web, these two tasks have received a considerable boost.

2.3 Overview of Information Systems

The general architecture of information systems can be divided into two distinct entities: (1) the backend consisting of database management systems (DBMS) comprising application logic, and (2) the frontend, which usually serves as the primary interface for user interaction (graphical user interface). Other solutions have been proposed in so-called tier approaches, where the number of different entities is either reduced for simple applications (all-in-one approaches) or drastically increased for complex applications (n-tier approaches) (Petersen 2001). For the purpose of this chapter, we will focus and discuss the two-tiered approach, which is often also referenced as the client–server architecture.

In information systems, the backend is often synonymous with the database, which the user accesses only indirectly (note that the discussed information systems of the IPK have a more direct access solution integrated). Primary data and metadata are stored and managed here. The DBMS is the software layer of the backend, and one of its tasks is to handle authorization and authentication and thus controls the granularity of data retrieval for specific user groups. For user updates or changes to records in the database, the DBMS is able to enforce constraints that ensure consistency rules are followed. Databases implement different data model and feature paradigms, and have evolved to support application scenario, with relational databases being the dominant class overall (Harrington 2016). The data is accessed indirectly either via application programming interfaces (API) or via special application logic through stored procedures, the specifics of indirect accesses are varied and going into detail here would go far beyond an overview of information systems.

In addition to information systems per se, so-called web-based information systems are playing an increasingly important role. In such systems, the front end is based on web technology. This means that the user interface is a web browser or is accessible via the WWW. Usually, the business logic of such a web application is implemented in a programming language suitable for the use case and deployed by assigning URLs to specific functions or methods. The programming language itself may implement or provide the required HTTP server, or a separate HTTP server such as Apache HTTP Server, Apache Tomcat, or Oracle WebLogic may be used upstream. Classically, business logic often communicates with the storage layer over an internal private network using protocols based on TCP/IP, sockets, or the file system. In complex web information systems, different persistence technologies are often used simultaneously in the storage layer, e.g., in web-based information systems that combine multiple databases in a single web application (Fig. 2.2). The data ingestion and management into the backend of scientific institutions is often realised by a laboratory information management system (LIMS). Its main purpose is to act as a sample management system, but recently data analysis functions and the ability to record digital laboratory documentation (also known as an electronic laboratory notebook or ELN) have also been integrated into some successful LIMS

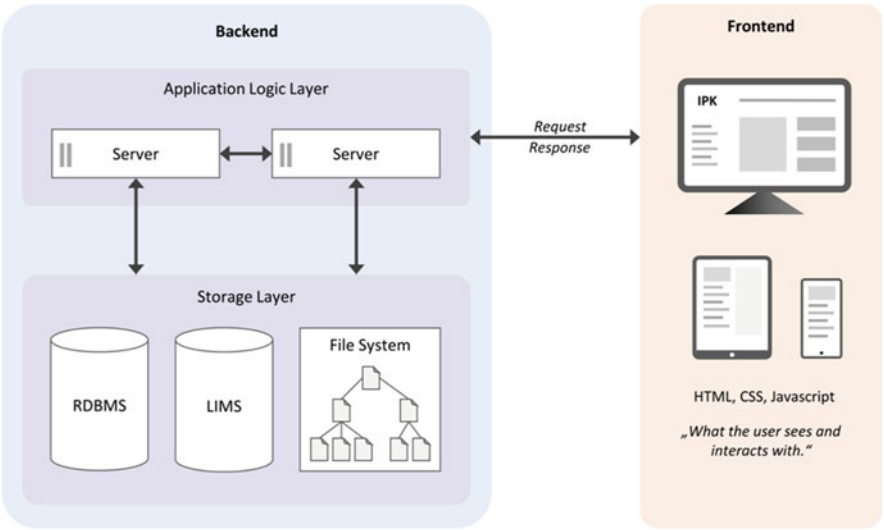

Fig. 2.2 Abstract architecture of web-based information systems

solutions. The actual LIMS implementation of the IPK Gatersleben is described in (Ghaffar et al. 2019).

A further component of backend technology is sensorics data which is mainly used to collect environmental data. The basic idea of sensor networks is based on the idea of the Internet of Things (IoT) (Madakam et al. 2015). Individual sensors are to be networked in a computational interconnected infrastructure. This concept is based on ubiquitous computing. This in turn describes a concept that moves away from the use of one end device to the use of many. This concept thus contrasts with dedicated, application-specific platforms that are designed and installed to combine data collection, storage, exchange and evaluation in one overall system. IoT goes far beyond the original concept of the internet. It is no longer just a network of different computers, but a network of all kinds of devices. Rather, the IoT is a network consisting of different objects that communicate over the Internet to collect and exchange data. This includes both actuators, as a component of a machine that is responsible for moving and controlling a mechanism, and sensors, which detect events or changes in its environment. Some examples of sensors are cameras, weather stations, ground sensors or airborne remote sensing, such as drones or satellites. Active elements are irrigation pumps, fans, lighting or even cooling or heating elements. Usually, both types are combined, like in agricultural machinery and greenhouse controls. These capabilities to build IoT networks are increasingly influencing the nature of experimentation. For example, the detection of phenotypes via sensors is being combined with targeted manipulation of the environment in the field of high-throughput plant phenotyping and breeding research (Fiorani and Schurr 2013; Watt et al. 2020). The concrete interaction of sensors and actors is a practical and technical challenge in terms of system integration that is not to be

underestimated. This is a practical hurdle, because infrastructures that span locations and organizations sometimes use highly heterogeneous interfaces and incompatible systems infrastructures. The homogenization of data formats is done by applying standards as described in Sect. 2.2. The homogenization of transmission protocols plays another central role here, as sensor data are continuous data streams. This affects, among other things, the protocols as well as data exchange formats and units. Application-specific network protocols are the backbone of IoT networks and are responsible for the communication of remote sensors. One of the commonly used network protocols is MQTT (Message Queuing Telemetry Transport). It is a lightweight protocol used to transport data between devices mainly on TCP/IP networks. It was jointly authored by Andy Stanford-Clark (IBM) and Arlen Nipper (Cirrus Link, then Eurotech) in 1999 (MQTT.org 2015). MQTT is an M2M (Machine to Machine) protocol best suited for the remote connections which require a "small code footprint" or in cases where the network bandwidth is limited, such as IoT devices. The publish-subscribe architecture of MQTT described in Obermaier (2018)) and illustrated in Fig. 2.3 is extremely lightweight compared to HTTP's request/response paradigm.

Where MQTT broker is the central component of the paradigm that acts as a server responsible for passing the messages between the publisher and subscribers. In case of an event, the publisher first transmits the data to a broker with a topic, and if a client requests data of a certain topic, the broker performs matching and then delivers messages accordingly.

Another important layer in information systems is the frontend. It is considered to be everything the user sees and interacts with directly. Especially in web-based

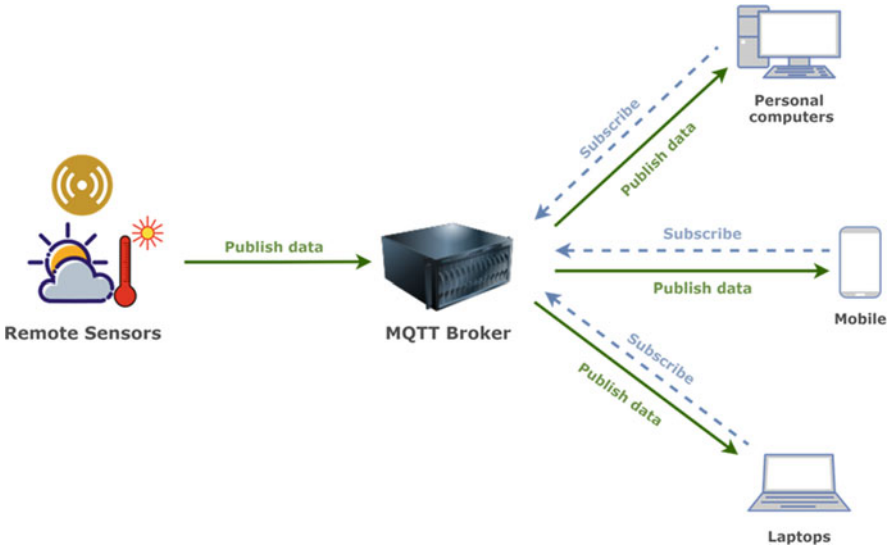

Fig. 2.3 Principle operation of the MQTT publish-subscribe architecture

information systems, the website rendered in the browser acts as the graphical user interface. The spectrum here ranges from the pure display of data, stored in the storage layer, to highly interactive "Rich Internet Applications" (RIA) (Fraternali et al. 2010), which very often also contain a large proportion of business logic in the form of Javascript. For example, the display of interactive diagrams always requires the use of business logic in the frontend code. In contrast, for the display of text, tables and static images, only the use of HTML and CSS is mandatory. In the age of mobile devices such as smartphones and tablets, the flexibility of the website layout plays an increasingly important role. The necessary flexibility results from the many different display sizes and page formats of mobile devices compared to traditional PC monitors. Therefore, the development is increasingly moving away from static, fixed layouts to so-called adaptive and responsive layouts that adjust as optimally as possible to the different display sizes and page formats.

2.4 Selected Data Management Information Systems

The following provides a general overview of some well-known information systems and data warehouses with a focus on plants developed and hosted at our research institute IPK Gatersleben. The description of each system includes the features and architecture, scope and general use cases. The web address where the system can be accessed is stated, as well as the supported data domains. Also explained is how data can be imported and exported and where the system fits into the Data Life Cycle.

2.4.1 The Genebank Information System (GBIS)

Globally, genebanks play an important role in the long-term conservation of plant genetic resources (Hoisington et al. 1999). They complement the conservation of biodiversity in farmers' fields and in nature. Besides the preservation of physical samples, data management is one of the most important tasks of a genebank and at the same time one of its greatest challenges (FAO 1997, 2010; Fowler and Hodgkin 2004; Weise et al. 2020). Well-structured documentation of all data and information available on a genebank accession is the basic prerequisite for genebanks to be used. A wide range of data must be taken into account.

The IPK genebank has been in existence for almost 80 years, but is partly based on even older collections, so that material from a period of almost 100 years is preserved. The focus of the documentation has continuously developed over this period, as have the technologies used for this purpose. Furthermore, a number of changes in organizational structures have taken place, and several generations of curators and scientists have maintained the material and constantly added further

parts to the collection. Continuous documentation is indispensable for both the preservation and the exploitation of the material.

The Genebank Information System (GBIS)[1] (Oppermann et al. 2015) is one of the central instruments for documentation and management in the IPK genebank. It was first introduced in 2006 and has been continuously developed ever since. The core of the GBIS is formed by an OnLine Transactional Processing (OLTP) system, which records the data produced in various genebank workflows. This data is compiled into an overall dataset that includes the following areas:

- Pure management data for conservation of collections.
  - Storage quantity and locations.
  - Growth and harvest management.
  - Germination rate, age of the samples, health tests.
  - Reporting and labelling.
- Data of legal significance.
  - Collection permits.
  - Correspondence, documentation of receipt.
- Data to assess the value of the resource.
  - Basic (passport) data.
  - Phenotypic observations.
  - Images of specimens, plants, fruits and seeds.
  - Comprehensive genetic data.

GBIS consists of two areas: (1) a public information and ordering system and (2) an internal system that serves data management and process support. An Oracle DBMS is used for data management; the various application components are implemented both as application server-based web applications and as standalone solutions. Figure 2.4 shows the architecture of the overall system.

From the original idea of documenting, cataloguing and describing plant genetic resources, genebank information systems are increasingly developing into instruments for scientific work and thus reflect the transformation of genebanks into bio-digital resource centres.

GBIS supports all steps of the data life cycle.

[1] https://gbis.ipk-gatersleben.de/

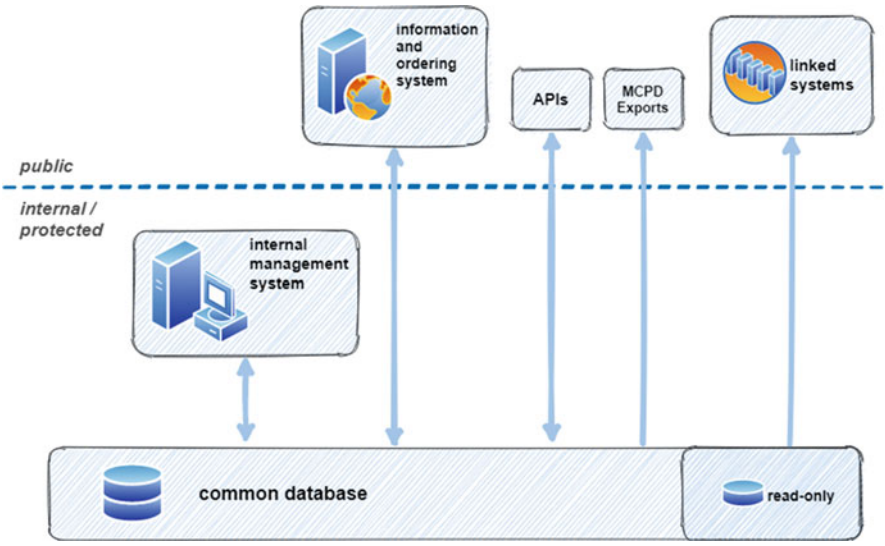

Fig. 2.4 Architecture of the genebank information system

2.4.2 The European Search Catalogue for Plant Genetic Resources (EURISCO)

Estimates put the number of genebanks worldwide at around 1800, with more than 600 in Europe (Engels and Maggioni 2012). Many genebanks have been in existence, in some cases for decades. Despite the introduction of IT support, especially in the late 1960s and 1970s, most genebanks remained largely isolated from each other. This did not change until the 1980s, when the first attempts were made to make information available across genebanks. It was then that the idea of Central Crop Databases (CCDB, Gass et al. 1997) was born. This idea consisted of strengthening cooperation between genebanks by networking the collections and also making genebank material more accessible to users as well as identifying possible duplicates between the individual collections. However, due to the low quality or lack of data, these goals could only be achieved to a limited extent (van Hintum 1997). One of the biggest difficulties in this context was that for a long time there were no uniform standards for the description and exchange of passport data. A standard that addressed this challenge is the Multi-Crop Passport Descriptors (MCPD). After the presentation of the first draft in 1997 (Hazekamp et al. 1997), the MCPD successively developed into a globally accepted and used standard (Alercia et al., Alercia et al. 2001, 2015). The emergence of MCPD as well as Darwin Core (Endresen and Knüpffer 2012; Wieczorek et al. 2012) represented milestones for

the development of international aggregator systems such as WIEWS,[2] EURISCO[3] or Genesys.[4] They enable the exchange of passport data between genebanks and these systems and thus allow a cross-genebank search for accessions of plant genetic resources.

One of the aggregator platforms mentioned is the European Search Catalogue for Plant Genetic Resources (EURISCO, Weise et al. 2017). This platform is operated within the framework of the European Cooperative Programme for Plant Genetic Resources (ECPGR)[5] and has been available online since 2003. The aim of EURISCO is to provide a central entry point for searching accession-specific passport data and phenotypic data on plant genetic resources accessions maintained in Europe. In addition, EURISCO assists its member countries in fulfilling national obligations, e.g. to the FAO. The majority of European ex situ collections are represented in EURISCO. A total of 43 countries are currently part of the EURISCO network. Each country compiles the data of its genebanks in a National Inventory and submits it to EURISCO on a regular basis. The MCPD standard is used for the passport data. Currently, more than two million genebank accessions from about 400 collections are documented in this way in EURISCO, covering more than 6700 genera and 45,000 species. In recent years, work has begun on depositing phenotypic observations collected on accessions in EURISCO in addition to the passport data. Unfortunately, there are no really widely accepted standards for the exchange of phenotypic data so far (Krajewski et al. 2015). This is complicated by the fact that observation values of genebank accessions were partly collected over long periods of time. Various initiatives to harmonize such data have existed since the 1970s, e.g. the IPGRI/Bioversity descriptor lists (IBPGR 1990; International Board for Plant Genetic Resources (IBPGR) and Commission of the European Communities (CEC) 1984; IPGRI et al. 2001), but they have never achieved general acceptance. More recent approaches aim at mapping different traits and methods onto each other using ontology terms, e.g. CropOntology (Shrestha et al. 2010, 2012), or to put a stronger focus on the description of the material used and the experiments conducted, e.g. MIAPPE (Ćwiek-Kupczyńska et al. 2016; Krajewski et al. 2015; Papoutsoglou et al. 2020). Altogether, this represents a particular challenge that has not yet been conclusively solved. EURISCO uses a minimum consensus approach for exchanging phenotypic data, which is limited to the absolutely necessary data fields.

The provision of data in EURISCO is done using a multi-tier system (Fig. 2.5). The data compiled in the National Inventories is imported into a central staging area through an upload tool. A series of data integrity checks are then performed, most of them at syntactic level, some also at semantic level. Automatically generated error reports help the data providers to successively correct data errors. After release by

[2] http://www.fao.org/wiews/

[3] http://eurisco.ecpgr.org/

[4] https://www.genesys-pgr.org/

[5] https://www.ecpgr.cgiar.org/

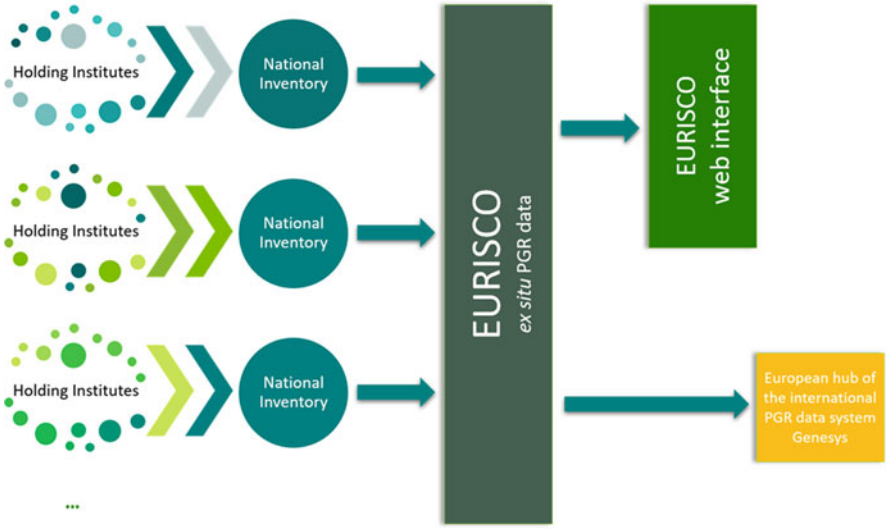

Fig. 2.5 Overview of the EURISCO architecture

the data providers, the new data is integrated into the overall EURISCO database. A web interface is available to the users of the system, which offers a variety of search, visualization and download options. Fuzzy searches are also supported here, for example in the case of scientific plant names and their synonyms (Kreide et al. 2019).

In addition, EURISCO forms the European hub of the international PGR information system Genesys. Passport data is regularly exchanged with Genesys, so that genebank accessions documented in EURISCO can also be found via the Genesys portal.

In terms of the Data Life Cycle, EURISCO can be assigned to the categories Preserve, Share and Reuse.

2.4.3 BARLEX

Sequencing and subsequent steps to reassemble the underlying genome sequence for complex plant species have been a lengthy and costly endeavour. Sequencing efforts in the species barley (*Hordeum vulgare* L.) were initiated more than a decade ago (Schulte et al. 2009). At that time, the state-of-the-art approach was to create a comprehensive physical map of overlapping BAC clones carrying small fragments of genome information (Ariyadasa et al. 2014; Schulte et al. 2011), sequence them using NGS technology (Steuernagel et al. 2009; Taudien et al. 2011), and then join the assemblies of the individual BAC clones with mate-pair reads (Beier et al. 2016).

The Barley Explorer (or BARLEX[6] for short) web application was developed to facilitate the process of joining these BAC assemblies (Colmsee et al. 2015). It showed interested users evidence of overlap between adjacent BAC assemblies and all available genomic data associated with each sequence contig. This information was presented in both tabular form and in an interactive graphical edge-node display.

Since its inception in 2015, BARLEX has evolved into the de facto hub for barley genomic sequence information (Beier et al. 2017). With the advent and adaptation of advanced sequencing and assembly techniques such as conformation capture sequencing (Lieberman-Aiden et al. 2009), incorporation of optical mapping (Staňková et al. 2016) or 10X Genomics linked reads (Mostovoy et al. 2016), the speed and accuracy of new complete pseudomolecule sequence assemblies have increased dramatically (Jiao and Schneeberger 2017). To date, new and updated reference barley genome assemblies have been released in 2012 (Mayer et al. 2012), 2017 (Mascher et al. 2017), 2019 (Monat et al. 2019), and 2021 (Mascher et al. 2021), with more than a dozen genotypes being sequenced at the moment to complement pan-genome sequencing efforts (Jayakodi et al. 2020, 2021). The pseudomolecule sequence, genomic scaffold structure, and molecular marker, repeat, and gene annotation (complemented by expression data) for these four different versions of the reference sequence are all available in BARLEX.

BARLEX is built on an Oracle relational database backend and consists of 57 tables, 17 materialized views, 37 stored procedures and more than 95 million rows of data. The web application is implemented with Oracle Application Express (APEX, formerly known as Oracle HTML DB) with custom Javascript procedures. Some of these Javascript procedures use the Cytoscape.js framework (Franz et al. 2016) which enables a graph-based interactive visualization. Additional functionality is supported by cytoscape-qtip, cytoscape-automove, cytoscape-cose-bilkent and cytoscape-context-menus, which help to make the user interface more intuitive and accessible. Tabular data within BARLEX can be exported in various predefined formats such as CSV and HTML. Please note that the download of gene or repeat annotations has been disabled in BARLEX and is distributed via links to long-term stable DOIs deposited at eDAL-PGP (Arend et al. 2016). The import of new data into BARLEX is done via semi-automatic import scripts by the BARLEX team. After manual curation of the data and transformation into the appropriate format, the data is fed into the database using an upload tool (Rutkowski 2005). This manual curation step includes both syntactic and semantic verification. Although ordinary users cannot modify the data via the web application, there is an option to leave feedback on all features and records so that administrators can be notified on feature requests and data inconsistency.

Many types of genomic data are represented in BARLEX, such as sequencing contigs (various technologies and methods), exome capture data, molecular marker data (array-based SNPs), expression data (from Iso-Seq and RNA-Seq), BLAST results, structural information about sequence composition, and sequence order

[6] https://barlex.barleysequence.org

and orientation in the finished pseudomolecules. With these data domains and the functions BARLEX supports, BARLEX covers the Analyse, Preserve, Share and Reuse fields in the Data Life Cycle.

2.4.4 BRIDGE

Although a wide diversity of landraces and PGRs are stored in genebanks, there has been little success in utilizing them and incorporating them into breeding programmes. One of the challenges here is the availability of information on molecular and phenotypic profiles of the entire seed stock. Apart from the fact that transferring beneficial alleles from PGRs to modern elite varieties is a challenge in itself (Wang et al. 2017), this availability is a prerequisite for incorporating PGRs into commercial plant breeding. Therefore, genebanks have begun to systematically categorize and catalogue their germplasm collections at both the molecular and phenotypic levels (Mascher et al. 2019; Romay et al. 2013). An example of one of these pioneering projects was carried out on the crop barley, where 22,626 accessions of the genebank hosted at the IPK Gatersleben were surveyed and analysed based on genotyping-by-sequencing (Milner et al. 2019). The resulting molecular profiles could now be combined for the first time with passport data, historical and newly collected phenotypic data to draw conclusions about the global barley diversity and to find interesting genes and loci for plant breeding. This information resource was adapted into the web portal BRIDGE[7] (König et al. 2020).

BRIDGE is both a data warehouse and exploratory data analysis tool for large-scale barley genomics. Through a unified collection manager for user-defined germplasm datasets, various analyses can be performed or visualized. One of the core features is the quick selection of collections either using the lasso selection tool on the provided graphical output or by setting different filters over the complete set based on passport data, phenotypic traits or molecular markers (SNPs). In addition, BRIDGE uses a concept known as "interactive brushing and linking", where changing parameters in one visualisation results in a direct response in other visualisations that are dynamically linked (Keim 2002). All of this supports the interactive user experience and enables rapid analyses of more than 9000 data points of phenotype data, visualisation of genetic diversity by PCA and t-SNE, or exploration of Manhattan plots to genome-wide association studies. Also integrated is an intuitive variant browser for the study of SNP data based on the GBS sequence data of 22,626 genotypes. Genotypic data can be exported in VCF files (Danecek et al. 2011) for custom collections of genotypes and specific genomic regions of interest, e.g. for a whole gene or single exons. The Java library "isa4j" (Psaroudakis et al. 2020) is used to realise a customised export of phenotypic data in the ISA-Tab format (Sansone et al. 2008, 2012). Based on the user's custom collections

[7] https://bridge.ipk-gatersleben.de

of genotypes, a ZIP-archive containing ISA-Tab formatted text files and additional phenotypic images is generated on the fly for download via the user's web browser.

The BRIDGE web portal utilises the client–server model as general architecture with REST-alike HTTP-APIs as the communication layer between client and server. HTML5, CSS3 and Javascript are used for client-side development. Groovy, Java and Python programming languages are used to implement the server-side counterparts of the HTTP-API. The web application framework "Grails" is used to implement all aspects except the server-side API of the integrated SNP browser. The server-side part of the SNP browser is implemented in Python using the Flask library. Well-established libraries like Numpy (Harris et al. 2020), Pandas (McKinney 2010), Zarr (Miles et al. 2020) and Scikit-learn (Pedregosa et al. 2011) are used for the performant handling of large SNP data matrices and scientific computing aspects like calculation of minor allele frequencies or principle component analysis.

All passport and phenotypic data are provided via the IPK-LIMS through project-specific logical relations to GBIS. Analysis result data like the outcome of GWAS or PCA is stored in standard tables in the Oracle RDBMS (Fig. 2.6). The import of analysis result data is performed via customised import scripts by using CSV files. Data of genomic diversity is imported by the conversion of VCF files to Zarr archives that are then used by the server-side part of the SNP browser. The VCF files can be optionally annotated by SnpEff (Cingolani et al. 2012). The import of

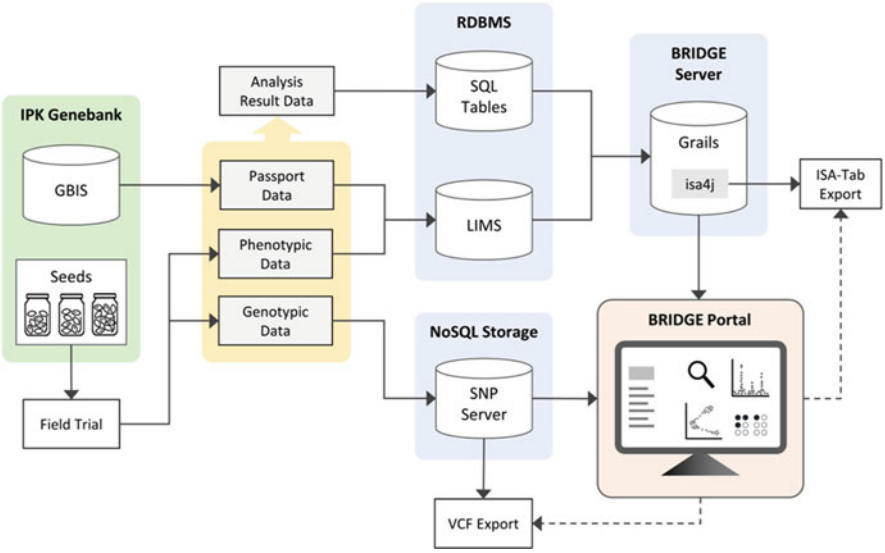

Fig. 2.6 Overview about the general architecture and data flow in BRIDGE

gene annotations is performed by directly using GFF3[8] files. Data that gets imported into the system is checked automatically for syntax errors. The responsibility for the general plausibility of the data belongs to the data provider who wants to present his project data in the portal. The process of data import is managed by the administrator of the web portal. As BRIDGE was designed to present the results of specific genebank genomics projects, it is currently not possible, nor is it intended, for end users to import and view their own data in the portal.

Regarding the Data Life Cycle, BRIDGE can be assigned to the categories Collect, Process, Analyse, Preserve, Share and Reuse.

2.4.5 e!DAL-PGP

The FAIR data principles are widely accepted by the scientific community for supporting long-term stable research data handling. Although established infrastructures such as the ELIXIR Core Data Resources and Deposition Databases provide comprehensive and stable services and platforms, a large quantity of research data is still inaccessible or at risk of getting lost. Currently several high-throughput technologies, like plant genomics and phenomics are producing research data in abundance, the storage of which is not covered by established databases.

The eDAL-PGP[9] (Plant Genomics and Phenomics) research data repository is a comprehensive infrastructure providing diverse datasets of plant-related research data. It has no general data type or data volume limitations, and therefore, it provides genomic sequences, phenotypic images, metabolite profiles and also research software and scripts. It started in productive mode in 2016 (Arend et al. 2016) and based on the previously developed JAVA-based eDAL infrastructure[10] (Arend et al. 2014), which follows an "infrastructure to data" (I2D) approach to provide an on-premise data management and publication system. This approach can in comparison to the common data publication-as-service model also feature a FAIR data publication culture, but it differs in costs and effort for establishment and maintaining (see Fig. 2.7).

The data publication-as-a-service model usually costs a fee, needs data property control and provides storage capacity limits. In contrast, the data publication premises model keeps data in-house and can use internal server and storage hardware by installation of the e!DAL software. The fully embedded data submission and review process allows to easily store and publish research data by using persistent DOIs. To make the data FAIRly available, e!DAL supports several relevant features (Arend et al. 2020).

[8] https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

[9] https://edal-pgp.ipk-gatersleben.de/

[10] https://edal.ipk-gatersleben.de/

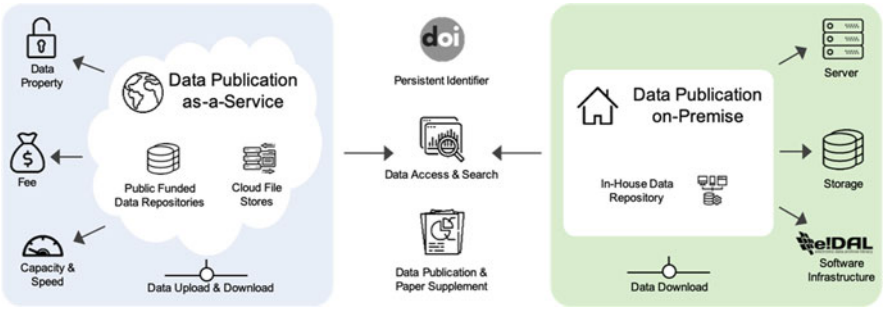

Fig. 2.7 Different data publication approaches

Providing machine-readable metadata, which are based on the DublinCore standard and are automatically embedded into the provided content pages of every published dataset, e!DAL guarantees that the contained research data is easily **FINDABLE** by common search engines. By using the well-established DOIs, the findability is further increased due to the diverse network and interactions of the DataCite services like ORCID or CrossRef. Furthermore, DOIs are persistent and guarantee the long-term stable **ACCESSABILITY** of published datasets. The DOI resolver provides simple access to the referenced datasets, e.g., in a research or data publication, even if the physical location of the underlying data changes over the time. On top of this, the e!DAL's web server takes care that the datasets are accessible via comprehensive content pages, which allow users to navigate through the dataset and download certain files or metadata. The content pages not only provide the metadata directly on the page, but also embed the metadata in the sources to provide the **INTEROPERABILITY** of the datasets. By using standardised schema (Guha et al. 2016) and format (Lanthaler and Gütl 2012) the information about relationship between datasets can be aggregated. The DublinCore is well-established and therefore e!DAL guarantees the long-term stable **REUSABILITY** of the datasets by collecting a minimal set of technical metadata, which are crucial to open and read the data files. The support of different licences makes it easy to clearly define by whom and how the datasets can be used.

The success of I2D Approach is shown by the constantly increasing number of datasets, accesses and downloads of the e!DAL-PGP instance. The comprehensive functionality of e!DAL as well as the simple installation and configuration, e.g. by using powerful and user-friendly infrastructures such as the ELIXIR AAI, are the reasons that in the meanwhile further institutional instances based on e!DAL were planned or already established. Nevertheless, even if scientists are getting more opportunities to exchange their research data within the community, the incentive is still quite low (Cousijn et al. 2019). The procedure of data publication and citation is in contrast to the established peer-review process for research articles not very common (Tenopir et al. 2015), which has of course cultural reason, but also technical limitations (Parsons et al. 2019). Beside the commercial Data Citation Index, also some open, community-driven initiatives like Make Data Count

(Cousijn et al. 2019) were developed to overcome these limitations and improve the incentives for researchers. Additionally, more and more publishers demand authors reference their research data as citations in the common reference list of their articles (SciData Editorial 2019). All these developments will help to increase the acceptance of research data as an important scientific asset and to establish a FAIR research data publication culture.

2.4.6 The IPK Weather Database: Collection and Provision of Meteorological Data

We encounter weather data every day and they often seem trivial. However, they are essential for interpreting the results of field trials, as the expression of traits can be weather-dependent (Philipp et al. 2018). The measurement of meteorological data represents a special type of data collection, as the data is continuously recorded over a very long period of time. As a result, the processes of the life cycle from the collect to the reuse of the data take place in parallel. Another special aspect is the change in data collection and processing methods.

The long tradition in recording meteorological data is accompanied by some changes in measurement intervals, sensor technology and data archiving. Manual recording of the values of analogue sensors on paper at fixed hours of the day is now replaced by continuous recording of electronic data in databases. This results in special requirements for statistical evaluation and error analysis (World Meteorological Organization (WMO) 2017).

Nationally and globally defined standards exist for the design of the measuring station and the data to be recorded, which in particular ensure the comparability of the measured values (Löffler 2012; World Meteorological Organization (WMO) 2018).

Meteorological observations have been recorded at the IPK since 1953. It is not difficult to conclude that these data are not primarily recorded digitally. The measurement results have been stored in databases only since 1993. For the period before 1993, at least the monthly values were subsequently captured and incorporated into the database. The result is that evaluations since 1993 are possible with a resolution accurate to the day, but analyses of the long-term measurement are only possible with a lower resolution (Fig. 2.8).

Today, data collection is done through an acquisition pipeline that stores, processes and aggregates the data collected by the data loggers to display and provide it to users in an appropriate way.[11] For this purpose, a series of plausibility checks are carried out on the raw data, and the time-based aggregations are calculated and saved from this cleansed dataset (Fig. 2.9).

[11] https://wetter.ipk-gatersleben.de/

	monthly	daily (7:00,12:00;21:00)	daily (every hour)	hourly
1953 – 03/1993	✓	✓ on paper	✗	✗
04/1993 – 12/1999	✓	✓	✗	✗
01/2000 – present	✓	✓	✓	✓

Fig. 2.8 Resolution of the measurement for the IPK weather station

Fig. 2.9 Acquisition pipeline

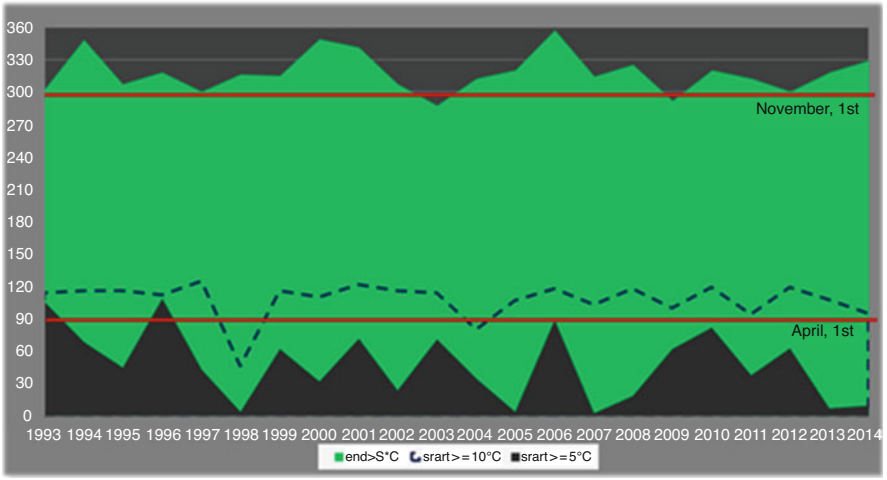

Fig. 2.10 Development of the growing season in Gatersleben 1993–2014

The data provided are not only used in the context of scientific experiments at IPK, but also serve as a basis for decisions on the conservation of biodiversity in the genebank.

Thus, the analysis of the meteorological data itself also offers insights into the climate development at the Gatersleben site, such as the development of the growing season (Fig. 2.10).

The above example shows the fluctuations in the start and duration of the growing season in the period from 1993 to 2014 as calculated. The recognizable variations from the threshold value of 1 April or 1 November influence the time at which traits are expressed. It also becomes clear that the calculation method influences

the result (start ≥5 °C or start ≥10 °C as dashed line). But it shows also a specific characteristic of meteorological data: The dataset is not finalized, but represents a daily extended data series on various meteorological parameters. This also means that conclusions drawn at an earlier point in time may have to be supplemented or revised for new studies.

With regard to the data life cycle, the IPK Weather Database covers the categories Collect, Process, Analyse, Preserve and Share.

2.4.7 Plant Phenotyping Portal

In addition to the IPK Weather Database, environmental data from high-troughput phenotyping facilitie can be collected too. This is done in greenhouses or growth chambers. Two important plant growth facilities are the Plant Cultivation Hall and LemnaTec greenhouses (Altmann 2020). Here, the environment can be controlled to various degrees. For data acquisition, the MQTT protocol is used and plays a crucial role in the communication within the interconnected sensoric infrastructure at IPK. The Plant Cultivation Hall and LemnaTec greenhouses have 498 and 130 soil and clima sensors respectively, which generate data every 5–10 min. Additionally, 13 environmental sensors are transferable from one facility to another. The environmental data is essential for contextual and statistical analysis, aiding in the improvements in the agricultural use cases when shared in standardised formats. The idea is to store the raw sensory data in an interoperable and reusable way (Memon 2020). Therefore, using Node-RED, a flow-based programming tool, the MQTT protocol is implemented to communicate the data between the vendor-specific sources and the database. The MQTT protocol transmits the data as messages. Hence, the data is enveloped in messages (Fig. 2.11) and published to the broker through the assigned MQTT topic. For a permanent recording the data, an authorised client subscribes to the topics that contain the relevant data and stores them in the database.

The topics are designed to contain the metadata related to the sensor data. For example, in Fig. 2.11, IPK_G.1300 is determined as the building where the sensor is located, followed by the room number, container, type of sensor (such as temperature, humidity or moisture), sensor's node id, and the sensor port, since a single sensor node may have multiple ports. Whereas the message body includes the captured sensor data.

```
Topic: IPK_G.1300/0.001/CONTAINER/TEMPERATURE/node_name15/4
Message: {
    "Value":"19.2",
    "Unit":"°C",
    "Timestamp":"2020-01-20T17:15:09.0"
}
```

Fig. 2.11 An example of MQTT published message (Memon 2020)

In order to permanently record sensoric data and make them accessible for downstream data analysis, they must be stored in databases (Stöbe 2019). Because sensor data is streaming data, i.e. continuously delivered, its archive can only be conducted by aggregation over windows as a discrete snapshot. Usually, the resolution of such windows differs from seconds to hours. This depends on the expected fluctuation rate of measured values. In respect to weather data, i.e. wind, temperature, humidity and solar radiation, aggregation over 5 min is common. By doing so, subscribed environmental data is averaged over 10 min and stored in a relational database. Its metadata, like sensor placements or locations, is maintained in LIMS. On top of the database backend, the web application "Plant Phenotyping Portal" was developed. It integrates the aggregated sensor data, the metadata, and the experimental setup with the goal of breaking down individual plants, the installed sensor, and related environmental measures over time. Figure 2.12 shows the interface of the application.

The application's interface allows users to download and view the sensory data between any two given time points for any available sensor(s). Furthermore, the application supports visualising the sensory data. For example, in Fig. 2.13, the chart above displays the temperature of different sensors, showing the sensor's operating status and the chart below presents the light intensity between specific durations.

The application uses the Oracle Application Express (APEX) framework for these visualizations and covers the Collect, Preserve, Share and Reuse parts of the Data Life Cycle.

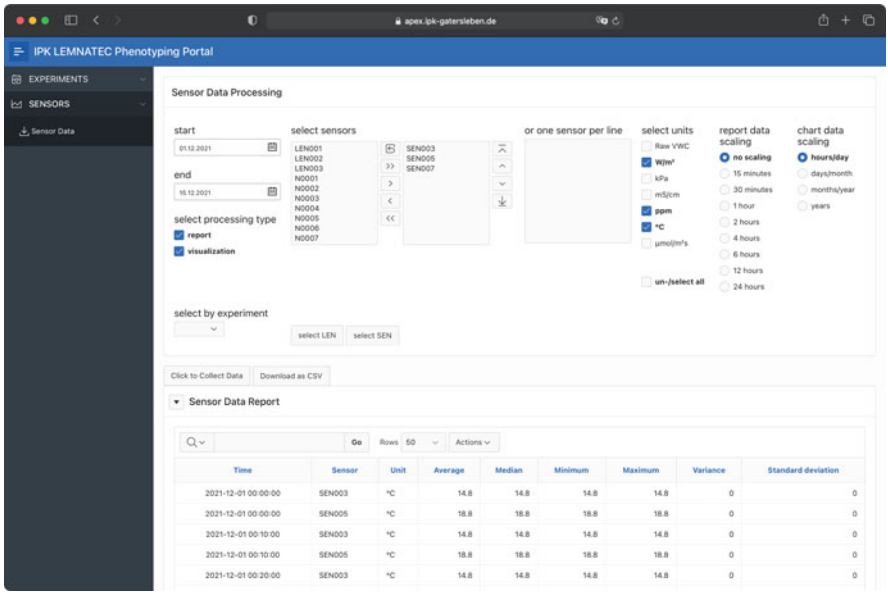

Fig. 2.12 User interface of the Plant Phenotyping Portal

Fig. 2.13 Visualisation of archived sensor data in a Web Information System

Table 2.1 Systems and Data Domains

	Plant genetic resources data	Genomic data	Genotyping data	Phenotyping data	Environmental data
GBIS	✔			✔	
EURISCO	✔			✔	
BARLEX		✔			
BRIDGE	✔		✔	✔	
e!DAL-PGP	✔	✔	✔	✔	
IPK Weather DB					✔
Plant Phenotyping Portal				✔	✔

2.5 Summary and Outlook

Data management and the applications described here are diverse and yet serve the purpose of preparing data under consideration of the FAIR principles and offering it to its users. The requirements and the functions of the individual system are of course closely coupled with the data domains covered. Table 2.1 gives an overview of the different combinations of data domains in the information systems that have been worked on at IPK over the last 20 years. Unsurprisingly, most of the systems presented are focused on plant genetic resources and phenotypic data, but more recently genomic, genotypic, and environmental data have increasingly been added.

Table 2.2 Systems and categories in Data Life Cycle

	Plan	Collect	Process	Analyse	Preserve	Share	Reuse
GBIS	✔	✔	✔	✔	✔	✔	✔
EURISCO					✔	✔	✔
BARLEX				✔	✔	✔	✔
BRIDGE		✔	✔	✔	✔	✔	✔
e!DAL-PGP					✔	✔	✔
IPK Weather DB		✔	✔	✔	✔	✔	
Plant Phenotyping Portal		✔			✔	✔	✔

Accordingly, the classification of the systems examined into the individual phases of the Data Life Cycle also varies. While all systems support the later *Preserve and Share* steps, the *Plan* phase is underrepresented (Table 2.2). This can be explained by the fact that most systems were not designed to collect new data and start the data collection process, but rather to document and present data in an appealing form and manner and to exchange it with the community.

In summary, the presentation of the information systems has shown for which data domains data management solutions have been developed at IPK in Gatersleben. These were developed in general independently of each other and have thus grown historically. However, it can be stated that all steps of the data life cycle are served by the systems.

Generally, the applications described do not stand alone, but are designed via various interfaces in such a way that interaction between information systems is possible. One such example is depicted here in Fig. 2.14, where the interconnections between the IPK Genebank, GBIS, BRIDGE, BARLEX and eDAL-PGP are illustrated. The IPK Genebank and its GBIS serve as a primary data and material resource for genebank genomics experiments and field trials. The phenotypic and genotypic data derived from experimental field trials is then fed into visualization and analysis web tools like BRIDGE and BARLEX, while phenotypic observations of regular genebank multiplication trials are directly integrated into the GBIS. The genotypic data in the form of SNP-matrices (VCF files) is also deposited in eDAL-PGP for FAIR-compliant long-term storage. DOI-based hyperlinks from the SNP-browser in BRIDGE to the corresponding datasets in eDAL-PGP allow the users to download the original VCF files to their personal computers or HPC-servers for their own analysis. Hyperlinks from the visualised gene features in the BRIDGE SNP-browser to BARLEX allow the users to retrieve further information about the barley genome and genes.

Important challenges for the future are, on the one hand, the consistent semantic interlinking of the various information systems specialised in their use cases via unique identifiers and, on the other hand, the creation of central entry points for data research and data analysis. Currently we are using the IPK LIMS system as a central repository to implement a unique management of identifiers of biological objects. To increase efficiency, it is also important to develop reusable generic

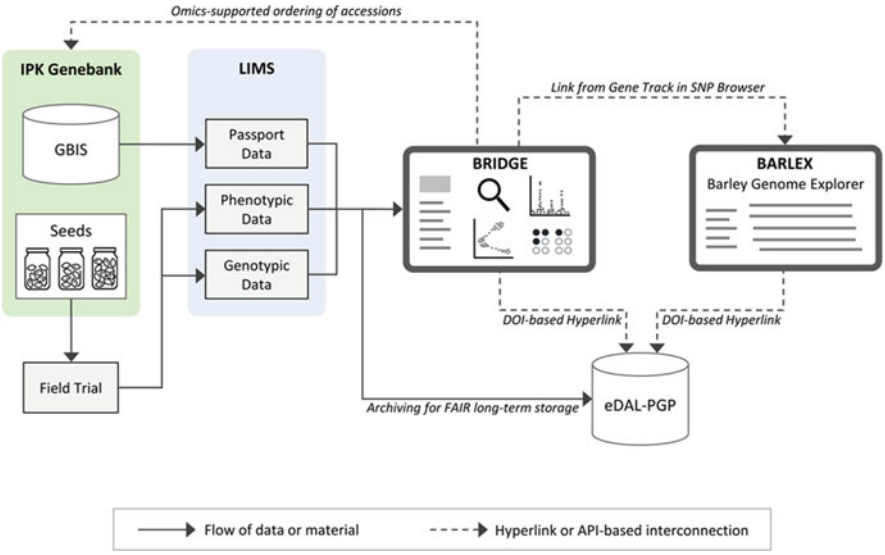

Fig. 2.14 Visualization of interconnections between IPK Genebank and its GBIS, BRIDGE, BARLEX and eDAL-PGP in the frame of genebank genomics experiments

software components for recurring tasks of interactive research data presentation and visualisation.

We have presented the approach of a research institution. It is obvious that an institute like the IPK Gatersleben does not exist autonomously. There are connections to cooperation partners all over the world. In order for the entire scientific community to be able to use the data, this data must be offered accordingly and thus be reusable. The foundation for knowledge discovery and innovation is good data management, because it allows data to be reused and new connections to other data to be formed by the community. One challenge is to make datasets not only understandable to humans but also readable and actionable by machines (Mons 2019). Open (non-binary) formats and richly annotated metadata are a prerequisite for this. However, in many areas of the life sciences, one or both of these requirements are not met, hindering both knowledge discovery and progress in general. The FAIR data principles (Wilkinson et al. 2016) are a start to making such a vision of the future a reality. To properly understand the FAIR principles, it is important to distinguish between two cases: First, FAIRification of existing data and FAIR-by-design, data created with FAIR principles in mind (Jacobsen et al. 2020). FAIRification of existing data is arguably the more challenging task to accomplish, as it requires updating data and metadata.

An example of a project focused on FAIRification of data is the ELIXIR (Crosswell and Thornton 2012) implementation study FONDUE. The task is to link available plant genotyping and phenotyping data using stable identifiers and to document those links in the repository metadata thus enabling search, retrieval

and reuse of such linked data. In this study, the main focus is on the so-called ELIXIR Core Data Resources (Drysdale et al. 2020), which are widely used in the life sciences and include such well-known repositories like the European Nucleotide Archive (Leinonen et al. 2011). The idea is to trigger a shift in thinking among data producers through this top-down approach by changing policies at key (genomic) data entry points. One obstacle to be overcome is that such further developments do not remain isolated cases and are adapted by other data providers and repositories. Only in this way can profound progress be made.

Many promising results have already been achieved with the approaches and data management systems described above. However, the integration of a wide variety of data is only at the beginning of the development. The long-term research goal of IPK Gatersleben is to develop into a bio-digital resource centre. For this purpose, a central entry point for accessing the IPK data needs to be established. Furthermore, the stored information about the biological objects should be provided with identifiers in such a way that traceability and integrability beyond the IPK institute boundaries are possible. These challenges will be the focus of activities for the bioinformaticians, data stewards and data scientists in the future.

Acknowledgements This work was supported by grants from the German Federal Ministry of Education and Research to, Uwe Scholz (SHAPE 2: FKZ 031B0884A, de.NBI: FKZ 031A536A) and Matthias Lange (AVATARS: FKZ 031B0770A) and from European Union's Horizon 2020 Research and Innovation Program to Matthias Lange (AGENT project: grant agreement no. 862613). The initial development of GBIS was jointly funded by the German Federal Ministry of Education and Research and the German Federal Ministry for Food and Agriculture. The development and maintenance of EURISCO are being funded by the European Cooperative Programme for Plant Genetic Resources.

References

Alercia A, Diulgheroff S, Metz T (2001) FAO/IPGRI multi-crop passport descriptors [MCPD]. Food and Agriculture Organization of the United Nations (FAO); International Plant Genetic Resources Institute, (IPGRI), Rome

Alercia A, Diulgheroff S, Mackay M (2015) FAO/bioversity multi-crop passport descriptors V. 2.1 [MCPD V. 2.1]. Food and Agriculture Organization of the United Nations (FAO); Bioversity International, Rome

Altmann T (2020) Forschungsbericht—Research Report—2018 - 2019. Leibniz-Institute of Plant Genetics and Crop Plant Research, Gatersleben, pp 78–81

Arend D, Lange M, Chen J, Colmsee C, Flemming S, Hecht D, Scholz U (2014) E!DAL—a framework to store, share and publish research data. BMC Bioinform 15:214. https://doi.org/10.1186/1471-2105-15-214

Arend D, Junker A, Scholz U, Schüler D, Wylie J, Lange M (2016) PGP repository: a plant phenomics and genomics data publication infrastructure. Database 2016:baw033. https://doi.org/10.1093/database/baw033

Arend D, König P, Junker A, Scholz U, Lange M (2020) The on-premise data sharing infrastructure e!DAL: Foster FAIR data for faster data acquisition. GigaScience 9:giaa107. https://doi.org/10.1093/gigascience/giaa107

Ariyadasa R, Mascher M, Nussbaumer T, Schulte D, Frenkel Z, Poursarebani N, Zhou R, Steuernagel B, Gundlach H, Taudien S, Felder M, Platzer M, Himmelbach A, Schmutzer T, Hedley PE, Muehlbauer GJ, Scholz U, Korol A, Mayer KFX, Waugh R, Langridge P, Graner A, Stein N (2014) A sequence-ready physical map of barley anchored genetically by two million single-nucleotide polymorphisms. Plant Physiol 164:412. https://doi.org/10.1104/pp.113.228213

Beier S, Himmelbach A, Schmutzer T, Felder M, Taudien S, Mayer KFX, Platzer M, Stein N, Scholz U, Mascher M (2016) Multiplex sequencing of bacterial artificial chromosomes for assembling complex plant genomes. Plant Biotechnol J 14:1511–1522. https://doi.org/10.1111/pbi.12511

Beier S, Himmelbach A, Colmsee C, Zhang X-Q, Barrero RA, Zhang Q, Li L, Bayer M, Bolser D, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Sampath D, Heavens D, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Houben A, Doležel J, Ayling S, Lonardi S, Langridge P, Muehlbauer GJ, Kersey P, Clark MD, Caccamo M, Schulman AH, Platzer M, Close TJ, Hansson M, Zhang G, Braumann I, Li C, Waugh R, Scholz U, Stein N, Mascher M (2017) Construction of a map-based reference genome sequence for barley, Hordeum vulgare L. Sci Data 4:170044. https://doi.org/10.1038/sdata.2017.44

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6:80–92. https://doi.org/10.4161/fly.19695

Colmsee C, Beier S, Himmelbach A, Schmutzer T, Stein N, Scholz U, Mascher M (2015) BARLEX—the barley draft genome explorer. Mol Plant 8:964–966. https://doi.org/10.1016/j.molp.2015.03.009

Cousijn H, Feeney P, Lowenberg D, Presani E, Simons N (2019) Bringing citations and usage metrics together to make data count. Data Sci J 18:9. https://doi.org/10.5334/dsj-2019-009

Crosswell LC, Thornton JM (2012) ELIXIR: a distributed infrastructure for European biological data. Trends Biotechnol 30:241–242. https://doi.org/10.1016/j.tibtech.2012.02.002

Ćwiek-Kupczyńska H, Altmann T, Arend D, Arnaud E, Chen D, Cornut G, Fiorani F, Frohmberg W, Junker A, Klukas C, Lange M, Mazurek C, Nafissi A, Neveu P, van Oeveren J, Pommier C, Poorter H, Rocca-Serra P, Sansone S-A, Scholz U, van Schriek M, Seren Ü, Usadel B, Weise S, Kersey P, Krajewski P (2016) Measures for interoperability of phenotypic data: minimum information requirements and formatting. Plant Methods 12:44. https://doi.org/10.1186/s13007-016-0144-4

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST (2011) The variant call format and VCF tools. Bioinformatics 27:2156–2158

Drysdale R, Cook CE, Petryszak R, Baillie-Gerritsen V, Barlow M, Gasteiger E, Gruhl F, Haas J, Lanfear J, Lopez R, Redaschi N, Stockinger H, Teixeira D, Venkatesan A, Elixir Core Data Resource Forum, Blomberg N, Durinx C, McEntyre J (2020) The ELIXIR Core data resources: fundamental infrastructure for the life sciences. Bioinformatics 36:2636–2642. https://doi.org/10.1093/bioinformatics/btz959

ELIXIR (2021) Research Data Management Kit. A deliverable from the EU-funded ELIXIR-CONVERGE project (grant agreement 871075) [WWW Document]. https://rdmkit.elixir-europe.org. Accessed 28 May 21

Endresen DTF, Knüpffer H (2012) The Darwin Core extension for genebanks opens up new opportunities for sharing genebank datasets. Biodivers Inform 8:12–29. https://doi.org/10.17161/bi.v8i1.4095

Engels JMM, Maggioni L (2012) AEGIS: a regionally based approach to PGR conservation. In: Maxted N, Dulloo ME, Ford-Lloyd BV, Frese L, Iriondo JM, Pinheiro de Carvalho MAA (eds) Agrobiodiversity conservation: securing the diversity of crop wild relatives and landraces. CABI, Wallingford, pp 321–326

FAO (1997) The State of the World's Plant Genetic Resources for Food and Agriculture. Food and Agriculture Organization of the United Nations, Rome

FAO (2010) The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Commission on Genetic Resources for Food and Agriculture, Food and Agriculture Organization of the United Nations, Rome

Fiorani F, Schurr U (2013) Future scenarios for plant phenotyping. Annu Rev Plant Biol 64:267–291. https://doi.org/10.1146/annurev-arplant-050312-120137

Fowler C, Hodgkin T (2004) Plant genetic resources for food and agriculture: assessing global availability. Annu Rev Env Resour 29:143–179. https://doi.org/10.1146/annurev.energy.29.062403.102203

Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD (2016) Cytoscape.Js: a graph theory library for visualisation and analysis. Bioinformatics 32:309–311. https://doi.org/10.1093/bioinformatics/btv557

Fraternali P, Rossi G, Sánchez-Figueroa F (2010) Rich internet applications. IEEE Internet Comput 14:9–12. https://doi.org/10.1109/MIC.2010.76

Gass T, Lipman E, Maggioni, L (1997) The role of Central Crop Databases in the European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR). In: Lipman, E et al. (eds) Central Crop Databases: Tools for Plant Genetic Resources Management, European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR); International Plant Genetic Resources Institute, Rome (Italy), pp 22–29

Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261. https://doi.org/10.1093/nar/gkh036

Ghaffar M, Schüler D, König P, Arend D, Junker A, Scholz U, Lange M (2019) Programmatic access to FAIRified digital plant genetic resources. J Integr Bioinform 16:20190060. https://doi.org/10.1515/jib-2019-0060

Guha R, Brickley D, Macbeth S (2016) Schema.org: evolution of structured data on the web. Commun ACM 59(2):44–51. https://doi.org/10.1145/2844544

Harrington JL (2016) Relational database design and implementation. Morgan Kaufmann, London

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2

Hazekamp T, Serwinski J, Alercia A (1997) Multicrop passport descriptors. In: Lipman E et al. (eds) Central Crop Databases: Tools for Plant Genetic Resources Management, European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR); International Plant Genetic Resources Institute, Rome (Italy), pp 40–44

Hoisington D, Khairallah M, Reeves T, Ribaut J-M, Skovmand B, Taba S, Warburton M (1999) Plant genetic resources: what can they contribute toward increased crop productivity? Proc Natl Acad Sci 96:5937–5943. https://doi.org/10.1073/pnas.96.11.5937

IBPGR (1990) Descriptors for Brassica and Raphanus. International Board for Plant Genetic Resources, Rome

International Board for Plant Genetic Resources (IBPGR), Commission of the European Communities (CEC) (1984) Plum descriptors. Rome

IPGRI, ECP/GR, AVRDC (2001) Descriptors for Allium (Allium spp.). International Plant Genetic Resources Institute, Rome; European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR), Asian Vegetable Research and Development Center, Taiwan

Jacobsen A, Kaliyaperumal R, da Silva Santos LOB, Mons B, Schultes E, Roos M, Thompson M (2020) A generic workflow for the data FAIRification process. Data Intell 2:56–65. https://doi.org/10.1162/dint_a_00028

Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, Vincent L, Ware D, Zapata F (2005) Plant ontology (PO): a controlled vocabulary of plant structures and growth stages. Comp Funct Genom 6:388–397. https://doi.org/10.1002/cfg.496

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, Ens J, Zhang X-Q, Angessa TT, Zhou G, Tan C, Hill C, Wang P, Schreiber M, Boston LB, Plott C, Jenkins J, Guo Y, Fiebig A, Budak H, Xu D, Zhang J, Wang C, Grimwood J, Schmutz J, Guo G, Zhang G, Mochida K, Hirayama T, Sato K, Chalmers KJ, Langridge P, Waugh R, Pozniak CJ, Scholz U, Mayer KFX, Spannagl M, Li C, Mascher M, Stein N (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588:284–289. https://doi.org/10.1038/s41586-020-2947-8

Jayakodi M, Schreiber M, Stein N, Mascher M (2021) Building pan-genome infrastructures for crop plants and their use in association genetics. DNA Res 28. https://doi.org/10.1093/dnares/dsaa030

Jiao W-B, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Genome Stud Mol Genet 36:64–70. https://doi.org/10.1016/j.pbi.2017.02.002

Keim DA (2002) Information visualization and visual data mining. IEEE Trans Vis Comput Graph 8:1–8. https://doi.org/10.1109/2945.981847

König P, Beier S, Basterrechea M, Schüler D, Arend D, Mascher M, Stein N, Scholz U, Lange M (2020) BRIDGE—a visual analytics web tool for barley Genebank genomics. Front Plant Sci 11:701. https://doi.org/10.3389/fpls.2020.00701

Krajewski P, Chen D, Ćwiek H, van Dijk ADJ, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP, van Oeveren J, Pommier C, Scholz U, van Schriek M, Usadel B, Weise S (2015) Towards recommendations for metadata and data handling in plant phenotyping. J Exp Bot 66:5417–5427. https://doi.org/10.1093/jxb/erv271

Kreide S, Oppermann M, Weise S (2019) Advancement of taxonomic searches in the European search catalogue for plant genetic resources. Plant Genet Resour Charact Util 17:559–561. https://doi.org/10.1017/S1479262119000339

Lanthaler M, Gütl C (2012) On using JSON-LD to create evolvable RESTful services. In: WS-REST '12: proceedings of the third international workshop on RESTful design, April 2012. pp 25–32. https://doi.org/10.1145/2307819.2307827

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G (2011) The European nucleotide archive. Nucleic Acids Res 39:D28–D31. https://doi.org/10.1093/nar/gkq967

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289. https://doi.org/10.1126/science.1181369

Löffler H (2012) Meteorologische Bodenmesstechnik (vormals: Instrumentenkunde), Leitfaden für die Ausbildung im Deutschen Wetterdienst Nr. 6. Selbstverlag des Deutschen Wetterdienstes, Offenbach am Main (Deutschland)

Madakam S, Ramaswamy R, Tripathi S (2015) Internet of things (IoT): a literature review. J Comput Commun 3:164–173. https://doi.org/10.4236/jcc.2015.35021

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang X-Q, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427–433. https://doi.org/10.1038/nature22043

Mascher M, Schreiber M, Scholz U, Graner A, Reif JC, Stein N (2019) Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. Nat Genet 51:1076–1081. https://doi.org/10.1038/s41588-019-0443-6

Mascher M, Wicker T, Jenkins J, Plott C, Lux T, Koh CS, Ens J, Gundlach H, Boston LB, Tulpová Z, Holden S, Hernández-Pinzón I, Scholz U, Mayer KFX, Spannagl M, Pozniak CJ, Sharpe AG, Šimková H, Moscou MJ, Grimwood J, Schmutz J, Stein N (2021) Long-read sequence assembly: a technical evaluation in barley. Plant Cell 33:1888–1906. https://doi.org/10.1093/plcell/koab077

Mayer KFX, Waugh R, Langridge P, Close TJ, Wise RP, Graner A, Matsumoto T, Sato K, Schulman A, Muehlbauer GJ, Stein N, Ariyadasa R, Schulte D, Poursarebani N, Zhou R, Steuernagel B, Mascher M, Scholz U, Shi B, Langridge P, Madishetty K, Svensson JT, Bhat P, Moscou M, Resnik J, Close TJ, Muehlbauer GJ, Hedley P, Liu H, Morris J, Waugh R, Frenkel Z, Korol A, Bergès H, Graner A, Stein N, Steuernagel B, Scholz U, Taudien S, Felder M, Groth M, Platzer M, Stein N, Steuernagel B, Scholz U, Himmelbach A, Taudien S, Felder M, Platzer M, Lonardi S, Duma D, Alpert M, Cordero F, Beccuti M, Ciardo G, Ma Y, Wanamaker S, Close TJ, Stein N, Cattonaro F, Vendramin V, Scalabrin S, Radovic S, Wing R, Schulte D, Steuernagel B, Morgante M, Stein N, Waugh R, Nussbaumer T, Gundlach H, Martis M, Ariyadasa R, Poursarebani N, Steuernagel B, Scholz U, Wise RP, Poland J, Stein N, Mayer KFX, Spannagl M, Pfeifer M, Gundlach H, Mayer KFX, Gundlach H, Moisy C, Tanskanen J, Scalabrin S, Zuccolo A, Vendramin V, Morgante M, Mayer KFX, Schulman A, Pfeifer M, Spannagl M, Hedley P, Morris J, Russell J, Druka A, Marshall D, Bayer M, Swarbreck D, Sampath D, Ayling S, Febrer M, Caccamo M, Matsumoto T, Tanaka T, Sato K, Wise RP, Close TJ, Wannamaker S, Muehlbauer GJ, Stein N, Mayer KFX, Waugh R, Steuernagel B, Schmutzer T, Mascher M, Scholz U, Taudien S, Platzer M, Sato K, Marshall D, Bayer M, Waugh R, Stein N, Mayer KFX, Waugh R, Brown JWS, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N, The International Barley Genome Sequencing Consortium, Principal investigators, Physical map construction and direct anchoring, Genomic sequencing and assembly, BAC sequencing and assembly, BAC-end sequencing, Integration of physical/genetic map and sequence resources, Gene annotation, Repetitive DNA analysis, Transcriptome sequencing and analysis, Re-sequencing and diversity analysis, Writing and editing of the manuscript (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716. https://doi.org/10.1038/nature11543

McKinney W (2010) Data structures for statistical computing in python. In: van der Walt S, Millman J (eds) Proceedings of the 9th Python in Science Conference, pp 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Memon JA (2020) Concept and implementation of homogeneous Sensorics infrastructure for the analysis of environmental factors in plant phenotyping (Masterabschlussarbeit). Universität Bielefeld, Bielefeld

Miles A, Jakirkham, Durant M, Bussonnier M, Bourbeau J, Onalan T, Hamman J, Patel Z, Rocklin M, Shikharsg, Abernathey R, Moore J, Schut V, Raphael D, de Andrade ES, Noyes C, Jelenak A, Banihirwe A, Barnes C, Sakkis G, Funke J, Kelleher J, Jevnik J, Swaney J, Rahul PS, Saalfeld S et al (2020) Zarr-developers/zarr-python: v2.5.0. Zenodo. https://doi.org/10.5281/zenodo.4069231

Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpffer H, Basterrechea M, König P, Schüler D, Sharma R, Pasam RK, Rutten T, Guo G, Xu D, Zhang J, Herren G, Müller T, Krattinger SG, Keller B, Jiang Y, González MY, Zhao Y, Habekuß A, Färber S, Ordon F, Lange M, Börner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N (2019) Genebank genomics highlights the diversity of a global barley collection. Nat Genet 51:319–326. https://doi.org/10.1038/s41588-018-0266-x

Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, Ens J, Li C, Muehlbauer GJ, Schulman AH, Waugh R, Braumann I, Pozniak C, Scholz U, Mayer KFX, Spannagl M, Stein N, Mascher M (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome Biol 20:284. https://doi.org/10.1186/s13059-019-1899-5

Mons B (2019) FAIR science for social machines: let's share metadata Knowlets in the internet of FAIR data and services. Data Intell 1:22–42

Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD (2017) Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science cloud. Inf Serv Use 37:49–56. https://doi.org/10.3233/ISU-170824

Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, Cao H, Schlebusch SA, Giorda K, Schnall-Levin M, Wall JD, Kwok P-Y (2016) A hybrid approach for de novo human genome sequence assembly and phasing. Nat Methods 13:587–590. https://doi.org/10.1038/nmeth.3865

MQTT.org (2015) 10th birthday party | MQTT [WWW document]. https://web.archive.org/web/20150315025826/. https://mqtt.org/2009/07/10th-birthday-party. Accessed 18 May 2021

Obermaier D (2018) MQTT 5—Die Neuerungen für das IoT-Standardprotokoll. JAXenter. https://jaxenter.de/mqtt-5-internet-of-things-protocol-74891. Accessed 18 May 2021

Oppermann M, Weise S, Dittmann C, Knüpffer H (2015) GBIS: the information system of the German Genebank. Database 2015:bav021. https://doi.org/10.1093/database/bav021

Papoutsoglou EA, Faria D, Arend D, Arnaud E, Athanasiadis IN, Chaves I, Coppens F, Cornut G, Costa BV, Ćwiek-Kupczyńska H, Droesbeke B, Finkers R, Gruden K, Junker A, King GJ, Krajewski P, Lange M, Laporte M-A, Michotey C, Oppermann M, Ostler R, Poorter H, Ramirez-Gonzalez R, Ramšak Z, Reif JC, Rocca-Serra P, Sansone SA, Scholz U, Tardieu F, Uauy C, Usadel B, Visser RGF, Weise S, Kersey PJ, Miguel CM, Adam-Blondon A-F, Pommier C (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytol 227:260–273. https://doi.org/10.1111/nph.16544

Parsons MA, Duerr RE, Jones MB (2019) The history and future of data citation in practice. Data Sci J 18:52. https://doi.org/10.5334/dsj-2019-052

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Petersen J (2001) Benefits of using the n-tiered approach for web applications. Benefits Using N-Tiered Approach Web Appl. https://web.archive.org/web/20060618183914/https://www.adobe.com/devnet/coldfusion/articles/ntier.html. Accessed 20 May 2021

Philipp N, Weise S, Oppermann M, Börner A, Graner A, Keilwagen J, Kilian B, Zhao Y, Reif JC, Schulthess AW (2018) Leveraging the use of historical data gathered during seed regeneration of an ex situ Genebank collection of wheat. Front Plant Sci 9:609. https://doi.org/10.3389/fpls.2018.00609

Psaroudakis D, Liu F, König P, Scholz U, Junker A, Lange M, Arend D (2020) isa4j: a scalable Java library for creating ISA-Tab metadata [version 1; peer review: 2 approved]. F1000Res 9. https://doi.org/10.12688/f1000research.27188.1

Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, McMullen MD, Holland JB, Buckler ES, Gardner CA (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14:R55. https://doi.org/10.1186/gb-2013-14-6-r55

Rutkowski T (2005) Konzeption und Implementation einer plattformunabhängigen Importsoftware für ORACLE Datenbanken (Projektarbeit). Fachhochschule Harz, Wernigerode

Sansone S-A, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Jones P, Lister A, Miller M, Morrison N, Rayner T, Sklyar N, Taylor C, Tong W, Warner G, Wiemann S (2008) The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?" OMICS. J Integr Biol 12:143–149. https://doi.org/10.1089/omi.2008.0019

Sansone S-A, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, Begley K, Booth T, Bougueleret L, Burns G, Chapman B, Clark T, Coleman L-A, Copeland J, Das S, de Daruvar A, de Matos P, Dix I, Edmunds S, Evelo CT, Forster MJ, Gaudet P, Gilbert J, Goble C, Griffin JL, Jacob D, Kleinjans J, Harland L, Haug K, Hermjakob H, Sui SJH, Laederach A, Liang S, Marshall S, McGrath A, Merrill E, Reilly

D, Roux M, Shamu CE, Shang CA, Steinbeck C, Trefethen A, Williams-Jones B, Wolstencroft K, Xenarios I, Hide W (2012) Toward interoperable bioscience data. Nat Genet 44:121–126. https://doi.org/10.1038/ng.1054

Schulte D, Close TJ, Graner A, Langridge P, Matsumoto T, Muehlbauer G, Sato K, Schulman AH, Waugh R, Wise RP, Stein N (2009) The international barley sequencing consortium—at the threshold of efficient access to the barley genome. Plant Physiol 149:142. https://doi.org/10.1104/pp.108.128967

Schulte D, Ariyadasa R, Shi B, Fleury D, Saski C, Atkins M, deJong P, Wu C-C, Graner A, Langridge P, Stein N (2011) BAC library resources for map-based cloning and physical map construction in barley (Hordeum vulgare L.). BMC Genomics 12:247. https://doi.org/10.1186/1471-2164-12-247

SciData Editorial (2019) Data citation needed. Sci Data 6:27. https://doi.org/10.1038/s41597-019-0026-5

Shrestha R, Arnaud E, Mauleon R, Senger M, Davenport GF, Hancock D, Morrison N, Bruskiewich R, McLaren G (2010) Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. AoB Plants 2010:plq008. https://doi.org/10.1093/aobpla/plq008

Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, Arnaud E (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the crop ontology developed by the crop communities of practice. Front Physiol 3:326. https://doi.org/10.3389/fphys.2012.00326

Staňková H, Hastie AR, Chan S, Vrána J, Tulpová Z, Kubaláková M, Visendi P, Hayashi S, Luo M, Batley J, Edwards D, Doležel J, Šimková H (2016) BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol J 14:1523–1531. https://doi.org/10.1111/pbi.12513

Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, Schulte D, Petzold A, Felder M, Graner A, Scholz U, Mayer KF, Platzer M, Stein N (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. BMC Genomics 10:547. https://doi.org/10.1186/1471-2164-10-547

Stöbe E (2019) Konzeption und Implementierung einer nachrichtenorientierten Sensorikinfrastruktur für eine Pflanzenphänotypisierung und einer adaptiven Bewässerungssteuerung (Masterarbeit). Hochschule Midweida, Fakultät: Angewandte Computer- und Biowissenschaften, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung, Mittweida

Taudien S, Steuernagel B, Ariyadasa R, Schulte D, Schmutzer T, Groth M, Felder M, Petzold A, Scholz U, Mayer KF, Stein N, Platzer M (2011) Sequencing of BAC pools by different next generation sequencing platforms and strategies. BMC Res Notes 4:411. https://doi.org/10.1186/1756-0500-4-411

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, Pollock D, Dorsett K (2015) Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS One 10:e0134826. https://doi.org/10.1371/journal.pone.0134826

van Hintum TJL (1997) Central crop databases—an overview. In: Lipman E et al. (eds) Central Crop Databases: Tools for Plant Genetic Resources Management, European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR); International Plant Genetic Resources Institute, Rome (Italy), pp 18–21

Wang C, Hu S, Gardner C, Lübberstedt T (2017) Emerging avenues for utilization of exotic germplasm. Trends Plant Sci 22:624–637. https://doi.org/10.1016/j.tplants.2017.04.002

Watt M, Fiorani F, Usadel B, Rascher U, Muller O, Schurr U (2020) Phenotyping: new windows into the plant for breeders. Annu Rev Plant Biol 71:689–712. https://doi.org/10.1146/annurev-arplant-042916-041124

Weise S, Oppermann M, Maggioni L, van Hintum T, Knüpffer H (2017) EURISCO: the European search catalogue for plant genetic resources. Nucleic Acids Res 45:D1003–D1008. https://doi.org/10.1093/nar/gkw755

Weise S, Lohwasser U, Oppermann M (2020) Document or lose it—on the importance of information management for genetic resources conservation in genebanks. Plants 9:1050. https://doi.org/10.3390/plants9081050

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: an evolving community-developed biodiversity data standard. PLoS One 7:e29715. https://doi.org/10.1371/journal.pone.0029715

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018

World Meteorological Organization (WMO) (2017) Challenges in the transition from conventional to automatic meteorological observing networks for long-term climate records. WMO-No. 1202, Geneva

World Meteorological Organization (WMO) (2018) Guide to instruments and methods of observation, Volume V: quality assurance and management of observing systems. WMO-No. 8, Geneva

Part II
Data/Database Integration

Chapter 3
Research Data Resources for Epidemiology

Louise Corti and Deborah Wiltshire

Abstract This chapter introduces the data landscape that enables access to data resources for researching in the field of epidemiology. The chapter covers a general introduction to methods, standards and infrastructures for curating, sharing and reusing data of interest to health and epidemiology researchers.

By making data available and accessible for reuse, as well as offering reproducibility for published findings, new potential research opportunities are opened up. As an example, the chapter focuses on long-running cohort studies that collect rich and unique information on specific populations, which can be reanalysed and reworked for new analyses. It demonstrates approaches for describing and sharing data, showing how datasets can be prepared with various 'levels' of access depending on disclosure risk, and how personal data can be shared through the use of the '5 Safes' protocol, using legal gateways and safe havens. It points to some of the challenges in data sharing and reproducibility, from ethical and confidentiality considerations to intellectual property concerns over who should gain access. A range of case studies of how these archived data resources have been accessed and used are presented. Finally, it shares some of the logistical and technology barriers around data sharing infrastructures and how cross-disciplinary interests can help bridge differences in approaches.

Keywords Data sharing · Biomedical data · Social science data archives · Secondary data analysis

L. Corti (✉)
Office for National Statistics, London, UK
e-mail: louise.corti@ons.gov.uk

D. Wiltshire
GESIS—Leibniz-Institut für Sozialwissenschaften, Mannheim, Germany

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_3

3.1 Introduction

This chapter introduces the data landscape that enables access to data resources for researching in the field of epidemiology. The chapter covers a general introduction to methods, standards and infrastructures for curating, sharing and reusing data of interest to health and epidemiology researchers.

By making data available and accessible for reuse, as well as offering reproducibility for published findings, new potential research opportunities are opened up. As an example, the chapter focuses on long-running cohort studies that collect rich and unique information on specific populations, which can be reanalysed and reworked for new analyses. It demonstrates approaches for describing and sharing data, showing how datasets can be prepared with various 'levels' of access depending on disclosure risk, and how personal data can be shared through the use of the '5 Safes' protocol, using legal gateways and safe havens. It points to some of the challenges in data sharing and reproducibility, from ethical and confidentiality considerations to intellectual property concerns over who should gain access. A range of case studies of how these archived data resources have been accessed and used are presented. Finally, it shares some of the logistical and technology barriers around data sharing infrastructures and how cross-disciplinary interests can help bridge differences in approaches.

3.2 History of Epidemiological Data Resources

Data collected from administrative sources, such as demographic information and medical records, and surveys represent a rich and unique resource that can be reanalysed and reworked for new analyses. By making data available for reuse, new potential avenues are opened up, enabling researchers to access data that they would not be able to collect themselves. Data archives can provide curation of precious resources, through digitising and preserving information, and assuring appropriate governance and access to resources.

Around the world, there are a number of well-established disciplinary-based data services, such as the UK Data Service and the US Interuniversity Consortium for Political and Social Research (ICPSR). These services have national remits bringing together expertise across a number of fields to make key national and international socio-economic datasets shareable, usable and sustainable. Organisations such as these have established services designed to meet the data and information needs of today's social science researchers and data analysts. Data acquisition and processing, quality assurance procedures; systematic resource discovery systems; value-added support materials; and web-based interfaces for data browsing, exploration and data download are all key features of successful service delivery.

As new forms of data come on stream, these present challenges for data quality assessment, data delivery and confidentiality.

Surveys of interest to the epidemiologist include government surveys, census data sources, academic surveys, national and international time series. In the past 10 years, attention has also turned to administrative or register data sources and other 'big' data, for example from medical devices, environmental monitoring and financial transactions.

Reusing existing data reduces respondent burden, enables data linkage and the creation of new datasets, informs published and provides transparency about method and opportunities to reproduce results (US National Academy of Sciences 2005). Furthermore, collecting and collating high-quality, reliable, representative data is expensive and technically demanding.

3.2.1 Aggregate Statistics

Many countries' governments publish national summary statistics on a broad range of indicators covering aspects of economic, social and well-being. In the UK, the Office for National Statistics (ONS) provides a range of up-to-date national statistics on business, trade and industry, the economy, employment and the labour market, and on people, population and community (ONS 2021). For health researchers, there is a large range of information on births and deaths, and health and social care. The UK's NHS England provides additional and detailed health-related information, used to inform debate, decision-making and research both within Government and by the wider community (NHS Digital 2021a). Many other countries provide similar resources on public websites.

Rich comparative global data are also collected by governments and non-governmental organisations. Statistical indicators covering economic and other indicators of countries' performance permit comparisons between countries and over time. The geographical scope is typically extensive. The World Bank, International Monetary Fund and United Nations provide access to data on topics covering national accounts, industrial production, employment, trade, demography, human development and other indicators of national performance and development. Nation state contributors follow guidance that enables data gathering which would be difficult to achieve without this level of authority and structured international cooperation. An example of comparable open aggregate statistics across many dimensions of the health domain is *The State of the World's Children United Nations Children's Fund (UNICEF)* through the UNdata Explorer (Unicef 2021).

3.2.2 Biomedical Surveys

Surveys provide both microdata (data records at the individual, household or organisation level) and aggregate data (summary statistics such as counts reported in tables in government publications or websites). Continuous surveys produce

accurate population estimates about social and economic behaviour and attitudes and offer great opportunities for time series analysis. An example of a repeated national survey of health in the UK is the Health Survey for England (HSE) (NHS Digital 2021b).

This survey monitors trends in the nation's health and care, by providing health-related information about adults and children living in private households in England. The survey includes core questions on: smoking, alcohol, general health, measurements such as height, weight and blood pressure, analysis of blood and saliva samples, question modules on specific topics that vary from year to year. Repeated surveys such as this permit comparison of groups over time, something that is not readily achievable retrospectively for reasons of both recall and mortality.

Secondary analysis can be undertaken on longitudinal data sources, where data are collected from the same individuals over a period of time. The research potential of a longitudinal dataset, be it a cohort study or a panel survey, increases as the study matures. These surveys are expensive to conduct but offer great potential to study and understand change in individuals' circumstance and health over time and across the life-course (Ruspini 2002).

A growing number of well-known longitudinal studies from across the world make data available for reuse. There also exist well-established birth cohort studies and studies relating to specific cohorts, such as children and those in later life, where data are available for research use. Some of the key studies are summarised in Table 3.1.

Longitudinal studies, such as the UK Millennium Cohort Study (MCS), which charts the conditions of social, economic and health advantages and disadvantages facing children born at the start of the twenty-first century uses an approach to consent that is explicit and consistent. From the outset, MCS sought informed parental consent (Shepherd 2012). Letters and leaflets sent in advance of the surveys summarise what participation in the survey will involve, and written consent is sought from parents for their participation (interview) and the participation of their child(ren) (e.g. assessments, anthropometric measurements, collection of oral fluids and saliva, linking of administrative data on education and health, teacher surveys). Where parents give consent to the participation of their child(ren) in one or more elements of a survey, the inclusion of the child(ren) requires their agreement and compliance. Parents were not asked to consent on behalf of the child, but were asked for their permission to allow the interviewer to speak to the child and ask for their consent to participate in each element.

Linking multiple sources of data can add power to the analytic potential of individual sources. Microdata from surveys can be linked to other microdata files directly through common identifiers or indirectly via probabilistic linkage. Common identifiers need to be coded in exactly the same way in both datasets. The internet also enables open, usually aggregate, data sources to be published through web interfaces and linked. Increasingly anyone can gain access via the internet to updated 'data feeds' which are drawn from a vast number of public data sources and updated in real-time, such as weather reports or current stock market share prices.

Table 3.1 Examples of long-running longitudinal and cohort surveys, available for research use

Country	Survey	Dates	Description	Topics	Host Institution and reference
UK	British household panel study (BHPS)	1991–2009	Household panel study that collects information about all household members from a representative sample of ~5000 households (around 10,000 individuals)	Household characteristics and conditions, demographics, socio-economic status, employment, physical and mental health, fertility, child development, ageing, life outcomes, attitude and health behaviours	Institute for Social and Economic Research, University of Essex Institute for Social and Economic Research (2021)
UK	Understanding society	2009	Household panel study that expanded upon the BHPS; around 40,000 households were included, equating to around 100,000 individuals. Includes ethnic minority and immigrant and ethnic minority boost samples. Includes the BHPS sample from wave 2	Household characteristics and conditions, deprivation, socio-economic status, current and historic employment, physical and mental health, biomarkers, fertility histories, child development, ageing, life outcomes, attitudinal data and health behaviours, migration	Institute for Social and Economic Research, University of Essex Institute for Social and Economic Research (2021)
UK: Britain	The Medical Research Council (MRC) National Survey of Health and Development (NSHD)	1946 to present	The oldest of the British birth cohort studies, collecting data from birth on the health and social circumstances of a representative sample of 5362 men and women born in March 1946	Child development, initial conditions, morbidity, life outcomes, ageing	Medical Research Council National Survey of Health and Development, University College London Medical Research Council National Survey of Health and Development (2021)
UK: Britain	1958 National Child Development Study (NCDS)	1958 to present	British birth cohort study collecting health, education, social and economic data from a representative sample of 17,415 individuals born in a particular week in 1958	Child development, initial conditions, family situation, morbidity, educational measures, socio-economic status, life outcomes, physical and mental health, biomarkers, ageing	Centre for Longitudinal Studies, Institute of Education, University of London Power and Elliott (2006)

(continued)

Table 3.1 (continued)

Country	Survey	Dates	Description	Topics	Host Institution and reference
UK: Britain	1970 British cohort study (BCS70)	1970 to present	British birth cohort study that follows around 17,000 individuals born in 1970, collecting health, education, social, family and economic data	Child development, initial conditions, morbidity, life outcomes, family and social background, physical and mental health, biomarkers, ageing	Centre for Longitudinal Studies, Institute of Education, University of London Elliott and Shepherd (2006)
UK	Millennium cohort study (MCS)	2000-	British birth cohort study of a representative sample of 18,819 individuals born in the UK in 2000–2001. Collects data on social and economic conditions, health outcomes and advantages. Allows generational change to be studied through comparison with earlier cohorts	Child development, initial conditions, family and social background, educational development and outcomes, employment	Centre for Longitudinal Studies, Institute of Education, University of London Connelly and Platt (2014)
UK: Avon	Avon Longitudinal Study of Parents and Children (ALSPAC)	1991 to present	Study of around 14,000 women who were pregnant between 1991–1992. Collects data on the environmental and genetic factors affecting health and development outcomes	Child development, initial conditions, morbidity, biological markers	Medical Research Council, University of Bristol University of Bristol (2021)
UK: England	English Longitudinal Study of Ageing (ELSA)	1998 to present England	A longitudinal study which focuses on later life, following a representative sample of 12,000 men and women aged 50 and over in England. The sample is drawn from those who have previously participated in the health survey for England	Morbidity, quality of life, medical conditions, cognitive function, psychosocial health, retirement	NatCen Social Research, University College London Department of Epidemiology and Public Health Steptoe et al. (2013)

UK: Britain	Health and Lifestyle Survey (HALS)	1984 to present	A sample of 9003 adults in Britain collecting data on health and lifestyle measures	Physiological measures, health beliefs and behaviours, diet, physical activity, causes of death, cancer	Social and Community Planning Research (SCPR); School of Clinical Medicine, University of Cambridge Cox (1995)
UK: London	Whitehall II	1985–2018	A cohort of 10,308 participants aged 35–55 years, of whom 3413 were women and 6895 men, was recruited from the British civil service in 1985	Ageing, morbidity, psychosocial health, socio-economic circumstances and health inequalities	University College London (2021)
USA: Oakland, California	Children of the great depression: Social change in life experience	1920	The first longitudinal study of a depression cohort, following 167 individuals born in 1920 from their elementary school through the 1960s	Economic situation	Institute of Human Development, University of Berkeley Elder (1974)
USA: Harvard	Harvard study of adult development	1937	A longitudinal study that tracked 268 white male Harvard sophomores in 1937 during the great depression for a period of 80 years; 1330 of their offspring in the 1960s/1970s	Physical and mental health, ageing	Harvard University Vaillant (2012)
USA	Panel study of income dynamics (PSID)	1968-	The study began in 1968 with a nationally representative sample of over 18,000 individuals living in 5000 families in the United States	Employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy, education	Institute for Social Research, University of Michigan Institute for Social Research (2020)
USA	Nurses' Health Study	1976-	The original study was established in 1976. The studies are now in their third generation with Nurses' Health Study 3 and include more than 280,000 participants	Risk factors for major chronic diseases in women	Nurses' Health Study Nurses' Health Study (2019)

(continued)

Table 3.1 (continued)

Country	Survey	Dates	Description	Topics	Host Institution and reference
USA	National Health and Nutrition Examination Survey (NHANES)	1960	Program of survey studies designed to assess the health and nutritional status of adults and children in the United States, started in the early 1960s. In 1999, the survey became a continuous program, examining a nationally representative sample of about 5000 persons each year	Health and nutritional status	National Centre for Health Statistics, Centers for Disease Control and Prevention National Centre for Health Statistics (2021)
Australia	Household, income and labour dynamics in Australia (HILDA)	2001-	This was the first representative household panel study in Australia, sampling 7682 households (13,969 individuals) and collecting data from all household members aged 15 years and over	Family dynamics, socio-economic measures, physical and mental health	Melbourne Institute of Applied Economic and Social Research
South Africa: Johannesburg-Soweto	Birth to 20 study	1990-	Cohort of babies, born in 1990 urban areas tracked; the first cohort born into a democratic South Africa. The original Birth to Ten study became the Birth to Twenty study in 2000. The children came to be known colloquially as Mandela's Children, because they were born in 7 weeks following Nelson Mandela's release from prison on 11 February 1990	Growth, health, well-being and educational progress. Environmental influences (poverty, migration and political violence), access to health services, nutrition, child care and growth and development	University of the Witwatersrand Richter et al. (2007)
Germany	The German socioeconomic panel (SOEP)	1984-	SOEP is one of the largest and longest-running multidisciplinary household surveys worldwide. Every year, approximately 30,000 people in 15,000 households are interviewed	Household composition, occupation, employment, earnings, health and life satisfaction	Deutsches Institut für Wirtschaftsforschung Deutsches Institut für Wirtschaftsforschung (2021)

3.2.3 Biomarkers

Biosocial researchers are keen to disentangle the relationship between our biology and our behaviours and experiences. Researchers seek to understand what impact our social and economic situation has on our biology. Experiences such as long-term unemployment or chronic stress lead to physiological responses and biological changes. Likewise, researchers seek to understand how our biology might predispose us to certain conditions or to certain behaviours such as substance abuse, and subsequently how this influences how we live and what inequalities we face (Hobcraft 2016).

While many longitudinal studies collect information about participants' physical and mental health, these questions are often self-reported measures. These are of great value, but are by their very nature subjective. With the rising interest in biosocial research, the demand for more objective measures of health and biology has grown in recent decades and increasingly a wide range of biomarkers are collected in key longitudinal studies.

Biomarkers are an objective measure of biological or pathogenic processes that allow us to measure health and disease much more objectively than through the self-report of survey participants. The combination of biomarkers and social survey data is an extremely powerful tool in health and biosocial research. A wide range of biomarkers exist and are collected in a number of ways, through physical measurements, saliva and blood samples and even the collection of milk teeth.

The collection of biomarkers in longitudinal studies are a particularly valuable source of data, as biomarker measurements may be repeated on multiple occasions within the same individual, allowing researchers to examine changes in key measures. The longitudinal design of these studies has the additional advantage of enabling us to establish temporal order. This is of crucial importance when looking to examine biological or physiological changes following key life events such as divorce or unemployment. Many of the longitudinal studies described in Table 3.1 collect biomarkers, and Table 3.2 provides a summary of key biomarkers in five UK longitudinal studies.

Collecting these measures typically requires a more stringent set of legal and ethics requirements. There are differences in the way that bio-medical research ethics committees might become involved in approval, or even at what level approval must be sought, i.e. national, regional and institutional levels. While in most countries obtaining informed consent in a written form is obligatory before a collection of dried blood can happen, legal and occupational restrictions in certain countries forbid 'unauthorised' persons to collect capillary blood (Schmidutz 2016). For example, in Austria, Poland and the Czech Republic in 2016, only medical doctors and nurses under supervision were allowed to collect blood via the fairly innocuous finger-prick method.

Table 3.2 Biomarkers in five key UK longitudinal studies

Biomarker	Application	Understanding Society	1958 National Child Development Study	1970 British Cohort Study	Millennium Cohort Study	English Longitudinal Study of Ageing
Physical measurements						
Height, weight, waist circumference Body fat percentage (bioelectrical impedance)	Used to measure BMI and excess body fat, obesity. Risk factor for major chronic conditions and social outcomes	Yes	Yes	Yes	Yes	Yes
Respiratory function (spirometry) (FVC, FEV, PF, FEV, FVC)	Indicates obstructive and restrictive respiratory diseases such as COPD	Yes	Yes			Yes
Diastolic/systolic blood pressure Resting pulse rate	Risk factor for stroke and heart and cardio-vascular disease	Yes	Yes	Yes		Yes
Grip strength	Indicates muscle strength, limits in physical function and disability	Yes		Yes		Yes
Digit length	Indicates in vitro exposure to hormones, linked to personality traits	Yes				
Accelerometry	Electromagnetic device worn on the wrist or hip to measure physical activity levels				Yes	
Blood samples	Blood analytes and DNA	Yes	Yes	Yes		Yes
Cholesterol and triglycerides	Measures fat in the blood, indicator of cardiovascular disease	Yes	Yes	Yes		Yes

Glycated haemoglobin HbA1c	Measures glucose intolerance, undiagnosed diabetes	Yes	Yes	Yes		Yes
C-reactive protein Fibrinogen	Measures inflammation and the immune system, acute or chronic stress	Yes	Yes			Yes
CMV seropositivity	Wear and tear on the immune system, chronic stress	Yes				
Haemoglobin Ferritin	Measures anaemia, can indicate poor nutrition, increases with age, can have significant negative health outcomes	Yes				Yes
Alkaline phosphatase, alanine aminotransferase, AST, GGT Albumin	Measures liver function, can indicate liver or muscle damage alcohol/drug use, obesity, other disease	Yes				
Creatinine eGFR	Measures kidney function, can indicate kidney disease, function decreases with age	Yes				
Testosterone	Hormone associated with stress processes, building muscle, ageing. Indicator for aggression	Yes				
Insulin-like growth factor 1 (IGF1)	Hormone associated with stress processes, building muscle, ageing. Indicator for diet, diabetes, cancer	Yes	Yes			Yes

(continued)

Table 3.2 (continued)

Biomarker	Application	Understanding Society	1958 National Child Development Study	1970 British Cohort Study	Millennium Cohort Study	English Longitudinal Study of Ageing
Dehydroepiandrosterone suffocate (DHEAS)	Hormone associated with stress processes, building muscle, ageing. Indicator for cardiovascular disease, muscle strength, cognition	Yes				Yes
Vitamin D	Important for healthy bones, immune system, muscle function, reducing inflammation		Yes			Yes
White blood cell count	Indicator of the immune system					Yes
Genetic data		Yes	Yes	Yes	Yes	Yes
Saliva samples			Yes	Yes	Yes	Yes
Cortisol	Hormone, indicates stress from infection, trauma, restricted diet		Yes			Yes
Hair samples						Yes
Cortisol	Hormone, indicates stress from infection, trauma, restricted diet					Yes
Milk teeth					Yes	

Source: Biomarkers, Genetics & Epigenetics | Understanding Society; The English Longitudinal Study of Ageing (ELSA) (elsa-project.ac.uk); CLS (ucl.ac.uk)

3.2.4 New Forms of Data

The arrival of 'big data' has changed social scientists' expectations, and the speed at which these data are being created has prompted a sense of urgency to capture and exploit these new sources of information. New sources of data deriving from official registrations, commercial, financial and administrative transactions, internet and social media, tracking and digital sensors, and aerial and satellite images have become available as data commodities for the social scientist (OECD 2013).

While they have brought technological and infrastructure challenges for data owners, data services and repositories, the panacea of big data has led to a focus on developing solutions for powerful analytics, such as predictive modelling and text mining, sometimes at the expense of questioning sustainability and reproducibility, ethics and data protection. As these sources were not initially collected for research, data management challenges and analysis can be complex. Laney (2001) famously tweeted a definition of big data that has stuck, known as the 3 Vs:

> 'Big data' is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

These challenges bring added complications to getting started with analysis of data, compared to more traditional datasets. In short, scaled up strategies and technologies for data management, data hosting and analysis need to be employed. While these may be more recent matters for the social sciences, they are certainly familiar within the natural sciences. GenBank, founded in 1979 as the Los Alamos Sequence Database, was established as a repository for biological sequences. Building on this innovation, the International Nucleotide Sequence Database Collaboration (INSDC) actively calls for deposits of DNA sequence data and collects and gives access to growing volumes of nucleotide and amino acid sequence data (Blaxter et al. 2016). Next-generation sequencing technologies, metagenomics, genome-wide association studies (GWAS), and the 1000 Genomes Project increase the volume and complexity of these sequence data collections (Siva 2008).

When it comes to ethical protocols for collecting new forms of data for use in research, informed consent that would be used to gather data from subjects in a survey may not be sufficient. In countries with privacy legislation in place, access to large-scale administrative data will have particular governance regimes and access arrangements set out in law, usually mediated by ethical approval for new research and adequate safeguards being put in place. In the UK, linking individual-level administrative records to survey responses requires respondents to give their explicit consent. In the example of the UK Millennium Cohort Study, for linking of health records to the study data, around 92% of cohort members' parents consented to linking survey data with birth register and/or hospital maternity data was obtained from 92% of the cohort mothers (Tate et al. 2006).

3.3 Curating Data for Reuse

Data archives or domain repositories are key to data access. Some of the major survey data archives are over 50 years old and have had to evolve and develop access services that meet the needs of today's users.

Early data archives to support the quantitative social sciences were established in both the United States and Europe in the 1960s. As collections of survey data grew and the number of similar archives came on stream, collaboration developed harmonised approaches to data storage, access and documentation. In the late 1970s, the Council of European Social Service Data Archives (CESSDA) was founded, which both promoted networks of data services for the social sciences and fostered cooperation on key archival strategies, procedures and technologies (CESSDA 2021a). The fruits of this collaboration have been more consistent tools, common standards, inter-service communication and formal structures for data sharing.

Of course, these early data archives predated the internet by decades. The gradual development of online data services has meant that from the mid-1990s onwards, many users interact primarily with a data service online. Online data delivery now incorporates data discovery, delivery, online analysis, visualisation and web-based training.

This has been in a context of a boom in online data publishing. Data sharing policies among research funders have driven exponential growth in open and restricted data repositories, hosting all kinds of research data. The international re3data.org registry indexes and provides extensive information about more than 2450 research data repositories (re3data.org 2021). 438 of these repositities are in the social and behavioural sciences.

With more institutional data repositories holding local data, comes a need for higher level discovery portals to enable data to be located. Research Data Australia is a one-stop shop portal for discovering hundreds of research data resources dispersed across Australia (Australian Research Data Commons 2021). Similarly, NARCIS in the Netherlands is a portal for the discovery of datasets and publications (Data Archiving and Networked Services 2021).

A growing number of academic publishers also play a role in ensuring that the data that underpins published findings in journals are available for reviewers, readers and reproducers. In the social sciences economics, political science and psychology have led the way. Journal policies expect research data to be made available upon request, submitted as supplemental material, or more formally deposited in a suitable or mandated domain or public repository. For example, Springer Nature journals mandate specific repositories for particular disciplines. For example their data journal, *Scientific Data* includes the UK Data Service ReShare and ICPSR openICPSR self-deposit systems (Springer Nature 2021).

It is not enough, however, to simply publish data as is. First, file and data formats may not be suited to numeric data extraction. For example, if the document has been saved as a pdf. Second, there may be insufficient documentation available to understand the data. Third, the published data may not be set up for longer-term

maintenance and research access. In the next section, we describe tried and tested curation methods for data that can be used to future proof data.

3.3.1 Curation Standards

The social data archiving community led the way in establishing and promoting common standards and shared good practice. Global data sharing activities are now quite mature, with fora like the Research Data Alliance (RDA) focusing on robust data infrastructure and shared data description methods across disciplines (Research Data Alliance 2021).

Data services with a remit to acquire, prepare and deliver data for researcher usually have a collection strategy, a pipeline where incoming data is processed for release, and maintenance and support activities. Data services typically select and appraise potential data collections against criteria designed to ensure that they are appropriate for re-use and long-term preservation. Both the UK Data Service and ICPSR have dedicated Collections Development Policies (UK Data Service 2021c; ICPSR 2015). Significant factors to account for when appraising and selecting data for acquisition include significance, uniqueness, usability, volume, formats, costs, and potential future use (UK Data Service 2021a).

Deposit or ingest is the process whereby data and related materials are transferred from data owners to a data repository. Data deposit agreements are agreed to enable data to be shared and establish the intellectual property and commercial ownership rights in the data, as well as any privacy concerns concerning personal data (UK Data Service 2021b).

Data services use bespoke in-house procedures to prepare data and documentation for online access (UK Data Service 2020; ICPSR 2021d). When data are acquired, the data service checks data integrity, missing values and anomalies or inconsistencies in the data. File formats are also examined to ensure they are in an optimal format for long-term preservation and dissemination. Data are then assessed for disclosure risk. This ensures that where the data are collected from human subjects, they have consented to data being collected on the basis of anonymity and cannot be identified from the data. Examples of potentially disclosive variables are geographic location, detailed occupation and industry, household size, exact age and any other variable which alone or in combination is unique. Where this is so, it may be necessary to group values to remove potentially identifiable values. For example age might be banded into categories and household size may be 1, 2, 3, 4, 5 and 6+. The amount of work of this type that is done depends on the data service's policies and resources.

Finally, the description of and documentation about the data is examined to ensure that there is sufficient context for onward use. Questionnaires, code books, interviewer instructions, and technical reports are required to interpret survey data. Original and subsequent publications resulting from use of the data are also captured and made available to users. These useful materials are bundled into one or multiple

user guides. Without this kind of documentation, it is difficult for potential users to determine whether a given dataset may be appropriate for their intended research and that they can feel confident to correctly interpret results produced. Nationally representative repeated surveys tend to produce very high-quality documentation, such as detailed technical reports. The Health Survey for England for example has exemplary documentation (Natcen 2014). Data Services endeavour to work with data creators in the early stages to ensure that good data management practices are adhered to and that high-quality documentation is produced and kept along the way (Corti et al. 2019).

To aid discovery of the data, a structured metadata record is created that captures core descriptive attributes of the study and resulting data. The Data Documentation Initiative (DDI) is a rich and detailed metadata standard for social, behavioural and economic sciences data, used by most social science data archives in the world (DDI Alliance 2021). A typical DDI record records mandatory and optional metadata elements relating to:

- *Study description* elements: information about the context of the data collection, scope of the study (e.g. topics, geography, time, data collection methods, sampling and processing), access information, information on accompanying materials, and provide a citation.
- *File description* elements: indicates data format, file type, file structure, missing data, weighting variables and software used.
- *Variable-level descriptions:* sets out the variable labels and codes, and question text where available.

One of the end points of the 'data ingest' process is converting the resulting package of data and documentation files to suitable user-friendly formats (e.g. for microdata, Stata, SPSS or delimited text formats) placing these on a preservation system and publishing them online.

Depending on the level of sensitivity and risk of disclosure, data are made available on a spectrum of access that requires different levels of safeguards. This is described under the Accessing Data section. A significant amount of human resource goes into preparing data in established data services. As the size or volume of the data increases, manual processes involved in data cleaning and preparation become unsustainable. Automated tools and QA pipelines are used to help assess and remedy problems in data.

Making data available is not the end of the data repository's work. Data must be maintained over time to ensure its continued usability. Data formats are updated as software changes and older formats become obsolete. Data updates may also become available as data depositors make corrections either in response to the discovery of errors, or in the light of improved estimates of population characteristics. By carefully maintaining data, future users benefit from a growing stock of historical data.

3.3.2 Curating New Forms of Data

Traditional data services have played a leading role in opening up access to digital social and economic data, new methods of access are required for more complex forms of data, where volume, quality, validity and reliability are likely to be challenging. Curating and hosting services must cope with incoming streams of real-time data and enable exploration and linkage of a variety of types of data assets. This requires adjustments to traditional practices of data management and storage, data publishing and tools for data analysis. However, a critical look at data provenance and trustworthiness, ethical and legal entitlements, data quality, structure and usability is still paramount. The joint 2018 Royal Society and British Academy report on governance for data managing and using data in the twenty-first century recognised that these new applications require robust governance (The Royal Society and British Academy 2018).

The provision of good metadata is useful for using administrative data sources, such as hospital records, where possible discoverable data dictionaries, including how derived measures have been created and quality statements (UK Statistical Authority 2015). In the public health domain, Gilbert et al. (2018) created useful guidance for information about linking data (GUILD) aimed at better understanding and reducing linkage error. Linkage errors can affect disadvantaged groups, which may, in turn, undermine evidence used for public health policy and strategy. Lack of information for linking primarily occurs due to the different processes used by various agencies along the 'data linkage pathway'; for example where there is no unique identifier across different datasets, or sometimes unreliable, quasi-identifiers, such as name, sex, date of birth and postcode, which might be used for linking. GUIDE advocates for information that could ideally be made available at each step by various data providers and linkers. This includes providing future analysts with reports on linkage accuracy and errors.

3.4 Finding Data

The internet offers huge capacity to discover useful data sources for research. An increasing amount of rich information about the public sector is available in many countries. By opening up their information for all to access, the innovation and economic potential of public sector information can be better harnessed. The OECD (2018) report on open government data noted that governments recognise that open data is re-used as a requirement for value creation; which in turn requires both improving data quantity and the capacity to identify high-value data to increase re-use.

The US government open data portal (Data.gov) was launched in 2009, shortly after President Obama launched his plans for government transparency. In April 2021 the site held just under 300,000 datasets (data.gov 2021; Madrigal 2009). In

the UK, the government's Open Data White Paper of 2012 set out standards for the timely release of open public sector data in standardised, machine-readable and open formats. It outlined what citizens, the public sector and businesses could expect from government and public services to harness the benefits of open data (Cabinet Office 2012). The UK government data portal holds thousands of open datasets and showcases how open government data have been used in apps and reports, and services (data.gov.uk 2021). Open UK health data statistical sources can also be freely accessed from the main government website (gov.uk 2021).

Other public data sources are made available via real-time data feeds, such as current weather reports or stock market share prices. The ability to create 'smart cities' relies on open data. NYC Open Data is a portal of hundreds of New York City public datasets made available by city agencies and organisations in an effort to improve the transparency and accountability of the city's government. Data on New York covers many domains that are relevant to the running of a major city . . . Health data include hospital facilities, health insurance enrolment, air quality and the Central Park Squirrel Census (The City of New York 2021). An example of published Linked Open Data is DBPedia which enables linking of structured information in Wikipedia entries to each other and beyond to other data sources (DBPedia 2021). In 2010 it claimed to hold more than 228 million entities.

Open data tend to vary hugely in their quality. Some offer no documentation, while others conform to approved metadata standards. Without long-term accessibility, and persistence of links, data sources can be there 1 day and gone the next. Longer term funding to maintain accessibility and timeliness of data is often a problem, where updating may not be possible. To address the quality of open numeric data, a number of certification systems have evolved to help establish the quality and robustness of open data systems. An example of such an awarding body is the Open Data Institute in the UK (Open Data Institute 2021). Certificates require the data publisher to provide evidence (in the form of a web page) that can demonstrate transparency for the processes and systems in place to manage and publish data. The evidence focuses on the need for detailed machine-actionable metadata as well as information on rights and conditions of use.

Curated data archives offer online data catalogues with links to access data, supporting documentation and guidance on how to use these resources. These resources have already been quality assured with good documentation provided, so the user can trust the data sources.

Examples of searchable online data catalogues for social scientists include the UK Data Service, in the US, ICPSR and various European countries' social data archives (ICPSR 2021a; UK Data Service 2021d). The Council for European Social Science Data Archives (CESSDA) hosts a federated catalogue that enables users to search for data collections across a range of European countries (CESSDA 2021b) and enables discovery of over thousands of datasets from a range of European countries. In the US, The Dataverse Project searches over around 100,000 data collections worldwide and includes results from a federated network of 'Dataverses' (The Dataverse Project 2021).

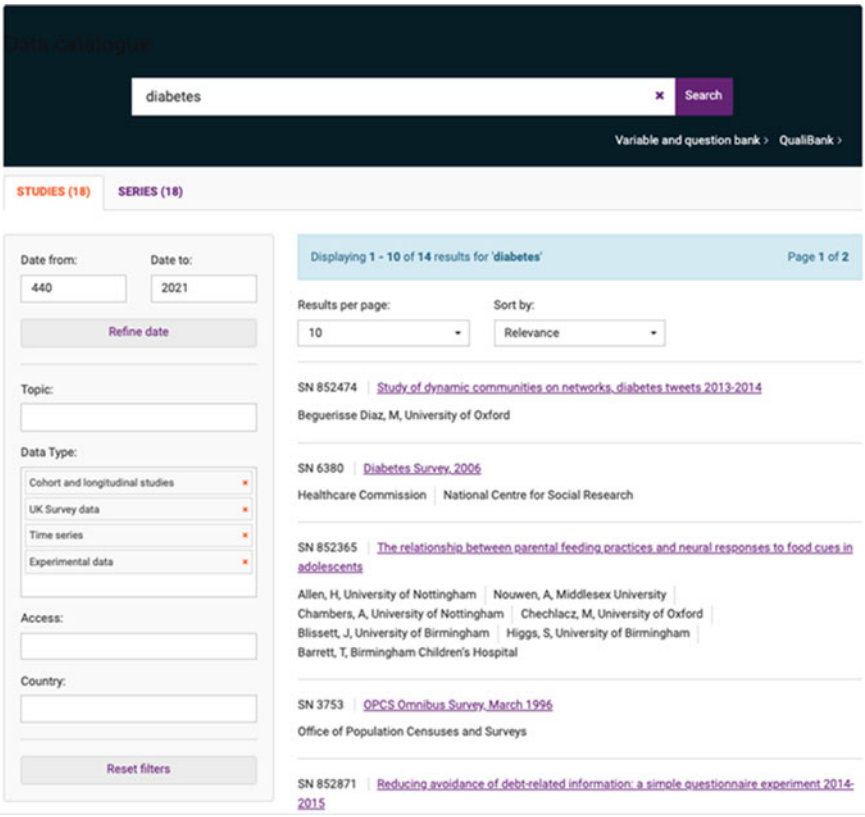

Fig. 3.1 Term search in the Discover catalogue, UK Data Service

'Discover' is the search tool for the UK Data Service's catalogue. Users can search and browse by subject, type of data, data producer and date of data collection. An example of a catalogue search on the word 'diabetes' is shown in Fig. 3.1. Catalogue records are indexed on search engines like Google, and so a Google search will also locate datasets. In our example, we look for recent UK surveys that hold information on diabetes.

Searching on the term 'diabetes' returns 14 hits, which can be filtered by facets on the left of the display window. Facets include data type, subject, date, country and dates. To limit our study to UK surveys, we restrict our search by selecting 'UK studies' in the data type facet, which now yields 9 results.

To view the catalogue entry for any of the studies in the results, the user clicks the linked title of the study in the results list. Figure 3.2 shows the catalogue record for the 2015 Health Survey for England.

The record includes an abstract, key information and documentation as well as download link where appropriate. The ease of access relates to where the data falls in terms of the access spectrum described in Table 3.3. The 2015 Health Survey

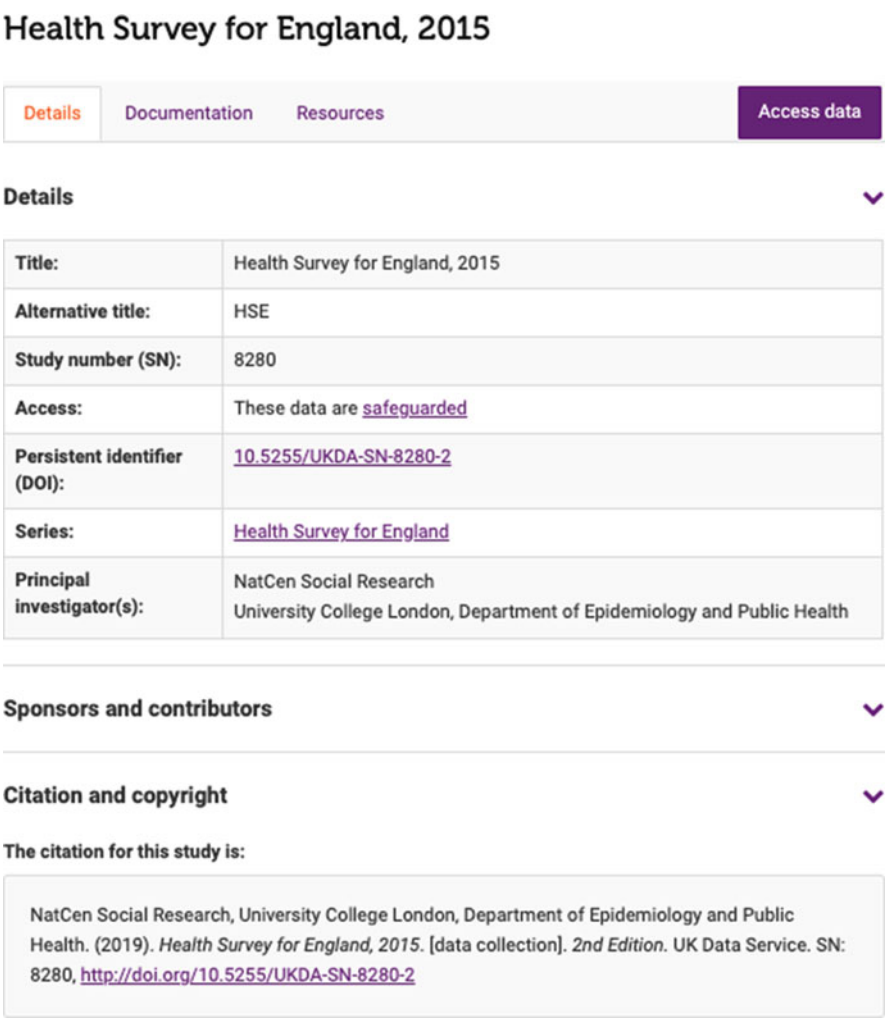

Fig. 3.2 Metadata record for a survey in the UK Data Service Discover catalogue

for England data files are 'safeguarded' and can be downloaded by all those who register with the service and agree to basic licence conditions. All access is free, because the UK Data Service is funded to provide free data access services and does not seek cost recovery.

Popular studies such as this are also available to 'Explore online' in Nesstar. Nesstar is the UK Data Service's online data browsing, analysis, subsetting and download tool that enables easy access to richly documented variables. Instant tabulation and graphing can be done (UK Data Service 2021e). Full question text, universe and routing information are typically displayed alongside variable name, code values and labels, and frequencies. Using Nesstar, a user can specify subsets

Table 3.3 Access policy: UK Data Service and ICPSR

Data sensitivity	ICPSR type	UK Data Service type	UK licence/agreement type	Access control
No identifying information	Open data	Open data	Open licence without any registration UK Open Government Licence (OGL) Creative Commons Attribution 4.0 International Licence (CC4.0)	Open, all uses allowed. Attribution/citation of data required
Identifying variables are treated, banded, aggregated or omitted	Public use files	Safeguarded data	End User Agreement	User agreement, user registered and authenticated, and, where appropriate, additional dataset specific conditions are agreed to
Deidentified sensitive data, such as geographic identifier or detailed occupational codes	Scientific use files	Controlled data	Bespoke secure access user agreement	User agreement, user registered and authenticated. User accreditation through training, project approval by a data access committee. Use within a restricted environment and scrutiny of research outputs

and download data tables in a range of formats. A frequency table is shown in Fig. 3.3 from the 2015 Health Survey for England, showing the wording and routing for the question, as well as the distribution of the variable. We can see that this dataset contains 8029 individuals who responded to the question.

The ability to browse data quickly is particularly useful when assessing whether a dataset might be appropriate for a research question. A researcher seeking to explore the characteristics of the subpopulation of those who had been made redundant in the previous 3 months might be concerned to be starting with such a small group as this.

In the US, a search on diabetes in the ICPSR catalogue brings up over 1316 results (ICPSR 2021a). Results can be filtered by subject, geography, data format, time period, restriction type and recency and as shown in Fig. 3.4.

ICPSR also has dedicated topical archives that are individually supported by government departments. Examples include the Health and Medical Care Archive (HMCA), the National Addiction & HIV Data Archive Programme (NAHDAP) and the Patient-Centered Outcomes Data Repository (PCODR) (ICPSR 2021b). ICPSR restricts access to its data to a paid membership in order to raise revenue necessary for its organisation.

Dataset: Health Survey for England, 2015

Health Survey for England, 1991-

Variable EverDi: Do you now have, or have you ever had diabetes?

LITERAL QUESTION

Do you now have, or have you ever had diabetes?

Values	Categories	N	
1	Yes	608	7.6%
2	No	7421	92.4%
-9	Refused	2	
-8	Don't know	3	
-1	Not applicable	5714	

SUMMARY STATISTICS

Valid cases	8029
Missing cases	5719
Minimum	1.0
Maximum	2.0

This variable is numeric

UNIVERSE

ASK ALL AGED 16+

Fig. 3.3 Health Survey for England 2015 frequency table for variable 'diabetes', Nesstar

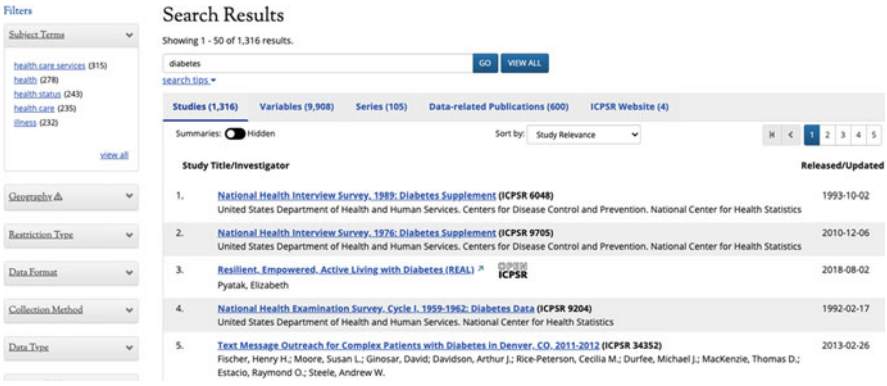

Fig. 3.4 Term search in the ICPSR data catalogue

3.5 Accessing Data

A robust data sharing philosophy ensures that access control measures are always proportionate to the kind of data, the level of confidentiality involved and the potential risk of disclosure. The guiding principle is that data are open where possible and closed when necessary. A data owner decides how to classify their data for appropriate access and licenses their data for onward use. Table 3.3 summarises US and UK licensing and access control safeguards in place.

The End User Licence has contractual force in law, in which users agree to certain conditions, such as not to disseminate any identifying or confidential information on individuals, households or organisations and not to use the data to attempt to obtain information relating specifically to an identifiable individual. Users can use data for research purposes or teaching and learning, but cannot publish or use them in a way that would disclose people or organisations' identities.

The 5 Safes framework is commonly used to permit trusted access to deidentified sensitive data (Office for National Statistics 2017). The safes represent safeguards on different dimensions:

- **Safe data:** data are treated to protect respondent confidentiality.
- **Safe projects:** researchers must justify why they need the data and demonstrate that their project fulfils a public good. Each project must be approved by the data owner.
- **Safe people:** through a signed Declaration and User Agreement researchers agree to the confidentiality obligations owed to the data and must successfully complete a standard face-to-face Safe Researcher Training (SRT) course.
- **Safe settings**: a Secure Lab (safe haven) offers remote access through a secure virtual private network via the researcher's own institutional desktop PC or via a Safe Room onsite at the UK Data Archive or a safe room with equivalent security.
- **Safe outputs**: researchers are trained to produce safe outputs during the SRT training and researchers must submit their outputs to trained staff to be checked for Statistical Disclosure Control before they can be released for publication.

Safe havens that operate the 5 Safes provide a confidential, protected virtual environment within which authorised researchers can access sensitive microdata. Examples in the UK are the ONS Secure Research Service (SRS) and the Secure Lab at the UK Data Service. Both the Trusted Research Environments (TRE) allow researchers to analyse the data remotely from their institution, on a central secure server, with access to familiar statistical software and office tools, such as Stata, R and Microsoft Office. Controlled data stored on encrypted machines, and no data travel over the network; instead the user's computer becomes a remote terminal, and outputs from the secure system are only released on approval. Remote access to restricted social and economic data has become more common and largely replaces the need to rely on researchers having to physically visit secure sites in person.

The formal protocols and agreements include: Data Protection Registration, accredited to the international standard for data security, ISO27001, Depositor Licences, and legally binding User Access Agreements. UKDA has in place technical and organisational secure data handling procedures, and all staff sign a non-disclosure agreement and are trained in data security.

3.5.1 Example 1: Accessing a Birth Cohort Survey Dataset for Research

In order to understand how a researcher might access longitudinal survey data to study child development, the example of the Millennium Cohort Study (MCS) is provided.

The MCS began its life in 2000 and is conducted by the Centre for Longitudinal Studies (CLS) at University College London (UCL). It aims to chart the conditions of social, economic and health advantages and disadvantages facing children born at the start of the twenty-first century. Data collection is/has been funded from a number of sources, including the Economic and Social Research Council (ESRC) and various government departments and agencies.

The study has been tracking the 'Millennium children' through their childhood years and plans to follow them into adulthood. The sample for the first MCS survey (MCS1) was drawn from all live births in the UK over approximately 12 months from 1 September 2000 in England and Wales and from 1 December 2000 in Scotland and Northern Ireland. It was selected from a random sample of electoral wards, disproportionately stratified to ensure adequate representation of all four UK countries, deprived areas and areas with high concentrations of Black and Asian families.

The first sweep (MCS1) interviewed both mothers and resident partners of infants included in the sample when the babies were 9 months old, and the second sweep (MCS2) was carried out with the same respondents when the children were 3 years of age. The third sweep (MCS3) was conducted in 2006, when the children were aged 5 years, the fourth sweep (MCS4) in 2008, when they were 7 years old, the fifth sweep (MCS5) in 2012–2013 at age 11 years and the sixth sweep in 2015/16 when the cohort members were 14 years old.

All available MCS survey data can be accessed through the UK Data Service, from the primary data collections from birth until the most recent data collection. The longitudinal family file contains basic information on outcomes by data collections and weights. Information on geography at the time of interview is also available, as is linked administrative data on educational outcomes. Comprehensive study and data documentation is freely available from the UKDS website and the CLS website too. A series of harmonised datasets are also available which combine key measures such as height, weight and BMI across five UK cohort studies. Use of the data is restricted to specific purposes after a simple but effective user registration. Use of the safeguard data requires an End User Licence to be signed. The controlled access data, such as fine grained geographical identifiers and linked administrative data, are restricted to use in the TRE.

DNA has been collected at age 14 years, from cohort members and resident biological parents, and access to genotyped data will follow the NCDS model using an independent Access Committee. CLS is not directly responsible for maintaining and updating the archives of biological samples and genetic information, but these are held in anonymised form by collaborating institutions.

MCS has established linkage with various geographic data (ward level (Administrative and CAS), Lower Super Output Area and Output Areas; Index of Multiple Deprivation Rural/Urban indicators have been linked at Lower Super Output Area) and consented linkage of educational records from the National Pupil Database (NPD) in England, Wales and Scotland. Work is ongoing on a variety of other linkages, such as health records and parent's economic records.

Collaboration with similar international studies like the 'Growing Up' studies (e.g. Ireland, Australia) offers comparable questions and opportunities for creating harmonised measures.

3.5.2 Example 2: Accessing Biomarker Data from a Longitudinal Survey

Understanding Society is one of the largest household panel studies in the world, collecting information annually from around 40,000 households. Biomarker data were collected during a health assessment carried out by a registered nurse in waves 2 and 3. During the health assessment visit, a wide range of physical measurements were taken from 20,000 eligible adults. A detailed list of the measurements taken can be found in Table 3.4 (Benzeval et al. 2016).

Many of the physical measurements collected are also found in many of the studies in Table 3.1, but one of the more unusual physical measurements in Understanding Society is digit length. This measurement was collected in the Understanding Society Innovation Panel. The Innovation Panel is a separate smaller sample of 1500 households, where new survey questions and methodologies are tested. In each wave, researchers can bid to have their experiments included in the annual questionnaire. In waves 6 and 7, the lengths of participants' index and ring

Table 3.4 Access routes for biomarker data in understanding society

Type of biomarker	Accessible via	Access level/application process
Physical/anthropomorphic measures	UK Data Service	End User Licence, downloadable after registering with the UKDS
Blood analytes	UK Data Service	End User Licence, downloadable after registering with the UKDS
Genetic and epigenetic data	European Genome-phenome Archive	Applications via EGPA, considered by Wellcome Trust Sanger Institute data access committee
Genetic and epigenetic data with survey data	Understanding Society health data team	Application and variable request forms, considered by understanding society
Further analysis of frozen blood samples	Professor Meena Kumari	Contact professor Kumari to discuss research proposal

Source: Accessing the data | Understanding Society

fingers were measured. These measurements are considered to be an indicator of exposure to hormones such as oestrogen and testosterone in utero (Jäckle et al. 2019). The level of hormone exposure has been linked with a number of personality traits (Fink et al. 2004).

In addition to the physical measurements, unfasted blood samples for which at least one biomarker was achieved were taken from 13,107 adults. Participants aged 16 and 17 years could consent to provide a blood sample although nurses were advised to discuss this with a parent as well. Those who were pregnant, had certain conditions or took anticoagulant medication were excluded. The blood samples were processed and analysed by Newcastle upon Tyne Hospitals NHS Foundations Trust (NUTH) under strict quality protocols and the data made available alongside the physical measurements and the main survey data (Benzeval et al. 2016). In addition to the blood analytes, genetic and epigenetic data were obtained and some samples frozen to allow for future analysis of DNA, plasma, serum and whole blood.

Biomarker data are made available to researchers who can demonstrate that their research is in the public interest. Similarly to social survey data, biomarker data are subject to different levels of access based on their sensitivity and disclosiveness. Access policies are discussed in more detail later in this chapter, but the access routes for the different types of biomarker data in Understanding Society are summarised in Table 3.4.

For the genetic and epigenetic data, researchers are required to complete an application process through either the European Genome-Phenome Archive or directly through the Understanding Society health data team. Researchers are required to submit a detailed research proposal which is then carefully considered by an expert panel who assess the public benefit and validity of the research as well as considering the potential for ethical issues and incidental findings.

3.5.3 Example 3: Accessing New Forms of Data in Biomedical Social Research

The growing availability of big data has led to a focus on developing solutions for powerful analytics combined with sustainable and secure ways of curating and delivering data. For research that demands large amounts of data and computationally challenging analyses, efficient storage, rapid exploration, visualisation and data linkage solutions at scale are needed. New architectures and infrastructures are needed as we move forward to accommodate increasing data types and volumes.

Medical records and hospital data will be updated on a regular basis, meaning that traditional repository solutions for downloading prepackaged bundles of data packages may no longer be useful. There are storage issues with potentially large amounts of data, and the limits of processing on a researcher's local PC. Access needs to be offered by bringing the users to the data, instead of users taking away data. This has not been the typical model for accessing survey resources, unless data

are protected in a safe haven. Slice and dice and aggregation methods are useful for reducing the size of data, for example, through aggregating or selecting by measures, geographical area or time period.

Examples of big data publishing platforms include scalable frameworks. Apache Hadoop is an open source framework that allows for the distributed processing of large datasets across clusters of computers using simple programming models (Apache Software Foundation 2021). Google BigQuery is Google's serverless, highly scalable, and cost-effective multicloud data warehouse (Google Cloud 2021). These solutions offer security infrastructures that can accommodate authentication (who you are), authorisation (what you're allowed to do) and encryption, all critical mechanisms to implement a robust data governance framework; vita for access to biomedical data that may be sensitive. Academic data centres in the social sciences are attempting to scale up their traditional repository services using Hadoop-type platforms (Bell et al. 2017).

Social science archives do not yet routinely provide access to their open data via APIs (Application Programming Interfaces). Yet, more and more public sector data sources are, such as the World Bank (World Bank 2021). In the health domain, the UK's NHS Digital provides a catalogue of APIs available to access a huge range of health indicators (NHS Digital 2021c).

When it comes to piecing together strands of data in the health field, there is an increased tension between safeguarding the privacy of peoples' information and reaping the benefits of research using powerful linked and matched sources. As the risk of identification increases so the need to enable safe and trusted access to linked data becomes essential, such as the use of the 5 Safes, described above. The UK's Data Ethics Framework sets out useful high-level principles for undertaking such data science ethically and has a useful workbook to accompany it, as set out in the case study below (Cabinet Office 2020).

In the USA, there has been a growing interest in using artificial intelligence (AI) to mine data from electronic health records (EHRs), and social media to predict an incapacitated person's preferences regarding their healthcare decisions. In the case of patients who do not have the capacity to make healthcare decisions, Lamanna (2018) proposes that AI can build on current tools that identify patient preferences, such as consenting to a given treatment, and can offer a step change in the power to predict these preferences. The computational work of his 'autonomy algorithm' inputs data about patients and derives as an output a confidence estimate for a patient's predicted healthcare-related decision. Ethical issues are raised by the use of the algorithm. First, machine confidence in a prediction does not mean that the person should choose the pathway. Second, as larger datasets become available and allow higher levels of predictive accuracy, should AI replace human decision-making, regardless of a patient's decision-making capacity?

With many more emerging big data algorithmic trials of this nature data, time will tell, whether this will happen or not.

3.6 Ways of Using Data

Smith (2008) reviewed the extent of secondary data analysis and quantitative methods more widely, in selected British education, sociology and social work journals. She found that while secondary analysis was not widespread in social work papers, 42% of the quantitative papers in education used secondary analysis compared with 75% of the quantitative papers in sociology. In economics, secondary analysis is core to most research practice. For epidemiology research, the long-term surveys described above have huge potential for insight, but the statistical methods of analysis require some expertise.

Types of uses of secondary data are summarised as follows (Corti and Thompson 2012).

3.6.1 Providing Description and Context

Data can help provide background for a study or contextualise a new study and its findings. Oyebode and Mindel (2013) reviewed government documents demonstrating the contribution of Health Survey for England data to every stage of the policy-making process in quantifying obesity in England.

3.6.2 Comparative Research, Restudy or Follow-Up

Comparative research can be undertaken across time or place. Comparison brings greater power to answer research questions, for example when data can be combined with data beyond its original sample or geographical limitations. Effort needs to be made to ensure that similar phenomena are compared when two or more separate studies are being used.

3.6.3 New Questions and Interpretations

This is the classic secondary analysis approach to resuing data, where new questions are asked of 'old data'. Walters (2015) used the four waves of data from the US National Longitudinal Study of Adolescent to Adult Health during the period 1994–2008, to see how school problems and anti-social attitudes in adolescent years affected adult criminal and substance abuse in early adulthood.

3.6.4 Replication or Validation of Published Work

While the scientific method is premised on replicability, most re-studies do not usually involve attempts to validate or undermine researchers' previous analyses. However, the pursuit of objective verification of results has demanded attention following more recent well-known cases of obviously fraudulent research in psychology, and the crisis of hidden results and publication bias in clinical trials reporting (Enserink 2012; Goldacre 2015a). The Reproducibility Project carried out independent replications of 100 studies in psychology, and preliminary results suggest that only 39 of the 100 key findings could be replicated (Baker 2015). In clinical trials, concerns about the concealment of results and publication bias have escalated, with journals like the *British Medical Journal* claiming that they will only publish trials that commit to sharing data on request (Loder and Grives 2015).

3.6.5 Research Design and Methodological Advancement

Well-documented descriptions of the research methods used in a former investigation can inform the design of a new study. Sampling methods, data collection and fieldwork strategies, and interview protocols are all used by study designers to follow the best practice. Tried and tested question wording used in national major surveys can be reused when designing local surveys to ensure comparability with national results. In instances where the information is available, researchers can exploit survey 'paradata' (data about how a survey was administered) to explore methodological issues, like non-response or interviewer effects.

3.6.6 Teaching and Learning

There is a need for students to engage with 'real' data, to obtain results which relate to the real world, and to tackle real data handling problems (Smith 2008). Real data is well suited to teaching substantive social science as well as facilitating the teaching of research methods and can really engage students. In the UK and US, efforts to improve statistical literacy among students of social science have created some useful resources to help students confront secondary data including those created by data services. These include user support and self-guided training and instructional materials, in the form of step-by-step guides, videos and short webinars (UK Data Service 2021f; ICPSR 2021c).

In the era of data-intensive research, researchers are increasingly seeing the benefits of working with data across disciplinary boundaries. As such, new ways of working and appreciating different data types and methods are needed. Skills for retrieving, assessing, manipulating, and analysing big data, as well as thinking

'algorithmically', become important. Training offers around the world are responding to this growing demand (Belmont Forum 2017; University of Cambridge 2015).

3.7 Research Examples Using Epidemiological Data

3.7.1 Example 1: Use of Birth Cohort Survey Data: Does Premature Birth Affect a child's Long-Term Health or Development?

Research question: Thanks to advances in modern medicine, the chance of babies surviving if they are born prematurely is high. However, there is a concern that children surviving such early births will suffer from ill-health and developmental effects. This research investigated whether a premature birth has these negative consequences on child development.

Data used: This research used data from the first (2001–2003), second (2003–2005) and third (2006) surveys of the Millennium Cohort Study (MCS) as well as a special dataset featuring birth registration and maternity hospital episode data (University of London, Institute of Education, Centre for Longitudinal Studies, 2017a, 2017b, 2017c). The MCS is a longitudinal survey of a cohort of around 19,000 children born across the UK between September 2000 and January 2002. Topics covered include family socio-economic background, the circumstances of pregnancy and birth, child health, child behaviour, childcare and parenting style.

Methods used: Data on gestational age were determined from the maternal report included in the first survey of the MCS and the data in the hospital records dataset. Groups of children born in one of four preterm gestational ages – early term (37–38 weeks), late preterm (34–36 weeks), moderately preterm (32–33 weeks), very preterm (32 weeks or less) – were compared with those children born at full-term (39–41 weeks). Logistic regression was conducted on each of the gestational age groups listed above. The analysis took into account the clustered study design of the survey.

Brief findings: The researchers discovered that the higher the prematurity, the greater risk of these ill effects. However, the differences between each group were small. Those children born late or moderately preterm were the most likely to have a higher disease burden at ages 3 and 5 years. Compared with full-term births, those born in late preterm or early term also had poorer health and educational outcomes at ages 3 and 5 years.

Source: Boyle et al. 2012.

3.7.2 *Example 2: Use of Biomarker Data: The Association Between Biological Health and Socio-Economic Position: Blood Analytes from Understanding Society*

The ability to access a combination of social survey data and biomarker data facilitates a wide range of research questions. Examples of research carried out using analytes from blood and saliva samples follow.

Research question: Biological health and socio-economic position are known to be associated, via the mechanism of the body's stress responses. Thus social disadvantage is considered to lead to higher biological health risk when other factors such as health behaviours and existing health conditions are controlled for. This study examines how to measure the biological changes that are related to socio-economic factors among different age groups.

Data used: This research used biomarker data collected via blood samples from Understanding Society to examine the association between social position and ageing. Understanding Society is a longitudinal household panel study that started in 2009 and has collected information on every member of around 40,000 households in the UK annually. It is a multi-topic study that includes the collection of a wide range of biomarkers.

Methods used: The team used key biomarkers from 9088 participants of the study to develop a Biological Health Score (BHS) which expanded upon the allostatic load (a long established measure of the wear and tear of key physiological systems due to stress responses) by including measures of liver and kidney function. Figure 3.5 shows how the BHS was created and the range of biomarkers that were included in the measure.

The score is calculated and interpreted as the higher the score, the greater the biological health risk. A total of 16 biomarkers were used and the study also included a range of covariates such as marital status, education, age, comorbidities, medication, health behaviours such as drinking and smoking, all of which were available in the Understanding Society data.

Brief findings: Differences in BHS were found for most of the covariates, for example smokers had higher BHS scores than non-smokers. When looking at measures of socio-economic positions, such as education, the study found that those in the lower education group had higher BHS scores with these associations stronger for the inflammatory and metabolic systems. Furthermore these associations were not explained by covariates such as health behaviours. The study concluded that there is an association between BHS and socio-economic position, with the greatest biological health risk found in those who were most disadvantaged.

Source: Karimi et al. 2019.

- The BHS is calculated using the distribution of the (n=16) biomarkers included in the study and targeting four physiological systems (as included in the allostatic load) and two organs:
- *Endocrine system* (n=1/2 biomarkers in women and men, respectively): dehydroepiandrosterone sulfate (DHEA-S) and testosterone (in men only).
- *Metabolic system* (n=4): glycated haemoglobin, high-density lipoprotein cholesterol (HDL), total cholesterol and triglycerides.
- *Cardiovascular system* (n=3): systolic and diastolic blood pressure, and pulses.
- *Inflammatory/Immune system* (n=3): C reactive protein, fibrinogen and insulin-like growth factor 1 (IGF-1).
- *Liver function* (n=3): alanine transaminase, aspartate transaminase and gamma glutamyltransferase.
- *Kidney function* (n=1): creatinine.
- For a given individual i , the BHS is calculated as the sum (across all 16 biomarkers) of binary variables indicating if that person belongs to the 'at-risk' quartiles of each biomarkers.

Fig. 3.5 Definition and calculation of the Biological Health Score (BHS)

3.7.3 Example 3: Use of Biomarker Data: Second Hand Smoking in Children: Saliva Samples from the Health Survey for England

Saliva samples also provide biomarker data and are often collected as part of biomarker and health surveys such as the Health Survey for England.

Research question: Second-hand tobacco smoke (passive smoking) affects children's lung function, subsequent lung function as adults and risk of chronic disease as adults. Children are considered at greater risk as they have faster respiratory rates, so take in proportionately more second-hand smoke than adults. In addition their developing organs are at greater risk from exposure to toxins and so exposure can lead to cancers in both childhood and adulthood, meningitis and cardiovascular disease.

Data used: The study used saliva samples collected from children aged 4–15 years as part of the Health Survey for England. The Health Survey for England is a repeated, annual survey which aims to survey the health of the population in England.

Methods used: From the saliva samples, measures of cotinine were obtained. Cotinine is a metabolite of nicotine and its presence is an indicator of recent exposure to tobacco and/or its smoke. It is considered the most useful biomarker in smoking-related studies due to its sensitivity to second-hand smoke exposure. It has a half-life in the body for 16–20 h and levels of 12 ng/mL or above indicate direct

smoking while levels between 0.1 ng/mL and below 12 ng/mL indicate exposure to second-hand smoke.

Brief findings: Exposure to second-hand smoke in children had decreased for boys. In 2011–2013 HSE data, 41% of boys found to have been exposed to tobacco smoke, by 2014–2015, this had decreased to 38%. For girls, the proportion exposed remained consistent across this period. As would be expected, exposure was higher among children living in households where at least one person smoked, but fewer than 9% of children lived in such households so higher exposure to second-hand smoke is confined to a decreasing proportion of the population.

Source: NHS Digital 2016.

3.8 Conclusion

Secondary analysis of biomedical data permits a range of valuable analyses to be undertaken quickly, effectively, transparently and with minimal respondent burden. Online access to data has simplified and speeded up access and digital formats enable users to easily consult rich documentation, explore and analyse data online and to make linkages between appropriate resources in a context of an increasingly complex data landscape.

The number of online data outlets has grown significantly over the past 5 years, but dedicated domain specific data services, like the UK Data Services and ICPSR have a role in helping set the high standard for high quality data publishing. As new and larger data types come on stream, so data services need to adapt, providing new platforms and new tools for selecting and querying data, alongside the traditional download of smaller datasets.

Perhaps the biggest challenge for established data services is in finding ways to describe effectively the underlying methods used to create these records, providing potential users with a fuller understanding of the provenance and meaning of readily available data. Here collaboration with microdata methodologists is beneficial, some of whom have already moved into this space.

This is a fast moving area with much to be resolved and at least as much potential for the researcher. However, we need to ensure that researchers and data services themselves are well equipped to deal with the challenges ahead, with a need for statistical, methodological and computational skills.

References

Australian Research Data Commons (2021) Research Data Australia. http://researchdata.ands.org.au/. Accessed 01 May 2021

Apache Software Foundation (2021) Apache Hadoop. https://hadoop.apache.org/. Accessed 20 May 2021

Baker M (2015) First results from psychology's largest reproducibility test. Nature News. 30 April 2015. http://www.nature.com/news/first-results-from-psychology-s-largest-reproducibility-test-1.17433. Accessed 01 May 2021

Bell, D, L'Hours H, Cunningham N et al (2017) Scaling up: digital data services for the social sciences. Case Study UK Data Service. https://www.ukdataservice.ac.uk/media/604995/ukds-case-studies-scaling-up.pdf. Accessed 20 May 2021

Belmont Forum (2017) Belmont Forum endorses curricula framework for data-intensive research. Belmont Forum news, Montevideo, Uruguay. http://www.belmontforum.org/news/belmont-forum-endorses-curricula-framework-for-data-intensive-research/. Accessed 01 May 2021

Benzeval M, Kumari M, Jones AM (2016) How do biomarkers and genetics contribute to understanding society? Health Econ 25:1219–1222. Accessed 01 May 2021. https://doi.org/10.1002/hec.3400

Blaxter B, Danchin A, Savakis B et al (2016) Reminder to deposit DNA sequences. Science:780. https://science.sciencemag.org/content/352/6287/780.1. Accessed 01 May 2021

Boyle EM, Poulsen G, Field DJ et al (2012) Effects of gestational age at birth on health outcomes at age 3 and 5 years of age: population based cohort study. Br Med J 344:e896. Accessed 01 May 2021. https://doi.org/10.1136/bmj.e896

Cabinet Office (2012) Open data white paper: unleashing the potential. UK government cabinet Office. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/78946/CM8353_acc.pdf. Accessed 01 May 2021

Cabinet Office (2020) Data Ethics Framework: Guidance. https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework. Accessed 20 May 2021

CESSDA (2021a) Members tools and services. Council of European Social Science Data Archives. https://www.cessda.eu/Tools-Services/For-Members. Accessed 01 May 2021

CESSDA (2021b) Data Catalogue. Council of European Social Science Data Archives. https://datacatalogue.cessda.eu. Accessed 01 May 2021

Connelly R, Platt L (2014) Cohort profile: UK millennium cohort study (MCS). Int J Epidemiol 43(6):1719–1725. Accessed 01 May 2021. https://doi.org/10.1093/ije/dyu001

Corti L, Thompson P (2012) Secondary analysis of archived data. In: Goodwin J (ed) SAGE secondary data analysis. Sage, London

Corti L, Van den Eynden V, Bishop L et al (2019) Managing and sharing research data: a guide to good practice, 2nd edn. Sage, London

Cox B (1995) Health and lifestyle survey: seven-year follow-up, 1991–1992: SN3279 [computer file]. UK data archive [distributor], Colchester. Accessed 01 May 2021

Data Archiving and Networked Services (2021) NARCIS: National Academic Research and Collaborations Information System. http://www.narcis.nl. Accessed 01 May 2021

Data.gov (2021) The home of the U.S. Government's open data. https://www.data.gov. Accessed 01 May 2021

Data.gov.uk (2021) Find open data. https://data.gov.uk. Accessed 01 May 2021

DBPedia (2021) DBPedia. https://www.dbpedia.org. Accessed 01 May 2021

DDI Alliance (2021) Document, discover and interoperate. http://www.ddialliance.org. Accessed 01 May 2021

Deutsches Institut für Wirtschaftsforschung (2021) German Socioeconomic Panel. https://www.diw.de/en/diw_01.c.615551.en/research_infrastructure__socio-economic_panel__soep.html. Accessed 01 May 2021

Elder G (1974) Children of the great depression: social change in life experience. University of Chicago Press, Chicago

Elliott J, Shepherd P (2006) Cohort profile: 1970 British Birth Cohort (BCS70). Int J Epidemiol 2006 Aug 35(4): 836–843. doi: https://doi.org/10.1093/ije/dyl174

Enserink M (2012) Diederik Stapel Under Investigation by Dutch Prosecutors. Science 2 October 2012. American Association for the Advancement of Science. http://news.sciencemag.org/scienceinsider/2012/10/diederik-stapel-under-investigat.html. Accessed 01 May 2021

Fink B, Manning JT, Neave N (2004) Second to fourth digit ratio and the 'big five' personality factors. Personal Individ Differ 37(3) Accessed 4 May 2021. https://doi.org/10.1016/j.paid.2003.09.018

Gilbert R, Lafferty R, Hagger-Johnson G et al (2018) GUILD: GUidance for information about linking datasets. J Public Health 40(1):191–198. https://academic.oup.com/jpubhealth/article/40/1/191/3091693. Accessed 20 May 2021

Goldacre B (2015a) Scientists are hoarding data and It's ruining medical research. BuzzFeed News. July 22, 2015. http://www.buzzfeed.com/bengoldacre/deworming-trials. Accessed 01 May 2021

Goldacre B (2015b) Scientists are hoarding data and It's ruining medical research. BuzzFeed News. July 22, 2015. http://www.buzzfeed.com/bengoldacre/deworming-trials. Accessed 01 May 2021

Google Cloud (2021) Google BigQuery. https://cloud.google.com/bigquery/. Accessed 20 May 2021

gov.uk (2021) Research and statistics: health. https://www.gov.uk/search/research-and-statistics?keywords=health&content_store_document_type=all_research_and_statistics&order=relevance. Accessed 01 May 2021

Hobcraft J (2016) ABCDE of biosocial science. Society Now 24. https://esrc.ukri.org/files/news-events-and-publications/publications/magazines/society-now/society-now-issue-24/. Accessed 01 May 2021

ICPSR (2015) ICPSR Collection Development Policy. http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/policies/colldev.html. Accessed 01 May 2021

ICPSR (2021a) Find data. https://www.icpsr.umich.edu/web/ICPSR/search/studies?q=diabetes. Accessed 01 May 2021

ICPSR (2021b) Thematic data collections. University of Michigan. https://www.icpsr.umich.edu/web/pages/ICPSR/thematic-collections.html. Accessed 01 May 2021

ICPSR (2021c) Teaching and learning. https://www.icpsr.umich.edu/web/pages/instructors. Accessed 01 May 2021

ICPSR (2021d) Guide to social science data preparation and archiving. https://www.icpsr.umich.edu/web/pages/deposit/guide. Accessed 01 May 2021

Institute for Social and Economic Research (2020) Understanding society. University of Essex. https://www.understandingsociety.ac.uk. Accessed 01 May 2021

Institute for Social and Economic Research (2021) British household panel study (BHPS). University of Essex. https://www.iser.essex.ac.uk/bhps. Accessed 01 May 2021

Institute for Social Research (2020) The panel study of income dynamics (PSID). University of Michigan. http://psidonline.isr.umich.edu. Accessed 01 May 2021

Jäckle A, Al Baghal T, Burton J et al. (2019) Understanding society the UK household longitudinal study innovation panel, waves 1-11, 6849_ip_waves1-11_user_manual_June_2019.v1 (understandingsociety.ac.uk). Accessed 02 May 2021

Karimi M, Castagné R, Delpierre C et al (2019) Early-life inequalities and biological ageing: a multisystem biological health score approach in understanding society. J Epidemiol Community Health 73:693–702. Accessed 26 May 2021. https://doi.org/10.1136/jech-2018-212010

Lamanna C (2018) Should artificial intelligence augment medical decision making? The case for an Autonomy Algorithm. AMA J Ethics 9(20):902–910. https://philpapers.org/rec/LAMSAI-3. Accessed 26 May 2021. https://doi.org/10.1136/bmj.h2373

Laney D (2001) 3D data management: controlling data volume, velocity, and variety' application delivery strategy. Meta Group Gartner Blogs. Accessed 6 Feb 2001

Loder E, Grives T (2015) The BMJ requires data sharing on request for all trials. BMJ 2015:350. http://www.bmj.com/content/350/bmj.h2373. Accessed 01 May 2021

Madrigal A (2009). Data.gov launches to mixed reviews. Wired. 21 May 2009. http://www.wired.com/2009/05/datagov-launches-to-mixed-reviews. Accessed 01 May 2021

Medical Research Council National Survey of Health and Development (2021) National Survey of Health and Development. http://www.nshd.mrc.ac.uk/. Accessed 01 May 2021

Natcen (2014) User guide for the health survey for England, 2014. UK Data Service http://doc.ukdataservice.ac.uk/doc/7919/mrdoc/pdf/7919_hse2014_user_guide.pdf. Accessed 01 May 2021

National Centre for Health Statistics (2021) About the National Health and Nutrition Examination Survey. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm. Accessed 01 May 2021

NHS Digital (2016) Health survey for England 2015 Children's smoking and exposure to other people's smoke. [ARCHIVED CONTENT] (nationalarchives.gov.uk) Accessed 01 May 2021

NHS Digital (2021a) Data and information. https://digital.nhs.uk/data-and-information. Accessed 21 May 2021

NHS Digital (2021b) Health Survey for England. https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england. Accessed 01 May 2021

NHS Digital (2021c) API catalogue https://digital.nhs.uk/developer/api-catalogue. Accessed 01 May 2021

Nurses' Health Study (2019) About Nurses' Health Study. https://www.nurseshealthstudy.org/about-nhs. Accessed 01 May 2021

OECD (2013) New data for understanding the human condition. https://www.oecd.org/sti/inno/new-data-for-understanding-the-human-condition.pdf. Accessed 01 May 2021

OECD (2018) OECD open government data report https://www.oecd.org/gov/digital-government/open-government-data.htm. Accessed 01 May 2021

Office for National Statistics (2017) The 'Five Safes'—Data Privacy at ONS. https://blog.ons.gov.uk/2017/01/27/the-five-safes-data-privacy-at-ons/. Accessed 01 May 2021

Office for National Statistics (2021) Main figures. https://www.ons.gov.uk/. Accessed 21 May 2021

Open Data Institute (2021) Certify your open data. https://certificates.theodi.org. Accessed 01 May 2021

Oyebode O, Mindel J (2013) Use of data from the health survey for England in obesity policy making and monitoring. Obes Rev 14(6):463–476. https://doi.org/10.1111/obr.12024

Power C, Elliott J (2006) Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol 35(1):34–34. Accessed 01 May 2021. https://doi.org/10.1093/ije/dyi183

Re3data.org (2021) Registry of research data repositories. http://www.re3data.org. Accessed 01 May 2021

Research Data Alliance (2021) About RDA. https://www.rd-alliance.org/about-rda. Accessed 01 May 2021

Richter L, Norris S, Pettifor J et al (2007) Cohort profile: Mandela's children: the 1990 birth to twenty study in South Africa. Int J Epidemiol 36(3):504–511. Accessed 01 May 2021. https://doi.org/10.1093/ije/dym016

Ruspini E (2002) Introduction to longitudinal research. Routledge, London

Schmidutz D (2016) Synopsis of policy-rules for collecting biomarkers in social surveys: field report on the collection of dried blood spot samples in SHARE. SERISS Deliverable. https://seriss.eu/wp-content/uploads/2016/12/SERISS_D6.10_Synopsis_DBS_Policy_Rules_2016.pdf. Accessed 20 May 2021

Shepherd P (2012) Millennium Cohort Study Ethical review and consent. https://cls.ucl.ac.uk/wp-content/uploads/2017/07/MCS-Ethical-review-and-consent-Shepherd-P-November-2012.pdf. Accessed 20 May 2021

Siva N (2008) 1000 genomes project. Nat Biotechnol 26:256. http://www.nature.com/scitable/content/1000-genomes-project-95670. Accessed 01 May 2021

Smith E (2008) Using secondary data in educational and social research. Oxford University Press, Oxford

Springer Nature (2021) Recommended data repositories. Sci Data. http://www.nature.com/sdata/data-policies/repositories. Accessed 01 May 2021

Steptoe A, Breeze E, Banks J et al (2013) Cohort profile: the English longitudinal study of ageing. Int J Epidemiol 42(6):1640–1648. Epub 2012 Nov 9. PMID: 23143611; PMCID: PMC3900867. Accessed 01 May 2021. https://doi.org/10.1093/ije/dys168

Tate R, Calderwood L, Dezateux C et al (2006) Mother's consent to linkage of survey data with her child's birth records in a multi-ethnic national cohort study. Int J Epidemiol 35:294–298. Accessed 01 May 2021. https://doi.org/10.1093/ije/dyi287

The City of New York (2021) NYC open data.Health https://data.cityofnewyork.us/browse?q=health. Accessed 01 May 2021

The Dataverse project (2021) The Dataverse project: Open source research data repository software. http://dataverse.org. Accessed 01 May 2021

The Royal Society and the British Academy (2018) Data management and use: Governance in the 21st century. https://royalsociety.org/topics-policy/projects/data-governance/. Accessed 01 May 2021

University College London (2021) Whitehall II history. UCL Department of Epidemiology and Public Health. https://www.ucl.ac.uk/epidemiology-health-care/research/epidemiology-and-public-health/research/whitehall-ii/background. Accessed 01 May 2021

Unicef (2021) The State of the World's Children. https://www.unicef.org/reports/state-worlds-children-2021. Accessed 01 May 2021

UK Data Service (2020) Document your data. University of Essex. https://www.ukdataservice.ac.uk/manage-data/document.aspx. Accessed 01 May 2021

UK Data Service (2021a) Selection and appraisal criteria. University of Essex. http://ukdataservice.ac.uk/media/455175/cd234-collections-appraisal.pdf. Accessed 01 May 2021

UK Data Service (2021b) Licence agreement. University of Essex. http://ukdataservice.ac.uk/deposit-data/how-to/regular-depositors/deposit. Accessed 01 May 2021

UK Data Service (2021c) Collection development policy. University of Essex. http://ukdataservice.ac.uk/media/398725/cd227-collectionsdevelopmentpolicy.pdf. Accessed 01 May 2021

UK Data Service (2021d) Discover. Data catalogue. University of Essex. https://beta.ukdataservice.ac.uk/datacatalogue/studies. Accessed 01 May 2021

UK Data Service (2021e) Nesstar catalogue. University of Essex. nesstar.ukdataservice.ac.uk. Accessed 01 May 2021

UK Data Service (2021f) User support materials. University of Essex. https://www.ukdataservice.ac.uk/use-data.aspx. Accessed 01 May 2021

UK Statistical Authority (2015) Administrative data quality assurance toolkit. https://www.statisticsauthority.gov.uk/archive/assessment/monitoring/administrative-data-and-official-statistics/quality-assurance-toolkit.pdf. Accessed 20 May 2021

University of Bristol (2021) Cohort profile. Avon longitudinal study of children and parents. http://www.bristol.ac.uk/alspac/researchers/cohort-profile/. Accessed 01 May 2021

University of Cambridge (2015) Preparing social scientists for the world of big data research features. 18 June 2015. http://www.cam.ac.uk/research/features/preparing-social-scientists-for-the-world-of-big-data#sthash.ahGTJzKB.dpuf. Accessed 01 May 2021

US National Academy of Sciences (2005) Expanding access to research data: reconciling risks and opportunities. National Academies of Science, Washington, DC

Vaillant G (2012) Triumphs of experience: the men of the Harvard Grant study. Belknap Press/Harvard University Press, Cambridge

Walters GD (2015) Criminal and substance involvement from adolescence to adulthood: precursors, mediators, and long-term effects. Justice Q 32:729–747

World Bank (2021) Data, for developers: about the API. https://datahelpdesk.worldbank.org/knowledgebase/topics/125589. Accessed 01 May 2021

Chapter 4
Data Warehousing of Life Science Data

Benjamin Kormeier and Klaus Hippe

Abstract Increasingly, scientist have begun to collect biological data in different information systems and database systems that are accessible via the internet, which offer a wide range of molecular and medical information. Regarding the human genome data, one important application of information systems is the reconstruction of molecular knowledge for life science data. In this review paper, we will discuss major problems in database integration and present an overview of important information systems. Furthermore, we will discuss the information reconstruction and visualization process based on that integrated life science data. These database integration tools will allow the prediction for instance of protein–protein networks and complex metabolic networks.

Keywords Data warehouse · Life science · Database integration

4.1 Introduction

The diverse research areas of molecular biology generate a variety of publicly available data stored in molecular biology databases. These databases are global via information systems and are mostly publicly available. In recent years, the number of molecular biology databases has increased exponentially. There are currently 1641 databases providing information from different categories (Rigden and Fernández 2020).

The importance of data integration has been known in bioinformatics for several years. Therefore, it is essential for scientists to analyze and process information from different and distributed systems. The molecular biological data has a high degree of semantic heterogeneity because the data comes from a series of experiments. In molecular biology, complex problems are tackled that rely on an immense and

B. Kormeier (✉) · K. Hippe
FH Bielefeld, University of Applied Sciences, Interaktion 1, Bielefeld, Germany
e-mail: bkormeie@techfak.uni-bielefeld.de

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_4

diverse amount of data. The number of databases and the data they contain are increasing steadily, which means that data distribution and high redundancy cannot be excluded. For these reasons, it is important to develop data warehouse systems for keeping consistent and non-redundant data.

4.1.1 Aims and Scope

The integration of life science and biological data from heterogeneous, autonomous, and distributed data sources is an important task in bioinformatics. The challenge is to integrate huge data sets regarding the large heterogeneity of the databases on the semantic and technical level (Kormeier 2010). Therefore, relevant integration approaches in the field of data warehouses as well as modeling and simulation software approaches will be introduced. We will focus on several widely used data warehouse approaches. Furthermore, some selected tools for modeling and visualizing of biochemical pathways will be presented in this review paper.

4.2 Molecular Database Integration

The integration of data sets from data sources with different heterogeneities is a challenge not only in the economy but also in research and science. Especially, in the life sciences, numerous biological datasets are experimentally generated, which have significant heterogeneities in various domains. The storage, deployment, and administration of these data are usually done by molecular biology databases. Usually, these databases are freely available, distributed worldwide, and linked together by explicit cross-references. There are also significant differences in the structuring of data, accessibility, and copyright. One aspect of bioinformatics is the implementation of applications with which help an effective data integration of molecular biology databases are made possible. The goal of the data integration is to realize a database that has a uniform data structure and provides all the necessary data from the data sources. The data sources usually have different schemas, which is why schema transformation and schema integration are necessary. After that, the actual integration of the data stocks from the respective data sources takes place. The data is analyzed and validated so that inconsistencies and duplicates are identified and eliminated. During this data cleanup, merging and completion of incomplete data sets can also be done. As a result of this data fusion, a complete data set is realized to provide more information than the original data records from the data sources. The resulting consistent and structured database enables an efficient and global view of all data sources from the data sources. However, the merging of databases from different data sources is linked to three basic problems that will be described in the following sections.

4.2.1 Distribution, Autonomy, and Heterogeneity

With the help of specific software solutions, the integration of data from different data sources is realized. Such systems usually have different integration architectures, which successfully overcome the three basic problems of data integration. The distribution, autonomy and heterogeneity of a data source represent these basic problems and are also described as an orthogonal dimension of data integration (Leser and Naumann 2007).

One problem that needs to be addressed in data integration is the global distribution of data sources. Usually, the databases are provided by different systems and are geographically distributed. Because of this different localization of the data, a distinction is made between the physical and the logical distribution. With the help of a materialized integration architecture, the problem of physical distribution can be overcome. The provision of metadata and data cleansing methods by the integration system enables the removal of the logical distribution.

The autonomy of a data source is usually unavoidable, because the responsible organization of a data source usually uses its own development strategies and technologies. The term autonomy in connection with the data integration means that the data source can autonomously decide on the provision, the access possibilities, and the copyright of the data. In addition, autonomy is responsible for different problems of heterogeneity. In Conrad (1997), the different types of autonomy are discussed in detail.

The main problem that needs to be addressed in data integration is heterogeneity. If two information systems do not provide identical methods, models, and structures for accessing the database, these are called heterogeneous. Different kinds of heterogeneity are defined according to (Leser and Naumann 2007) as follows: technical heterogeneity, syntactic heterogeneity, data model heterogeneity, structural heterogeneity, schematic heterogeneity, semantic heterogeneity (Kormeier 2010). Autonomy is primarily responsible for heterogeneity, but distribution can also create heterogeneity. It is possible to force specific properties to be homogenous by restricting autonomy of a data source. This can be achieved by standards in exchange formats, interfaces, and protocols.

4.2.2 Approaches of Database Integration

The development of an integrated database system is a complex task, particularly, when a large number of heterogeneous databases have to be integrated. Hence, an elaborate blueprint of the architecture of the system is essential. However, another non-trivial problem is the availability of databases that should be integrated. Generally, there exists two architectures for integration. They are divided into materialized integration and virtual integration (Kormeier 2010). The main difference between the two integration architectures is the location of the relevant databases

during integration. A materialized integration architecture is a central and persistent database and copies all the necessary data from the data sources into the database. In contrast, a virtual integration architecture does not have such a database and therefore does not copy any data. Therefore, the integrated and homogenous data set of a virtual integration architecture only exists virtually and has to be realized again for all requests. However, there are also hybrid architectures that have materialized and virtual data sets.

Different approaches of database integration have been frequently discussed and reviewed since the beginning of the millennium. The most important are the following three approaches besides data warehouses:

- Hypertext navigation systems. HTML frontends linked to molecular biological databases.
- Federated database systems and mediator-based systems. Virtual integration does not store any data in a global schema. Federated systems integrate multiple autonomous database systems into a virtual single federated database. Usually, each database is interconnected via a computer network. The databases may be geographically decentralized. In comparison to federated database systems, multi-database systems do not have a global schema, rather these systems interactively generate queries for several databases at the same time (Kormeier 2010).

4.2.3 Data Warehouses (DWH)

In this section, we want to have a closer look at data warehouses. Data warehouses are one of the widely used architectures of materialized integration. Usually, data warehouses are used in the field of information management. In particular data analysis, data mining and long-term storage of business intelligence in companies are the major advantages of data warehouse systems. In bioinformatics data, warehouses are usually used for data integration (Kormeier 2010). There is no consistent definition of the DWH term. While different consortia such as the OLAP Council are trying to standardize the DWH term, the first definition was given by Inmon (1996):

> *A data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data support of management's decision.*

Only through the data warehouse process can a data warehouse system accomplish the various issues. This dynamic process is responsible for the acquisition,

storage and analysis of the data. The data warehouse process can be divided into the following four steps:

1. In the first step, the component extraction, transformation and loading are used, which are summarized under the term ETL components (Günzel and Bauer 2009). This step is called the ETL process and is responsible for extracting the data sets from the data sources and transforming them. Furthermore, the ETL process is responsible for loading the structured data into the DWH.
2. The persistent storage of data in DWH will be realized in the next step. However, some analyzes or specialized applications do not need all the data, so that can be realized in the so-called data marts.
3. These data marts represent a specific view of the DWH and are created in the third step.
4. The analysis and evaluation of the databases take place in the last step. The results are then provided to the different applications.

The key benefits of the materialized integration architecture are efficient data cleansing, unrestricted query capabilities, and good query performance. The disadvantage of this integration architecture may under certain circumstances be the timeliness of the database. However, this aspect always has to be considered in the context of the respective analysis or question, because not every topic needs up-to-date data. The relevance of the data is particularly important for complex analyzes of the financial markets. In the context of molecular biology research, updating the database every quarter is sufficient. The data sources are usually molecular biology databases and their updating is usually done every quarter.

4.3 Related Data Integration Approaches

In this chapter, relevant integration approaches in the field of data warehouses will be introduced. Furthermore, related visualization approaches for molecular networks and life science data will be discussed.

4.3.1 Data Integration Approaches

In the literature, data integration approaches in bioinformatics are divided into the following classes (Leser and Naumann 2007):

- Indexing systems: SRS (Sequence Retrieval System) (Etzold et al. 1996), Entrez (Kaps et al. 2006), and BioRS (Wheeler et al. 2004; Maglott et al. 2007).
- Multi-databases: OPM (Object Protocol Model) (Chen and Markowitz 1995), DiscoveryLink (Haas et al. 2001), and BioKleisli (Davidson et al. 1997).

- Ontology-based integration: TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) (Stevens et al. 2000), ONDEX (Köhler et al. 2006), and CoryneRegNet (Pauling et al. 2012).
- Data warehouse: Atlas (Shah et al. 2005), BioWarehouse (Lee et al. 2006), Columba (Trissl et al. 2005), Biozon (Birkland and Yona 2006), Booly (Do et al. 2010), JBioWH (Vera et al. 2013), Unison (Hart and Mukhyala 2009), and SYSTOMONAS (Choi et al. 2007). However, under certain aspects, CoryneRegNet, and ONDEX can also be assigned to this category.

Due to the already mentioned advantages (e.g., performance, availability of data, and simple conception), the data warehouse technology has become established in bioinformatics. Most of the applications were developed for specific molecular-biological questions, which means that they could not be used in other projects and their questions, or only through extensive extensions of the respective software solution. Atlas, BioWarehouse, Columba, ONDEX, and CoryneRegNet use the data warehouse technique for data integration, whereas CoryneRegNet and ONDEX provide a web service. Atlas, BioWarehouse, and ONDEX provide a software infrastructure for data integration, rather Columba, CoryneRegNet. They provide a web interface and therefore they are directly useable (Kormeier 2010). In addition, the database of many systems is out of date or no longer available, so important information is not available to the user. In particular, the complexity and flexibility of the respective software as well as the attitude of the project financing are responsible for it. In recent years, a plenty of systems have been implemented and made available to the user.

4.4 Data Warehouse for Life Science Data Warehouse

In the previous sections of this chapter, several principles and approaches for database integration and network visualization were introduced. A couple of the principles of the introduced integration systems are well suited to be used within the database integration system that will be presented. Particularly the functions of the software toolkit BioDWH (Töpel et al. 2008) will be illustrated. Furthermore, DAWIS-M.D. 2.0 (Hippe et al. 2011), a web-based data warehouse approaches based on the BioDWH integration toolkit, will be described.

4.4.1 BioDWH: A General Data Warehouse Infrastructure for Life Science Data Integration

BioDWH is an open source software toolkit, which can be used as a general infrastructure for integrative bioinformatics research and development. The advantages of the approach are realized by using a Java-based system architecture and

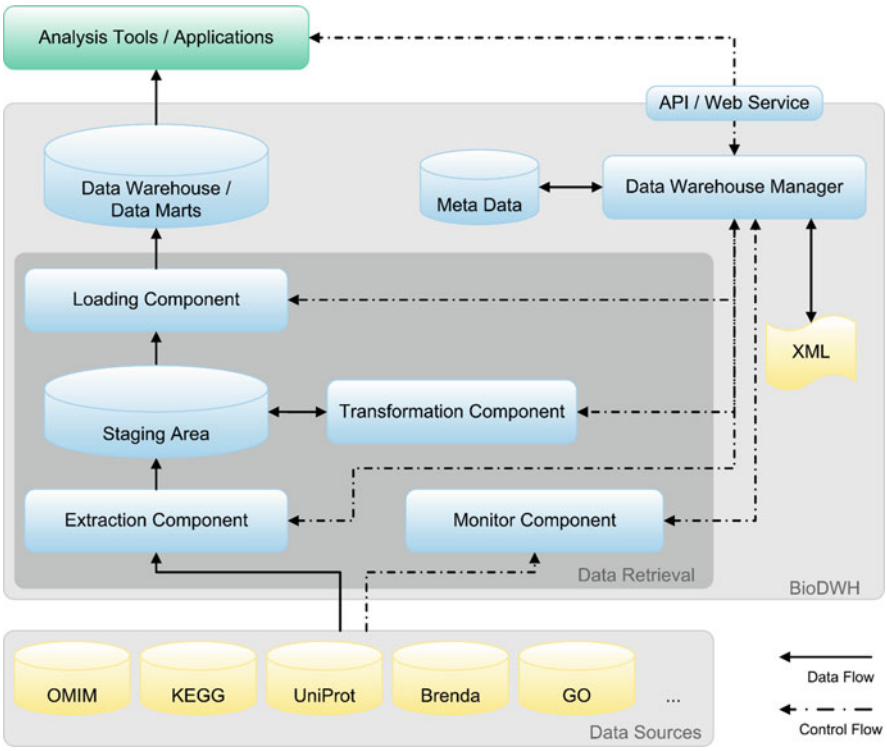

Fig. 4.1 Schematic illustration of the BioDWH system architecture following the general data warehouse design (Hippe 2014)

object-relational mapping (ORM) technology (Kormeier 2010). Figure 4.1 shows the reference architecture of a data warehouse system. This architecture is the foundation of the system architecture of BioDWH. Basically, the system consists of the Data Retrieval module, the Data Warehouse Manager, and a Graphical User Interface (GUI). The user is able to control the infrastructure via GUI and XML configuration files. Core of the system is the Data Retrieval component that is composed of Loading, Transformation, Extraction Component, i.e., the Parser (ETL), Monitor component, and the Staging Area. The parser library provides a large number of ready-to-use-parsers for biological and life science databases which are available, such as UniProt, KEGG, OMIM, GO, Enzyme, BRENDA, OMIM, Reactome, iProClass, and more. Using the BioDWHParser interface, it is easy to create own tailored parser. To achieve independence from the RDBMS, a persistence layer is necessary. Therefore, a well-engineered object-relational mapping framework called Hibernate was used as a persistence layer. Hibernate performs well and is independent from manufacturers like MySQL, PostgreSQL, or Oracle. Thus, the Hibernate framework fits perfectly into the infrastructure of the BioDWH application.

The system is realized as a Java-based open source application that is supported on different platforms with an installed Java Runtime Environment (JRE). Nowadays, Java is very popular and usually installed on most of the computers. Additionally, Java is available on most platforms such as Windows, Linux, and MacOS. Thus, Java applications have a high degree of platform independence. Moreover, Java applications over flexible software solutions that can be provided to a large audience. In this way the software solutions can become widely used (Kormeier 2010).

Another feature of BioWH is an implemented easy-to-use Project Wizard that supports the user or administrator to configure a DWH integration process in four steps. No additional knowledge in database systems or computer science is necessary. The whole configuration starting from database connection settings, via parser configuration to monitor configuration, is supported by the graphical user interface. In background the BioDWH infrastructure is running with multiple threads which means it is possible to run several download processes, uncompress processes, or integration processes in parallel (Kormeier 2010). Finally, a logging mechanism watches the integration process and starts a simple recovery process to guarantee a consistent state of the data warehouse.

4.4.2 DAWIS-M.D. 2.0: A Data Warehouse System for Metabolic Data

One of the major challenges in bioinfomatics is the integration and management of data from different sources and their presentation in a user-friendly format. DAWIS-M.D. 2.0 is a platform-independent data warehouse information system for metabolic data. The information system integrates data from 13 widely used life science databases (KEGG, EMBL-Bank, Transfac, Transpath, SCOP, JASPAR, EPD, UniProt, HPRD, GO, BRENDA, ENZYME, and OMIM). The information of integrated databases is divided into 13 various biological domains (Compound, Disease, Drug, Enzyme, Gene, Gene Ontology, Genome, Glycan, Pathway, Protein, Reaction, Reactant Pair, and Transcription Factor), which are available via the graphical user interface of the web application. The data warehouse architecture (Fig. 4.2) provides a platform-independent web interface that can be used with any common web browser. The system enables intuitive search of integrated life science data, simple navigation to related information as well as visualization of biological domains and their relationships. To ensure maximum up-to-dateness of the integrated data the BioDWH data warehouse infrastructure including a monitor component is used. The persistence layer of DAWIS-M.D. 2.0 uses the ORM technique, whereby the application layer is independent from database layer. Thereby, it is possible to support different database management systems. The DAWIS-M.D. 2.0 data warehouse incorporates the advantages of a navigation and informational system and builds a bridge to the network editor approach VANESA

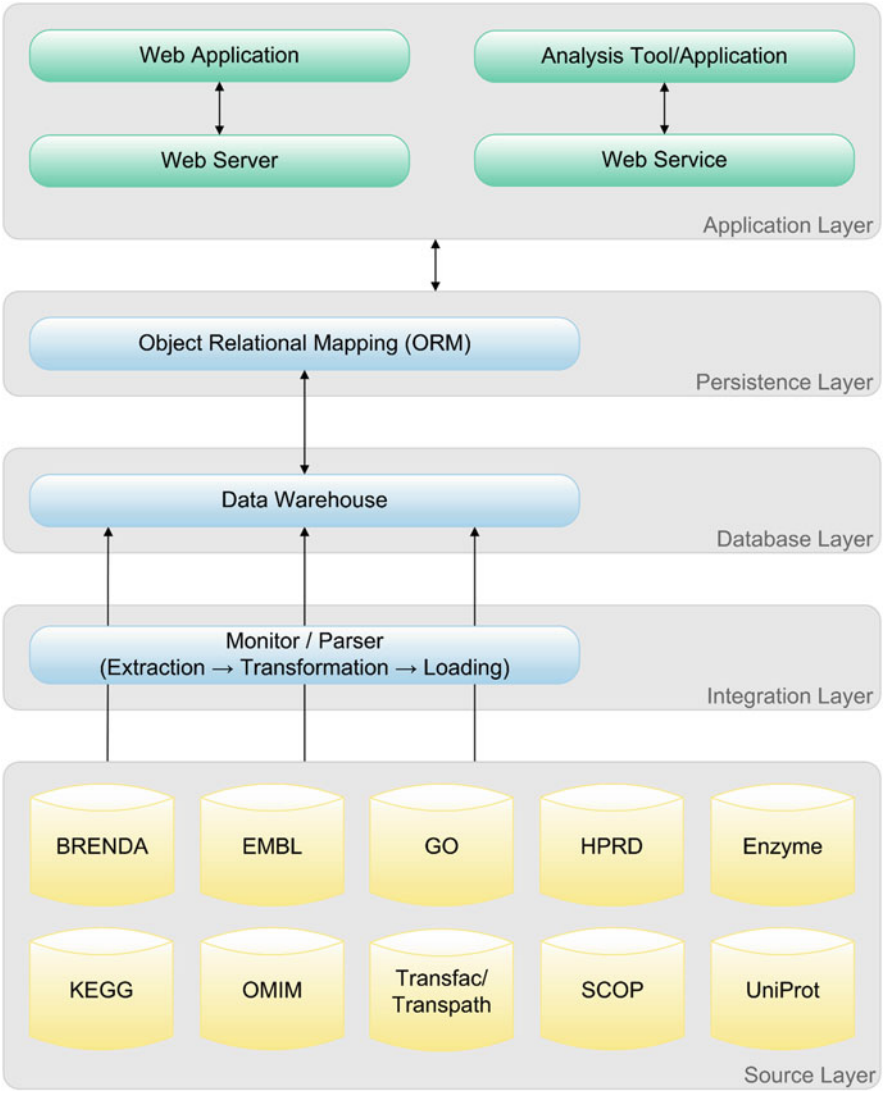

Fig. 4.2 Schematic representation of the DAWIS-M.D. n-layer system architecture from the original heterogeneous data sources to the web application layer (Hippe 2014)

(Brinkrolf et al. 2014). Hence, it is possible to browse through the integrated life science data and bring the information into a modeling and visualization environment. Therefore, it is easy for the scientists to search information of interest, find relationships and interactions between different biomedical domains and bring them for editing, manipulation, and analyzing directly into the VANESA network editor. Finally, the scientists gain a better understanding of complex biological

problems and are able to develop new theoretical models for further experiments. DAWIS-M.D. 2.0 is available at https://agbi.techfak.uni-bielefeld.de/DAWISMD/ (Hippe et al. 2011).

4.5 Summary

Different research domains of life sciences by different experimental methods generate an immense and diverse amount of data. Usually, such data is stored in database systems. For a comprehensive and efficient usage of the data, it is necessary to integrate the distributed and heterogeneous data and provide them for further analysis to the researcher. Moreover, the user needs to be supported by applicable tools for navigation within the integrated data sets that support an efficient and precise processing of the data. The number of molecular databases has been continuously increasing in the last decade (Töpel et al. 2008). Today, approximately 1641 publicly available databases and information systems for life science data are listed in the NAR catalogue (Rigden and Fernández 2020). This is mainly due to technological progress and computer-aided laboratory automation.

Transparency, integrity, semantic correctness, and non-redundancy are classical requirements of integration and therefore very important. However, other requirements gain importance in life science data integration, such as an efficient access to the increasing amount of data which should be, but is not always up-to-date. Furthermore, solutions for complex and changing schemata in life science data are required. Hence, the challenge was to combine diverse and multiple data and to bring them into a homogenous, consistent state. The new system should be flexible and also applicable in general for any other project. For that purpose, BioDWH data warehouse software kit is developed as a Java-based open source application for building life science data warehouses using common relational database management systems. By using the object-relational mapping (ORM) technology, it is no longer necessary to select the local database management system based on the restrictions of the integration software. BioDWH provides a number of ready-to-use parsers to extract data from public life science data sources and to store the information in a data warehouse. The integration process is supported by an easy-to-use graphical user interface that makes it possible to integrate any supported database in a few steps into a local database (Töpel et al. 2008).

DAWIS-M.D. 2.0 is a publicly available web-based system that integrates data from 13 different biomedical databases and divided the integrated data from the different data sources into 13 biomedical domains (Hippe et al. 2011). This data warehouse information system provides an integrated and consistent view of large-scale biomedical data. Additionally, relationships and interactions between multiple data sets and biomedical domains are identified and displayed (Janowski 2013). The advantages of the DAWIS-M.D. 2.0 application are the usability, performance, high level of platform independence, and wide range of life sciences information and biological knowledge (Hippe et al. 2011). Furthermore, the system is connected by

a "remote-control" to the VANESA network editor to easily visualize and analyze biological networks from data of interest.

Software solutions that provide visualization, analysis services, and an information management framework are in high demand among scientist as already discussed. It is not surprising that many groups over the world have contributed to the task of developing such software frameworks. Therefore, a DWH system to search integrated life science data and simple navigation called DAWIS-M.D. 2.0 as a base for a modeling and visualization system called VANESA were implemented (Hippe et al. 2011).

In conclusion, in this chapter, we presents a powerful and flexible data warehouse infrastructure BioDWH that can be used for building project-specific information systems, such as DAWIS-M.D. 2.0. Finally, the system was the basis for network modeling and pathway reconstruction in different scientific projects. The presented applications are in use since more than one decade within several projects as well as in ongoing in-house projects.

Acknowledgments This work could not have been carried out without the support of many people at Bioinformatics/Medical Informatics Department of Bielefeld University.

References

Birkland A, Yona G (2006) BIOZON: a system for unification, management and analysis of heterogeneous biological data. BMC Bioinform 7(1):70

Brinkrolf C, Janowski SJ, Kormeier B, Lewinski M, Hippe K, Borck D, Hofestädt R (2014) VANESA—a software application for the visualization and analysis of networks in system biology applications. J Integr Bioinform 11(2):239

Chen IMA, Markowitz VM (1995) An overview of the object protocol model (opm) and the opm data management tools. Inf Syst 20(5):393418

Choi C, Munch R, Leupold S, Klein J, Siegel I, Thielen B, Benkert B, Kucklick M, Schobert M, Barthelmes J, Ebeling C, Haddad I, Scheer M, Grote A, Hiller K, Bunk B, Schreiber K, Retter I, Schomburg D, Jahn D (2007) SYSTOMONASan integrated database for systems biology analysis of pseudomonas. Nucleic Acids Res 35:D533537

Conrad S (1997) Föderierte Datenbanksysteme—Konzepte der Datenintegration. Springer, Berlin

Davidson SB, Overton GC, Tannen V, Wong L (1997) BioKleisli: a digital library for biomedical researchers. Int J Digit Libr 1(1):36–53

Do L, Esteves F, Karten H, Bier E (2010) Booly: a new data integration platform. BMC Bioinform 11(1):513

Etzold T, Ulyanov A, Argos P (1996) SRS: information retrieval system for molecular biology data banks. Methods Enzymol 266:114128

Günzel H, Bauer A (2009) Data-Warehouse-Systeme. dpunkt.verlag, Heidelberg

Haas LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC (2001) Discoverylink: a system for integrated access to life sciences data sources. IBM Syst J 40(2):489511

Hart RK, Mukhyala K (2009) Unison: an integrated platform for computational biology discovery. In: Pacific symposium on biocomputing, pp 403–414

Hippe K (2014) Identifikation von potenziellen Transkriptionsfaktorbindestellen in Nukleotidsequenzen basierend auf einem Data-Warehouse-System. Bielefeld University, Bielefeld

Hippe K, Kormeier B, Janowski SJ, Töpel T, Hofestädt R, DAWIS-M.D. (2011) 2.0—a data warehouse information system for metabolic data. In: Proceedings of the 7th International Symposium on Integrative Bioinformatics

Inmon WH (1996) Building the data warehouse. Wiley, Indianapolis

Janowski SJ (2013) VANESA—a bioinformatics software application for the modeling, visualization, analysis, and simulation of biological networks in systems biology applications. Bielefeld University, Bielefeld

Kaps A, Dyshlevoi K, Heumann K, Jost R, Kontodinas I, Wolff M, Hani J (2006) The BioRS(TM) integration and retrieval system: an open system for distributed data integration. J Integr Bioinform 3(2)

Köhler J, Baumbach J, Taubert J, Specht M, Skusa A, Rüegg A, Rawlings C, Verrier P, Philippi S (2006) Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics 22:13831390

Kormeier B (2010) Semi-automated reconstruction of biological networks based on a life science data warehouse. Bielefeld University, Bielefeld

Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DW, Tenenbaum JD, Karp PD (2006) BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinform 7:170

Leser U, Naumann F (2007) Informationsintegration. dpunkt Verlag, Heidelberg

Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez gene: gene-centered information at NCBI. Nucleic Acids Res 35(suppl 1):D26–D31

Pauling J, Röttger R, Tauch A, Azevedo V, Baumbach J (2012) Coryneregnet 6.0—updated database content, new analysis methods and novel features focusing on community demands. Nucleic Acids Res 40:D610–D614

Rigden DJ, Fernández XM (2020) The 2021 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Res 49(D1):D1–D9. https://doi.org/10.1093/nar/gkaa1216

Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF (2005) Atlas—a data warehouse for integrative bioinformatics. BMC Bioinform 6:34

Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, Brass A (2000) TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 16:184185

Töpel T, Kormeier B, Klassen A, Hofestädt R (2008) BioDWH: a data warehouse kit for life science data integration. J Integr Bioinform 5(2):93

Trissl S, Rother K, Müller H, Steinke T, Koch I, Preissner R, Frömmel C, Leser U (2005) Columba: an integrated database of proteins, structures, and annotations. BMC Bioinform 6:81

Vera R, Perez-Riverol Y, Perez S, Ligeti B, Kertész-Farkas A, Pongor S (2013) JBioWH: an open-source Java framework for bioinformatics data integration. Database 2013:bat051

Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD et al (2004) Database resources of the National Center for biotechnology information: update. Nucleic Acids Res 32(suppl 1):D35–D40

Chapter 5
Automation in Graph-Based Data Integration and Mapping

Marcel Friedrichs

Abstract Data integration plays a vital role in scientific research. In biomedical research, the OMICS fields have shown the need for increasingly larger datasets, like proteomics, pharmacogenomics, and even newer fields like foodomics. In 2019 Nucleic Acids Research counted 1637 databases, accounting only for a fraction of all data sources available online. Data integration efforts need to process large amounts of heterogeneous data from different file formats ranging from simple files to complex relational databases and increasingly graph databases. Aside from data formats, availability is another obstacle. Whether files are available for direct download, need a user account, or are available only through an application programming interface (API). Keeping data sources up-to-date is important to make use of the latest discoveries in the respective fields, retrieve error corrections, and potentially mitigate issues with other data sources referencing newly added entities. Finally, all data sources provide information on certain entities and in most cases make use of specific identification systems. In the best case, data sources provide cross-references to other data sources. In order to generate robust mappings between all required data sources, identifiers of good quality need to be selected forming new connections between the entities. All of these vital steps and issues of data integration and mapping benefit from automation and are in most parts able to be fully automated. Workflow systems and integration tools are capable of automating different elements of the aforementioned steps and require varying levels of computer science skills. This chapter describes these issues, and the potential of the fully automated, graph-based data integration and mapping tool BioDWH2 is explored.

Keywords Data warehouse · Data integration · Graph database · Software tools

M. Friedrichs (✉)
Faculty of Technology, Bioinformatics/Medical Informatics Department, Bielefeld University, Bielefeld, Germany
e-mail: mfriedrichs@techfak.uni-bielefeld.de

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_5

5.1 Introduction

Data integration plays a vital role in scientific research analyses. Advancements in biomedical research gave rise to the OMICS fields starting with genomics, transcriptomics, and proteomics. The list of new OMICS fields has increased dramatically with additions such as pharmacogenomics, foodomics, and antibody-omics. Each of these fields requires its own data, experiments, and new databases increasing the overall complexity of available data sources and effort needed to keep information up-to-date. In 2018 Imker conducted a survey of published databases in the Nucleic Acids Research (NAR) database issues and concluded that 1700 databases were covered in 25 years (Imker, 2018). The 2021 NAR database issue added 90 new resources and with updates and removals now count 1641 databases (Rigden and Fernández, 2020). The Online Bioinformatics Resources Collection (OBRC) contained 1542 bioinformatics databases and other resources (Chen et al., 2007) which has grown to 2417 as of July 2021. These numbers only represent the resources added to common registries resulting in a likely larger number of databases available online.

For the use-case of medical information systems, multiple OMICS levels are relevant in drug therapy safety (Kapoor et al., 2016; Qian et al., 2019). Where previously the main focus of analyses were interaction networks of drugs, diseases, and side effects, the growing opportunities of molecular information in the clinical context (Krier et al., 2016; Sanderson et al., 2019) add new OMICS fields in the form of genes, variants, pathways, RNA regulation, and many more. Examples would be the "PharmGKB" (Whirl-Carrillo et al., 2012), "DrugCentral" (Avram et al., 2020), "DrugBank" (Wishart, 2006), and "OMIM" (Online mendelian inheritance in man, 2021) databases. The integration and mapping of this information could provide an in-depth understanding of individual patient cases and reduce adverse drug reactions toward personalized medicine.

This growing complexity increases lead time of research projects as users need to analyze data sources with heterogeneous file formats, availability, and information schemata. Much of these issues benefit from integration pipelines and tools which are easy to use and take care of data warehousing and mapping tasks. With the growing adoption of graph databases and formats (Fabregat et al., 2018; Hassani-Pak et al., 2016; Shoshi et al., 2018; Yoon et al., 2017), the transformation of heterogeneous data sources into a common graph data structure is beneficial in representing complex and highly connected biological information. While data warehousing solutions provide users with all data sources in a single database, the information is still loosely coupled. Most data sources provide identification systems or external references for their data. However, changes in referenced data sources are not immediately propagated and might lead to loss of information, and data sources need to be constantly updated. Finally, automated mapping techniques are important in building tightly coupled relationships between data sources in a data warehouse. While these mapped relationships may never cover all available

information, they build a starting ground for research analyses and enable the discovery of new and potentially meaningful information.

This chapter describes the problems and solutions of data integration and information mapping and closes with a possible solution using the open-source BioDWH2 tools.

5.2 Data Integration and Mapping

Different data integration approaches have been developed in the past decades. As with many architectural problems, each comes with their own set of advantages and disadvantages (Schwinn and Schelp, 2005). The approaches differ in a multitude of aspects, such as heterogeneity, distribution, access, redundancy, technology, and more. This section will first look at federated databases, data warehouses, and data lakes under the aforementioned aspects. Afterwards, the role of mapping strategies for these approaches is explored.

5.2.1 Federated Database System

Arguably the simplest approach to implement is federated database systems (FDBS). A FDBS consists of multiple, independent component databases which are directly accessed by the FDBS. There are no restrictions on the location or technology of the component databases. The only exception is that the FDBS needs access via any means of local or remote communication. The access may be restricted using credentials which need to be stored in the federated database management system. FDBS can be divided into loosely and tightly coupled systems.

Loosely coupled FDBS give the user direct access to the component database schemas. The advantage is a minimal overhead in administration of the database system and new schema additions of the component databases are directly available to the user. However, the users need to understand and process the schema and heterogeneity of the component databases themselves which may result in redundant work.

Tightly coupled FDBS mitigate this problem by introducing schema transformations and views on the component databases. Heterogeneous information from different component databases can be normalized and provided to the user for direct use. Additionally, selecting and filtering the raw data into qualitative subsets is possible by providing schema views. This increases the administrative overhead of the FDBS as changes in the components need to be updated. If a user needs specific information from a component, the transformations and views may need to be changed by the FDBS administrator. The main benefit is that these transformations need to be done only once and not for each user.

Using component databases directly has the obvious advantage that no data has to be stored locally by the FDBS. New information is directly available and no update strategy, except for schema changes, needs to be employed. In the early days of FDBS one of the downsides was performance as sending queries and results via the internet was slow. With the increasing internet speed worldwide this problem is less relevant today. Another issue is availability. A FDBS is not protected against component databases being unavailable due to maintenance, outages, and more. Finally, all queries are sent directly to component databases outside of the FDBS control. Sensible information such as patient data may be sent in the queries and therefore need to adhere to privacy and security regulations, which may be complicated in a FDBS setting.

Federation regained popularity in recent years in the field of plant breeding with the BrAPI (Breeding API) project (Selby et al., 2019). Researchers worldwide can provide plant breeding data via a standardized application programming interface (API) which then can be used in a federated system. An advantage of the API standard is the reduced need for schema transformation on the FDBS side.

5.2.2 Data Warehouse

Data warehouses (DWH) are in contrast to FDBS central databases of integrated data. Heterogeneous data sources are parsed and all the information is stored in one central database. If necessary, the information is transformed to match the used database system or the central database schema. In case the data warehouse is created for a specific purpose, the data may also be filtered or further processed. This process is often referred to as ETL (extract, transform, load). Figure 5.1 visualizes this process.

The integration in a central database has the advantage of independence from third parties and network connections to component databases. Outages will affect either all or no data in the central database and the data warehouse administration can implement preventative measures. The central integration comes at a cost.

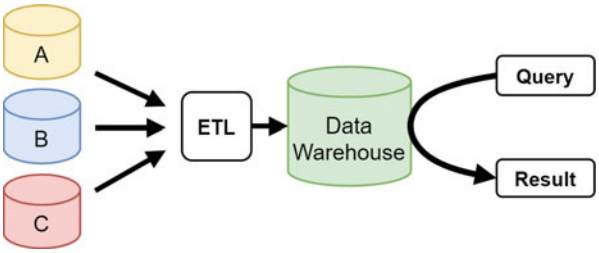

Fig. 5.1 Heterogeneous data sources A, B, and C are integrated into a central data warehouse by means of an ETL (extract, transform, load) process. Queries are performed directly on the data warehouse which has a single schema for all data

Hardware for data storage needs to be available and data sources need to be integrated on a semi-regular basis when updates are available. Data sources and their data formats need to be understood and suitable integration pipelines developed. Mapping the data source schema to the central database schema is comparable to the tightly coupled FDBS approach and changes to the source schemas need to be updated as well. Privacy and security aspects are easier with data warehouses because sensible information can stay inside a controlled network environment like a hospital for example.

5.2.3 Data Lake

A relative new approach is the so-called data lakes (Khine and Wang, 2018). Originating from the field of big data and machine learning, data lakes differ from data warehouses in several key aspects. First, all data from any data source is dumped as-is or with as little transformation as possible into the data lake. This can be structured data such as relational databases, documents such as PDFs, or even raw data such as sensor readouts. The principal idea is that the use of the data is unknown beforehand or may change in the future. Therefore, all data are equally important and should neither be modified, nor filtered. Queries are then performed on the data lake and the heterogeneous information transformed during query execution. Figure 5.2 visualizes this process.

For big data applications using machine learning (ML) algorithms this approach is of great interest, because many modern ML algorithms extract features automatically from heterogeneous and large amounts of data without prior knowledge. However, when writing traditional queries for data analysis, data lakes may impose an even greater burden on the user, similar to loosely coupled FDBS. While the idea of collecting all data possible and having them ready anytime is daunting, this has several downsides. First, even as storage space is getting cheaper over time, data lakes will require a lot of space because all the information is stored. Secondly, different data require different storage solutions. Data lakes often consist of a multitude of subsystems including relational, graph, and document databases

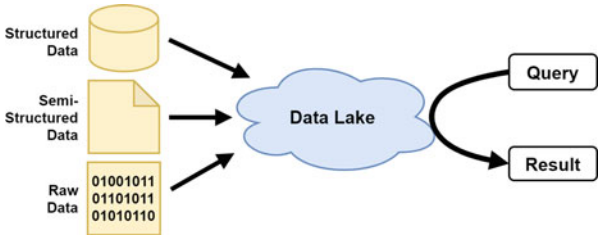

Fig. 5.2 Data lakes consist of structured, semi-structured, and raw data. Queries are performed directly on the data lake and information are transformed in the query processing

as well as key-value stores. The administrative overhead in maintaining all of these systems is larger than a singular database system. Lastly, queries need to handle all types of heterogeneous data. For example SQL queries are tailored to relational databases, but are not well suited for graph database. Query plan optimization is a complicated task for data lakes in order for queries to execute in a reasonable time frame.

5.2.4 Data Mapping

Data integration and analysis require some form of data mapping to connect entities from heterogeneous data sources. In the case of integration, FDBS and data warehouses can use mappings for schema transformation and linking or merging entities together. Data lakes store data as-is and therefore mapping entities are shifted to query execution of subsequent data analyses. Mapping helps in connecting entities and relationships between data sources in order to gain a new data quality. New insights can be generated if mapping connects previously disconnected information clusters.

Mapping can be performed on a variety of information. This includes names, synonyms, identification systems, or more specific entity properties. For example chemical structures can be represented as IUPAC International Chemical Identifier (InChI) identifiers. These InChI ids can then be used to map similar structures. Name and synonym mappings in general are more error-prone than other methods. Depending on the context names may be used for different entities or the words of the name are ordered or cased differently than in other data sources. Additionally, different languages may further complicate the mapping process.

One of the most common mapping methods are identification systems. Almost all data sources define their own identifiers for entities and sometimes even relationships. Examples are the DrugBank identifier “DB00122” for the chemical Choline or the HGNC id “HGNC:5962” for the gene IL10. Databases can provide cross-references to other databases using these identifiers making them especially suited for mapping between data sources. However, not all identification systems should be used to merge entities as being the same. Depending on the scope of the database or identification system information may be represented as a singular entity where other databases provide more granular entities of the same kind. A selection of suitable identification systems can therefore drastically improve the mapping result.

Multiple strategies exist on how mapped entities should be handled. Entities can either be merged together destructively into a singular entity or relationships between these entities can be created non-destructively marking them as mapped. Here, we will explore a hybrid solution by introducing a mapping layer for entities and relationships using only identification systems. The example uses terminology of graph structures but can be transferred to other systems as well.

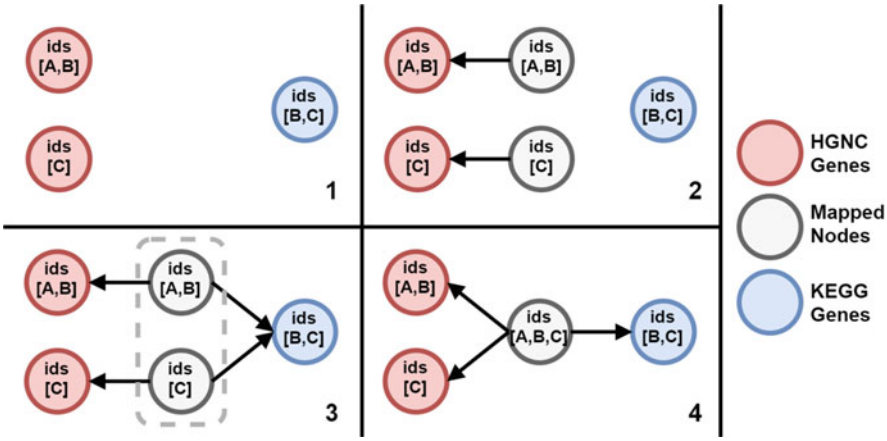

Fig. 5.3 Node mapping example for Gene nodes from two data sources. (1) Mapping operates on a single graph with all data sources merged. (2) Nodes of the first data source are mapped. As no identifiers overlap, two mapping nodes are created and connected to the source nodes. (3) Nodes from the second data source are mapped. This results in an identifier overlap between two mapping nodes. (4) The result is a single mapping node as the two mapping nodes are merged

Nodes of interest are mapped into the mapping layer as visualized in Fig. 5.3. This process takes each individual node and generates a node mapping description. Identifiers from suitable identification systems as well as names and synonyms are collected in the mapping description. If mapping nodes with overlapping identifiers exist, they are collected and collapsed into a singular mapping node. Identifiers and names are merged using standard sets. If none is matched, a new mapping node is created from the mapping description. Finally, an edge is introduced from the source node to the respective mapping node. This process is repeated for all nodes building up the basis for the mapping layer.

Mapping of direct relationships (edges) or more complex relationship paths across multiple nodes is handled similar to the node mapping. For each data source, edge paths of interest need to be defined. A path is comprised of a series of node labels which are connected by an edge label and edge direction. The edge direction can be either forward, backward, or bidirectional and is important to prevent paths going backward where needed. The first and last node labels of the path are required to be used in the node mapping process before, so that their mapped nodes already exist. These edge paths can then be mapped as an edge in the mapping layer between the two mapping nodes. A trivial path of length one being mapped is visualized in Fig. 5.4.

However, meaningful relationships between nodes may involve a more complex path of edges. As paths get longer, the time a mapping process takes will increase accordingly as all possible paths are searched for using depth-first search starting from all existing nodes with the first node label. A path example of length three is visualized in Fig. 5.5.

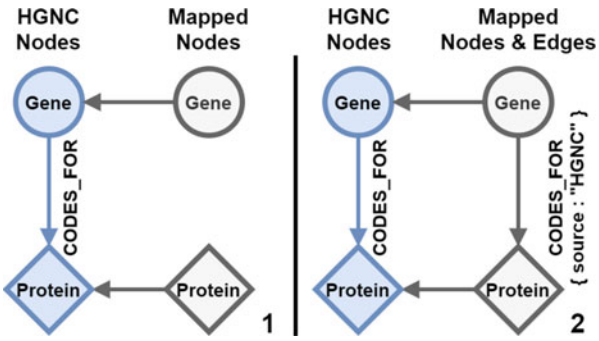

Fig. 5.4 Trivial edge mapping between two mapped data source nodes. (1) The HGNC data source provides two nodes Gene and Protein in blue which are connected with a CODES_FOR edge. Both are connected to their respective mapping node in grey. (2) A new edge with the mapped CODES_FOR label is created between the mapping nodes

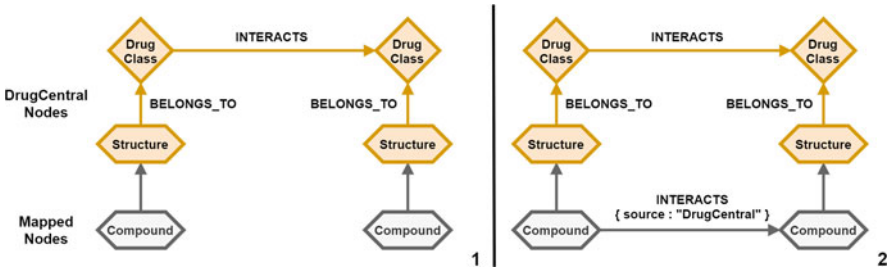

Fig. 5.5 Path mapping of four data source nodes and three edges. (1) Two Structure nodes in orange from the same data source are both associated with a respective DrugClass node. These two DrugClass nodes are linked with an INTERACTS edge. Both Structures are connected to their respective mapping node in grey. The path of length three is matched and provided to the path mapping. (2) A new edge with the mapped INTERACTS label is created between the mapping nodes

5.3 BioDWH2

As shown before, a multitude of problems and techniques exist in the field of data integration and mapping. The BioDWH2 tool presented here solves multiple of these issues while being as easy and automated as possible (Friedrichs, 2021). BioDWH2 is implemented as a modular open-source Java program that is easily extensible with new data source modules. For BioDWH2 to be run, an existing installation of the Java Runtime Environment (JRE) 8 is required. The goal is the automation of data warehouse creation for research projects. A data warehouse solution was chosen for its simplicity in user accessibility and better privacy and security control.

Where data warehouses usually filter data for specific purposes, BioDWH2 uses the unbiased approach of data lakes by integrating all information from each data source where possible. This allows for generated databases to be usable as broadly as possible. Graph database structures were chosen for their high flexibility in large amounts of data and relationships.

5.3.1 BioDWH2 Workspace

As more and more heterogeneous data sources are required for a certain task, the amount of files to be handled gets increasingly complex. Therefore, a fixed schema of managing source, as well as intermediate files in a folder structure is crucial. BioDWH2 takes care of this task by introducing the concept of workspaces. Workspaces allow users to create as many physically separate data warehouse projects as needed. A strict folder structure simplifies research data management. With all sources and intermediate processing steps in a central location, workspaces are easy to compress, backup, and transfer if necessary. The workspace provides a sub-folder structure for each data source containing the source files and metadata information stored in a JSON file. Metadata include the current source version stored, file names, and flags whether the updater, parser, and exporter of the data source finished successfully.

5.3.2 Architecture

BioDWH2 is developed using a modular architecture and the factory method pattern. This allows for new data source modules to be added and maintained without modification of the core project. An architectural overview is visualized in Fig. 5.6.

Every modular architecture needs a core project containing the abstract base classes for the implementing modules. The BioDWH2-Core component provides these base classes as well as a graph data structure and many additional utilities. Networking utilities for example help in communication with HTTP and FTP requests. Dependencies for popular file format libraries, as well as custom implemented file format parsers help data source modules load common formats and simplify the implementation process. These include Open Biological and Biomedical Ontology (OBO), CSV, structure-data file (SDF), and many more.

Data source modules are slim java modules with the BioDWH2-Core as a dependency. They implement the abstract ETL classes of the core for their respective data source. This includes an updater, parser, graph exporter, and mapping describer. This ensures a streamlined implementation process for new modules and reduces the maintenance effort.

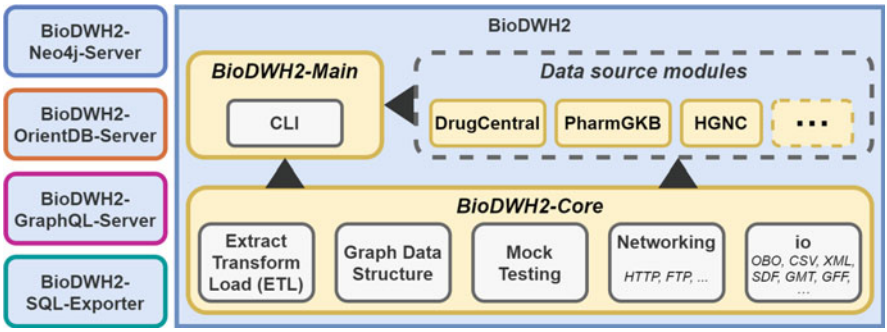

Fig. 5.6 BioDWH2 is built using a modular architecture. The core provides the general program flow and utilities. Data source modules are built on top of the core and implement the abstract ETL process. The main module brings the core and all data source modules together for execution. Additional server and exporter tools complement BioDWH2 for access and analysis needs. These include a Neo4j-, GraphQL-, and OrientDB-Server as well as an SQL exporter

The third component of the architecture is BioDWH2-Main. This java module references the core and all data source modules in the BioDWH2 project. Additional third-party data source modules are included as jar files using the java runtime classpath. The main component provides a simple command-line interface (CLI) as the primary interaction point for the end-users. All tasks such as creating and updating workspaces are performed using this CLI.

5.3.3 Program Flow

BioDWH2 follows the data warehouse paradigm as outlined in Sect. 5.2.2 with the addition of a subsequent mapping step. This results in a strictly defined program flow. Every BioDWH2 project will follow the steps as visualized in Fig. 5.7. As projects are created as workspaces, the creation of a workspace and configuration of used data sources is always the first step. Subsequently, the status of a workspace can be queried or the integration process started as often as necessary.

The integration process itself is split into five tasks and can be repeated whenever a new version of a data source or data source module has become available. As data source modules need to load their respective raw data files, each respective updater implementation checks for the newest version online and downloads them to the workspace if necessary. Once downloaded, the raw data files need to be parsed and exported into the BioDWH2 internal graph data structure by each data source module. The graph data structure is a simple file-based directed property graph model comprised of nodes and edges. Custom unique and non-unique index structures for edge and node properties enable fast queries for existing data. Nodes hereby represent entities such as genes or proteins. Edges represent entity relationships such as a gene codes for a specific protein.

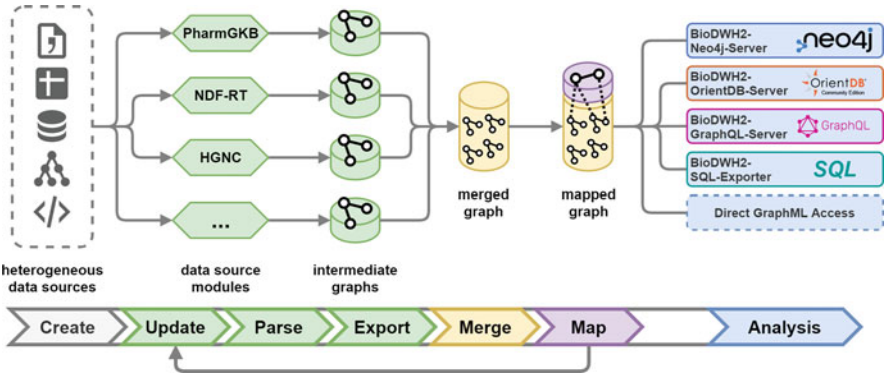

Fig. 5.7 Complete overview of the BioDWH2 data flow. Heterogeneous data sources are updated, parsed, and exported via the data source modules. The resulting intermediate graphs are then merged and mapped into one graph. This graph may then be accessed for analysis using different platforms

The internal graph data structure is stored in each data sources directory. Additionally, each graph is also exported in Graph markup language (GraphML) format (Brandes et al., 2013) for easier access. GraphML was chosen for its simple structure and widespread adoption and interoperability. As the data sourceś graph schema may not be known by the user beforehand, a meta graph visualization and statistic is generated for each graph. The number of nodes and edges per label are exported in tabular format to a text file. The visualization is generated as a static image and interactive HTML page.

After the update, parse, and export steps for each data source the resulting intermediate graphs are collected and merged into a single graph. To distinguish nodes and edges from each data source, their labels are prefixed with their respective data source moduleś ID. This supports the user in writing distinct queries during analysis as well as the mapping process in associating nodes with data sources. The merged graph represents the first data warehouse stage of BioDWH2 containing all requested data sources. As described before, a meta graph and associated statistics are generated and the graph is exported in GraphML format.

The final step of the integration process is the generation of the mapping layer. This meta-layer creates new nodes and edges from common entities and relationships as provided by the data source modules. The mapping itself is based on the description in Sect. 5.2.4. Each data source module provides an implementation of a "MappingDescriber." This describer is able to tell the core mapping process which node labels and edge paths in the data sourceś graph are of interest. Each of these nodes and paths are then queried and presented to the describer individually. Where applicable, the describer then provides a mapping description which is used to create the meta-layer nodes and edges. If multiple entities from different data sources were mapped to the same meta-node, these data sources are now interconnected.

This implementation allows for an automated mapping of data warehouses with any number of sources and only limited by the descriptions provided by the data source modules.

5.3.4 Database Access

The BioDWH2 tool covers the whole integration and mapping process, but provides no analysis capabilities. Every graph in the process is exported to the workspace in GraphML format. These files could be used directly; however, this may not be feasible especially for large graphs. To provide users with easy-to-use analysis capabilities multiple complementary tools are available. As every user might have personal preferences, license restrictions, or technological restrictions, the following database systems were selected as choices and more may be added in the future. Each tool uses the mapped graph databases from a workspace to either provide the data directly, or export a platform specific database. The BioDWH2-Neo4j-Server allows for the creation of a Neo4j graph database as well as running a Neo4j server and browser embedded in the tool itself. No setup of a Neo4j server is needed and queries can be run using the Cypher query language directly in the user's web browser. This allows for a frictionless usage of BioDWH2 for users already familiar with the Neo4j ecosystem. Analogously the BioDWH2-OrientDB-Server tool creates an OrientDB graph database (https://www.orientdb.org) and provides an embedded OrientDB server and browser. GraphQL (https://graphql.org) despite the name is primarily a query language for APIs. However, it is possible to define a schema definition for property graphs such as the BioDWH2 graph data structure. The BioDWH2-GraphQL-Server is currently in development, to provide a GraphQL endpoint for analysis queries, which directly operate on the workspace database. Finally, if users may want to use their graph database on common web servers for which only SQL databases are available, the BioDWH2-SQL-Exporter can be used to transform a workspace graph into a relational SQL database dump. A complete overview of the data flow is visualized in Fig. 5.7 with access to the data using the aforementioned tools.

5.4 Summary

The integration and mapping of heterogeneous data sources is an important first step for scientific data analyses. A multitude of integration paradigms and common problems create a learning curve for researches new in the data integration field. This can delay research projects and shift attention away from subsequent data analyses. Therefore, the automation of integration and mapping tasks is important in reducing this barrier and bringing research projects to analyses faster.

The BioDWH2 suite of tools intends to help users with these issues. As every user has different needs or approaches to data integration and analyses, distinct workflow steps allow for more use-cases and reach a broader audience. For newly started research projects, the final mapping layer might be a good starting point in interconnecting data sources of interest and getting an overview of the data. However, it is always possible to use the merged graph of all data sources or even individual data source graphs directly if those are more fitting for a project. In being as broadly usable as possible and supporting multiple platforms and tools for analysis, BioDWH2 can help in reducing time and effort needed for research projects and prevent common data integration mistakes for inexperienced users.

5.5 Availability

The BioDWH2 tools are free to use and available at https://github.com/BioDWH2. BioDWH2 is developed to be usable out of the box without any prerequisites except the Java Runtime Environment (JRE) version 8.

References

Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen DT, Curpan R, Halip L, Bora A, Yang JJ, Knockel J, Sirimulla S, Ursu O, Oprea TI (2020) DrugCentral 2021 supports drug discovery and repositioning. Nucleic Acids Res 49(D1): D1160–D1169

Brandes U, Eiglsperger M, Lerner J, Pich C (2013) Graph markup language (GraphML)

Chen YB, Chattopadhyay A, Bergen P, Gadd C, Tannery N (2007) The online bioinformatics resources collection at the university of Pittsburgh Health Sciences library system–a one-stop gateway to online bioinformatics databases and software tools. Nucleic Acids Res 35(Database):D780–D785

Fabregat A, Korninger F, Viteri G, Sidiropoulos K, Marin-Garcia P, Ping P, Wu G, Stein L, D'Eustachio P, Hermjakob H. (2018) Reactome graph database: efficient access to complex pathway data. PLoS Comput Biol 14(1):e1005968

Friedrichs M (2021) BioDWH2: an automated graph-based data warehouse and mapping tool. J Integr Bioinform 18(2):167–176

Hassani-Pak K, Castellote M, Esch M, Hindle M, Lysenko A, Taubert J, Rawlings C (2016) Developing integrated crop knowledge networks to advance candidate gene discovery. Appl Translat Genomics 11, 18–26

Imker HJ (2018) 25 years of molecular biology databases: a study of proliferation, impact, and maintenance. Front Res Metrics Anal 3:18

Kapoor R, Tan-Koi WC, Teo YY (2016) Role of pharmacogenetics in public health and clinical health care: a SWOT analysis. Eur. J. Hum. Genet. 24(12):1651–1657

Khine PP, Wang ZS (2018) Data lake: a new ideology in big data era. ITM Web Conf 17:03025

Krier JB, Kalia SS, Green RC (2016) Genomic sequencing in clinical practice: applications, challenges, and opportunities. Dialogues Clin Neurosci 18(3):299–312

Online Mendelian Inheritance in Man, OMIM® (2021) Mckusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). https://omim.org. Accessed: 2021-01-24

Qian T, Zhu S, Hoshida Y (2019) Use of big data in drug development for precision medicine: an update. Expert Rev Precision Med Drug Dev 4(3):189–200

Rigden DJ, Fernández XM (2020) The 2021 nucleic acids research database issue and the online molecular biology database collection. Nucleic Acids Res 49(D1):D1–D9

Sanderson SC, Hill M, Patch C, Searle B, Lewis C, Chitty LS (2019) Delivering genome sequencing in clinical practice: an interview study with healthcare professionals involved in the 100,000 genomes project. BMJ Open 9(11):e029699

Schwinn A, Schelp J (2005) Design patterns for data integration. J Enterp Inf Manag 18(4):471–482

Selby P, Abbeloos R, Backlund JE, Salido MB, Bauchet G, Benites-Alfaro OE, Birkett C, Calaminos VC, Carceller P, Cornut G, Costa BV, Edwards JD, Finkers R, Gao SY, Ghaffar M, Glaser P, Guignon V, Hok P, Kilian A, KÖnig P, Lagare JEB, Lange M, Laporte MA, Larmande P, LeBauer DS, Lyon DA, Marshall DS, Matthews D, Milne I, Mistry N, Morales N, Mueller LA, Neveu P, Papoutsoglou E, Pearce B, Perez-Masias I, Pommier C, Ramírez-González RH, Rathore A, Raquel AM, Raubach S, Rife T, Robbins K, Rouard M, Sarma C, Scholz U, Sempéré G, Shaw PD, Simon R, Soldevilla N, Stephen G, Sun Q, Tovar C, Uszynski G, Maikel V (2019) BrAPI—an application programming interface for plant breeding applications. Bioinformatics 35(20):4147–4155

Shoshi A, Hofestädt R, Zolotareva O, Friedrichs M, Maier A, Ivanisenko VA, Dosenko VE, Bragina EY (2018) GenCoNet—a graph database for the analysis of comorbidities by gene networks. J Integr Bioinform 15(4):20180049

Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE (2012) Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 92(4):414–417

Wishart DS (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34(90001):D668–D672

Yoon BH, Kim SK, Kim SY (2017) Use of graph database for the integration of heterogeneous biological data. Genomics Inform 15(1):19

Chapter 6
DaTo: An Integrative Web Portal for Biological Databases and Tools

Yincong Zhou, Ralf Hofestädt, and Ming Chen

Abstract DaTo is a collection of published online biological databases and tools, started to offer service since 2011 and it has been continuously upgraded since then. In the latest version, there are 36,639 resources. DaTo offers a user-friendly interface and provides extensible URL-related comments, such as URL status, Geo location, and the authorship. A graphical interactive web browser was embedded into DaTo front-end to facilitate the research of ontology-based semantic similarity relationships among tools and databases. Using DaTo, the geographic location, health status, and journal associations were evaluated based on the historical development of bioinformatics tools and databases in the past 20 years. Besides, a specific collection of biological databases and tools can be generated. OverCOVID (http://bis.zju.edu.cn/overcovid) is such a sub-database to contain SAR-Covid-related bioinformatics resources. The updated version of DaTo is accessible via http://bis.zju.edu.cn/dato/.

Keywords Biological database · Biological tool · Text mining · Geographic network · Bioinformatics

6.1 Introduction

Thousands of online databases and data analysis tools have been developed for life science research to deal with the exploration of biological datasets. Although some of these methods have been published in special journals, such as Nucleic Acid Research (NAR) database issues (https://www.oxfordjournals.org/nar/database/c) or Webserver Issues (https://academic.oup.com/nar/issue/48/W1), others are still

Y. Zhou (✉) · M. Chen
Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, China
e-mail: yczhou@zju.edu.cn; mchen@zju.edu.cn

R. Hofestädt
Department of Bioinformatics and Medical Informatics, Faculty of Technology, Bielefeld University, Bielefeld, Germany

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_6

Table 6.1 A list of bioinformatics resource collections

Collection name	Description	Reference
BLD	For bioinformatics research organized within categories familiar to a biologist	Brazas et al. (2012)
BIRI	A public online searchable index of bioinformatics resources developed at the biomedical informatics group	de la Calle et al. (2009)
OMICtools	An informative directory for multi-omic data analysis	Henry et al. (2014)
OReFiL	An online resource finder for Lifesciences	Yamamoto and Takagi (2007)
JIBtools	The official bioinformatics tool list for the Journal of Integrative Bioinformatics	Hofestädt et al. (2013)
bio.tools	Bioinformatics Tools and Services Discovery Portal	Ison et al. (2015)
Database Commons	A catalog of biological databases	

scattered throughout the Internet and a large amount of literature. Internet searches via Bing, google or similar general search engines do not exclusively index online biological resources. Therefore, it is difficult to extract useful information. Researchers cannot find the appropriate tools or databases for their specific purpose due to the huge number of resources and the lack of a complete list of these resources (Chen et al. 2007). To deal with these problems, several groups have collected online life-science-related and bioinformatics-related resources, and provide search function, which can be obtained through the Internet (Table 6.1).

BLD is a catalog of URIs to bioinformatics resources, including databases and tools based on recommendations from experts in the field (Brazas et al. 2012).

BIRI uses keywords and sentence structures to identify related terms through custom patterns (de la Calle et al. 2009).

BLD and BIRI divide the items into subcategories based on research topics, making it impossible to return resources corresponding to specific search terms. For example, no results will be returned for a query of the word "rRNA" in BLD as well as BIRI.

OMICtools (https://omictools.com/) is a community-based search website (Henry et al. 2014). It bridges the gap between researchers and developers, and brings together an active worldwide user community, linking expert curators who submit and classify tools, to users who enhance the interface by providing feedback and comments. It can provide high-quality service due to the maintenance and upgrade of its team. However, his commercialization severely restricted the common use of users.

OReFiL (http://orefil.dbcls.jp/) is known as the only online collection that returns latest and inquiry related online resources based on peer-reviewed publications. But OReFiL not only focuses on the collection of online tools, but also includes all online resources which is relevant for a certain keyword. In addition, some of the

returned resources either lack an accurate title, description, or contain unrelated items. Another problem is that OReFiL cannot return all resources related to the search term. Taking "miRNA" as an example, the search results are limited to 500, and many items not related to miRNA are returned, some of which even link to PNG images that have nothing to do with miRNA (Yamamoto and Takagi 2007).

JIBtools (https://jib.tools) is a collection of tools lists curated by a specific editor who is responsible for a specific expert field (Hofestädt et al. 2013). This approach depends largely on the singleton's motivation to provide and update the list of tools.

Bio.tools (https://bio.tools) provides a manually curated list of online resources. Any researcher is able to register to the website and is allowed to add extra entries (Ison et al. 2015). However, no automatic workflow of this process is included, and the registration process is quite complex, as many different key terms must be registered into the system.

Another biological data repository for researchers which have to be mentioned is Re3data (Pampel et al. 2013) (http://re3data.org). Re3data is a comprehensive collection of biological data repositories available through Internet, which lists more than 1500 research data repositories. It also supports browsing by subject, content type and country, and offers an API for researchers. However, the stability of the website is insufficient, and it is easy to freeze when users search for biological resources.

Biosharing (McQuilton et al. 2016) (https://biosharing.org) is another repository on interconnected data standards, databases, and policies. It consists of 671 data standards, 831 databases and 85 policies, and visualized them with different label. On the other hand, biosharing only showed these data in static pages, which makes user hard to analyze the trend of biological data sources because of a lack of temporally and spatially dynamic module.

Database Commons (https://bigd.big.ac.cn/databasecommons) is a catalog only for biological databases. It allows users to easily access a comprehensive collection of public biological databases containing different data types and across different organisms. However, it only collects databases and does not provide tool indexes.

Moreover, none of them provide the opportunity to visualize the geographical locations of the database- or tool-institutes. However, a vast number of geographic information-related methods in other fields exist nowadays, which are making use of the underlying Google Maps API to visualize specific scientific aspects in relation to their geographical location. One recent example: it was used to provide geographic information concerning the health status of the Great Barrier Reef (Nim et al. 2015). Under such considerations, we developed DaTo, an automatic approach to collect, curate and index through a large collection of different databases and tools.

Comparing to other platform, DaTo adds a brand new dimension to the analysis and visualization of important online biological resources. By using a Google Maps based approach, online resources can be localized to specific countries, cities even institutes. Therefore, it is possible to figure out which research-related online resources are geographically close to the home institutes, along with which online resources are well developed or maintained at a certain university and which ones might have to be extended in the future. Also, it is possible to located cooperation

partners in the neighborhood, providing customized services required for local research. Depending on the research area, it might be important to cooperate with close-by institutions. Moreover, we have analyzed the health status of their web links, as well as the impact of the respective publications' journals, countries, and years.

6.2 Methodology and Implementation

6.2.1 Data Collection

We use the keyword "HTTP/FTP" to search in the PubMed database and retrieve the MEDLINE format results from the eultis provided by PubMed. We also integrate other related resources, such as Bioconductor, Bioconda, GitHub, as well as subscriptions to journals such as Bioinformatics, Nucleic acids research and Database to enlarge DaTo. Finally, there are 36,639 records in DaTo ranging from 1982 to 2020.

In-house scripts are applied to get the structured data from MEDLINE format results. And we used Europe PMC (https://europepmc.org/) to get the text-mined terms like chemical, organism if full text is available in Europe PMC. We got the normalized authors and institutions from Web of Science, mapped the authors to MAG (https://academic.microsoft.com/) to get each unique author identifier, mapped the institutions to Global Research Identifier Database (https://www.grid.ac/) to catalog the research organizations. The citation counts where self-citing is removed were also retrieved from Web of Science. URL domain information was parsed to generate sub-features; for example, IP address, location. Then keyword matching strategy was used to tag articles as tools, databases or web-servers and we got the language content of tools in the similar way. The final structured data are deposited into MongoDB and we plan to update DaTo regularly.

6.2.2 Implementation

DaTo has been keeping updating with state-of-art methods and technologies to improve the user experience. The methods and techniques that used in this version are listed in the following table (Table 6.2).

DaTo features a user-friendly query interface, providing comprehensive annotations for each result, such as the description of the resources, the abstract of the original literature, the link to the corresponding PubMed entries and corresponding webpage (Fig. 6.1). DaTo provides a user-friendly search interface and multilayer annotations for each result, such as resource descriptions, original document abstracts, corresponding PubMed entries, and links to corresponding website.

Table 6.2 Technologies used in each part of DaTo

Part	Name	Link
Web Front	React	https://github.com/facebook/react
	Next.js	https://nextjs.org
	D3.js	https://d3js.org
	Vis.js	https://visjs.org
	deck.gl	https://deck.gl
	Leaflet	https://leafletjs.com
	Highcharts	https://www.highcharts.com/
Database	MongoDB	https://www.mongodb.com
	Elasticsearch	https://www.elastic.co
	Neo4j	https://neo4j.com
NER	bioNerDS	http://bionerds.sourceforge.net
	BERT	https://github.com/google-research/bert

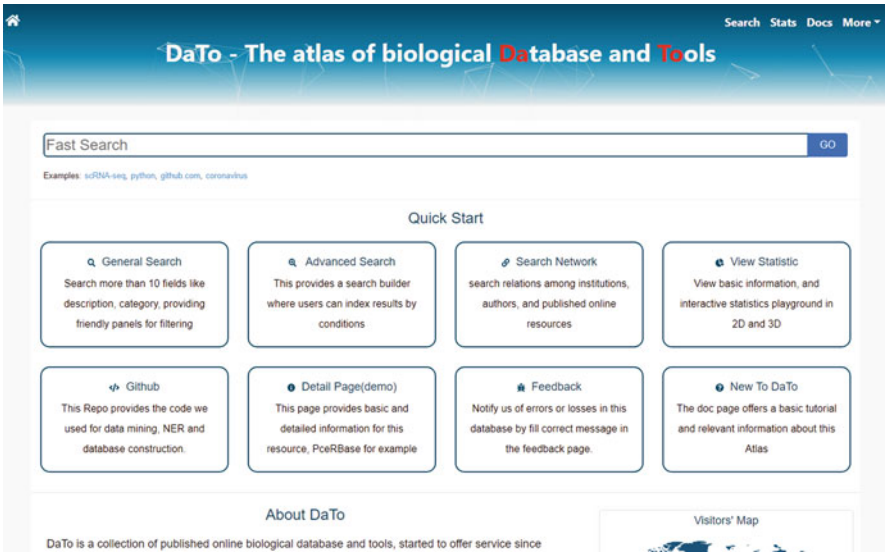

Fig. 6.1 DaTo web interface. On this home page, multiple quick entries like "advanced search" are provided to facilitate users to easily find the information they are interested in

6.3 Functionality and Usage

6.3.1 Geographic Location

A graphical interaction network browser has been embedded into DaTo web front-end, which enables researchers to explore of the connection between the tools and databases based on the similarity of MeSH term. To facilitate effectively investigate the international geographic distribution of the hosts and to allow users to intuitively and accurately perceive geographic locations, we tracked the affiliation of all first

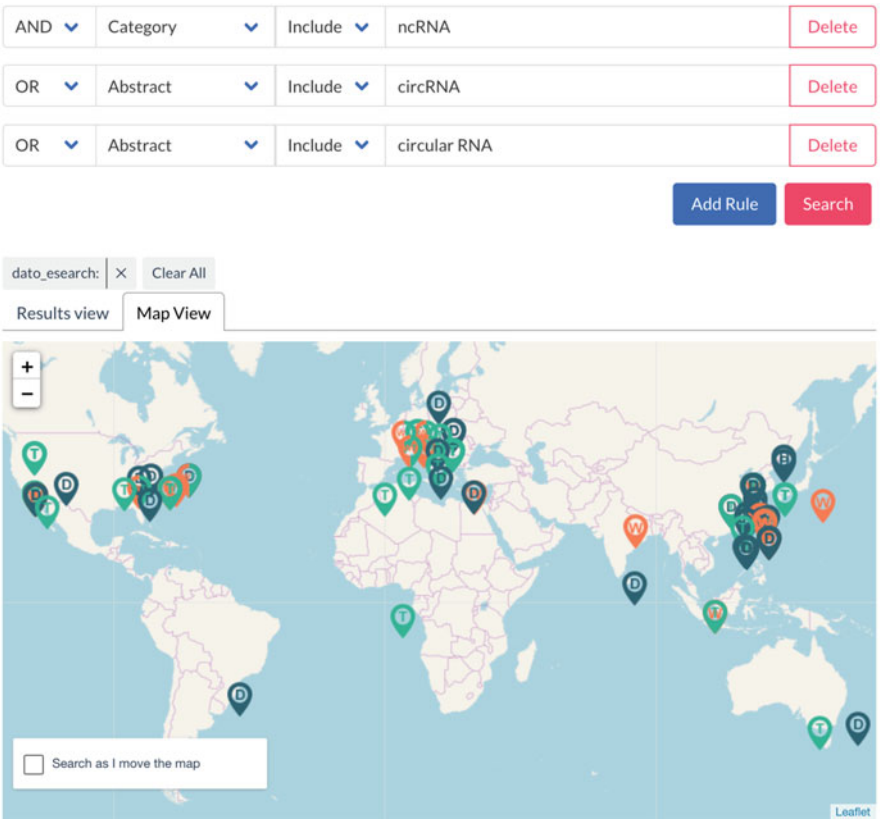

Fig. 6.2 Advanced search with search builder, the demo shows the search category as ncRNA, and the abstract contains the keywords of circRNA or circular RNA

authors and website IP addresses to reveal their location: area code, city, latitude, longitude, ISP and organization as well as country code and country name. DaTo adopts IP2Location for the geographic location of the URLs and Google Map API to display them. And the search results will be displayed on the leaflet map to show the geographic location of each record (Fig. 6.2).

6.3.2 Search Builder

We provide a range of filtering features, such as time, record type or journal name. Search results are sorted in three ways, "Highest Score," "Recently Publicated," and "Best Match" in default (Fig. 6.3).

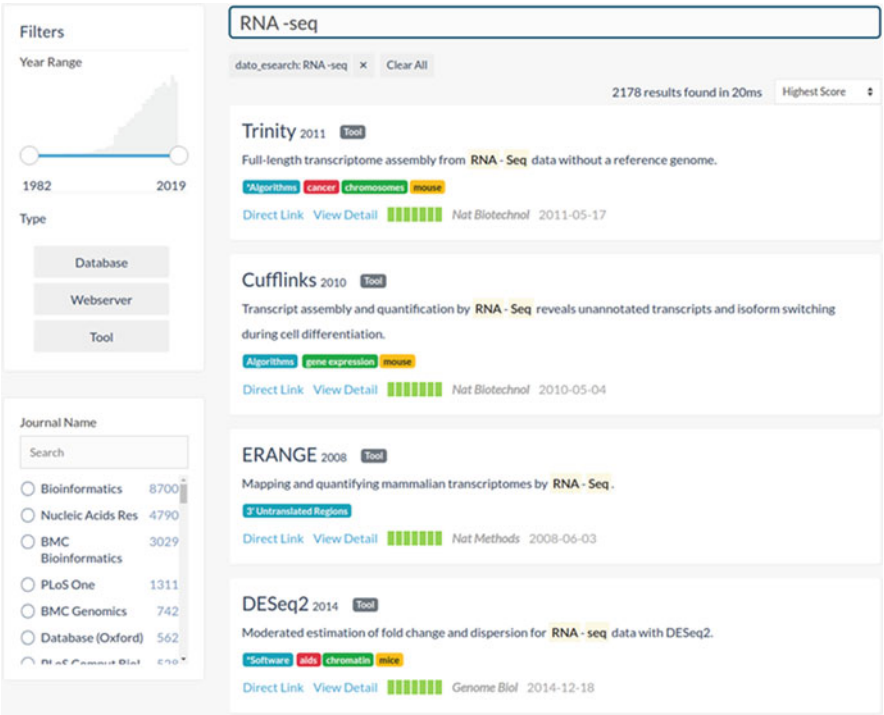

Fig. 6.3 Search Interface

6.3.2.1 Network Section

By analyzing the connection between authors and institutes around the world, we developed a graph database which composed of author, institute, journal, and their relationship. As of January 1, 2021, DaTo contains 35,221 papers published by 4905 research institutes from 83 countries and regions. This allows users to easily explore the cooperation of author–author, author–institute, institute–institute, and global cooperation trends. A demo usage of network section Cooperation between Bielefeld University and Zhejiang University (Fig. 6.4).

6.3.2.2 Statistics Section

Statistics consists of two parts: basic statistics and playground. We provide a series of information, such as the change in the number of publications over time, the publication status of various countries, etc. The distribution of bioinformatics resources globally over time can help researchers to analyze the historical development of bioinformatics databases and tools from geographical perspective. It is September 1, 1994, the first biological host appeared in Italy (Pongor et al. 1994). By 2000, the

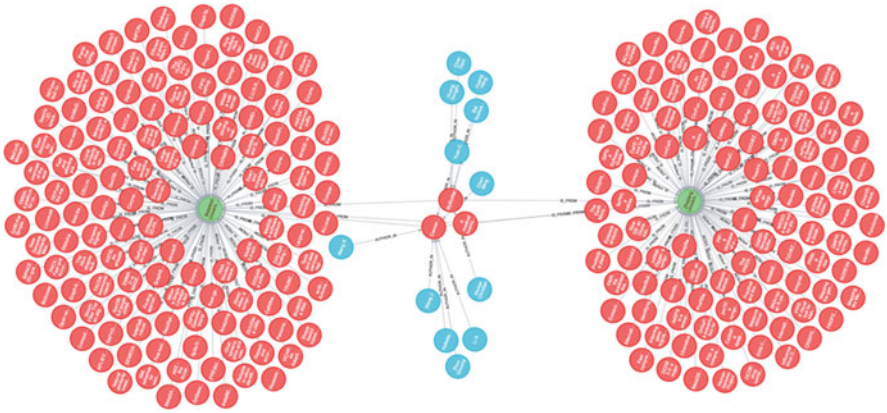

Fig. 6.4 Cooperation between Bielefeld University and Zhejiang University

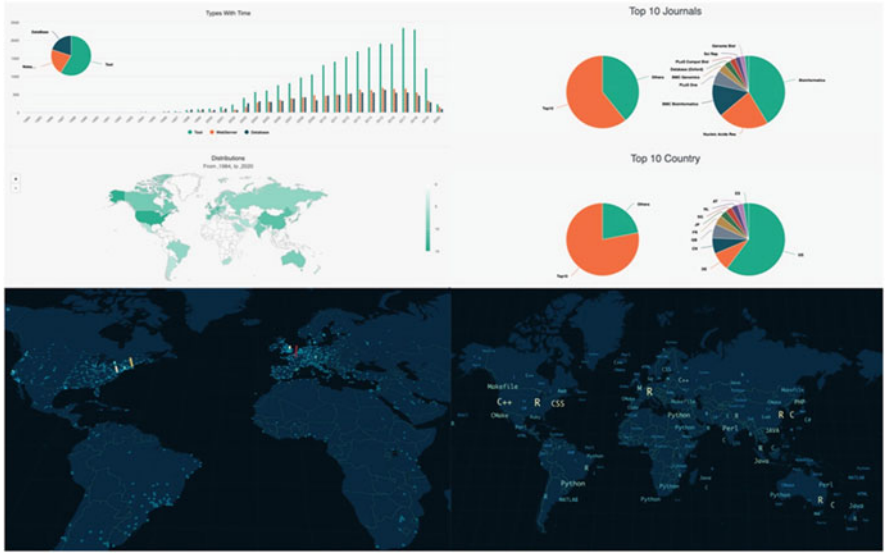

Fig. 6.5 Statistics review

amount of biological online resources had increased more than 800, with hosts all over North America, Europe, and East Asia. In the new century, biological databases and tools have developed rapidly, especially in emerging countries such as China, Russia, India, Brazil, and South Africa. Currently, biological resource hosts are distributed on six continents, most of which are located in the United States and the European Union (Fig. 6.5).

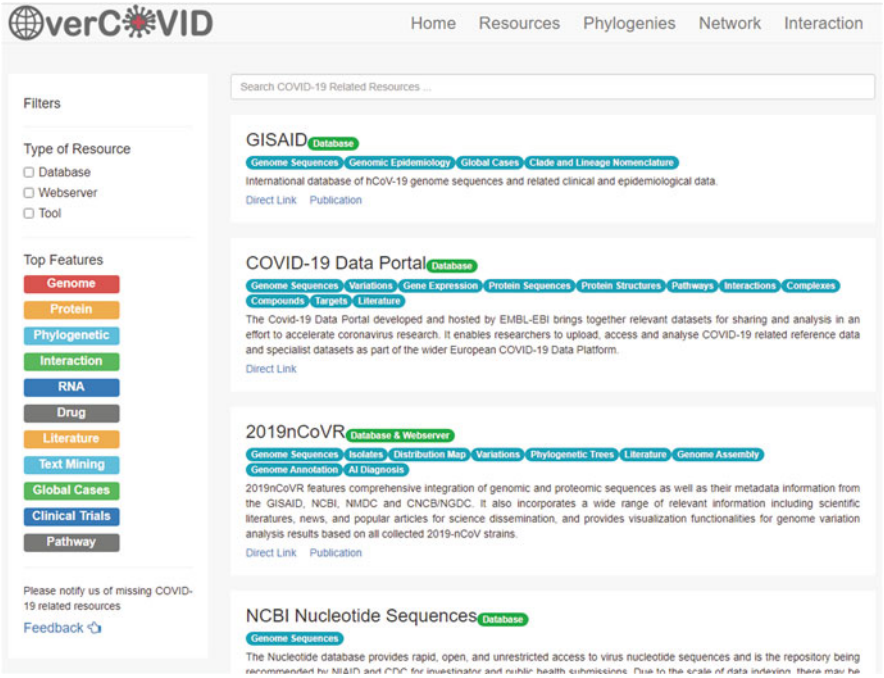

Fig. 6.6 The web portal OverCOVID

6.3.2.3 OverCOVID

The web portal OverCOVID (Ahsan et al. 2021a) is provided to share bioinformatics resources and information that may contribute to research advances (Fig. 6.6). Based on the collected databases, relationships and/or associations can be identified. For instance, Virus–Host Protein Interactions, Human Protein–Protein Interactions, ncRNA-associated Interactions, Drug–Target or Drug–Protein or Drug–Gene Interactions, and Drug Side Effects (Ahsan et al. 2021b).

6.4 Conclusion

Nowadays, biological sciences are generating more data than ever. Through continuous data accumulation and technology upgrading, we have made DaTo, a comprehensive and efficient online resource for biological researchers. As a constantly updated database, DaTo not only focuses on collecting bioinformatics resources, but also system analysis of the bioinformatics resources, which is much valuable for both experimental biologists and computational biologists. Through the tracking information and meta-information provided by this atlas, DaTo constructed

a storyboard of published biological databases and tools. Besides, a specific collection of biological databases and tools can be generated like OverCOVID due to modularization and precise classification of DaTo, we believe this will have more extensions in the future.

Acknowledgments This work was partially supported by the National Natural Science Foundation of China [No. 31571366], and CSC & DAAD (PPP program No. 57136444).

References

Ahsan MA, Liu Y, Feng C, Chen M (2021a) OverCOVID: an integrative web portal for SARS-CoV-2 bioinformatics resources. J Integr Bioinform 18:9–17

Ahsan MA, Liu Y, Feng C, Zhou Y, Ma G, Bai Y, Chen M (2021b) Bioinformatics resources facilitate understanding and harnessing clinical research of SARS-CoV-2. Brief Bioinform:bbaa416

Brazas MD, Yim D, Yeung W, Ouellette BF (2012) A decade of web server updates at the bioinformatics links directory: 2003–2012. Nucleic Acids Res 40:W3–W12

Chen Y-B, Chattopadhyay A, Bergen P, Gadd C, Tannery N (2007) The online bioinformatics resources collection at the University of Pittsburgh Health sciences library system—a one-stop gateway to online bioinformatics databases and software tools. Nucleic Acids Res 35(suppl 1):D780–D785

de la Calle G, García-Remesal M, Chiesa S, de la Iglesia D, Maojo V (2009) BIRI: a new approach for automatically discovering and indexing available public bioinformatics resources from the literature. BMC Bioinform 10(1):1

Henry VJ et al (2014) OMICtools: an informative directory for multi-omic data analysis. Database 2014:bau069

Hofestädt R, Kormeier B, Lange M et al (2013) JIBtools: a strategy to reduce the bioinformatics analysis gap. J Integr Bioinform 10(1):226

Ison J, Rapacki K, Ménager H et al (2015) Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res 44:D38–D47

McQuilton P, Gonzalez-Beltran A, Rocca-Serra P et al (2016) BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database 2016:baw075

Nim H, Done T, Schreiber F, Boyd S (2015) Interactive geolocational and coral compositional visualisation of Great Barrier Reef heat stress data. In: Big data visual analytics (BDVA). IEEE, Piscataway, NJ, pp 1–7

Pampel H, Vierkant P, Scholze F et al (2013) Making research data repositories visible: the re3data.org registry. PLoS One 8(11):e78080

Pongor S, Hátsági Z, Degtyarenko K et al (1994) The SBASE protein domain library, release 3.0: a collection of annotated protein sequence segments. Nucleic Acids Res 22(17):3610

Yamamoto Y, Takagi T (2007) OReFiL: an online resource finder for life sciences. BMC Bioinform 8(1):1

Chapter 7
The Use of Data Integration and Knowledge Graphs in Modern Molecular Plant Breeding

Bjoern Oest Hansen, Jan Taubert, and Thomas Thiel

Abstract Feeding nearly ten billion people by 2050 requires a year-on-year yield increase of major field crops of about 1–2%, while arable land will likely decrease due to urbanization and climate change. On the other hand, developing a new crop variety traditionally can take up to 10–12 years. To speed up molecular breeding new ways of harnessing breeding information, including state-of-the-art statistical methods and predicting candidate genes as targets for breeding from massive amounts of data are required. As most of the necessary data is still buried in thousands of public and proprietary databases, siloed in legacy systems or can only be found in spreadsheets, novel approaches in data integration to overcome these challenges are needed. Here we describe our approach of using workflow-driven data integration and knowledge graphs in an industrial application at one of the world's leading plant breeding companies.

We adopt state-of-the-art statistical approaches for plant breeding and apply them on public and in-house generated and expert-curated data from different data domains that date back to more than a decade. For this we use a customized instance of the open-source Galaxy computational platform and analyze breeding data in a workflow-driven approach. We also shed some light on the challenges of in-house deployment of open-source tools in an industrial application, as well as ensuring software quality and coding standards for own developments.

We apply knowledge graphs in knowledge discovery use-cases to show some benefits of handling ontology-enriched in-house data as a structured graph. Here it is possible to extract information related to connections, communities in the data, infer new edges, or look for complex patterns across the graph and to perform tasks that would have been highly complex and time consuming on a silo-based data information system.

Nevertheless, the challenge of ever-increasing data in breeding information remains and necessitates the combination of different approaches to continuously drive value from data.

B. O. Hansen · J. Taubert (✉) · T. Thiel
KWS Saat SE & Co. KGaA, Einbeck, Germany
e-mail: Jan.Taubert@kws.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_7

Keywords Data integration · Knowledge graph · Workflow · Galaxy · Ontologies · Open-source · Plant breeding · Industry

7.1 Introduction

Ensuring sustainable food supply for an increasing world population of nearly ten billion people by 2050 (see United Nations, Department of Economic and Social Affairs, https://www.un.org/development/desa/en/news/population/world-population-prospects-2019.html) requires significant progress in plant breeding and farming practices across the whole world. Climate change and the scarcity of arable land are set to impact food production in the foreseeable future. As part of the EU Green Deal, the Farm to Fork strategy sets out ambitious goals for agriculture in the coming years. These goals (see European Commission, Farm to Fork Strategy Action Plan 2020, https://ec.europa.eu/food/sites/food/files/safety/docs/f2f_action-plan_2020_strategy-info_en.pdf) include a reduction in the use of chemical plant protection by 50%, a reduction in the use of fertilizers by 20%, and the use of organic farming practices on at least 30% of farmland in the EU by 2030. On the other hand, developing a new crop variety traditionally can take up to 10–12 years.

To speed up plant breeding with the use of molecular technologies, new ways of harnessing breeding information, including state-of-the-art statistical methods and predicting candidate genes as targets for breeding from massive amounts of data are required. As most of the necessary data is still buried in thousands of public and proprietary databases, siloed in legacy systems or can only be found in spreadsheets, novel approaches in data integration to overcome these challenges are needed. Here we describe our approach of using workflow-driven data integration and knowledge graphs in an industrial application at one of the world's leading plant breeding companies KWS (see Box 7.1).

We adopt state-of-the-art statistical approaches for plant breeding and apply them to public and in-house generated and expert-curated data from different data domains that date back to more than a decade. For this we use a customized instance of the open-source Galaxy (Blankenberg et al. 2010) computational platform and analyze breeding data in a workflow-driven approach. The use of open-source software in the industry requires paying attention to the associated license terms and how such software is integrated into an industry application context. Furthermore, to ensure a high quality of own developed functionality a staged code quality and release process has been implemented with the goal to ensure high productivity in routine data analysis applications.

The challenges of integrating data across different types, from different years and different domains (e.g., genotypic, and phenotypic data) can then be addressed using workflows in Galaxy. Proprietary tools providing data from several in-house data

silos, together with applying public analysis tools are demonstrated in a genome-wide association study (GWAS) use case. Furthermore, the results of the GWAS study can be embedded in a wider context using data from a knowledge graph database.

Graphs are among the most flexible formats for a data structure. In a graph, information is described as a network of nodes and links between them, rather than tables with rows and columns. Both the nodes and edges can also have attributes assigned to them. Graph-based systems are easier to expand, as their schemas are not as strict as classical relational databases. In knowledge discovery research, this is a huge advantage. The term Knowledge graph was coined by Google in 2012, even though the topic itself has been around for longer. Though there is no formal definition of a knowledge graph, it is often described as a semantically enriched graph, supported by ontologies for standardizing the semantics. This allows for machine-readable meaning to be integrated with the data. By handling data as a structured graph, other benefits appear, it is possible to extract information related to connections, communities in the data, infer new edges, or look for complex patterns across the graph. It also becomes possible to perform tasks that would have been highly complex and time consuming on a silo-based data information system.

This combination of highly automated workflow-driven processing of genotypic and phenotypic data in plant breeding applications combined with a flexible exploration of the surrounding context of such results using knowledge graphs is supporting the decision-making of breeders at KWS. With better decision-making, plant breeding can improve the genetic potential of all crops to tackle the challenges of climate change, reduction of inputs, zero(low) chem ag and organic farming practices with the goal to provide the best seeds to our customers, the farmers.

Box 7.1 About KWS

About KWS

KWS is one of the world's leading plant breeding companies. With the tradition of family ownership, KWS has operated independently for more than 160 years. It focuses on plant breeding and the production and sale of seed for corn, sugar beet, cereals, potato, rapeseed, sunflowers, and vegetables. KWS breeding programs aim to offer every farmer—whether they use conventional or organic farming methods—targeted varieties and solutions to fit their operational needs, while also optimally tailored to the climatic conditions and specific geological conditions of their respective regions. This is the basis for efficient and productive agriculture. KWS uses leading-edge plant breeding methods. 5700 employees represent KWS in more than 70 countries.

Source: https://www.kws.com

7.2 Methods and Implementation

7.2.1 Deploying the Galaxy System in an Industry Application Context

The open-source Galaxy system developed by the Galaxy Project (Blankenberg et al. 2010, https://galaxyproject.org/) is a web-based platform for accessible, reproducible, and transparent computational research. The software is licensed (see https://galaxyproject.org/admin/license/) under the Academic Free License version 3.0 and images and documentation are licensed under the Creative Commons Attribution 3.0 (CC BY 3.0) License, which in principle allows deploying the Galaxy system in an industry application context (see Sect. 7.2.2). The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University. According to the Galaxy Project website (accessed January 2021) the Galaxy system is characterized by:

- **Accessible:** programming experience is not required to easily upload data, run complex tools and workflows, and visualize results.
- **Reproducible:** Galaxy captures information so that you do not have to; any user can repeat and understand a complete computational analysis, from tool parameters to the dependency tree.
- **Transparent:** Users share and publish their histories, workflows, and visualizations via the web.
- **Community centered:** Our inclusive and diverse users (developers, educators, researchers, clinicians, etc.) are empowered to share their findings.

The Galaxy system (Blankenberg et al. 2010) is publicly available at https://usegalaxy.org. As an important free and publicly accessible resource, it cannot offer encrypted data transfer and data storage and scalability. For most applications in an industry context, data integrity, data security and know-how protection are major concerns. Therefore, the preferred way would be to run your own Galaxy instance either on-premises or in your private cloud environment. This provides additional possibilities of closely integrating the Galaxy system with other in-house data resources, compute environments and storage systems.

Depending on the importance of the Galaxy system for data analysis needs at the company and resulting requirements to provide the system as a service to users, a more sophisticated setup than a single Galaxy system instance can be chosen. From our experience deploying the Galaxy system in an industry application context at a major plant breeding company, we recommend a setup that involves three Galaxy instances (Table 7.1): A test instance serves mainly for early testing by in-house users as well as the establishment and fine-tuning of Galaxy workflows. The productive Galaxy instance hosts tools and workflows suitable for routine operation with respect to fault tolerance and performance optimization. A third

Table 7.1 Recommended setup with three Galaxy instances

ID	Description	Users	Features	Updates
dev	Galaxy development system to follow main Galaxy branch closely and test/develop new Galaxy platform features	Galaxy in-house infrastructure team and some script developers	All required Galaxy tools installed, following the latest tool versions, possibly local new tool development	Very frequently, all Galaxy releases
test	Galaxy test system for Galaxy tools, not for testing Galaxy platform features	Above and certain Galaxy test users	All required Galaxy tools installed via Galaxy tool-shed, usually latest tool versions	Frequently, might skip minor releases
prod	Galaxy production system for routine high-performance data analysis workflows	All Galaxy users	Stable versions of required Galaxy tools for productive workflows installed via tool-shed, availability monitoring	Only major releases, maintenance window for updates

Table 7.2 Categories of development related to the Galaxy system

Category	Examples	Contributors	Distribution
Galaxy platform	Add new authentication mechanisms to Galaxy (e.g., OKTA), add more interfaces to compute cluster (e.g., IBM LSF)	Galaxy in-house infrastructure team and community developers	Submission to main Galaxy branch after community review
Public tools	Fixes to publicly available Galaxy tools (public tool-shed)	Script developers (internal and external)	In accordance with public tool owner
Proprietary tools	Specific tools for routine data analysis workflows (e.g., genomic selection)	Internal script developers (e.g., Biostatisticians)	Non-public, company confidential

Galaxy instance (Galaxy dev) serves for testing new Galaxy releases and in-house tool development.

As the Galaxy platform supports the management and installation of Galaxy tools via the Galaxy tool-shed, a local tool-shed instance is used to provide proprietary tools to the Galaxy instances. The Galaxy public tool-sheds are integrated to use and update publicly available tools for Galaxy. This allows for the clear separation of development in three major categories, see Table 7.2.

Another advantage of using a local tool-shed is the integration possibility into continuous code integration and deployment pipelines (CI/CD). Modern source code management platforms, like GitLab (see https://about.gitlab.com/) facilitate the setup of CI/CD pipelines which upon code submission to a tool repository manage the assembly, testing and deployment of changes to proprietary Galaxy tools to in-house Galaxy instances via the Galaxy tool-shed automatically. This automation ensures a high quality of the tools by performing unit testing and integration testing, as well as convenience for the script developers, which do not have to manually deploy tools to the Galaxy instances anymore. Provision of multiple Galaxy instances with different levels of productivity allows fine-tuning of CI/CD pipelines with respect to code quality and release speed for in-house Galaxy tool developments. Overall, this approach results in quality improvements, time savings and faster availability of features for the users of routine data analysis workflows (Fig. 7.1).

To be able to scale routine data analysis to multiple compute nodes beside the main Galaxy instance, a high-performance compute cluster is used. Galaxy schedules the jobs and submits these into different queues of the cluster. The cluster queue is determined by which tool should be run. For new tools, this mapping will be updated once a tool is in production usage. Factors like the number of CPUs or memory used on average by the tool will determine which queue it will be assigned to. For low memory consuming and quick running tools, e.g., data upload, a queue with a high priority is used so that the user will get tool run results almost immediately. For long running and high memory consuming tools, a lower priority queue is chosen, so that the impact of these long analyses on the overall Galaxy performance is mitigated. However, the users are made aware that such analysis might not finish when expected depending on the average cluster load. To achieve transparency, the status of the compute cluster is reported to the users on the Galaxy starting page (Fig. 7.2).

As an open-source software, the development of Galaxy largely depends on an active Galaxy community. To follow the latest developments timely, we also actively participate in the development of the Galaxy platform by submitting work items ("issues") for the public Galaxy repository, which are then being integrated into future Galaxy versions. Additionally, we contribute with own code submissions to the general Galaxy platform (Afgan et al. 2018).

7.2.2 Implications of Open-Source Licenses on the Use of Open-Source Software in the Industry

The use of free and open-source software (FOSS) in the industry is steadily increasing, driven not just by in most cases the absence of a license fee, but also by the highly innovative character of some FOSS products and packages, especially when it comes to addressing scientific challenges. However, besides these advantages

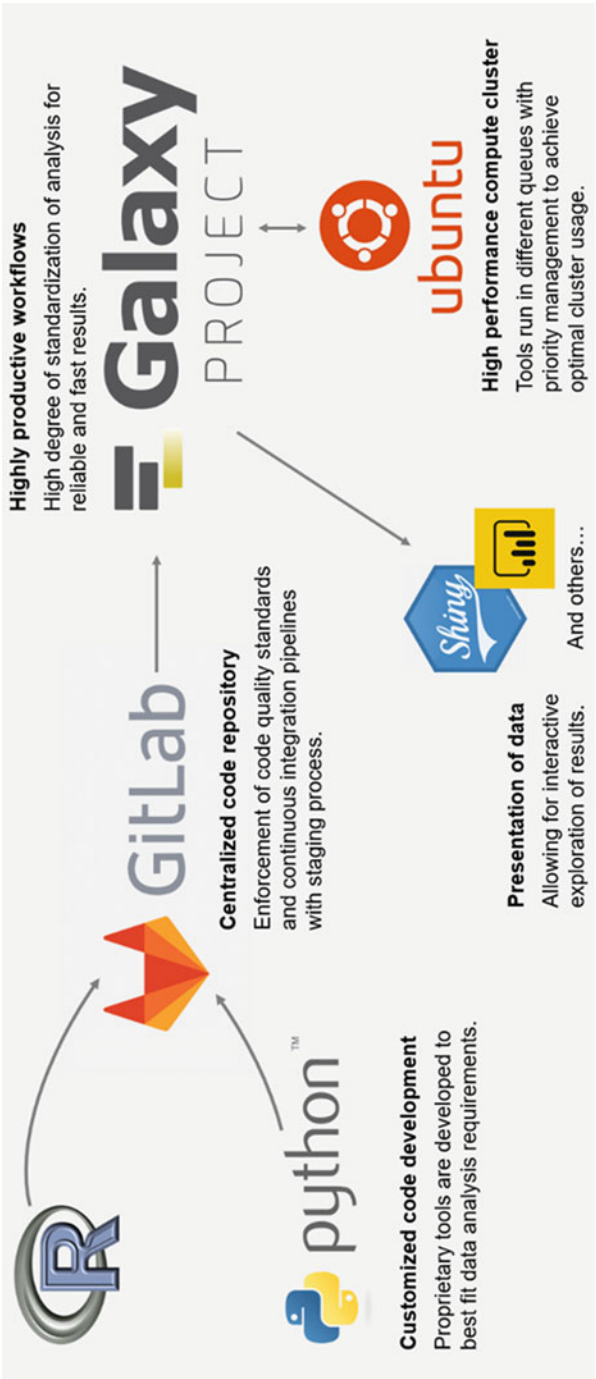

Fig. 7.1 Overview of components of an example on-premises Galaxy deployment

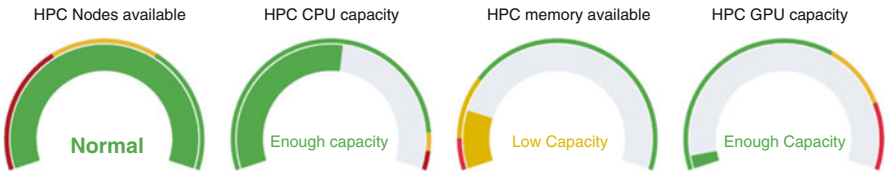

Fig. 7.2 Grafana-based reporting of HPC resources on local Galaxy starting page

there might come some pitfalls with associate open-source licenses, which need to be looked at closely. The Open-Source Initiative (OSI, https://opensource.org/) lists at least 104 OSI approved licenses, which might exist in different versions, qualities or have even been already retired. An open-source license ensures not just access to the source code, but in most cases also the free redistribution of the program, rules concerning derived work, proper acknowledgment of the code authors, some limitation of liability and the further distribution of the license itself. Here different license types include different rights and duties.

As a rule of thumb, just using unmodified FOSS programs in most cases can be seen as uncritical in industry, except for programs with licenses like GNU Affero General Public License (AGPL) or licenses explicitly restricting commercial use. Attention needs to be paid when modifying or further distributing FOSS programs with certain licenses, especially those with "Copyleft" clauses. As the GNU project supported by the Free Software Foundation states (https://www.gnu.org/licenses/copyleft.en.html): "Copyleft is a general method for making a program (or other work) free (in the sense of freedom, not "zero price"), and requiring all modified and extended versions of the program to be free as well." Such requirement can in some cases develop a viral character on additions made to FOSS programs and in the case of some licenses even on patents held by a company.

As know-how and intellectual property protection is a major concern in most industry, such cases need to be dealt with great care and attention. One prominent example is the violation of the GNU GPL license of the Linux kernel as part of the FRITZ!Box router operation system by AVM (https://avm.de) in 2011, which led to a lawsuit and finally to the distribution of AVM modifications under the same license conditions (see https://fsfe.org/activities/avm-gpl-violation/avm-gpl-violation.en.html). In this case, competitors of AVM could have benefitted from insights gained from the released source code.

Table 7.3 gives without any warranty of correctness or liability some examples of open-source licenses together with a "traffic light" indication of the perceived criticality of their usage in industry. In any case, it is advisable to get an expert opinion on the legal implications of each license in combination with the intended use. In general, to avoid possible future complications with open-source licenses in industry applications try avoiding strong Copyleft licenses (e.g., AGPL, GPL). If at all necessary, use unmodified libraries and executables called via "exec" or "fork" in the case of GPL, which are not distributed or bundled together with an industry application. In some cases, it is also possible that FOSS is available under different

Table 7.3 Examples of open-source licenses

Identifier	Name	Link	Usage
AFL 3.0	Academic Free License	https://opensource.org/licenses/AFL-3.0	Ok
AGPL v3	GNU Affero General Public License	https://opensource.org/licenses/AGPL-3.0	Critical
Apache 2.0	Apache License by Apache Software Foundation	https://opensource.org/licenses/Apache-2.0	Ok
BSD	Berkeley Software Distribution (3-clause)	https://opensource.org/licenses/BSD-3-Clause	Ok
CC0	Zero/public domain	https://creativecommons.org/share-your-work/public-domain/cc0/	Ok
CC-BY-NC	Creative Commons (CC)	https://creativecommons.org/licenses/by-nc/3.0/de/	Check
EPL	Eclipse Public License	https://opensource.org/licenses/EPL-2.0	Check
FreeBSD	FreeBSD License (BSD 2-clause)	https://opensource.org/licenses/BSD-2-Clause	Ok
GPLv2	GNU General Public License	https://opensource.org/licenses/GPL-2.0	Check
GPLv3	GNU General Public License	https://opensource.org/licenses/GPL-3.0	Check
IPL	IBM Public License	https://opensource.org/licenses/IPL-1.0	Check
ISC	Internet Software Consortium	https://opensource.org/licenses/ISC	Ok
LGPL v2	GNU Lesser General Public License	https://opensource.org/licenses/LGPL-2.0	Check
LGPL v3	GNU Lesser General Public License	https://opensource.org/licenses/LGPL-3.0	Check
MIT	MIT License by Massachusetts Institute of Technology	https://opensource.org/licenses/MIT	Ok
MPL	Mozilla Public License	https://opensource.org/licenses/MPL-2.0	Check
Ruby	Ruby License	https://www.ruby-lang.org/en/about/license.txt	Ok

licenses, here choose the less restrictive one, e.g., LGPL vs. GPL. For compliance reasons, it is also advisable to document the use of open-source licenses in industrial software applications. License finder tools exist (e.g., https://github.com/pivotal/LicenseFinder), which can be integrated into a continuous software build process to identify the licenses of the used software libraries.

Nevertheless, one of the original intentions of open-source licenses was that contributions by other parties help to improve the software overall for every user of the software. This principle should still be upheld even when FOSS is used in industry applications. Also, with industry application development it is possible to contribute to say more general features of a particular FOSS program, which do not constitute a competitive advantage and could be released to the general public. However, it is advisable to check the effort required and the acceptance of industry contributions to a particular FOSS program before contributing source code back to the original project. There are many examples of large companies like IBM, Google, Facebook, and many others making extensive contributions to FOSS packages. Advantages of contributing directly to FOSS packages include industry requirements becoming part of main FOSS releases and thus updates of FOSS packages require fewer modifications when deployed for industry applications.

7.2.3 Ensuring Software Quality and Code Standards for In-House Galaxy Tool Development

For highly productive data analysis workflows within the Galaxy system, it is important to ensure a high level of software quality and code standards for in-house developed functionality. Such functionality might not always be developed by professional software engineers, but also by biostatisticians, bioinformaticians, researchers, or even breeders themselves. This diversity of potential sources of customs tools to be integrated into the Galaxy system made it necessary to define a common set of rules or guidelines to which software quality adheres to:

- Increased process security (e.g., correct interpretation of analysis results).
- Similar end-user experience across several tools.
- Easy code transition between different developers (similar code structure, documentation, examples, and tests).
- Easier and faster to extend or refactor.
- Lower technical debt.

Basically, there are three code quality levels as part of these guidelines proposed (Table 7.4), which increase in requirements needed to be fulfilled by the respective software tool. Only tools with code quality level 1 (in some cases) and code quality level 2 (usually) should be considered for integration into the Galaxy system.

Table 7.4 Proposal of three levels of code quality

Level	Developers	Usage	Requirements
0	Only yourself	Prototypic, experimental, maybe throw-away code for one-time use	No recommendations
1	More than one developer	Tool used on a regular basis (more than once a week) Used by a low number of other users	Clear code structure and use of version control system Documentation of functions and potentially associated files Contains minimal working examples Passes Galaxy integration tests Preliminary performance evaluation
2	More than one developer and user support / software stewardship	An integral part of productive workflows for many users At least one other application relies on the correct operation	Standardized code structure which adheres to code style and/or templates Use of version control with CI/CD pipelines for Galaxy integration Documentation of the tool in a standard format (including external packages) Unit tests and integration tests with high coverage as part of CI/CD pipelines Realistic performance data for a variety of use-cases and test data

Only tools with code quality level 2 should be deployed to Galaxy productive instances (see Sect. 7.2.1). Tools with code quality level 1 can be deployed to Galaxy test instances for a limited number of users.

7.2.4 Ontologies for Structuring and Representing of Biological Knowledge

Ontologies are a framework for representing knowledge across a domain, in a format that is shareable and reusable. The goal is to provide standardization and structure, however standardization of terms in a domain is not enough for a successful ontology, adaptation is as important. One popular language for defining ontologies is

the Web Ontology Language (OWL, McGuinness and Van Harmelen 2004), which is built upon the resource description framework (RDF, Lassila and Swick 1998).

In the life science area, the "The Open Biological and Biomedical Ontologies (OBO) Foundry" (Smith et al. 2007) is a group of people working together to develop and maintain ontologies related to the field. They define principles for ontology development. More than 150 ontologies follow their guidelines.

In agribusiness, some of the important ones are the Agronomy Ontology (Jonquet et al. 2018), Plant Ontology (Bruskiewich et al. 2002, www.plantontology.org), Gene Ontology (Ashburner et al. 2000, www.geneontology.org), Crop Ontology (Shrestha et al. 2012, www.cropontology.org), Environment Ontology (Buttigieg et al. 2013, www.environmentontology.org) and Plant Trait Ontology (Arnaud et al. 2012, www.planteome.org). It is important to understand the structure of the ontology when working with it, for example the Gene Ontology, which is developed to describe the function of a gene product and contains three distinct graphs, one for functional domain: Cellular Component (where in the cell is the gene product active), Molecular Function (what is the specific function of the gene product), Biological Process (in what process is the gene product active). All of them are Directed Acyclic Graphs (DAG), which means that the edges in the graph have a direction, but there are no cycles: the direction is always one way. Standardized schemas are recommended whenever possible. Schema.org and in this case bioschemas.org would be a good place to start.

The relationships of ontological terms also encode knowledge, and they contain rules on how to traverse the relationships, for example on a hieratical structure it is possible to apply the "true path" rule, meaning that if something is annotated with a child term, all the parent terms are also implicit assigned. This could for example be if you have taken a sample from "vascular leaf," then you have indirectly also taken a sample from "leaf." This can be utilized when integrating data on multiple levels, for example, one dataset is measured in *vascular_leaf* "PO_0009025" and the other with *leaf* "PO_0025034", one can easily identify that *vascular_leaf* is a subterm of *leaf* and you can generalize to the nearest common ancestor. The same would be if we were to integrate *non_vascular_leaf* and *vascular_leaf*, the nearest common also be leaf, an example of this structure can be seen in Fig 7.3.

7.2.5 Using Knowledge Graphs for Linking Information Together

Recently another method for data integration has gained popularity, the knowledge graph. Have you ever asked a question on Google? Or used Alexa, SIRI, or Cortana? Then you most likely have been taking advantage of a knowledge graph, maybe without even knowing. The concept has existed since the 1980s but got traction when Google introduced their Knowledge Graph in a blog post in 2012. They described it as *"... we've been working on an intelligent model — in geek-speak, a 'graph' — that understands real-world entities and their relationships to one another: things, not strings."*.

A graph or a network as it is often called when referring to the practical use like social networks, or information networks, is a representation of data as entities with connections between them. In mathematical terms, an entity is a node or vertex, and a connection is an edge. A collection of nodes or vertices V together with a collection of edges E form a graph G = (V, E).

Graphs can be directed or undirected, for example, a network representing the co-occurrence of proteins in a cell is undirected, whereas a social network like Twitter is directed since the following is not reciprocal. The edges can also be unweighted or weighted, meaning that each edge has a weight assigned based on its importance.

Graphs are often visualized by drawing a point or circle for every vertex and drawing a line between two vertices if they are connected by an edge. If the graph is directed, the direction is indicated by drawing an arrow. Likewise, the weight of the edge is often represented by the thickness of the line between the vertices. Graphs allow the mathematical field of graph theory to be used when analyzing them. This could, for example, be looking at the number of connections for a node also known as the degree, or finding the shortest path between two nodes. Googles build their business around their Page Rank algorithm (Page et al. 1999), which identifies important websites among a network of websites linking to each other, which could also be seen as identifying the importance of a node based on its connections.

Emerging in the area of semantic web knowledge graphs are now seen widespread usage across many fields. There is no formal definition of a knowledge graph, though attempts have been made, one is by Ehrlinger and Wöß (2016), which define *"A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge."*; others are less strict and consider everything that is semantically connected in a graph to be a knowledge graph (Fig. 7.3).

A knowledge graph is a representation of a knowledge domain and its logic, using a graph. It can be seen as a network of nodes of information and edges connecting them instead of tables with rows and columns. By that, people and machines can benefit from a dynamically growing semantic network of facts about things and can use it for data integration, knowledge discovery, and in-depth analyses.

It allows companies and research institutes to utilize knowledge more efficiently. In the industry, the enterprise knowledge graph is nothing more than a graph containing a precise model of business processes, with which relevant questions, facts, and events can be analyzed more quickly. Adding more information to a knowledge graph increases its value. A lot of the work originated based on the semantic web idea (Berners-Lee et al. 2001) of creating computer readable connections between data on the internet. A human being can easily distinguish how a hyperlink relates on page with another, and what the reason for the link is, a computer cannot as easily do this. To deal with this a set of specifications that are widely used also within knowledge graphs were developed, including the Resource Description Framework (RDF) Core Model, the RDF Schema language (RDF schema), the Web Ontology Language (OWL) and last the SPAQRL query language to query data in RDF format.

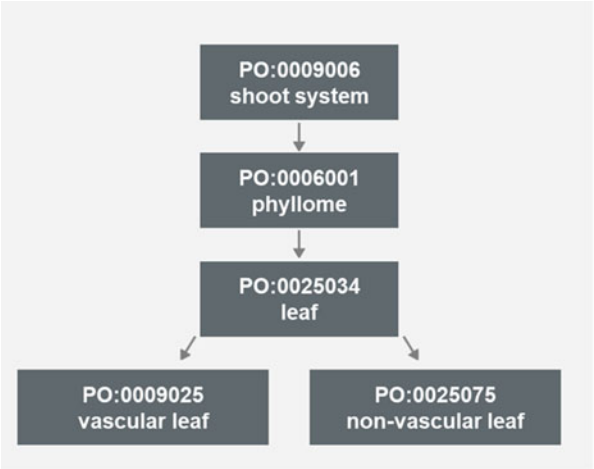

Fig. 7.3 Example of the direct children and ancestors of the ontological term "Leaf"

7.2.6 Representing Data in a Structured Format

Relational databases have been the de-facto industry standard for storing data since the 1960s. They store structured data in tables with defined columns and rows containing this data. RDBMS requires users to adhere to a schema of the data and structure their data and applications according to this.

Graphs are among the most flexible formats for data structure. In a graph, information is described as a network of nodes and links between them, rather than tables with rows and columns. Both the nodes and edges can also have attributes assigned to them. Graph-based systems are easier to expand, as they often are schemeless. It is still recommended to adhere to a schema, but it gives the flexibility of extending the schema when new data or connections arrive. There is usually not an optimal way of best modeling your data, it all depends on your question. Therefore, one should be prepared to evolve the data schema as the data and experience evolve.

Data can be modeled as graphs in multiple ways. One approach is to use the RDF standard. RDF stands for Resource Description Framework and it is a W3C standard for data exchange in the Web and is built using the existing web standards of XML and URI. It is used for describing data using relationships between objects. RDF connects data as triples, a triple is a statement about data consisting of three parts, the subject, predicate, and object. An example could be the Cellulose synthase A catalytic subunit 8 from the plant Arabidopsis thaliana, it has the id Q8LPK5 in the Uniprot (UniProt Consortium 2019) protein database. Uniprot offers API access to their data as triples. The connection between Q8LPK5 and Arabidopsis could be represented as

<http://purl.uniprot.org/uniprot/Q8LPK5>
<http://purl.uniprot.org/core/organism>
<http://purl.uniprot.org/taxonomy/3702>

The predicate in this case is from the Uniprot internal schema and is of the type organism. The definition of organism in this case is "The organism in which a protein occurs." The Subject is our protein of interest, and the organism is then defined in the Object, and is a taxonomy id referring to Arabidopsis thaliana. Another example is

<http://purl.uniprot.org/uniprot/Q8LPK5>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://rdf.ebi.ac.uk/resource/ensembl.transcript/AT4G18780.1>

Where the predicate is #seeAlso from a schema provided by W3, it links according to the specification, a resource to another "that might provide additional information about the subject resource." As can be seen here, it is possible to mix URIs from different sources, one is an internal Uniprot URI, and the other is referring to one from W3. The URI serves to standardize the context and meaning, by creating a schema and definition for the connections. Just because a database is schemaless, does not mean that it should be used without schemas, it just gives the flexibility to change and expand as needed. A good place to look for public schemas is schema.org.

The other option for storing data in a graph, is the Labeled Property Graph (LPG), in LPG you have a set of nodes and edges. Both nodes and edges have a unique ID and can contain key-value pairs to characterize them.

Both are valid approaches for building a knowledge graph, and which one fits best need to be evaluated for a given use case, based on many variables, such as what questions we want to be able to answer, which infrastructure is available, what do we need regarding performance and analytics capabilities. It is also important to remember that just as not all data types fit well in relational databases, so is it also that not all fit well in graphs, there is no one-size-fits-all solution for all needs.

7.2.7 Building Your Own Knowledge Graph

When starting to think about implementing a knowledge graph in a business, it is important first to identify a need and what questions you want to answer, and how they add value to the business. Then start with a minimum viable product to demonstrate the value. Always keep stakeholders closely informed to ensure that buy-in is created.

First step is usually to gather and process relevant datasets as well as identifying necessary taxonomies, ontologies and controlled vocabularies that would serve best in achieving the goal. It is beneficial in the beginning to identify datasets that do not change often, as well as keeping size in mind. This minimizes the need to spend too

much effort on updating data and scaling infrastructure. Generally, start small and then grow when enough interest and buy-in has been created, and more resources are made available.

It is important to clean the data before uploading, remove invalid entries, adjusting dates to be in the same format etc. Then it is time to get an overview of your data and design your semantic data model using ontologies etc. on how to use data together. There is no best fit all model, it all depends on the questions you want to answer. Often the data model will evolve with your knowledge graph.

Integrate data loading with Extract Transform and Load (ETL) tools to ensure quality and consistency when moving data from one system or format to the graph, Generate Semantic metadata to make it easier to find, and reuse data. This usually goes hand in hand with a strategy for FAIR data (Wilkinson et al. 2016) (see Box. 7.2).

Augment your graph via reasoning analytics and text analysis. Enrich your data by extracting new relationships from text, apply inference algorithms to the graph to identify hidden relationships, and extend your knowledge graph with information from the graph itself. For example, degree or betweenness of nodes. It is also possible to train models to evaluate if a connection is missing, or if it is added wrong, this kind of use-case for machine learning can be beneficial especially when manual data entry has been part of the process. In the end your graph will now have more data than the sum of its constituent datasets. Lastly, set up procedures to maintain and continuously load data into the graph to keep it alive.

7.2.8 Identifying Use-Cases for Applying a Knowledge Graph-Based Approach

Identifying the ideal proof of concept use case should not be difficult, a lot of organizations have already demonstrated the effectiveness. Some inspiration for popular use-cases across industries

- Recommender systems: discovering related data and content.
- Semantic data catalogs: agile data integration and improving FAIRness of the data within the organization.
- 360 views of customers, products, employees, users etc.
- Knowledge discovery: intuitive search and analytics using natural language.

One important cornerstone to identify suitable use-cases is an active survey of potential business problems among colleagues of different areas inside your organization. Solving these business problems should generate a certain value for the company, which exceeds the costs of implementation of such knowledge graph. In our experience, workshops with a good mixture of domain experts and data experts are beneficial to identify the questions to be answered on a solid data foundation. Agile approaches, for example Event Storming (Brandolini 2013), help to reduce the discrepancy for a common understanding between domain experts

and data. Factors like data availability, quality and governance are other important factors that influence this decision. Additionally, potential use-cases could be rated by the number of people they impact. Reaching a larger audience from the beginning can help creating buy-in from more people as well as reaching people with novel use-cases.

Box 7.2 FAIR Principles
What is FAIR principles?

Findable

Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

Accessible

Once the user finds the required data, she/he needs to know how they can be accessed, possibly including authentication and authorization.

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

Reusable

The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Source: https://www.go-fair.org/fair-principles/

7.3 Use-Cases

7.3.1 Using Galaxy Workflows for Ad-Hoc Data Analysis on Integrated Data

Over the years, Galaxy became more and more integrated into our research software infrastructure. We utilize specific in-house developed Galaxy tools to provide input data from different data domains such as genetic, phenotypic, OMICs data as well as genetic and genomic map data that are analyzed in different Galaxy workflows to support breeding decisions. The most common datatype utilized by public Galaxy tools is the tab-separated format. In order to use the general-purpose Galaxy tools, but also provide certain data format constraints and rule sets for specific in-house tools and user guidance, e.g., for workflow definition, we follow the somewhat pragmatic approach to define in-house data types based on the Galaxy tabular data format (see Fig. 7.4).

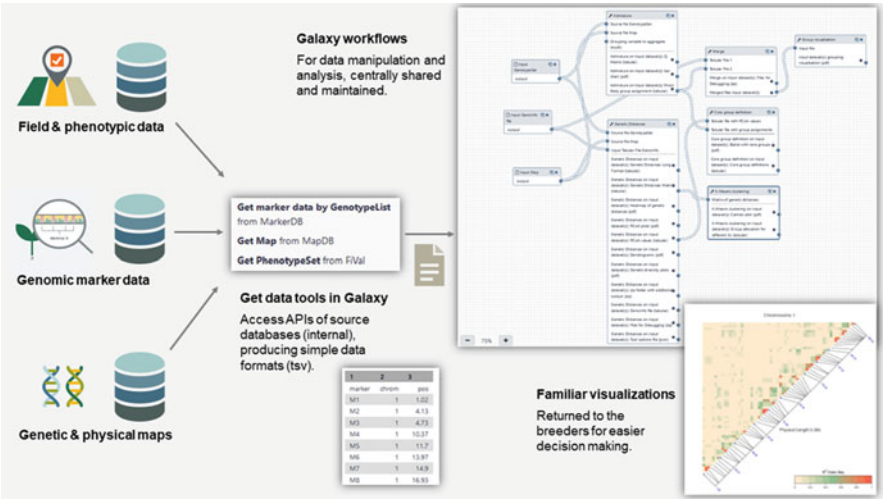

Fig. 7.4 Example of data used in Galaxy ad-hoc integration approach

More specifically, for each input data domain further analyzed in Galaxy workflows, we provide custom Galaxy tools ("Get Data" tools) that serve as connectors to other systems providing interfaces to specific integrated data. Those tools typically serve as entry points to ad-hoc analyses as part of Galaxy workflows executed in a self-service manner by our breeders. The output file(s) of the different Get Data tools are based purely on the tabular Galaxy data type for the reasons mentioned before but are specific to each data domain. This allows us to implement specific format validators on formatting and content. Additionally, this reduces errors for other in-house developed tools that depend on this specific input data, both within workflows but also for stand-alone tool runs inside Galaxy.

The genetic marker data is formatted as a named matrix (marker × genotype) which contains unphased biallelic SNP chip array data (AA, AT, AC, etc.) of the genotypes, whereas sporadic missing data is encoded as NA. Phenotypic data is encoded similarly in a genotype x trait matrix containing quantitative trait data. All SNP markers are cross-linked across different reference sequences within the different crops, thus allowing a precise location of trait-reference genome association.

Using the rich Galaxy API, we then transfer result data into downstream applications for storage and combined analysis of historic data.

To ease integration of data across sources and minimize errors, it is important that the data in each source accessed by our tools, follow the same standards and utilizes the same vocabularies and ontologies. Especially when combining with historical data, this can often be a challenge. To connect multiple heterogeneous data sources, a knowledge graph can be an advantage in ensuring that data is aggregated correctly. It allows heterogenous data to be connected with standardized machine-readable links and allows computational traversal between data sources identifying links between them and serves as a guide on where data could be aggregated.

7.3.2 *Knowledge Graphs to Enrich Genome-Wide Association Studies (GWAS) Data*

GWAS is a common approach to accelerate genomics-assisted plant breeding by detecting the genetic basis of phenotypic variation (e.g., traits of interest) on population scale based on many individuals (Tibbs Cortes et al. 2021). If certain genetic variations, usually Single Nucleotide Polymorphisms (SNPs), are found to be significantly more frequent in individuals expressing the desired trait compared to individuals that do not, the SNPs are said to be statistically associated to the trait of interest. These SNPs can serve as powerful pointers to genomic regions to assist in the selection of favorable plants for breeding and further used to support identification of candidate genes possibly involved in a certain trait. To further streamline and automate the knowledge generation in molecular breeding, we developed custom-made downstream web applications for specific approaches such as GWAS and provide APIs that allow feeding expert revised GWAS data into knowledge graphs.

In-house computed GWAS are undertaken in Galaxy on integrated genetic and phenotypic data in a way that allows traceability of the results. We then provide breeders a web-based platform to access computed results from Galaxy and store GWAS results alongside additional relevant information about the genetic material and other data in our in-house GWAS database. Finding a marker or a candidate gene is challenging. First scientists need to inspect large amounts of heterogeneous data to obtain a list of candidate genes, which then needs to converge to a ranked prediction of the most likely candidate(s) involved in the trait of interest. GWAS relies on all phenotypic data being described the same way, and accessible in the same format.

Often the SNPs cannot explain all the phenotypic variation. One reason for this is that GWAS relies on a strict P-value threshold of the SNPs after adjustment for false discovery rate to avoid false positives. This can partly be overcome by larger population sizes, however that is both costly and not always feasible. Another option is to bring in extra data to enhance it and add evidence to weaker SNPs. This could for example be gene co-expression networks, protein-protein interactions, gene regulation, protein domain information, functional information from homologues in other species, metabolic pathway information, or supporting evidence from literature.

GWAS is often applied to analyze complex traits such as resilience to drought (as opposed to monogenetic traits that follow strictly Mendelian inheritance). Associated SNPs might be distributed across many genes addressing one or more metabolic pathways. Here, trait expression can only be explained by a concerted action of multiple genetic factors that are often to a varying degree influenced by non-genetic factors such as environmental factors. To identify these, it can also be beneficial to bring in auxiliary information as described before.

A knowledge graph linking this data together with the relevant identifiers and synonyms can speed up the process of integrating this data, as well as augmenting

the results with other data sources. Enabling our researchers to get a better overview of relevant information and take information-based decisions. In the end the estimates of success need to be updated for the models being used, a feedback loop needs to be in place, for updating with experimental results on the predictions. If we want to augment the data with external sources, we need to be able to find ways of integrating this data. Sometimes there is no direct link between data points. For instance, if we want to add environmental information to our analysis for identifying candidate genes for a given trait. If the individual data points are linked, we can traverse the graph, using a graph algorithm such as Dijkstras shortest path (Dijkstra 1959). Collected data could include temperature measured at a location, a plant with a given mutation has been grown on that location during a specific time. That plant shows a particular phenotype. It is then possible to find the nodes of data where aggregation can take place to be able to connect these data and conclude about the temperature phenotype relationship.

7.3.3 Knowledge Graphs to Augment Metabolite Analysis

Plants produce a variety of small chemicals or metabolites, this could, for example, be stress hormones, measuring these metabolites is an essential part to understand more of how a given variety of plant responds. Analyzing and interpreting metabolite measurements can be time-consuming. This is a great example where knowledge graphs can assist us in making sense of the information, by augmenting the data we get out Measurement IDs, which can be matched with the corresponding metabolite, its name, synonyms, composition. It can also be linked to previous knowledge, such as literature and previous measurements.

This makes interpretation easier. At the same time, it can also be linked to internal costs of measuring, how long time does a measurement take, and what is the capacity for measuring. This can then be taken directly into context as a cost/benefit when analyzing data and deciding which metabolites are generating the most value by measuring. Questions like “What is the most optimal composition of measurements we can achieve for a given price if we want to predict a certain outcome?” can be answered. An example is seen in Fig. 7.5, a peak has been assigned with PN_10824, This can be difficult to interpret for a scientist, since this is not directly obvious what it refers to. Though, if that id was linked together with other information, it would be easy to see that it was abscisic acid, and it is a hormone that has been shown to be involved in regulating root growth. By saving the time the scientist has to spend looking for this information, and at the same time ensuring that all scientists have the same information available, we can increase efficiency and take better and more informed decisions.

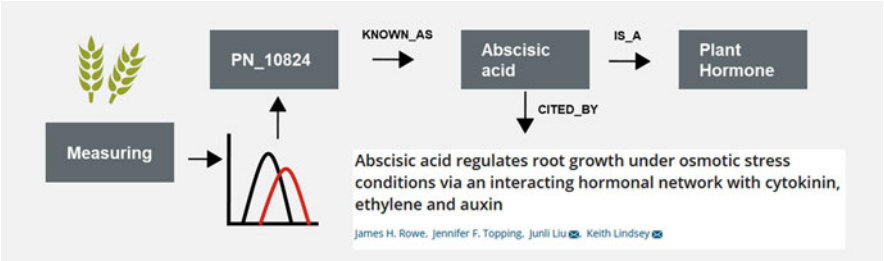

Fig. 7.5 An example of how data could be related to augment measurements of metabolites

7.4 Discussion and Conclusion

Solving the many challenges related to feeding the worlds growing population will be complex. It will involve many people and a lot of cross-disciplinary research to understand the interplay among plants, environment, people, logistics, and many other areas. Being able to handle and integrate large amounts of heterogeneous data from many sources will be an integral part of solving this challenge.

Standardized and reproducible research can help us speed up this process, by minimizing the number of errors and maximizing the utilization of the data generated. The development of the FAIR was an important step in the right direction. Novel analytical methods that can take advantage of larger and more complex datasets in the analysis are being developed. This is particularly true for machine learning, where methods such as Graph Neural Networks allow for the analysis of complex knowledge graphs. Developing more complex models and analyses could enable researchers to reach their conclusions faster with more precision. Standardized workflows and data integration is an important part of this. Since the methods are only as good as the data that goes in.

Open-source tools, standardized vocabularies and knowledge graphs are an integral part of the processes at KWS to solve these challenges. Enabling plant breeders and scientists to deliver better outcomes, storing information and as basis for improved decision-making for the future, to learn from and improve upon.

7.4.1 The Challenge of Increasing Data

It has been estimated that in 2015–2016 more data were created than in the preceding 5000 years of human history, and that amount increased so in 2017 alone, a similar amount was created. To be able to generate value, having information is not enough, the context for the information is important to be able to translate this into actionable insights and knowledge.

The data landscape in most tech businesses constantly grows more complex due to new technologies producing new measures and new tools for analyzing data. Some of it is structured data such as measurements from sensors or transactions in banks, but large amounts are unstructured, such as images, documents, relationships. A lot of knowledge is lost, due to a lack of context for the data.

One way of adding context to data is to connect it with other data. Data integration is the process of combining data from different data sources into a single, unified view. However, integrating data is one of the most time-consuming parts of a data scientist's work life.

Many enterprises suffer from data being locked in silos, making integration difficult due to different data models, descriptors, nomenclature, or unstructured data. This in the end prevents an optimal utilization of the accumulated knowledge inside an organization.

7.4.2 Combination of Approaches Needed

Data silos are a trait of many larger organizations, however, silos are a big hurdle toward many business-critical processes, for example, app development, data science, analytics, reporting and compliance. Implementing efficient enterprise data management can both decrease costs and increase performance and generate additional value for organizations and customers. There is no solution that fits all, but creating standardized pipelines and workflows, and keeping file formats as simple as possible are good rules of thumb. Providing data integration pipelines in a system like Galaxy, not only saves time for the user when they need to run an analysis, it also ensures reproducibility.

It is critical to knowledge discovery to be able to integrate different sources of data because it allows different information about the same entity to be related in new ways. A big challenge is synonymic naming and syntactically different identifiers. In a biological setting, this could be gathering different data that describe the same biological entity (e.g., gene, transcript, protein, etc.). Using ontologies can aid in the automatic integration and aggregation of data from multiple sources and ensure that data is reusable across departments. Data by itself for example in a data lake is not knowledge and has limited usage. Using graphs another layer of context can be added to the data when integrating it, this, in the end, gives more information than the sum of its parts, since the features of relationships, for example, node degree has been shown to be highly predictive as well.

Adding semantic or self-descriptive links and features to the data allows both computers to read it, but also makes onboarding of new staff members and exploratory data analysis easier since it is possible to read directly what a given piece of data represents. One way of dealing with this is to use an integration layer between the data sources and the end view. The integration layer will then be queried using for example Cypher or SPARQL, to then get the results from underlying data sources, the query will be translated into the query language of each data source. The integration layer is based on Ontologies and structured vocabularies to identify how

data should be mapped. This allows the utilization of machine learning and enables researchers to reach their conclusions faster with more precision. Standardized workflows and data integration is a crucial part of this.

References

Afgan E, Baker D, Batut B, Van Den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A (2018) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46(W1):W537–W544

Arnaud E, Cooper L, Shrestha R, Menda N, Nelson RT, Matteis L, Skofic M, Bastow R, Jaiswal P, Mueller L, McLaren G (2012) Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes. In: International Conference on Knowledge Engineering and Ontology Development, vol 2. SciTePress, Setúbal, pp 220–225

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25(1):25–29

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284(5):34–43

Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol 89(1):19–10

Brandolini A (2013) Introducing event storming. goo.gl/GMzzDv. Accessed 8 Jul 2017

Bruskiewich R, Coe EH, Jaiswal P, McCouch S, Polacco M, Stein L, Vincent L, Ware D (2002) The plant ontology (TM) consortium and plant ontologies. Comp Funct Genom 3(2):137–142

Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE (2013) The environment ontology: contextualising biological and biomedical entities. J Biomed Semantics 4(1):1–9

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1(1):269–271

Ehrlinger L, Wöß W (2016) Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS) 48:1–4

Jonquet C, Toulet A, Arnaud E, Aubin S, Yeumo ED, Emonet V, Graybeal J, Laporte MA, Musen MA, Pesce V, Larmande P (2018) AgroPortal: a vocabulary and ontology repository for agronomy. Comput Electron Agric 144:126–143

Lassila O, Swick RR (1998) Resource description framework (RDF) model and syntax specification–W3C Recommendations. Technical report, World Wide Web Consortium. https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

McGuinness DL, Van Harmelen F (2004) OWL web ontology language overview. W3C Recommend 10(10):2004

Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Stanford InfoLab, Stanford, CA

Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, Arnaud E (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the crop ontology developed by the crop communities of practice. Front Physiol 3:326

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25(11):1251–1255

Tibbs Cortes L, Zhang Z, Yu J (2021) Status and prospects of genome-wide association studies in plants. In: The plant genome, p e20077

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47(D1):D506–D515

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3(1):1–9

Part III
Integrative Data Analysis

Chapter 8
Integrative Data Analysis and Exploratory Data Mining in Biological Knowledge Graphs

Marco Brandizi, Ajit Singh, Jeremy Parsons, Christopher Rawlings, and Keywan Hassani-Pak

Abstract Modern life sciences are based on large amounts of data in many different formats, which model in many different ways a wide variety of interrelated species and phenomena at multiple scales. In this chapter, we show how to integrate and make sense of this wealth of data through digital applications that leverage knowledge graph models, which are ideal to flexibly connect heterogeneous information. Furthermore, we discuss the benefits of this approach when applied to data sharing practices, which maximise the opportunities to reuse integrated data for novel analysis and digital applications. Knetminer, a genetic discovery platform that leverages knowledge graphs built from molecular biology data sources, will be used as a significant use case of the described concepts.

Keywords Knowledge graph · Exploratory data mining · Network visualisation · SPARQL · Cypher · Jupyter

8.1 Introduction

In the past 20 years, the life sciences have become increasingly data-driven. In 2001, genomics took a leap forward with the announcement of the official completion of the human genome sequencing project, which cost hundreds of millions of dollars and decades of work (November 2018). By 2016, two milestones had been reached: the cost of sequencing a human genome fell to less than one thousand dollars and the work could be completed in just 2 days, hence more than one million human genomes have now been sequenced worldwide (Stephens et al. 2015). Similarly, during the same period a huge wealth of life science-related information of all

M. Brandizi · A. Singh · J. Parsons · C. Rawlings · K. Hassani-Pak (✉)
Rothamsted Research, Harpenden, UK
e-mail: marco.brandizi@rothamsted.ac.uk; ajit.singh@rothamsted.ac.uk; jeremy.parsons@rothamsted.ac.uk; chris.rawlings@rothamsted.ac.uk; keywan.hassani-pak@rothamsted.ac.uk

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_8

kinds has been produced. As another significant example, the incredible speed with which COVID-19 vaccines and therapies have recently been developed was possible thanks to mankind's ability to collect, share and analyse vast amounts of data in collaborative ways (Hutson 2020). Working with data from the life sciences, however, has multiple challenges, mainly due to the fact that life is a very complex phenomenon, which occurs at multiple scales and spans an extensive network of interactions (Regenmortel 2004). As a consequence, in life science, information sharing, integration and collaboration have become paramount (Figueiredo 2017; Check 2013; Wise et al. 2019).

In this chapter, we focus on sharing and analysing biological knowledge by means of knowledge graphs and exploiting data standards. This is a powerful approach that is particularly suited to explorative analysis of integrated heterogeneous data. Getting a quick overview of what is known about a given subject, or making new discoveries by linking usually unrelated areas of research are two key advantages offered by the adoption of standardised knowledge graphs.

The chapter is organised as follows. In Sect. 8.1 we give an overview of the many digital resources that are available to address research questions in life sciences. Section 8.2 presents an example of knowledge graph-based investigation into data from studies of the genetic factors that influence wheat crop yields. This will be based on the KnetMiner platform which is being developed by the authors as part of the Designing Future Wheat (DFW) project (Designing Future Wheat [Internet] 2021). KnetMiner provides a set of tools for building and exploring knowledge graphs built from a wide range of data sources related to molecular and functional genomics. In Sect. 8.3 we show how exploiting published knowledge graphs programmatically can be an additional way to explore life science data and we give an overview of the best practices and computational solutions for doing so. Conclusions from the use case and related bioinformatics resources are presented in Sect. 8.4.

8.1.1 The Landscape of Data-Driven Research in Life Science

It is useful to consider the range, types and origins of data currently driving life sciences research. The development of high-throughput technologies, which became increasingly affordable and hence scalable in the early 2000s, led to the production of large amounts of measurements concerning many kinds of biological phenomena (Lightbody et al. 2019; Yang et al. 2020). Here, we offer different axes by which the data-driven life science research and practice can be classified. Having an overview of the diversity of digital resources and ways to use them in the life sciences helps understanding how such resources are integrated in integrative methods such as those bases on knowledge graphs.

Technology-Based Axis Determining the genetic sequence of an organism has been one of the first available high-throughput technologies (Heather and Chain 2016). Computational methods to compare genes by their sequence similarity, such as

BLAST, evolved together with these technologies (Altschul 1997; Chowdhury and Garai 2017), along with statistical pattern discovery methods like data clustering, hidden Markov models or support vector machines (Bang et al. 2010). Gene expression research (Barah 2021) studies the quantitative changes in gene transcripts (which encode proteins) that result from changes due to cell/tissue development or as a result of treatments or environmental stresses. The methods used for these studies are related to gene sequencing and are enabled by high-throughput technologies, initially using microarrays, more recently through the more modern RNA-Seq (Mantione et al. 2014). Techniques like mass spectrometry (Watson and Sparkman 2007) allow for tracking protein abundance and activities independently from the genome or transcriptome. Many gene sequencing technologies are based on bioluminescence, one of many forms of imaging techniques (Shorte and Frischknecht 2007). These technologies contribute to the production of a wealth of useful data, which are often published in public databases. In biomedicine, significant advances in medical informatics have been made to track a variety of patient and clinical trial data, which are often collected from imaging, sequencing and other high-throughput equipment (Baumgartner et al. 2016). The recent development of life sign sensors, wearable devices and Internet of Things will generate even more data for medical research and healthcare. In other life science fields, both these general biomolecular technologies and more domain-specific ones are being developed. For example, image-based plant phenotyping platforms, multi-spectral imaging from UAVs, satellite telemetry and agricultural machine sensors are all examples of equipment used in agronomy and ecology research (Li et al. 2014; Beluhova-Uzunova and Dunchev 2019).

Phenomena and Scale Axis Living systems can be considered as a network of interacting and dynamic processes, which happen at many different scales. Molecular biology concerns mostly the molecular and chemical processes that happen at the subcellular level. In addition to the data resources that have been developed to capture genetic and genomic information, databases are also available that report relevant cell lines collected for research purposes or to map the repertoire of cells present in multicellular organisms (Forbes et al. 2017). At a higher level, many biobanks have been developed to collect human and other animal tissues from various organisms and conditions (Mayrhofer et al. 2016; Gostev et al. 2012). Similarly, plant biology data resources are available and are used by academic and industry researchers working in plant and crop genetics (Horler et al. 2018). Studying cohorts of individuals or entire populations is another useful technique in biological research. Software solutions are available for quick access (e.g., by means of query federation) to multiple repositories of clinical data and clinical trials (Murphy et al. 2007), many of which work together with standards and solutions for digital healthcare. Statistics and AI-based methods of machine learning have been applied to population genetics to understand gene functions by means of so-called genome-wide associations studies (GWAS (Yang et al. 2021; Nicholls et al. 2020)). Resources such as environment microbiology catalogues (Choi et al. 2017; Schüngel et al. 2013) or knowledge bases to support agronomic field trials and data-driven

ecology (Perryman et al. 2018; Arnaud et al. 2020) are examples of data-driven life science extended to ecosystems up to the planetary scale.

Digital Paradigms and Analysis Methods Axis Depending on the phenomenon under investigation, the data technology and the research purpose, many different data, mathematical and computational models are available to undertake data-driven science. Over the years, the life sciences have adopted a wide range of approaches including chemistry-based models (Demir et al. 2010; Degtyarenko et al. 2007), the physics of protein structures (Ausiello et al. 2008), systems theory (Dada and Mendes 2011), interactions in cell populations (Germain et al. 2011) and interactions in whole ecosystems (Meyer 2016). Statistics is a fundamental tool to compute estimations in experiments like clinical trials and field trials. Advanced synthesis methods like PCA, clustering, stochastic models and logistic models can be used to summarise the main characteristics of large datasets (Bang et al. 2010). In recent years, techniques from the discipline of artificial intelligence (AI), such as neural networks and machine learning, have gained enormous popularity (Tang et al. 2019; Liakos et al. 2018), since these techniques can be easily adapted to a wide variety of problems, especially as ever more data were available to tune the AI model parameters and ensure their predictive accuracy.

8.1.2 Data and Knowledge Representation in Life Science

The many approaches mentioned above both influence and are influenced by data models and data formats used both in bioinformatics or in the wider data sciences community. These have also changed over the history of computer science. Until the 1970s, flat data file formats were widely used in most computational applications. Flat formats for molecular sequence data such as FASTA (Mills 2014) are still used for particular data representations. These formats have an ad-hoc structure, which isn't based on any general syntax, and usually they represent uniform entities, such as a list of persons or a list of genes. In the 1980s, the relational model, and the SQL query language that is used to query it, have been widely studied, developed, adopted and standardised (Polding 2018). Nowadays, relational databases are widely used in applications where predefined data schemas can be defined and do not change significantly over time. In such a situation, this model can be very efficient, both in terms of space and time. Delimiter-based file formats, such as CSV (A Comparison of Serialization Formats [Internet] 2019), can be used to represent the relational model, when proper conventions are adopted to relate records in these files. Data marts are views on relational databases that are used to allow programmatic access to data. For instance, in life science, they support applications like Ensembl, a reference repository for gene information (Kinsella et al. 2011). Over the years, relational databases have been complemented by so-called NoSQL solutions (Corbellini et al. 2017; Sharma et al. 2016), where one is not constrained to predefined and rigid schemas. NoSQL systems have been developed

together with data formats like XML and JSON (A Comparison of Serialization Formats [Internet] 2019), which allow for the composition of trees of data items. These data formats are often used to realise networked services to exchange and process data in a distributed way. Application programming interfaces (API) based on web services, are a prominent example of those services (Surwase 2016; Brito et al. 2019).

Contemporary data management approaches are strongly influenced by semantics and knowledge representation needs. In fact, since digital computers are unrelated to the cognitive abilities of animals like humans, representing the real world by means of data and formal representations of their meaning is a fundamental step towards automating knowledge processing. Data semantic representations vary in a scale of expressivity and inference power (McGuinness 2005): from simple dictionaries of entity and relation types which are easy to use, though less practical for automated deduction from existing data, right up to the complex disciplines for the formalisation of knowledge by means of formal logics, abstract algebra, and derived computational tools (McGuinness and Van Harmelen 2004).

8.1.3 The Property Graph Paradigm

This chapter focuses on integrative data analysis based on knowledge graphs. The property graph data model, a particular way to represent knowledge graphs, is very flexible when heterogeneous information has to be put together. It provides a simple yet powerful knowledge representation paradigm, which, additionally, permits the characterisation of knowledge semantics at different levels of formalisation and expressivity (Zhang 2017). Figure 8.2 (top) clarifies what this means: nodes like the gene GL1 or the protein P27900 and (oriented) relationships between nodes are the basic building blocks of the model. Both nodes and relationships can have at least one associated type (Gene, Article, encodes, mentions), which makes it easy to characterise a given set of "instances" according to a data scheme. A scheme can be a rich relation network of types, and property graphs can be used to describe this as well, for the graph model is flexible enough to accommodate either a simple schematisation such as a list of types (i.e., a vocabulary or a code list), or a formal ontology based on first-order logic (McGuinness and Van Harmelen 2004). In the bottom diagram of Fig. 8.2, showing the RDF approach to knowledge graph representation (see Sect. 8.3.1), common schematisation mechanisms are used: Gene and Protein are declared as subclasses of BioMolecular Entity and the endpoint types (domain and range) for the "encodes" are specified. Not only do these simple schema elements allow for documenting "instance" data (e.g., how you're supposed to use the "encodes" relationship), they enable the inference of new knowledge. For instance, when data are loaded in a system that declares "GL1 encodes P27900," the system can deduce new knowledge like: GL1 is a Gene, P27900 is a Protein and both are biomolecular entities. While this implicit knowledge appears trivial to imply for a human being, computers need to encode

even the simplest logic to "understand" the meaning of data. Moreover, automated reasoning can be much more advanced than this basic example (Description Logics 2014).

In addition to being a good base for formal knowledge representation, property graphs can be used to report "fuzzy" information as well. For example, in the figure at issue, "mentions" has an attached "score" attribute. This suggests that the reported relations are not certain (as it is usually required by plain first-order logics and ontology languages that are based on it), but an estimation computed by some software (reported via "source"). It is possible to go much beyond this basic example, by enriching an initial property graph with predicted links, node and relation properties, based on AI and machine learning methods. Well-known examples of such kind of enrichment are described in (Gabrilovich and Usunier 2016).

The term "knowledge graph" is less well-defined than "property graph" (Ehrlinger and Wöss 2016). To summarise, it usually refers to property graphs representing non-trivial knowledge about a particular subject, often mixing collected data, schemas of various types, and inferred knowledge. Knowledge graphs intended this way show the power of graph-based data models with data integration tasks. In fact, data collected from many diverse data sources can be merged together by adopting common identifiers (e.g., data about P27900 might come from both Ensembl and UniProt, the merge is ensured by both sources using the same protein ID), and can be given a semantic description by mapping data to common schemes.

8.2 Using KnetMiner and Other Resources to Investigate Wheat Yield

To illustrate the different aspects of modern data integrative exploration and analysis in molecular biology, we shall consider an example of an investigation from crop science. Similar examples of more specialised investigations can be found in (Hassani-Pak et al. 2021; Adamski et al. 2020). Increasing agricultural productivity in a sustainable way is one of the UN's Sustainable Development Goals (SDG U 2019) and it is considered fundamental to ensure the nutritional needs of an increasing world population can be satisfied without an unbearable negative impact on the environment. Rice and wheat are the world's two leading food crops, together they serve as a staple food for almost half of humanity (Yang et al. 2021; Ling et al. 2013; Nadolska-Orczyk et al. 2017). Severe climate instability and emerging diseases pose a major threat to crop production and yield. A large number of research projects, pre-breeding and breeding programs are being funded to identify novel genes and to improve crop traits such as stress tolerance and improved crop yield. However, identification of candidate genes and experimental validation, from

lab to greenhouse to field, is a slow process that can last from years to decades. Following a wrong lead wastes significant effort, time and money.

With the availability of an increasing number of crop genomes, including: data from multiple "omic" layers, high-quality phenotypic data from large replicated field trials, in conjunction with the wealth of other information types from model and non-model species; we should now be in a position to accelerate targeted gene discovery to validation pipelines. However, marker and gene discovery for important agronomic traits remains challenging. Firstly, in many research organisations and breeding companies the expertise and software technologies to analyse the volume and variety of un-integrated data are simply missing. The second major complexity is the fact that most agriculturally important phenotypes are nearly "omnigenic," i.e., underpinned by highly complex and interconnected networks, with "core" genes explaining only small fractions of the genetic variance (Boyle et al. 2017). Approaches that integrate the interconnected networks within and across 'omic layers may be the only way to progress beyond the "stamp collecting" phase of cataloguing single marker-trait associations.

KnetMiner is a software package developed at Rothamsted Research that can embrace this complexity and exploit the considerable additional information that can be obtained by integrating the complementary plant genomes, genotype, phenotype and multi-omics data into a curated and machine mineable data model. This integration can enable the development of systematic approaches to find genes that are beneficial to crop performance and when perturbed through potential interventions such as gene editing approaches, have a positive impact on the overall biological outcome without producing negative side effects.

In the following use case, we demonstrate how an integrated approach can support scientists to overcome some of these challenges in gene discovery and knowledge mining. We start our investigation with a list of rice genes known to be associated with yield (Nadolska-Orczyk et al. 2017). We have identified the corresponding orthologs in wheat using the BioMart interface of Ensembl Plants (Kinsella et al. 2011) and we will describe an iterative approach to search for gene-trait linkages using KnetMiner for wheat and a knowledge graph composed of over 40 distinct datasets (Hassani-Pak et al. 2016). The KnetMiner web application (Hassani-Pak et al. 2021) can be searched with a list of genes derived, for example, from a data-driven analysis or literature review. In this case, we are interested in exploring a set of genes that have been associated with yield in rice, but their role in wheat has not yet been well characterised. We enter the list of gene identifiers initially found as described above into the Gene List box in KnetMiner and press the search button to obtain knowledge that is linked to the listed genes. Using the Evidence View resulting from this search, we can find the terms that are enriched in our gene list.

Such a gene set enrichment analysis is usually performed for single GO or pathway annotation datasets. In our integrated graph approach, we can test the enrichment of any network node present in the gene network, e.g., traits, phenotypes and diseases. Having developed a preliminary understanding of the key processes, we now intend to zoom in and ask more specific questions about our genes of interest

in relation to certain keywords. As the basis for the second search iteration, we decide to use the highly enriched terms from the first search with other yield-related keywords:

> *"slow growth" OR "dwarf stature" OR "SUMO Activating Enzyme Complex" OR "after spraying mutant plants" OR "coleoptile length" OR "growth" OR "cell division" OR "cell proliferation" OR "inflorescence" OR "carbohydrate metabolism" OR "photosynthesis" OR "grain number" OR "grain weight" OR "grain hardness" OR "spikelets" OR "tillers" OR "photoperiod" OR "vernalization" OR "vernalisation"*

Now we can search again in KnetMiner, by using the same gene list and the keywords above. This time, KnetMiner scores every gene based on their relevance to the input keywords and presents the results in *Gene View*. The top-scoring genes include RGL2, which is known to be involved in the plant development processes regulated by the plant hormone gibberellin, or BRI1, which encodes a receptor of a brassinosteroid hormone and is relevant in cell division, growth and elongation.

This can be confirmed by investigating the knowledge graph presented in the *Network View* (Fig. 8.1, top) resulting from our search. Powered by the knetmaps.js library (Singh et al. 2018), this view shows how search keywords are related to searched genes and other relevant entities in a graph-like and intuitive visualisation. From this, other interesting details that we can find from it are:

- The uncharacterised gene TRAESCS2A02G081900 is co-expressed (i.e., it actually produces the molecule(s) that it encodes) in WGCNA_global_module_1, a group of genes obtained from clustering experimental data about gene expression analysis.
- Similarly, TRAESCS5A02G238400, TRAESCS5B02G236900, TRAESCS5D02G245300 are expressed in WGCNA_global_module_12. Both clusters are annotated with terms like inflorescence, embryogenesis, growth, seed development, cell division.
- TRAESCS2A02G081900 is also expressed in another cluster, Meiosis-Study_Module_1, which is annotated with several terms about photosynthesis.
- Moreover, the transcript of this gene (i.e., the molecular entity it encodes), is in a cluster of proteins with similar sequence, in which the Q0D3B6 UniProt protein is annotated with ontology terms such as heading date, flowering, photoperiodism (the plant's ability to detect the day duration and respond with behaviours like initiating the flowering).
- Genes like AT4G39400 (BRI1) or AT1G14920 (GAI), from the well-studied Arabidopsis model organism, which, in addition to being named like some of the genes mentioned above and having analogous functions, are related by many other genes scored as relevant to the search terms.

To enable explainability of the results and support quicker decision making for end-users, the provenance of each node and edge, along with other properties needs to be visible to the user. KnetMiner displays this information in the Info Box with hyperlinks to the original data sources (e.g., PubMed, UniProt, Ensembl). By investigating the auto-generated knowledge networks by KnetMiner, expert

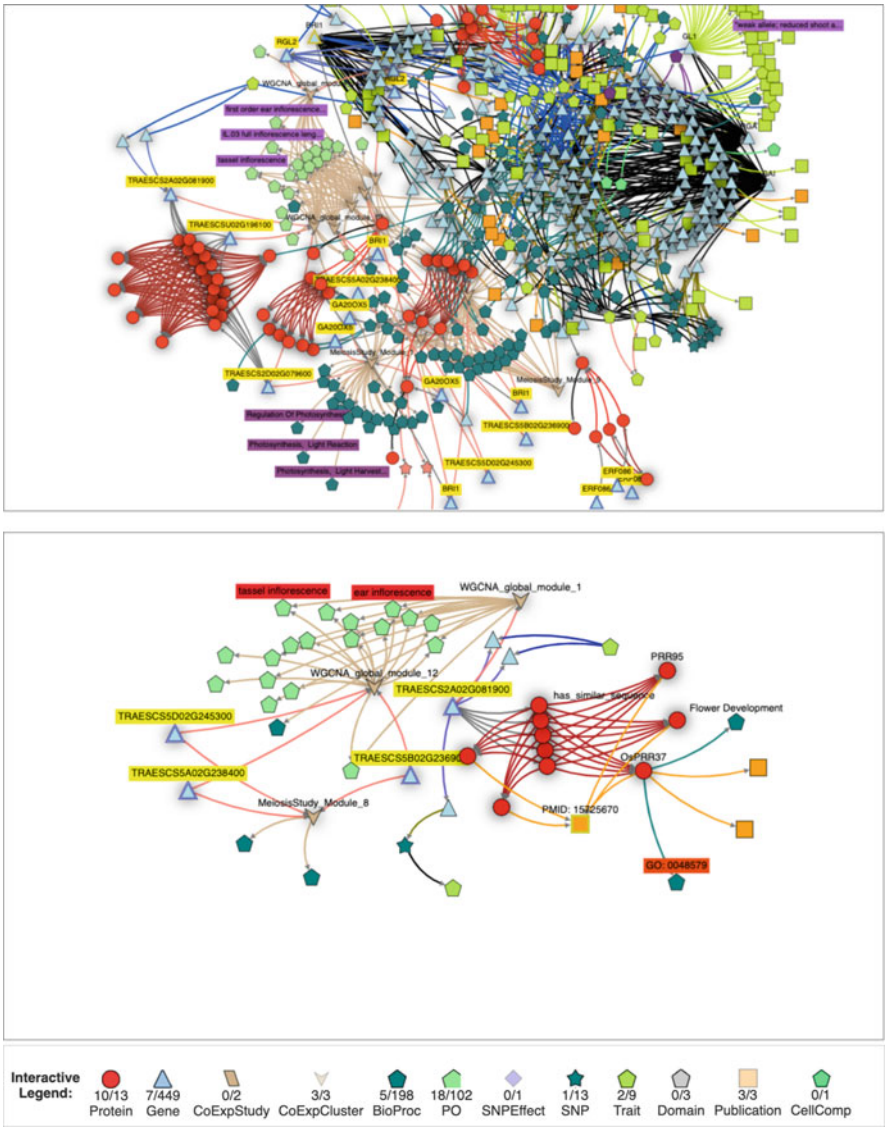

Fig. 8.1 KnetMiner in use Top: a first search shows a set of information related to the list of genes in Table 8.1. Bottom: a more specific search using the uncharacterised genes and a list of more specific keywords identified at the previous step

users can focus on making judgements about the quality and certainty of the presented evidence and quickly developing interesting biological stories. We may notice that a number of genes from our gene list were not linked to any of the keywords we provided. Nevertheless, we can open their knowledge graphs and

Table 8.1 An integrative data analysis use case

ENSEMBL ID	Accession	Name	Description
TraesCS3A02G245000	Os01g0718300	D61	development through controlling cell division and elongation
TraesCS3B02G275000	Os01g0718300	D61	development through controlling cell division and elongation
TraesCS3D02G246500	Os01g0718300	D61	development through controlling cell division and elongation
TraesCS3D02G401400	Os01g0883800	20ox2	Similar to GA C20oxidase2
TraesCS3B02G439900	Os01g0883800	20ox2	Similar to GA C20oxidase2
TraesCS3A02G406200	Os01g0883800	20ox2	Similar to GA C20oxidase2
TraesCS4D02G040400	Os03g0707600	SLR1	DELLA repressor protein, Gibberellin signaling
TraesCS4A02G271000	Os03g0707600	SLR1	DELLA repressor protein, Gibberellin signaling
TraesCS4A02G466700	Os03g0707600	SLR1	DELLA repressor protein, Gibberellin signaling
TraesCS4B02G043100	Os03g0707600	SLR1	DELLA repressor protein, Gibberellin signaling
TraesCS2A02G116900	Os07g0669500	FZP	transition from spikelet to floret meristem, Determination of panicle branching and
TraesCS2B02G136100	Os07g0669500	FZP	transition from spikelet to floret meristem, Determination of panicle branching and
TraesCS2D02G118200	Os07g0669500	FZP	transition from spikelet to floret meristem, Determination of panicle branching and
TraesCS2A02G081900	Os07g0695100	Hd2	Pseudo response regulator, Heading date, Long-day repression
TraesCSU02G196100	Os07g0695100	Hd2	Pseudo response regulator, Heading date, Long-day repression
TraesCS2D02G079600	Os07g0695100	Hd2	Pseudo response regulator, Heading date, Long-day repression
TraesCS5D02G245300	Os09g0457900	ERF102	AP2/ERF transcription factor, Regulation of the internode elongation
TraesCS5A02G238400	Os09g0457900	ERF102	AP2/ERF transcription factor, Regulation of the internode elongation
TraesCS5B02G236900	Os09g0457900	ERF102	AP2/ERF transcription factor, Regulation of the internode elongation

Rice genes known to be involved in the plant yield (from Nadolska-Orczyk et al. 2017). Wheat orthologues computed via Ensembl Plants BioMarts (Kinsella et al. 2011). Wheat genes which are not annotated with known functions in bold

explore their functions and roles, for example, we can see links to terms such as heading date, regulation of long day (a form of photoperiodism), development of the plant internode (part of the plant growth process), AP2/ERF transcription factors (involved in growth processes). Based on this, we decide to issue a new refined search in KnetMiner: the list of uncharacterised genes, plus the keywords in which these appear to be mostly related to them: *"inflorescence" OR "embryogenesis" OR "grain" OR "fruit formation" OR "photoperiod" OR "heading date" OR "long-day" OR "internode"*.

We find the results in Fig. 8.1 (bottom), where we can notice that TRAESCS2A02G081900 encodes similar protein sequences from rice, which are also mentioned in the publication PUBMED:15725670 (Murakami et al. 2005). Significantly, this is titled: *Circadian-associated rice pseudo response regulators (OsPRRs): insight into the control of flowering time*. Among these proteins, we

find that OsPRR37 is one of the rice-orthologous genes from which we started our search and both this protein and PRR95 are related to photoperiodism. The gene at issue is also expressed in the experimentally-obtained cluster of co-expressed genes WGCNA_global_module_1, which is annotated with the terms: inflorescence, stem internode, and the Plant Ontology term: hypocotyl (PO:0020100), which is about the plant development. The other genes are linked to inflorescence and grain development.

In summary, by integrating many types of information in one place along with tools to search the connected data efficiently, we can formulate sophisticated search queries and use an exploratory approach to get insights into complex biological mechanisms.

The richer the knowledge graph is, the more valuable it is. As an example, by searching the EMBL-EBI's Gene Expression Atlas (GXA (Papatheodorou et al. 2020)), we find that the genes identified in the example above are expressed in conditions like outer pericarp, leaf development, inflorescence, fruit formation, pollination, which is coherent with our findings. In addition to providing further confirmation of our findings, this suggests that integrating gene expression data could be another important source for gene prioritisation and could be considered for the integration in KnetMiner knowledge graphs in the future.[1] In Sect. 8.3, we show how the knowledge graph approach combined with the adoption of data standards ease this kind of integration between different datasets and services.

8.3 The Benefits of Data Sharing Practices

In the previous example, all the applications we have used (Ensembl, KnetMiner, and GXA) are based on "raw", machine-readable data, which have the potential to do much more than the software tools or use cases that they were originally designed for. Essentially, data made available in this form can be reloaded by other applications, especially by programs written by the bioinformaticians to realise novel data analysis algorithms or visualisations. This allows for reusing data in new and novel, unexpected ways, which can also reduce the cost/benefit ratio in generating and maintaining them (The Principles of Good Data Management [Internet] 2014; Miksa et al. 2019; Wilkinson et al. 2016). Machine-readable formats are valuable both in the case of data with restricted access and under open-access licences that allow for maximal dissemination and reuse (Jaakkola et al. 2014; Murray-Rust 2008). However, in several contexts, a movement for open data has emerged in recent years, as part of a larger set of ideas for the "openness" of knowledge and intangible intellectual productions, which include computer software (Weber 2009) and scientific research (Bartling and Friesike 2014; Koepsell 2010). While releasing data to the public is not always possible due to issues like

[1] This shortened URL can be used to see the GXA visualisation: https://tinyurl.com/ye3fq8mk

patient privacy or business confidentiality (Anderson 2007; Wiseman et al. 2019), open data models are considered particularly important for publicly-funded research (Schade et al. 2015; Molloy 2011) and, in general, information produced by the public administration (Attard et al. 2015). Furthermore, open access to literature, data and software have the potential to improve reproducibility, a crucial aspect of evidence-based scientific research.

The KnetMiner platform itself is based on these ideas about sharing: the data that we walked through in the use case above are mostly imported from well-known sources, often maintained by organisations dedicated to collecting experimental data and biological knowledge from around the world and making them available both through end-user applications and in forms like: CSV files, JSON-based web APIs, or knowledge graph formats (see Sect. 8.3). KnetMiner datasets leverage these machine-readable data by means of the KnetBuilder tool (Hassani-Pak et al. 2021; Taubert and Köhler 2014) (formerly named Ondex), a framework based on the idea of defining data processing workflows, which use plug-ins such as data importers for various formats, graph data transformations like identifier-based merging and data discovery based on text mining techniques. Similar data integration and workflow frameworks exist, both for biology and many other fields (Leipzig 2016). For instance, Galaxy (Afgan et al. 2018) is another tool based on composing data imports and transformations into workflows, which is mainly based on a web user interface. In contrast, Snakemake (Köster and Rahmann 2018) and Nextflow (Di Tommaso et al. 2017) are mainly based on a command line user interface, which make them suitable for developers and for running data workflows in cluster and cloud architectures, thus exploiting the high parallelism and computing performance of such platforms.

Sharing data is much more difficult to realise than discuss or promise, due to both technical problems and social factors. Recently, years of research and practice on this issue have led to establishing a set of good data sharing principles, or FAIR principles (i.e., Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016)), to guide data producers, publishers and other stakeholders. Here, we want to present these principles, in the order inspired by our experience with the Knetminer data. With datasets having the complexity that our datasets have, a first important step is to represent data according to the Interoperability principle: ideally, it should be possible to use data in all the needed applications without requiring changes like format or schema conversions. To make a very simple example, minimal interoperability for representing genes and gene properties like international symbol, description, position in the chromosome, could be ensured by adopting a CSV format and establishing certain column names for such properties. In the case of graphs, Fig. 8.2 shows examples of analogous use of standard types and properties (e.g., bioschema:Gene, schema:description). Similarly, organisations like Ensembl publish sets of reference identifiers for genes (e.g., the many TraesXXX identifiers mentioned above), which are actively reused around the world, so that two data files using the same ID can easily be merged together. Establishing shared formats, data models, type names and identifiers is about setting data standards, a fundamental way to realise interoperability. In Sect. 8.3, we describe how these ideas have

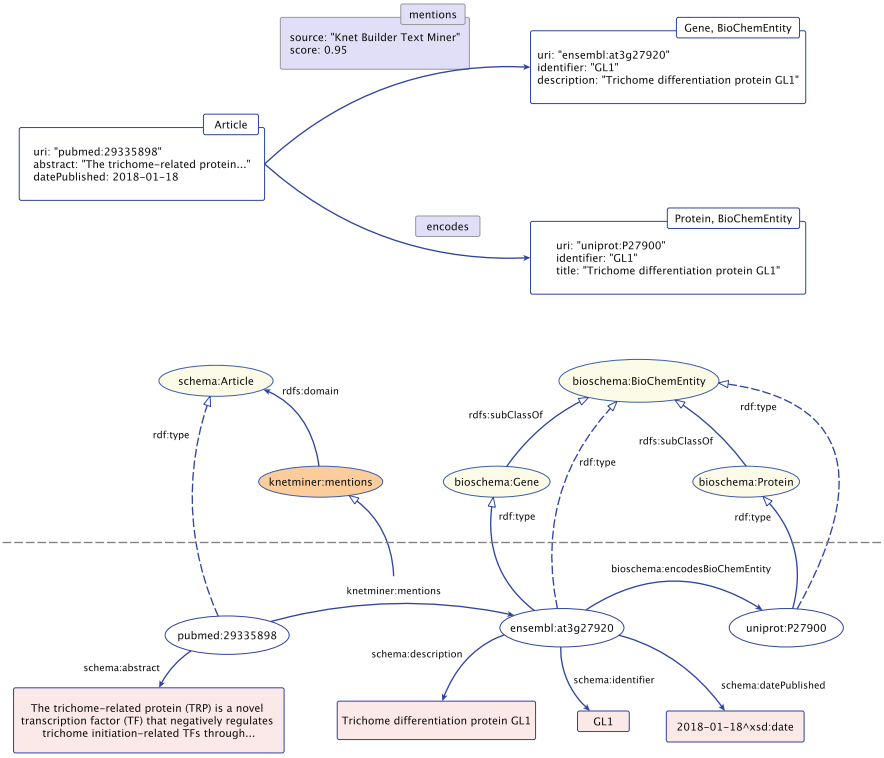

Fig. 8.2 Top: a simple instance of a knowledge graph, represented by means of the property graph model (PG). Node elements are characterised by type labels (Article, Gene) and knowledge can flexibly be connected by means of relationships (encodes, mentions). Both nodes and relationships can have key/value attributes. Bottom: How the RDF and Semantic Web standards (SW) model knowledge graphs. The graph model is more granular than the PG (node properties are additional graphs), nodes are based on resolvable web URIs (e.g., ensembl:atg27920, a shortened URI, resolves to the graph of its outgoing direct links). The SW has explicit support for data schematisation (upper side), including automatic reasoning features (dashed links can be computationally inferred). PG is easier with representing relationship's properties (non-native mechanisms like reifications are available in RDF (Thakkar 2020))

been much extended by the linked data community. Standardised and interoperable data can be made Accessible by using technologies that are standard themselves. Originally designed to publish human-readable documents over the Internet, the world wide web, and in particular, the HTTP protocol, have been enormously successful at sharing both human-readable documents and raw data documents. In particular, generalised forms of web addresses, the Uniform Resource Identifiers (URIs (Antoniou 2008)), can be used to identify digital resources worldwide and, at the same time, to provide access to machine-readable data about the entities that the resources describe. For example, a URI can be used to identify a data file, a scientific paper, or a protein, and data about these entities can be obtained by "resolving"

the URI, e.g., making an HTTP request to it, which returns the CSV data file, a JSON document of metadata descriptors about the paper, or an RDF-encoded graph document about the protein (see Sect. 8.3). Once data are made accessible, making them *Findable* is another important data sharing principle. This relies on dataset descriptors as well, which provide us with information on the whole dataset. For example, thanks to services like Google Dataset Search (Brickley et al. 2019) or DataCite (Brase 2009), it is possible to search datasets based on metadata-specified criteria such as what the data contents are about, when they were created, who are its publishers. Finally, data that are easy to find, access and interoperate with other data, become more *Reusable*. Reusability is also favoured by clear licences about their usage conditions, independently of whether the data are open or restricted. Licence details can be synthesised as standardised metadata, so that software applications can automatically make decisions on how to use datasets (Rodrıguez-Doncel et al. 2013).

8.3.1 Contributions of the Linked Data Community to the Data Sharing Principles

Before the seminal paper that popularised the FAIR acronym, many other efforts have been made to apply one or more of the same data principles, both in life science and other fields. A very prominent one has been the Semantic Web technologies and the community of linked data that was born out of them (Mountantonakis and Tzitzikas 2019). The term Semantic Web was made popular in 2001, by a seminal paper co-authored by Tim Berners Lee, who previously invented the World Wide Web approach (Antoniou 2008; Berners-Lee et al. 2001). Its main idea is to share data by leveraging graphs and the existing web technology and principles. As shown in Fig. 8.2 (bottom), the basic building block of the approach is the RDF data model, where knowledge is organised as graphs of triples, with each triple linking an entity to a property value or another entity, by means of a typed link. The triple is alternatively described with the language metaphor, i.e., the outgoing node is like the subject of a statement, the typed link is like a predicate and the destination value or entity is like the predicate's object. Both the (non-value) nodes and the predicates always consist of URIs. This is a specific Semantic Web characteristic, which extends the Web to realise universal data identifiers, which additionally offer an accessibility mechanism to join and explore data about a given resource. Namely, the URI uniprot:P27900 in Fig. 8.2 (a possible abbreviation of https://www.uniprot.org/uniprot/P27900), in addition to identifying the GL1-encoded protein, can be resolved via HTTP, resulting in a set of RDF triples, offering more data about the protein at issue (like the RDF triples in the figure).

It is considered best practice to ensure that URIs like this one are backed by a data publisher, who can use a proper URI prefix (usually called name space) as a reference source for the data they maintain and share on line. In the acronym

RDF (Resource Description Framework), this mechanism is what has originated the name "resources" for RDF nodes and predicates. Clearly, in order to make all of that possible, RDF needs to have a format to encode the conceptual data model and indeed, common serialisations exist for RDF (Meindertsma 2019), including the API-compatible JSON-LD format.

The "Semantic" part of the Semantic Web is built on top of the RDF layer, by means of RDF representations of schema languages like RDF-Schema and OWL (Antoniou 2008). In particular, OWL is the standard way in which formal, very expressive and semantically-rich computational ontologies are used to characterise the meaning in the Semantic Web and linked data world. OWL is based on Description logic, a kind of first-order logic designed to define set membership propositions in a way that allows for automatic reasoning. For instance, in the OBI ontology, sophisticated OWL definitions can be used to automatically infer the nature of a biomedical experiment and the characteristics of its components (Bandrowski et al. 2016). Other, less formal schematisations exist in the linked data arena. For example, the SKOS vocabulary can be used to define thesauri and taxonomies of terms in RDF, as it has been done for the agri-food vocabulary AGROVOC (Caracciolo et al. 2013).

Over the years, the Semantic Web has been the technological base of the linked data community, which has promoted data sharing based on RDF, the Semantic Web and related principles (Mountantonakis and Tzitzikas 2019). This has led to significant projects and data publications in many fields (Avila-Garzon 2020).

8.3.2 New Directions for Linked and Graph Data

While linked data projects are still of primary importance nowadays, the approach has clear limits, recently highlighted by new directions that the world is taking to solve data sharing needs. The data-hungry nature of artificial intelligence requires means to generate and collect data, no matter how good they are from the point of view of sharing principles like the FAIR principles. The emergence of data lakes (Che and Duan 2020) or cloud-based data management frameworks (Holmes 2015) are examples of such trends. Similarly, artificial intelligence is influencing the idea of automating the classification and schematisation of raw data, with fewer concerns for more manually curated efforts (Gabrilovich and Usunier 2016). Another related issue is that many people who have to deal with such data, like data journalists, biodata curators or web developers, may find the graph data models difficult to learn and use, thus, they may prefer to interact with data in more familiar ways, such as downloadable tabular files or JSON-based APIs. Related to that, recently GraphQL is emerging (Brito et al. 2019) as a standard for querying graph data in a more accessible way by means of simple data object templates in the JSON format, which return results that are very similar to the initial templates. Due to the increasing popularity of such approaches, systems that allow for GraphQL access to linked data (Taelman et al. 2018) and graph databases (Lyon 2021) are suitable

for reconciling graph-encoded data to modern data access technologies. Similarly, mapping RDF data onto property graph databases like Neo4j is another promising approach (Thakkar 2020; Brandizi et al. 2018a).

Regarding data schematisation, while OWL-based ontologies remain an advanced data modelling approach for specialised applications (Smith et al. 2007), a major problem with them is that they require very specific expertise and that it is very difficult to do OWL-based data integration across a wide content arena which potentially could encompass the whole World Wide Web. Due to these issues, complementary approaches are emerging to develop lightweight ontology-like schemas, which are designed primarily for applications like improving results from search engines or integrating very heterogeneous data generated by non-expert end-users. A prominent example of this is schema.org (Guha et al. 2016), in the field of life sciences, the bioschemas extension is being defined to specifically represent data from this domain (Gray et al. 2017).

8.3.3 Applying Data Sharing Principles in KnetMiner

In this section, we show a concrete application of the topics discussed above using data behind instances of the KnetMiner application. As already described above, KnetMiner is a web application that can be deployed as a particular instance over a given dataset. A dataset is generally a collection of different data resources selected as important for research on a given organism or a given subject, such as the interaction between an organism and its main pathogens. The data resources in a dataset are integrated from multiple, well-known sources, into the form of a knowledge graph, namely a property graph. For historical reasons, the current native format that KnetMiner is using is based on XML, and an XML basic schema that essentially encodes the main entities of a property graph. This format is called Ondex XML, or OXL, due to Ondex (Taubert and Köhler 2014), the suite for data integration that is used to map various data formats onto an OXL property graph. In recent years, our group has started to publish OXL data following FAIR principles and using the interoperable linked data technologies described above. A first step for that consists of mapping the entities in OXL onto a simple, application-oriented ontology based on OWL. This is the conceptual basis for converting from OXL to RDF via a dedicated tool (Brandizi et al. 2018a, b), using a generic library for converting Java objects to RDF (java2rdf [Internet] 2021).

The RDF obtained this way is published in multiple endpoints based on graph databases. One graph database is Virtuoso (Brandizi et al. 2018b), which has direct support for RDF and SPARQL, the Semantic Web standard query language for RDF. This way, the data can be queried and explored via SPARQL either through a web browser or from a client program (Brandizi 2020). We have decided to support both the RDF and Neo4j endpoints due to the complementary sets of advantages that they have (Brandizi et al. 2018b). As mentioned above, Neo4j is a property graph system, which offers ease of use, good performance and Cypher, the property

graph query language that is particularly compact and simple to learn and apply. A notable use of our Neo4j endpoint is in a KnetMiner component: the semantic motif traverser, which explores known graph patterns to find entities relevant to genes during KnetMiner queries (Fig. 8.3). A typical application of the RDF/SPARQL is integrating data, which is eased by tools like TARQL (Tarql 2020) and paradigms like shared URIs. In (Brandizi 2020) we discuss the use of the Jupyter framework (Perkel 2018) to benefit from both the SPARQL and Cypher access. In order to maximise data interoperability, we curate manual mappings from our RDF data to equivalent or broader data types defined in standard ontologies most commonly used

Fig. 8.3 Querying and using graph data. Top: a SPARQL example, the Semantic Web standard to query RDF data, based on a graph pattern syntax (e.g.,?study matches any study node having an identifier and a title). The query matches data that were integrated from both KnetMiner and EBI GXA (middle). Bottom: a simplified case of the Cypher queries (the Neo4j query language for property graphs) that KnetMiner uses to relate a gene to relevant entities, which is at the basis of application results described in Sect. 8.2

in life science and other fields. This is a typical linked data approach, which, for instance, allows for querying multiple datasets referring to the type bioschema:Gene and seamlessly obtaining results from KnetMiner and other data using or mapping the same OWL class.

This is part of a broader ongoing project of ours named AgriSchemas (Rothamsted Research, UK 2019), which mainly aims to map common data in the agri-food domain to the Bioschemas standard, contributing to the latter with possible specific extensions, developing reusable tools for realising the mapping-driven conversions. In Fig. 8.3, we show how this can be exploited starting from GXA-achieved results presented in Sect. 8.2, where a particular GXA experiment (accession E-MTAB-3103) was found to be involved in the expression of our candidate genes. In the figure, GXA data and KnetMiner data are queried using the SPARQL language to compare conditions in which genes are expressed in that experiment (GXA data) with publications that mention the same genes (KnetMiner data). The query is against the data from the two datasets that we have integrated using the AgriSchemas approach, the results integrate thanks to that and they can be explored in a unified view. For the future, we plan to further extend the project with front-end components, which will be based on the GraphQL standard and where tasks like rendering user interfaces will be automatically driven by the common definitions in Bioschemas types. This will maximise the reusability of such components by relying on data standardisation.

8.4 Conclusions

Modern biology is increasingly a data-intensive science. Historically, Biology was relatively data-poor and dominated by reductionist approaches, where phenomena involving a very small set of actors and interactions (e.g., one or few genes and one or few phenotypes) were considered one by one. Since the advent of high-throughput technologies, in the early 2000s, it has been practical to take a very large number of measurements, like the expression of tens of thousands of genes at a time allowing for the identification of the main components affected by experimental factors by means of statistical testing, data visualisations and data mining. More recently, a trend has been emerging to integrate masses of data referring to multiple phenomena at different scales, as well as multiple organisms or entire ecosystems. Being designed mostly to flexibly unify such diversity of data, the combination of knowledge graphs, linked data and graph databases are powerful tools to perform integrative data exploration and analysis, in life science as well as other disciplines. Furthermore, graph models are effective at representing data schematisation and semantics in standardised and interoperable ways, which is a fundamental aspect of sharing data and knowledge according to the FAIR principles. In turn, data sharing maximises data usefulness, by feeding a virtuous cycle where new discoveries are made by integrating existing data and new interesting data are shared from new knowledge and its encoding in machine-readable formats. In particular, this

is relevant to artificial intelligence, which nowadays is mostly based on machine learning and thus is very data-demanding. In this area, semantic representation has still to offer contributions to machine learning methods, as it is emerging, for instance, from works regarding the use of hybrid approaches to build knowledge graphs and predict new knowledge (Gabrilovich and Usunier 2016; Reese et al. 2021).

Acknowledgments This work was supported by the UKRI Biotechnology and Biological Sciences Research Council (BBSRC) through the Designing Future Wheat ISP (BB/P016855/1), the FAIR BBR (BB/S020020/1) and DiseaseNetMiner TRDF (BB/N022874/1). CR and KHP are additionally supported by strategic funding to Rothamsted Research from BBSRC. We acknowledge all the past and present members of the KnetMiner Bioinformatics team at Rothamsted for their scientific inputs and software contributions, especially: Joseph Hearnshaw, Martin Castellote, and Richard Holland.

References

A Comparison of Serialization Formats [Internet] (2019). https://blog.mbedded.ninja/programming/serialization-formats/a-comparison-of-serialization-formats/. Accessed 11 May 2021

Adamski NM, Borrill P, Brinton J, Harrington SA, Marchal C, Bentley AR et al (2020) A roadmap for gene functional characterisation in crops with large genomes: lessons from polyploid wheat. Elife 9:e55646

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M et al (2018) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46:W537–W544

Altschul S (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anderson JG (2007) Social, ethical and legal barriers to E-health. Int J Med Inform 76:480–483

Antoniou G (2008) A semantic web primer, 2nd edn. MIT Press, Cambridge, MA

Arnaud E, Laporte MA, Kim S, Aubert C, Leonelli S, Cooper L et al (2020) The Ontologies Community of Practice: an initiative by the CGIAR Platform for Big Data in Agriculture. SSRN Electron J. https://www.ssrn.com/abstract=3565982. Accessed 11 May 2021

Attard J, Orlandi F, Scerri S, Auer S (2015) A systematic review of open government data initiatives. Gov Inf Q 32:399–418

Ausiello G, Gherardini PF, Marcatili P, Tramontano A, Via A, Helmer-Citterich M (2008) FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures. BMC Bioinform 9:S2

Avila-Garzon C (2020) Applications, methodologies, and technologies for linked open data: a systematic literature review. Int J Semant Web Inf Syst 16:53–69

Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC et al (2016) The ontology for biomedical investigations. PLoS One 11:e0154556

Bang H, Zhou XK, van Epps HL, Mazumdar M (eds) (2010) Statistical methods in molecular biology [Internet]. Humana Press, Totowa, NJ. http://link.springer.com/10.1007/978-1-60761-580-4. Accessed 2021 May 10

Barah P (2021) Gene expression data analysis: a statistical and machine learning perspective. Gene Expression Data Analysis, S.l.

Bartling S, Friesike S (2014. Accessed 9 May 2021) Opening Science [Internet]. Springer International, Cham. https://doi.org/10.1007/978-3-319-00026-8

Baumgartner C, Beckmann JS, Deng H-W, Shields DC, Wang X (eds) (2016) Application of clinical bioinformatics, 1st edn. Springer, Dordrecht
Beluhova-Uzunova RP, Dunchev DM (2019) Precision farming–concepts and perspectives. Probl Agric Econ
Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284:34–43
Boyle EA, Li YI, Pritchard JK (2017) An expanded view of complex traits: from polygenic to Omnigenic. Cell 169:1177–1186
Brandizi M (2020) The Power of Standardised and FAIR Knowledge Graphs [Internet]. KnetMiner. https://knetminer.com/cases/the-power-of-standardised-and-fair-knowledge-graphs.html
Brandizi M, Singh A, Hassani-Pak K (2018a) Getting the best of linked data and property graphs: rdf2neo and the KnetMiner use case. SWAT4LS
Brandizi M, Singh A, Rawlings C, Hassani-Pak K (2018b) Towards FAIRer Biological Knowledge Networks Using a Hybrid Linked Data and Graph Database Approach. J Integr Bioinforma [Internet]. De Gruyter. https://www.degruyter.com/view/journals/jib/15/3/article-20180023.xml. Accessed 2 Sep 2020
Brase J (2009) DataCite—a global registration agency for research data. In: 2009 Fourth International conference on cooperation and promotion of information resources in science and technology, pp 257–261
Brickley D, Burgess M, Noy N (2019) Google Dataset Search: building a search engine for datasets in an open web ecosystem. In: World Wide Web Conference [Internet]. ACM, San Francisco, CA, pp 1365–1375. Accessed 12 May 2021. https://doi.org/10.1145/3308558.3313685
Brito G, Mombach T, Valente MT (2019) Migrating to GraphQL: a practical assessment. In: 2019 IEEE 26th Int Conf Softw Anal Evol Reengineering SANER [Internet]. IEEE, Hangzhou, pp 140–150. https://ieeexplore.ieee.org/document/8667986/
Caracciolo C, Stellato A, Morshed A, Johannsen G, Rajbhandari S, Jaques Y et al (2013) The AGROVOC linked dataset. Seman Web 4:341–348
Che H, Duan Y (2020) On the logical design of a prototypical Data Lake System for biological resources. Front Bioeng Biotechnol 8:553904
Check HE (2013) Geneticists push for global data-sharing. Nature 498:16–17
Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R et al (2017) Strategies to improve reference databases for soil microbiomes. ISME J 11:829–834
Chowdhury B, Garai G (2017) A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics 109:419–431
Corbellini A, Mateos C, Zunino A, Godoy D, Schiaffino S (2017) Persisting big-data: the NoSQL landscape. Inf Syst 63:1–23
Dada JO, Mendes P (2011) Multi-scale modelling and simulation in systems biology. Integr Biol 3:86
Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A et al (2007) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36:D344–D350
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 28:935–942
Description Logics (2014) IEEE Intell Syst 29:12–19
Designing Future Wheat [Internet] (2021) Designing. Future Wheat. https://designingfuturewheat.org.uk/. Accessed 20 May 2021
Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017) Nextflow enables reproducible computational workflows. Nat Biotechnol 35:316–319
Ehrlinger L, Wöss W (2016) Towards a definition of knowledge graphs. Semant Posters Demos SuCCESS 48:2
Figueiredo AS (2017) Data sharing: convert challenges into opportunities. Front Public Health 5:327
Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J et al (2017) COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 45:D777–D783

Gabrilovich E, Usunier N (2016) Constructing and mining web-scale knowledge graphs. ACM, pp 1195–1197. http://dl.acm.org/citation.cfm?doid=2911451.2914807. Accessed 22 Feb 2018

Germain RN, Meier-Schellersheim M, Nita-Lazar A, Fraser IDC (2011) Systems biology in immunology: a computational modeling perspective. Annu Rev Immunol 29:527–585

Gostev M, Faulconbridge A, Brandizi M, Fernandez-Banet J, Sarkans U, Brazma A et al (2012) The BioSample database (BioSD) at the European bioinformatics institute. Nucleic Acids Res 40:D64–D70

Gray AJ, Goble C, Jimenez RC (2017) Bioschemas: from potato salad to protein annotation. Springer, Berlin

Guha RV, Brickley D, Schema MS (2016) Org: evolution of structured data on the web. Commun ACM 59:44–51

Hassani-Pak K, Castellote M, Esch M, Hindle M, Lysenko A, Taubert J et al (2016) Developing integrated crop knowledge networks to advance candidate gene discovery. Appl Transl Genom 11:18–26

Hassani-Pak K, Singh A, Brandizi M, Hearnshaw J, Parsons JD, Amberkar S et al (2021) KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. Plant Biotechnol J:pbi.13583

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107:1–8

Holmes A (2015) Avoiding big data antipatterns [Internet]. https://www.slideshare.net/grepalex/avoiding-big-data-antipatterns. Accessed 12 May 2021

Horler R, Turner A, Fretter P, Ambrose M (2018) SeedStor: a germplasm information management system and public database. Plant Cell Physiol 59:e5

Hutson M (2020) Artificial-intelligence tools aim to tame the coronavirus literature. Nature

Jaakkola H, Mäkinen T, Eteläaho A (2014) Open Data: opportunities and challenges. In: Proc 15th Int Conf Comput Syst Technol [Internet]. ACM, New York, NY, pp 25–39. Accessed 7 Mar 2018. https://doi.org/10.1145/2659532.2659594

java2rdf [Internet] (2021) EBI BioSamples Database Project. https://github.com/EBIBioSamples/java2rdf. Accessed 12 May 2021

Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G et al (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database 2011:bar030

Koepsell D (2010) Back to basics: how technology and the open source movement can save science. Soc Epistemol 24:181–190

Köster J, Rahmann S (2018) Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 34:3600–3600

Leipzig J (2016) A review of bioinformatic pipeline frameworks. Brief Bioinform:bbw020

Li L, Zhang Q, Huang D (2014) A review of imaging techniques for plant phenotyping. Sensors 14:20078–20111

Liakos K, Busato P, Moshou D, Pearson S, Bochtis D (2018) Machine learning in agriculture: a review. Sensors 18:2674

Lightbody G, Haberland V, Browne F, Taggart L, Zheng H, Parkes E et al (2019) Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief Bioinform 20:1795–1811

Ling H-Q, Zhao S, Liu D, Wang J, Sun H, Zhang C et al (2013) Draft genome of the wheat A-genome progenitor Triticum urartu. Nature 496:87–90

Lyon W (2021) Fullstack GraphQL applications with GRANDstack [Internet]. Manning Publications. https://books.google.co.uk/books?id=DbsKzgEACAAJ

Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM et al (2014) Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. Med Sci Monit Basic Res 20:138–142

Mayrhofer MT, Holub P, Wutte A, Litton J-E (2016) BBMRI-ERIC: the novel gateway to biobanks: from humans to humans. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 59:379–384

McGuinness DL (2005) Ontologies come of age. Spinn semantic web bringing world wide web its full potential. The MIT Press, pp 171–194
McGuinness DL, Van Harmelen F, others. OWL web ontology language overview. W3C Recomm 2004;10:2004
Meindertsma J (2019) What's the best RDF serialization format? [Internet]. Ontola.io. http://ontola.io/blog/rdf-serialization-formats/. Accessed 12 May 2021
Meyer K (2016) A mathematical review of resilience in ecology. Nat Resour Model Wiley Online Libr 29:339–352
Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. PLoS Comput Biol 15:e1006750
Mills L (2014) Common File Formats. Curr Protoc Bioinforma [Internet]. https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bia01bs45. Accessed 11 May 2021
Molloy JC (2011) The open Knowledge Foundation: open data means better science. PLoS Biol 9:e1001195
Mountantonakis M, Tzitzikas Y (2019) Large-scale semantic integration of linked data: a survey. ACM Comput Surv 52:1–40
Murakami M, Matsushika A, Ashikari M, Yamashino T, Mizuno T (2005) Circadian-associated rice pseudo response regulators (OsPRRs): insight into the control of flowering time. Biosci Biotechnol Biochem 69:410–414
Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC et al (2007) Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside. AMIA Annu Symp Proc:548–552
Murray-Rust P (2008) Open Data Sci Ser Rev 34:52–64
Nadolska-Orczyk A, Rajchel IK, Orczyk W, Gasparis S (2017) Major genes determining yield-related traits in wheat and barley. Theor Appl Genet 130:1081–1098
Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP (2020) Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. Front genet. Frontiers 11:350
November J (2018) More than Moore's mores: computers, genomics, and the embrace of innovation. J Hist Biol 51:807–840
Papatheodorou I, Moreno P, Manning J, Fuentes AM-P, George N, Fexova S et al (2020) Expression atlas update: from tissues to single cells. Nucl Acids Res Oxford Acad 48:D77–D83
Perkel JM (2018) Why Jupyter is data scientists' computational notebook of choice. Nature 563:145–146
Perryman SAM, Castells-Brooke NID, Glendining MJ, Goulding KWT, Hawkesford MJ, Macdonald AJ et al (2018) The electronic Rothamsted archive (e-RA), an online resource for data from the Rothamsted long-term experiments. Sci Data 5:180072
Polding R (2018) Databases: Evolution and Change [Internet]. https://medium.com/@rpolding/databases-evolution-and-change-29b8abe9df3e
Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S et al (2021) KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. Patterns 2:100155
Regenmortel MHVV (2004) Reductionism and complexity in molecular biology: scientists now have the tools to unravel biological complexity and overcome the limitations of reductionism. EMBO Rep 5:1016–1020
Rodrıguez-Doncel V, Suárez-Figueroa MC, Gómez-Pérez A, Poveda-Villalón M (2013) Licensing patterns for linked data. In: Proc 4th Int Workshop Ontol Patterns Appear
Rothamsted Research, UK (2019) AgriSchemas and FAIR-ification of DFW Data [Internet]. https://www.slideshare.net/mbrandizi/agrischemas-progress-report. Accessed 12 May 2021
Schade S, Granell C, Perego A (2015) Coupling public sector information and public-funded research data in Europe: a vision of an open data ecosystem. In: Information and communication technologies in public administration: innovations from developed countries. CRC, London, pp 275–298

Schüngel M, Stackebrandt E, Bizet C, Smith D (2013) MIRRI—the microbial resource research infrastructure: managing resources for the bio-economy. EMBnet J 19:5
SDG U (2019) Sustainable development goals. Energy Prog Rep Track SDG 7
Sharma S, Shandilya R, Patnaik S, Mahapatra A (2016) Leading NoSQL models for handling big data: a brief review. Int J Bus Inf Syst 22:1
Shorte SL, Frischknecht F (eds) (2007) Imaging cellular and molecular biological functions: with 13 tables. Springer, Berlin
Singh A, Rawlings CJ, Hassani-Pak K (2018) KnetMaps: a BioJS component to visualize biological knowledge networks. F1000Res 7:1651
Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W et al (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25:1251–1255
Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ et al (2015) Big Data: astronomical or genomical? PLoS Biol 13:e1002195
Surwase V (2016) REST API modeling languages-a developer's perspective. Int J Sci Technol Eng 2:634–637
Taelman R, Vander Sande M, Verborgh R (2018) GraphQL-LD: linked data querying with GraphQL. In: ISWC 2018 17th International Semantic Web Conference, pp 1–4
Tang B, Pan Z, Yin K, Khateeb A (2019) Recent advances of deep learning in bioinformatics and computational biology. Front Genet 10:214
Tarql: SPARQL for Tables—Tarql—SPARQL for Tables: Turn CSV into RDF using SPARQL syntax [Internet]. https://tarql.github.io/. Accessed 1 Sep 2020
Taubert J, Köhler J (2014) Molecular information fusion in Ondex. In: Approaches in Integrative Bioinformatics. Springer, Berlin, pp 131–160
Thakkar H (2020) A survey of approaches for supporting data interoperability between RDF and property graph databases [Internet]. http://harshthakkar.in/wp-content/uploads/Semantics_Seminar_Report_2020_HT_RDF-PG.pdf
The Principles of Good Data Management [Internet] (2014) IGGI (Intra-governmental Group on Geographic Information). http://cedadocs.ceda.ac.uk/1085/
Watson JT, Sparkman OD (2007) Introduction to mass spectrometry: instrumentation, applications, and strategies for data interpretation. Wiley, Hoboken, NJ
Weber S (2009) The success of open source. Harvard University Press, Cambridge, MA
Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3
Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E et al (2019) Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today 24:933–938
Wiseman L, Sanderson J, Zhang A, Jakku E (2019) Farmers and their data: an examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming. NJAS Wagening J Life Sci 90–91:100301
Yang W, Feng H, Zhang X, Zhang J, Doonan JH, Batchelor WD et al (2020) Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. Mol Plant 13:187–214
Yang Y, Aduragbemi A, Wei D, Chai Y, Zheng J, Qiao P, et al (2021) Large-scale integration of meta-QTL and genome-wide association study discovers the genomic regions and candidate genes for yield and yield-related traits in bread wheat [Internet]. https://www.researchsquare.com/article/rs-342038/v1
Zhang ZJ (2017) Graph databases for knowledge management. IT Prof 19:26–32

Chapter 9
Exploring Plant Transcription Factor Regulatory Networks

Ranran Yu and Dijun Chen

Abstract Transcription factors (TFs) are key nodes of gene regulatory networks that specify plant morphogenesis and control specific pathways such as stress responses. TFs directly interact the genome by recognizing specific DNA sequence, in terms of a complex system to fine-tune spatiotemporal gene expression. The combinatorial interaction among TFs determines regulatory specificity and defines the set of target genes to orchestrate their expression during developmental switches. In this chapter, we provide a catalog of plant-specific TFs and a comprehensive assessment of whether genome-wide analyses have so far been used for identifying potential direct target genes for each TFs. We further construct comprehensive TF-associated regulatory networks in the model plant *Arabidopsis thaliana* using genome-wide datasets from our ChIP-Hub database (http://www.chiphub.org/). We discuss how to dissect the network structure to identify potentially important cross-regulatory loops in the control of developmental switches in plants.

Keywords Transcription factor · Gene regulatory network · Genome-wide analysis · Plant

9.1 Introduction to Plant Transcription Factors

9.1.1 Overview of Transcription Factor-Mediated Gene Regulation

Plant evolution involves genetic responses to biotic and abiotic stresses. To react quickly to external changes, plants have formed complex signal networks. After receiving external information, plants can regulate target genes and their products

R. Yu · D. Chen (✉)
State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, China
e-mail: dijunchen@nju.edu.cn

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_9

on different levels (including transcriptional control, RNA processing control, RNA transport and localization control, translation control, mRNA degradation control, protein activity control) to make adaptive changes through these signal networks.

Transcription is the initial step, and it will affect subsequent steps. Transcription can also regulate the tissue-specific expression of genes and the response of gene expression to specific signals. It can thus affect the differentiation of organs and plant adaptation to the environment.

The transcriptional control of gene expression occurs mainly through cis-acting elements, and trans-acting factors also play an important role in response to stresses. Cis-acting elements are DNA sequences that regulate genes along the same nucleotide chain and usually lack the ability to encode protein. Trans-acting factors refer to proteins that regulate gene expression on different nucleic acid chains. The gene encoding this type of protein is not on the same chain as the nucleic acid chain that recognizes and binds to it.

The cis-regulatory elements of gene expression in transcriptional control are important in plant development and stress responses. Promoters and enhancers are two kinds of such cis-regulatory elements. Gene expression in transcriptional control is regulated by promoters, where gene transcription initiates, and more distal enhancers, which control temporal and spatial activity (Lenhard et al. 2012). Despite sharing some features, promoters and enhancers have historically been considered to be distinct classes of regulatory elements (Mikhaylichenko et al. 2018). Promoters are DNA sequences that can be recognized or bound by RNA polymerases to initiate transcription. The most common basal promoter element is located around the transcription start site. Enhancers are regions of the genome that can enhance the expression of particular genes linked to them after binding to a specific protein. Because chromatin has a special spiral structure, even if the enhancer and the gene are located far apart in sequence, there is a chance that they will come into contact with each other. Most enhancers are far away from the target genes and appear either upstream or downstream of those genes.

In eukaryotes, RNA polymerases are responsible for transcription, and RNA polymerase II is the most active (Fig. 9.1; Adcock and Caramori 2009). However, RNA polymerases have no special affinity for the promoter and are unable to complete transcription alone. Transcription requires the participation of numerous transcription factors (TFs) and co-transcription factors. They form a complex with RNA polymerase II to allow transcription to be initiated in the correct location. TFs have a modulated structure, including a DNA-binding domain (DBD), a trans-activating domain (TAD), and an optional signal sensing domain (SSD). Among these, TAD contains binding sites to other proteins, and these binding sites typically have active functions (AFs).

TFs tightly control where and when the nearby target gene is expressed by binding to the DNA (Kummerfeld and Teichmann 2006). TFs, as trans-acting factors, can also specifically interact with cis-regulating elements of eukaryotic genes to control chromatin and transcription, forming a complex system that guides expression of the genome. TFs are important in response to stresses such as insect attack and drought (Rushton et al. 2010).

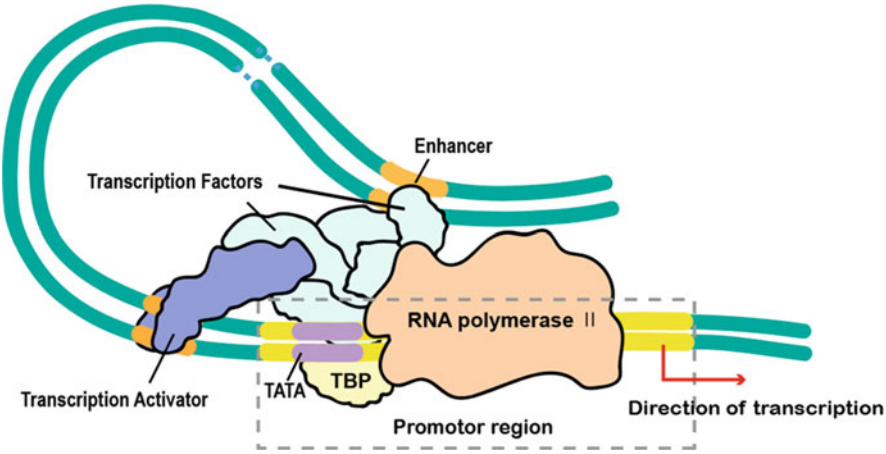

Fig. 9.1 Schematic diagram of the work of transcription factors

The recognition of short DNA sequences by TFs is a key step in transcription. Detailed analysis of different TFs shows that they have a modular structure in which specific regions of the molecule are responsible for binding to DNA, while other regions stimulate or inhibit transcription (Nuruzzaman et al. 2013). TFs therefore contain DNA-binding domains that recognize specific sequences within the promoter and some specific regions of the genes they regulate.

DNA-binding domains read genomic DNA sequences in three basic ways: base readout, indirect readout, and shape readout. Base readout means that TFs recognize a given nucleotide sequence by means of hydrogen bonding, hydrophobic interactions, or formation of salt bridges between amino acid side chains and accessible portions of base pairs. Indirect readout is connected with the main interaction with TFs and the DNA phosphoric acid backbone. Shape readout means that TF recognizes and combines the shape characteristics of DNA.

Once recognized and bound in the manner described above, TFs promote the binding of RNA polymerase to DNA. However, they not only promote the binding of RNA polymerase and DNA but also catalyze the modification of histones. This function is accomplished by direct action or by recruitment of other proteins with specific catalytic activity.

TFs regulate the amount of gene products (RNA and proteins) in cells by controlling transcription rates, and they are also regulated. Most TFs do not work alone. To complete gene transcription, a series of transcription factors must be bound to the DNA regulatory sequence. This collection of transcription factors, in turn, recruits mediating proteins, such as cofactors, for efficient recruitment of pre-initiation complexes and RNA polymerases.

TFs with the same type of DNA-binding domain, referred to as TFs from the same family, tend to have more similar DNA-binding specificities than TFs belonging to different families. Variations in DNA-binding specificity do occur in

the same family and are due to changes in specific residues of the DNA binding domain.

9.1.2 Plant-Specific Transcription Factor Families and Function

TFs are generally defined as proteins that directly bind to target gene promoters and regulate the expression of target genes in a sequence-specific manner. The DNA binding domain (DBD) in the sequence of TFs largely determines the sequence specificity of its binding to the cis-element DNA in the upstream promoter region of the gene (Weirauch et al. 2014).

DBDs are evolutionarily conserved and are the main basis for distinguishing different transcription factor families. Generally, TFs can be classified into specific families based on the types of DBDs contained in their sequence. For example, the family of Ethylene insensitive-like (EIL) TFs regulating plant growth and development all contain the Ein domain (Riechmann et al. 2000).

Although some transcription factor families have a one-to-one correspondence according to DBD, there are some families with more a more complex correspondence. Some TFs contain two or more DBDs. Therefore, the number of DBDs is often used to distinguish between different transcription factor families, such as the MYB transcription factor family. These TFs are classified according to the number of repeats in the MYB-DNA-binding structure domain. Those with only one DBD are in the Myb-related family, while those with two or more are in the Myb family.

Transcription factors (TFs) play crucial roles in almost all biological processes. Most Arabidopsis TFs belong to large families with similar DBD structures. In this chapter, we focus on some of TF families.

9.1.2.1 bZIP TF Family

The basic region/leucine zipper (bZIP) TFs family control important processes in all eukaryotes. The bZIP TFs have a basic region (BR) that binds DNA and a leucine zipper region (ZR). The bZIP domain consists of two structural features located on a continuous alpha helix. Plant bZIP proteins preferentially bind to DNA sequences with ACGT cores (Jakoby et al. 2002). Several studies have demonstrated the interaction between the bZIP DNA-binding motif and the yeast transcriptional activator GCN4 (Ellenberger et al. 1992).

In plants, bZIP is the primary regulator of many developmental and physiological processes, including morphogenesis, seed formation, and abiotic and biological stress responses. The regulation of the expression pattern of the bZIP gene and its changes in activity often contribute to the activation of the signaling pathways and regulatory networks of different physiological processes.

For example, in the salicylic acid (SA)-mediated signaling pathway triggered by attack from a biotrophic pathogen, one class of bZIP proteins that is linked to biotic stress responses comprises the TGA (TGACGTCA cis-element-binding proteins) and can interact with the non-expresser of pathogen-related (PR) genes (NPR1), which is a key component in the SA defense signaling pathway activates the expression of SA-responsive genes. BZIP transcription factors associated with pathogen defense can also recognize a variety of cis-elements in the promoters of their target genes (Alves et al. 2013).

9.1.2.2 bHLH TF Family

The bHLH (Basic helix-loop-helix) TFs are widely distributed in eukaryotes and have been characterized in non-plant eukaryotes. In mammalian systems, considerable structural, functional, and phylogenetic analyses have been performed. This is the second largest family in plants after the MYB family.

The bHLH TFs family is defined by the BHLH signature domain, which consists of 60 amino acids and has two functionally distinct regions. The HLH (helix-loop-helix) region, located at the end of the C-terminal, is a dimerized region, which consists mainly of hydrophobic residues and forms two amphoteric helices separated by a circular region of variable sequence and length. The core DNA sequence motif recognized by the bHLH proteins is a consistent hexanucleotide sequence known as the E-box (5′-CANNTG-3′). The identities of the two central bases determine the different types of E-boxes. One of the most common is the palindrome G-box (5′-CACGTG-3′). Certain conserved amino acids within the basic region of the protein provide recognition of the core consensus site, while other residues within the region dictate the specificity of specific types of E-boxes (Toledo-Ortiz et al. 2003).

Like most transcription factors, bHLH can regulate gene expression through interaction with specific motifs in target genes. Functionally, bHLH transcription factors are widely involved in plant growth and metabolism, including photomorphogenesis, light signal transduction, and secondary metabolism. They are also involved in plant response to adversity (Sun et al. 2018).

9.1.2.3 MYB TF Family

The MYB family of transcription factors (TFs) is named for its conserved MYB domain and is present in all eukaryotes. The first Myb gene was an “oncogene” from the avian myeloblastic disease virus, v-Myb. Many vertebrates contain three genes associated with v-Myb, c-Myb, a-Myb, and b-Myb, and similar genes have been identified in insects, plants, fungi, and slime molds. The structures and functions of MYB transcription factors in plants are highly conserved compared with animals and yeasts (Li et al. 2015).

The proteins encoded are critical for controlling proliferation and differentiation of multiple cell types and share conserved MYB DNA-binding domains. This domain usually consists of three imperfect repeats (R), each of which forms a helix-turn-helix structure of 53 amino acids.

MYB proteins can be classified according to the number of MYB repeats (1–4). The three replicates in c-Myb are R1, R2, and R3, respectively. The repeats of other MYB proteins are classified according to their similarity to R1, R2, or R3. Plant MYB proteins are mainly divided into three categories: R2R3-MYB, which has two adjacent repeats; R1R2R3-MYB, with three adjacent repeats; and a heterogeneous group collectively known as MYB-associated proteins, which usually contain an MYB repeat sequence.

Phenotypic analysis and dissection of mutants with interesting phenotypes revealed the function of 125 R2R3-MYB genes in *Arabidopsis thaliana*. The R2R3 MYB gene controls many aspects of plant secondary metabolism, as well as the characteristics and fate of plant cells (Stracke et al. 2001).

9.1.2.4 WRKY TF Family

WRKY transcription factors are a large family of transcriptional regulators in plants and form integral parts of signaling webs that modulate many plant processes. The defining feature of WRKY TFs is their DNA binding domain. This is named after the nearly invariant WRKY amino acid sequence at the N-terminus (Rushton et al. 2010).

Studies on WRKY transcription factors show that members of this multigene family play a role in transcriptional reprogramming related to plant immune responses. WRKY TFs are also involved in many processes, including embryogenesis, seed coat and trichome development, anthocyanin synthesis and hormone signaling (Pandey and Somssich 2009).

Although their DNA-binding domains are highly conserved, the overall structure of WRKY proteins is highly diverse and can be divided into distinct groups. All known WRKY proteins contain either one or two WRKY domains. They can be classified according to the number of WRKY domains and the characteristics of the zinc finger-like motif. WRKY proteins with two WRKY domains belong to group I, while most proteins with one WRKY domain belong to group II. Generally, the WRKY domains of group I and group II members have the same Cys2-His2 zinc-finger motif. The single finger motif of a small subset of WRKY proteins is distinct from that of group I and II members. Instead of a C2–H2 pattern, their WRKY domains contain a C2–HC motif. Owing to this distinction, they have been assigned to the newly defined group III (Eulgem et al. 2000).

9.1.2.5 AP2/ERF Family

The APETALA 2/ethylene-responsive element-binding factor (AP2/ERF) family is a large family of plant-specific transcription factors, which includes the following major subfamilies: the AP2, RAV, ERF, and dehydration-responsive element-binding protein (DREB) subfamilies (Mizoi et al. 2012).

The AP2/ERF family is a large group of transcription factors containing AP2/ERF type DNA-binding domains, whose members are encoded by 145 loci in *Arabidopsis thaliana*. This domain was first identified as a repeated motif within the Arabidopsis homeotic gene APETALA 2 (AP2) involved in flower development and a similar domain was found in *Nicotiana tabacum* ethylene-responsive element-binding proteins (EREBPs) (Sakuma et al. 2002). The AP2/ERF superfamily is defined by the AP2/ERF domain, which consists of about 60–70 amino acids and is involved in DNA binding. These proteins are involved in the transcriptional regulation of biological processes related to growth and development, as well as in response to environmental stimuli.

Genes in the AP2 family are involved in the regulation of developmental processes such as flower development, leaf epidermal cell properties, and embryo development. Many proteins in the ERF family are involved in different functions of cellular processes, such as hormonal signal transduction, abiotic stress, regulation of metabolism and developmental processes (Nakano et al. 2006).

9.1.2.6 NAC TF Family

NAM, ATAF, and CUC (NAC) transcription factors constitute a large protein family. NAC TFs are plant-specific TFs involved in development and abiotic and biological stress responses (Nakashima et al. 2012). This protein family contains a highly conserved N-terminal DNA-binding domain and a variable C-terminal domain.

NAC was originally derived from the names of three proteins, no apical meristem (NAM), ATAF1-2, and CUC2 (cup-shaped cotyledon), that contain a similar DNA-binding domain (Fang et al. 2008). Many of them, including Arabidopsis CUC2, are involved in plant development. Some NAC genes are upregulated during injury and bacterial infection, while others mediate viral resistance (Nakashima et al. 2012).

In Arabidopsis, drought induces the production of NAC transcription factors. Overexpression of three NAC genes (ANAC019, ANAC055 and ANAC072) in *Arabidopsis thaliana* (At) improved plant stress tolerance and altered the expression of drought, salinity and low-temperature stress-induced genes (Hu et al. 2006). Overexpression of AtNAC2 leads to altered lateral root development and increased salt tolerance (He et al. 2005).

Individual transcription factors may be involved in more than one biological process. Here, we only list some of the functions of transcription factors in some families according to the literature. Note that the full function of this transcription factor is not shown (Table 9.1).

Table 9.1 Part of the function of some transcription factors

Function	Name	TF family	References
Biotic			
Anthocyanin biosynthesis	GL3, TT8	bHLH	Nesi et al. (2000), Zhang et al. (2003)
	PAP1 (vegetative tissues)	MYB	Serna and Martin (2006)
	CPC (negative regulator)	MYB	Serna and Martin (2006)
Proanthocyanidin biosynthesis	TT8	bHLH	Baudry et al. (2004)
Iron homeostasis	ORG2, ORG3	bHLH	Feller et al. (2011)
Regulation of iron uptake	FIT	bHLH	Feller et al. (2011)
Metal homeostasis, auxin-conjugate metabolism	ILR3	bHLH	Feller et al. (2011)
Glucosinolate biosynthesis	HAG2, HIG1, ATR1, HAG3/PMG2. HAG1/PMG1	MYB	Dubos et al. (2010)
Phenylpropanoide pathway	PFG3, PFG1, PFG2, PAP2, PAP1, TT2	MYB	Feller et al. (2011)
Flavonol biosynthesis (all tissues)	PFG2, PFG1, PFG3	MYB	Feller et al. (2011)
Seed coat differentiation	GL3, EGL3, TT8, MYC1	bHLH	Gonzalez et al. (2009)
Fruit differentiation	IND	bHLH	Feller et al. (2011)
Cell fate	FLP, WER, GL1, NOK, MIXTA	MYB	Song et al. (2009)
Programmed cell death	XND1	NAC	Zhao et al. (2008)
Cell cycle, pre-mRNA splicing and transcriptional regulation of cyclins	CDC5	MYB	Burns et al. (1999); Lin et al. (2007)
Cell-cycle regulation;	XAL1	MADS	Tapia-López et al. (2008)
Circadian clock	CCA1, LHY	MYB	Lu et al. (2009)
Development			
Fertilization	UNE12, UNE10	bHLH	Feller et al. (2011)
Early embryo development	MEE8	bHLH	Feller et al. (2011)
Embryo sac development	AGL23	MADS	Colombo et al. (2008)
Central cell and endosperm development	AGL62, AGL80	MADS	Kang et al. (2008), Portereiko et al. (2006)

(continued)

Table 9.1 (continued)

Function	Name	TF family	References
Root hair formation and development	LHW, MYC1, GL3, EGL3, RHD6, RSL1–5	bHLH	Feller et al. (2011)
	WER, TRY, CPC, ETC1, ETC2, ETC3, TCL1, MYBL2	MYB	Feller et al. (2011)
	NAC1	NAC	Guo et al. (2005)
	ANR1, FYF, XAL1	MADS	Zhang and Forde (1998), Nawy et al. (2005), Tapia-López et al. (2008)
Fruit development	ALC (dehiscence), SPT	bHLH	Rajani and Sundaresan (2001)
	GOA, AGL15 (maturation)	MADS	Prasad et al. (2010), Harding et al. (2003)
Carpel margin development	SPT	bHLH	Feller et al. (2011)
Transmitting tract and stigma development	HEC1, HEC2, HEC3	bHLH	Gremski et al. (2007)
Anther development	AMS	bHLH	Feller et al. (2011)
	TDF1, MS35	MYB	Dubos et al. (2010)
Stamen development	DUO1, TDF1, MS35, BOS1	MYB	Dubos et al. (2010)
Axillary meristem regulation/lateral organ separation	AS1(leaves), LOF1, RAX3, RAX2/BIT1, RAX1	MYB	Dubos et al. (2010)
	AGL6	MADS	Kim et al. (2005)
Hypoctyle elongation	LAF1 (far red light-mediated phyA signaling)	MYB	Yang et al. (2009a, b)
	RAX2/BIT1 (blue light-mediated CRY1 signaling)	MYB	Dubos et al. (2010)
Embryogenesis/seed maturation	PGA37	MYB	Dubos et al. (2010)
	AGL15	MADS	Heck et al. (1995)
Petal epidermis cell shape	MIXTA	MYB	Perez-Rodriguez et al. (2005)
Seedling hypocotyl elongation (far-red light)	LAF1	MYB	Feller et al. (2011)
Shoot morphogenesis and leaf patterning	AS1	MYB	Feller et al. (2011)
Leaf senescence	NTL4, NTL9, VNI2, OR/RD, AtNAP	NAC	Nuruzzaman et al. (2013)
	RAV (positive)	AP2/ERF	Woo et al. (2010)

(continued)

Table 9.1 (continued)

Function	Name	TF family	References
Pollen maturation and tube growth	AGL65	MADS	Adamczyk and Fernandez (2009)
Carpel and ovule development; periodic lateral root formation	SHP1, SHP2, STK	MADS	Liljegren et al. (2000), Moreno-Risueno et al. (2010), Pinyopich et al. (2003)
Transition to flowering (activator)	XAL1, FYF, AGL71, AGL72, SOC1	MADS	Tapia-López et al. (2008), Dorca-Fornell et al. (2011), Smaczniak et al. (2012)
Transition to flowering (repressor)	AGL15 (with AGL18)	MADS	Smaczniak et al. (2012)
Transition to flowering (activator)	AGL18 (with AGL15)	MADS	Smaczniak et al. (2012)
	FLC, MAF1-4, SVP	MADS	Michaels and Amasino (1999), Ratcliffe et al. (2001), Hartmann et al. (2000)
	AGL17, AGL19, AGL24, MAF5, AGL28, AGL6, FYF	MADS	Han et al. (2008), Schönrock et al. (2006), Michaels et al. (2003), Ratcliffe et al. (2003), Yoo et al. (2006, 2011), Smaczniak et al. (2012)
Seed pigmentation and endothelium development	ABS	MADS	Smaczniak et al. (2012)
Seed development	PHE1	MADS	Köhler et al. (2003)
Sepal and petal longevity	AGL15	MADS	Smaczniak et al. (2012)
Flower organ senescence and abscission	FYF	MADS	Chen et al. (2011)
Positively regulate floral organ identity	AP2	AP2/ERF	Dinh et al. (2012)
Number and distribution of stomata	AGL16	MADS	Kutter et al. (2007)
Stomata development	ICE1, SCRM2	bHLH	Nadeau (2009)
Meristem identity specification	CAL, FUL, AP1	MADS	Kempin et al. (1995), Smaczniak et al. (2012)
Negatively regulate plant development	TINY, RAP2.4, RAP2.6	AP2/ERF	Sun et al. (2008), Lin et al. (2008), Zhu et al. (2010)
Negatively regulate ABA signaling during seed germination	RAV	AP2/ERF	Kagaya et al. (1999)

(continued)

Table 9.1 (continued)

Function	Name	TF family	References
Stress			
Cold acclimatization response and freezing tolerance	ICE1, SCRM2	bHLH	Feller et al. (2011)
	LOV1	NAC	Yoo et al. (2007)
	CBF1	AP2/ERF	Yang et al. (2009a, b)
ABA signaling	MYC2, AtAIB	bHLH	Abe et al. (2003)
	NTL8	NAC	Kim et al. (2008)
	ABI4, TINY, ORA47	AP2/ERF	Bossi et al. (2009), Sun et al. (2008)
Light signaling	MYC2, PIF1/PIL5, PIF3, PIF4, PIF5/PIL6	bHLH	Abe et al. (2003), Feller et al. (2011)
Gibberellin signaling	PRE1-5, PIF1/PIL5, PIF3, PIF4, PIF5/PIL6	bHLH	Feller et al. (2011)
JA signaling	MYC2	bHLH	Abe et al. (2003)
	BOS1	MYB	Baldoni et al. (2015)
	ORA47	AP2/ERF	Chen et al. (2016)
GA signaling	NTL8	NAC	Kim et al. (2008)
Shade avoidance response	PAR1	bHLH	Carretero-Paulet et al. (2010)
Drought-stress response	BOS1	MYB	Baldoni et al. (2015)
	AhNAC2, RD29A, RD29B, RAB18, ERD1, COR47, COR15a, KIN1, AREB1, CBF1, NTL6	NAC	Liu et al. (2011)
	ERF53, RAP2.4, RAP2.4A	AP2/ERF	Lin et al. (2008), Cheng et al. (2012)
Phosphate starvation response	PHR1	MYB	Hosoda et al. (2002), Bustos et al. (2010)
Pathogen infection	BOS1	MYB	Baldoni et al. (2015)
	NIT2, ATAF2	NAC	Huh et al. (2012)
Wounding	BOS1	MYB	Baldoni et al. (2015)
Salt tolerance	NTL8	NAC	Kim et al. (2008)
Positively regulate hypoxia tolerance	ERF71/HRE2, ERF72/RAP2.3, ERF74/RAP2.12, ERF75/RAP2.2	AP2/ERF	Lee et al. (2015), Welsch et al. (2007), Gasch et al. (2016)

9.1.3 Bioinformatic Resources for Plant Transcription Factors

A list of bioinformatics databases for plant TFs are shown in the Table 9.2.

Table 9.2 Database for the classification of transcription factor families

Database	URL	Plant species
AGRIS	http://arabidopsis.med.ohio-state.edu/AtTFDB/	*Arabidopsis thaliana*
RARTF	http://rarge.gsc.riken.jp/rartf/	*Arabidopsis thaliana*
DATF	http://datf.cbi.pku.edu.cn/	*Arabidopsis thaliana*
DRTF	http://drtf.cbi.pku.edu.cn/	Rice
DPTF	http://dptf.cbi.pku.edu.cn/	Poplar
TOBFAC	http://compsysbio.achs.virginia.edu/tobfac/	Tobacco
wDBTF	http://wwwappli.nantes.inra.fr:8180/wDBFT/	Wheat
soyDB	http://casp.rnet.missouri.edu/soydb/	Soybean
SoybeanTFDB	http://soybeantfdb.psc.riken.jp/	Soybean
RiceSRTFDB	http://www.nipgr.res.in/RiceSRTFDB.html	Rice
STIFDB	http://caps.ncbs.res.in/stifdb	Arabidopsis, rice
IT3F	http://jicbio.nbi.ac.uk/IT3F/	Arabidopsis, rice
GRASSIUS	http://grassius.org/summary.html	Corn, rice, sorghum, sugar cane
LegumeTFDB	http://legumetfdb.psc.riken.jp/	Soybean, root, tribulus alfalfa
TreeTFDB	http://treetfdb.bmep.riken.jp/index.pl	Jatropha, papaya, cassava, poplar
GramineaeTFDB	http://gramineaetfdb.psc.riken.jp	Brachypodium, rice, sorghum, maize
PlnTFDB	http://plntfdb.bio.uni-potsdam.de/v3.0/	Multiple species
PlantTFDB	http://planttfdb.cbi.pku.edu.cn/	Multiple species
DBD	http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Home	Multiple species

9.2 Methods for Genome-Wide Identification of Transcription Factor Binding Sites

From the introduction, we know that transcription factors (TFs) are sequence-specific DNA-binding proteins that regulate gene expression in organisms. They recognize specific sequences in the DNA and bind together to accomplish their functions. Based on the manner of TF recognition, many methods for finding TFBS have been developed.

9.2.1 Experimental Methods for Identifying TFBSs

Over the past decade, next-generation sequencing (NGS) technologies, such as ChIP-Seq and DAP-Seq, have provided better technical support for exploring the

landscape of TFs. This has led to the creation of many databases for transcription factor binding sites (TFBS) deposition and profiling, including TRANSFAC, JASPAR, UniPROBE, and SwissRegulon. Both of them have substantially boosted our understanding of interactions between TF and TFBS.

Here, we address some experimental methods for identifying TFBSs to explain how TFBS are identified experimentally (Lai et al. 2019).

9.2.1.1 ChIP-seq

Chromatin immunoprecipitation sequencing (ChIP-Seq) is a technique for the profiling of genome-wide TFBSs bound by a given TF. Combined with chromatin immunoprecipitation and large-scale parallel sequencing, ChIP-seq can map genome-wide TFBSs in vivo.

A standard ChIP-seq protocol is as follows. First, chromatin-containing TF–DNA complexes are isolated from sample tissues which are treated with a crosslinking reagent and subjected to nuclei purification. Then the DNA fragments associated with a TFs are enriched and the chromatins in them are cut into smaller pieces by sonication. In the next step, the DNA–protein complex is immunoprecipitated using antibodies specific to the protein. Finally, the crosslinks are reversed, freeing the DNA for analysis and ultimately determining the sequence that binds to the protein (Park 2009).

ChIP-seq offers a number of advantages. For example, its resolution can reach the level of a single nucleotide and is limited only by the alignability of reads to the genome. In the process of hybridization using ChIP-chip, there may be cross hybridization between incomplete matched sequences, resulting in false positives. This increases the signal noise. However, ChIP-seq can be unaffected by these noises. It generates a more accurate list of protein binding sites and transcription factors and allows in vitro high-throughput identification of TF binding specificity, which helps predict TFBS in genome sequences.

9.2.1.2 DAP-seq

Although many of the advantages of ChIP-seq have already been mentioned, the main disadvantages of Chip-seq are cost and availability. DNA affinity purification sequencing (DAP-seq) has been recently developed.

Compared to ChIP-seq and its variants, DAP-Seq can be performed at a lower cost and in a high-throughput manner. It is an alternative to ChIP-seq and expresses transcription factor proteins in vitro, interacts with genomic DNA and sequences the DNA bound to the recombinant protein. It also analyzes the gene sites bound to the protein as well as the specific DNA motif. Because there is no need for specific antibodies and the fixation of DNA and protein, the technical requirements of the analysis are reduced compared with ChIP-seq. DAP-seq is a high-throughput TF binding site discovery method, using genomic DNA in vitro, which enables rapid

identification of target genes directly bound downstream by transcription factors (Bartlett et al. 2017).

DAP-seq has some limitations. DAP-seq lacks the chromatin background of cells because this is an in vitro interaction experiment. Controls therefore need to be introduced in the analysis, and DAP-seq should be combined with other methods. In addition, the experimental design and data analysis must consider that many transcription factors may show different DNA binding characteristics in the presence of cofactors.

Despite these limitations, the combination of DAP-seq in vitro and ChIP-seq in vivo is an information-rich method for modeling TFBS and predicting TF binding in vivo.

9.2.1.3 ATAC-seq

In general, TFs preferentially bind to TFBSs in the depleted region of the nucleosome (NDR), where chromatin is more accessible. This has been confirmed by large-scale studies, in which most of the active cis-elements are present in NDR in different species. Thus, obtaining datasets of chromatin accessibility is a necessary step in the research process. For this research, the best current methods are DNase-seq, MNase-seq, FAIRE-seq, and ATAC-seq.

Among these, ATAC-seq is popular because it requires no sonication or phenol-chloroform extraction like FAIRE-seq, no antibodies like ChIP-seq, and no sensitive enzymatic digestion like MNase-seq or DNase-seq. The method needs a minimum number of input samples, and its preparation can be completed in less than 3 h, which is faster than the alternative methods.

The Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq) methodology relies on the use of hyperactive translocase Tn5 to construct a library and uses DNA transposable enzyme technology to analyze chromatin accessibility. Tn5 is a prokaryotic transposase that endogenously functions through a “cut and paste” mechanism.

Under normal circumstances, ATAC-seq can be divided into three independent components: cell lysis, transposition, and amplification. Crosslinking generally reduces library creation efficiency, and therefore, some studies recommend starting with fresh unfixed cells for maximum sensitivity (Buenrostro et al. 2015).

9.2.2 Computational Frameworks for Modeling TFBSs

Using computer methods to find TFBS is also common. A quantitative TFBS model representing TF-DNA-binding affinity allows accurate de novo prediction of given TF binding sites. These models can be calculated from a set of known TFBSs. Here, we focus on the most widely used and representative TFBS modeling algorithms (Lai et al. 2019).

The most commonly used representative model is the Position Weight Matrix (PWM). To construct the PWM matrix, one first needs to obtain the Position Frequency Matrix (PFM), which is the number of occurrence of four nucleotides at each position (Fig. 9.2; PFM from JASPAR). As can be seen from PFM, in the first position A appears 201 times, C 201 times, G 396 times, and T 201 times. Similarly, the frequency of each position is converted into a frequency to obtain the position probability matrix (PPM).

On the basis of the PPM, the PWM can be obtained using the formula $W_{j,k} = \log_2(M_{j,k}/b_k)$ to obtain the correction. In the above formula, b is the background probability, which, in this case, is 0.25 (assuming the same amounts of ATCG in the genome), where $M_{j,k}$ is the probability of base k in position j. The PWM matrix can be visualized in the way of TFBS logo (Fig. 9.3), which indicates the preference of TF binding at each location.

Although PWM is widely used, it is based on the assumption that each location in the same TFBS is independent of the binding affinity of the other locations.

Frequency matrix

A [	201	372	283	23	999	0
C [	201	152	72	350	0	999
G [	396	321	72	523	0	0
T [	201	152	571	102	0	0

Fig. 9.2 Position Frequency Matrix (PFM) of HYH

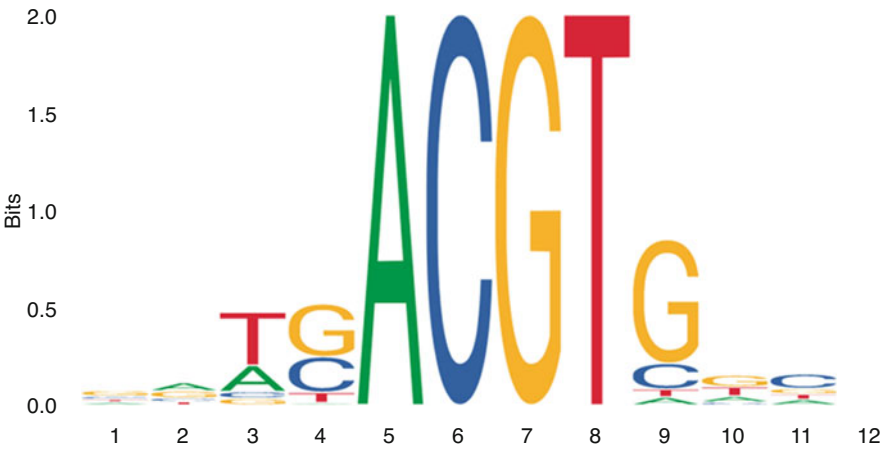

Fig. 9.3 Motif presentation of HYH

Therefore, PWM cannot describe the dependence between bases. With these issues in mind, newer models, such as the binding energy model (BEM), dinucleotide weight tensor (DWT) and TF Flexible Model (TFFM), were developed.

Some studies have found that highly conserved DNA-binding domains can bind different sequences, and this could not be explained by the base readout and indirect readout. For example, the TF paralogs and glucocorticoid receptors share only one third of their sequence, but they can bind similar DNA motifs through a set of the same amino acids. The shape of the DNA helps TFs recognize TFBSs. For this reason, some algorithms integrate Hi-C data (high-throughput/esolution idea conformation capture), such as *DeFine*.

For TFBS, we typically only focus on its core motif. However, although relatively little sequence information is provided by the flanking regions, they do work jointly with the core motif to determine the shape feature. *SelexGLM* is a model that takes flanking region data into consideration (Table 9.3).

9.3 Constructing and Dissecting Transcription Factor-Associated Regulatory Networks

Transcription factors can help us understand some of the workings of the signaling network—for example, which genes it is regulated by and which genes it affects. Using data in the ChIP-Hub (a database containing ChIP-seq data for many plant transcription factors), we analyzed previously described TF to obtain such a TFs regulatory network map. Each node in the map represents a transcription factor, and the line between two nodes represents the interaction between two transcription factors (Fig. 9.4; left). In this network, we also divided each transcription factor (node) into different categories based on how closely they are connected (Fig. 9.4; right).

We identified 22 transcription factors involved in flower development and showed their regulatory network in red (Fig. 9.5). This provided a pattern describing the transcription factors related to flower development. Although the network appears incomplete (perhaps because only some of the transcription factors described in the literature have been described), we can see what other transcription factors affect these transcription factors.

This biological process is very complex. As mentioned earlier, one transcription factor may be involved in multiple biological processes. For example, transcription factors such as XAL1 and FYF, which are associated with root development, are also involved in flower development.

Table 9.3 Some computational frameworks for modeling TFBSs

TFBS modeling methods	Description	Features integrated	Web server or source code
PWM (position weight matrix)	PWMs are normalized representations of the position-specific log-likelihoods of a nucleotide's probability to occur at each position in a sequence	NA (not applicable)	NA
BEM (binding energy model)	BEM introduces energy parameters of adjacent nucleotides to the binding affinity quantification	Dependencies (adjacent positions) and binding affinity data	http://stormo.wustl.edu/TF-BEMs/
TFFM (TF Flexible Model)	TFFM model integrates a Markov model to take dependencies and background into account	Dependencies (adjacent position) and background	http://cisreg.cmmt.ubc.ca/cgi
DWT (dinucleotide weight tensor)	DWT is a regulatory motif model that incorporates arbitrary pairwise dependencies for TFBS prediction	Dependencies between all positions	http://dwt.unibas.ch/fcgi/dwt
DeFine	DeFine quantifies TF–DNA-binding affinity and facilitate evaluation of functional noncoding variants in the genome based on deep learning algorithms	Integrate Hi-C data	http://define.cbi.pku.edu.cn
SelexGLM	SelexGLM incorporates core motif flanking region for TFBS binding quantification	Core motif flanking region	https://www.bioconductor.org
DFIM (Deep Feature Interaction Maps)	DFIM estimates pairwise interactions between features (such as nucleotides or subsequences) in any input DNA sequences by a neural network	Dependencies between all positions, interaction between motifs, core motif flanking region, and chromatin accessibility	https://github.com/kundajelab/dm.

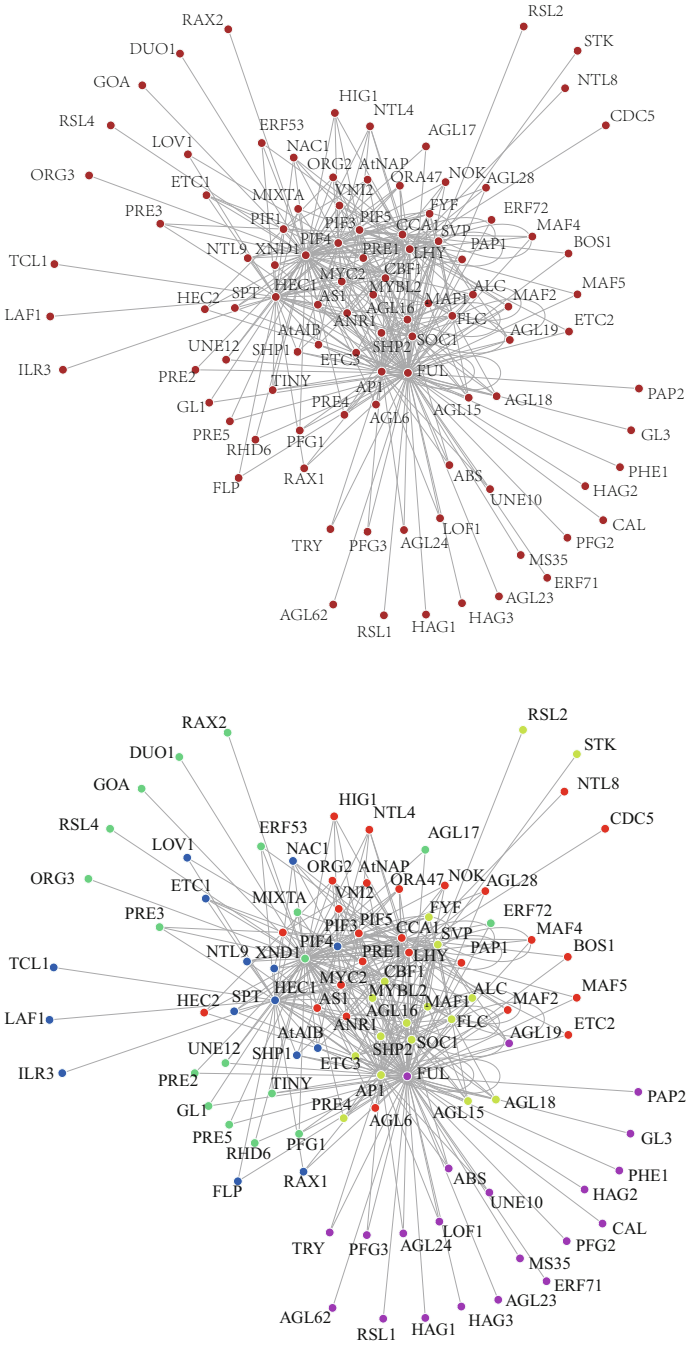

Fig. 9.4 Transcription factor regulatory network

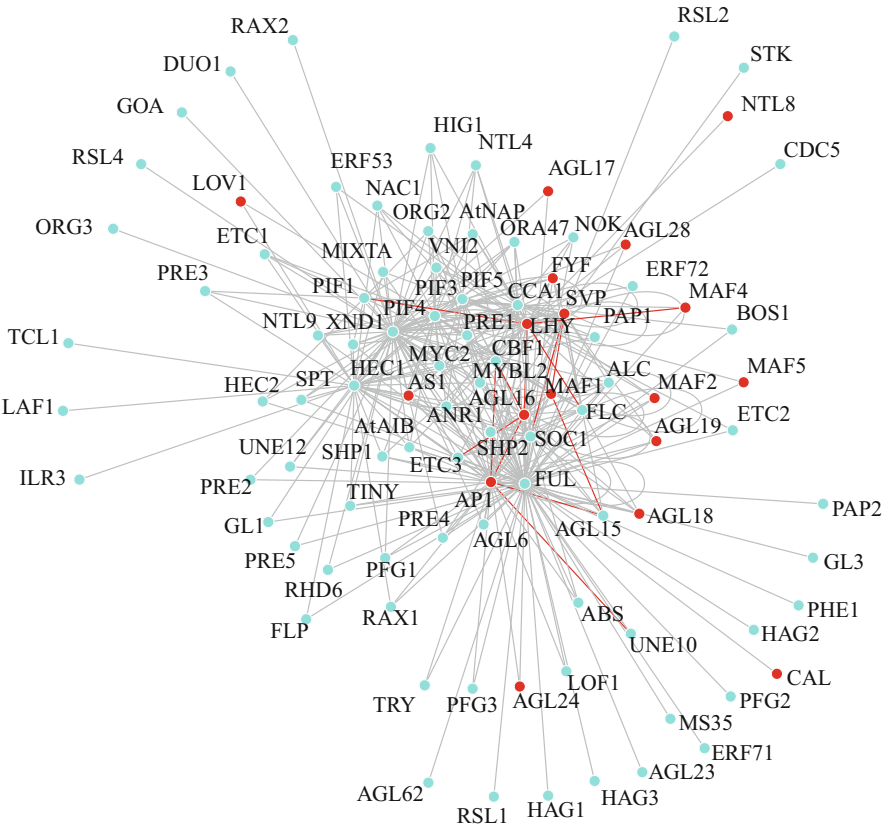

Fig. 9.5 Floral-associated transcription factor regulatory networks

9.4 Concluding Remarks

Transcription factors constitute the key nodes of the gene regulatory network. It is involved in biological processes such as plant morphological change and stress response, and affects the final results. We can use the existing data to build gene regulatory networks to assist research. In this chapter, we introduce some of the major transcription factor families in plants. Some simple networks were constructed using transcription factors and ChIP-hub data mentioned in the literature. All these are important means to assist the study of plant.

References

Abe H, Urao T, Ito T, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. Plant Cell 15(1):63–78. https://doi.org/10.1105/tpc.006130

Adamczyk BJ, Fernandez DE (2009) MIKC * MADS domain heterodimers are required for pollen maturation and tube growth in Arabidopsis 1[W][OA]. Plant Physiol 149(4):1713–1723. https://doi.org/10.1104/pp.109.135806

Adcock IM, Caramori G (2009) Transcription factors. In: Asthma and COPD (Second Edition) Basic Mechanisms and Clinical Management. National Heart and Lung Institute, Imperial College School of Medicine, London, UK. 373–380. https://doi.org/10.1016/B978-0-12-374001-4.00031-6. Accessed 30 Jan 2009

Alves MS, Dadalto SP, Gonçalves AB, De Souza GB, Barros VA, Fietto LG (2013) Plant bZIP transcription factors responsive to pathogens: a review. Int J Mol Sci 14:7815. https://doi.org/10.3390/ijms14047815

Baldoni E, Genga A, Cominelli E (2015) Plant MYB transcription factors: their role in drought response mechanisms. Int J Mol Sci 16:15811. https://doi.org/10.3390/ijms160715811

Bartlett A, O'Malley RC, Huang SSC, Galli M, Nery JR, Gallavotti A, Ecker JR (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat Protoc 12:1659. https://doi.org/10.1038/nprot.2017.055

Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B, Lepiniec L (2004) TT2, TT8, and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in Arabidopsis thaliana. Plant J 39(3):366–380. https://doi.org/10.1111/j.1365-313X.2004.02138.x

Bossi F, Cordoba E, Dupré P, Mendoza MS, Román CS, León P (2009) The Arabidopsis ABA-INSENSITIVE (ABI) 4 factor acts as a central transcription activator of the expression of its own gene, and for the induction of ABI5 and SBE2.2 genes during sugar signaling. Plant J 59(3):359–374. https://doi.org/10.1111/j.1365-313X.2009.03877.x

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ (2015) ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol 109:21.29.1. https://doi.org/10.1002/0471142727.mb2129s109

Burns CG, Ohi R, Krainer AR, Gould KL (1999) Evidence that Myb-related CDC5 proteins are required for pre-mRNA splicing. Proc Natl Acad Sci U S A 96(24):13789–13794. https://doi.org/10.1073/pnas.96.24.13789

Bustos R, Castrillo G, Linhares F, Puga MI, Rubio V, Pérez-Pérez J, Solano R, Leyva A, Paz-Ares J (2010) A central regulatory system largely controls transcriptional activation and repression responses to phosphate starvation in arabidopsis. PLoS Genet 6(9):e1001102. https://doi.org/10.1371/journal.pgen.1001102

Carretero-Paulet L, Galstyan A, Roig-Villanova I, Martínez-García JF, Bilbao-Castro JR, Robertson DL (2010) Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in Arabidopsis, poplar, rice, moss, and algae. Plant Physiol 153(3):1398–1412. https://doi.org/10.1104/pp.110.153593

Chen MK, Hsu WH, Lee PF, Thiruvengadam M, Chen HI, Yang CH (2011) The MADS box gene, FOREVER YOUNG FLOWER, acts as a repressor controlling floral organ senescence and abscission in Arabidopsis. Plant J 68(1):168–185. https://doi.org/10.1111/j.1365-313X.2011.04677.x

Chen HY, Hsieh EJ, Cheng MC, Chen CY, Hwang SY, Lin TP (2016) ORA47 (octadecanoid-responsive AP2/ERF-domain transcription factor 47) regulates jasmonic acid and abscisic acid biosynthesis and signaling through binding to a novel cis-element. New Phytol 211(2):599–613. https://doi.org/10.1111/nph.13914

Cheng MC, Hsieh EJ, Chen JH, Chen HY, Lin TP (2012) Arabidopsis RGLG2, functioning as a RING E3 ligase, interacts with AtERF53 and negatively regulates the plant drought stress response. Plant Physiol 158(1):363–375. https://doi.org/10.1104/pp.111.189738

Colombo M, Masiero S, Vanzulli S, Lardelli P, Kater MM, Colombo L (2008) AGL23, a type I MADS-box gene that controls female gametophyte and embryo development in Arabidopsis. Plant J 54(6):1037–1048. https://doi.org/10.1111/j.1365-313X.2008.03485.x

Dinh TT, Girke T, Liu X, Yant L, Schmid M, Chen X (2012) The floral homeotic protein APETALA2 recognizes and acts through an AT-rich sequence element. Development 139(11):1978–1986. https://doi.org/10.1242/dev.077073

Dorca-Fornell C, Gregis V, Grandi V, Coupland G, Colombo L, Kater MM (2011) The Arabidopsis SOC1-like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems. Plant J 67(6):1006–1017. https://doi.org/10.1111/j.1365-313X.2011.04653.x

Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L (2010) MYB transcription factors in Arabidopsis. Trends Plant Sci 15:573. https://doi.org/10.1016/j.tplants.2010.06.005

Ellenberger TE, Brandl CJ, Struhl K, Harrison SC (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: crystal structure of the protein-DNA complex. Cell 71:1223. https://doi.org/10.1016/S0092-8674(05)80070-4

Eulgem T, Rushton PJ, Robatzek S, Somssich IE (2000) The WRKY superfamily of plant transcription factors. Trends Plant Sci 5:199. https://doi.org/10.1016/S1360-1385(00)01600-9

Fang Y, You J, Xie K, Xie W, Xiong L (2008) Systematic sequence analysis and identification of tissue-specific or stress-responsive genes of NAC transcription factor family in rice. Mol Genet Genomics 280:547. https://doi.org/10.1007/s00438-008-0386-6

Feller A, MacHemer K, Braun EL, Grotewold E (2011) Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant J 66:94. https://doi.org/10.1111/j.1365-313X.2010.04459.x

Gasch P, Fundinger M, Müller JT, Lee T, Bailey-Serres J, Mustropha A (2016) Redundant ERF-VII transcription factors bind to an evolutionarily conserved cis-motif to regulate hypoxia-responsive gene expression in Arabidopsis. Plant Cell 28(1):160–180. https://doi.org/10.1105/tpc.15.00866

Gonzalez A, Mendenhall J, Huo Y, Lloyd A (2009) TTG1 complex MYBs, MYB5 and TT2, control outer seed coat differentiation. Dev Biol 325(2):412–421. https://doi.org/10.1016/j.ydbio.2008.10.005

Gremski K, Ditta G, Yanofsky MF (2007) The HECATE genes regulate female reproductive tract development in Arabidopsis thaliana. Development 134(20):3593–3601. https://doi.org/10.1242/dev.011510

Guo HS, Xie Q, Fei JF, Chua NH (2005) MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for Arabidopsis lateral root development. Plant Cell 17(5):1376–1386. https://doi.org/10.1105/tpc.105.030841

Han P, García-Ponce B, Fonseca-Salazar G, Alvarez-Buylla ER, Yu H (2008) AGAMOUS-LIKE 17, a novel flowering promoter, acts in a FT-independent photoperiod pathway. Plant J 55(2):253–265. https://doi.org/10.1111/j.1365-313X.2008.03499.x

Harding EW, Tang W, Nichols KW, Fernandez DE, Perry SE (2003) Expression and maintenance of embryogenic potential is enhanced through constitutive expression of AGAMOUS-Like 15. Plant Physiol 133(2):653–663. https://doi.org/10.1104/pp.103.023499

Hartmann U, Höhmann S, Nettesheim K, Wisman E, Saedler H, Huijser P (2000) Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. Plant J 21(4):351–360. https://doi.org/10.1046/j.1365-313X.2000.00682.x

He XJ, Mu RL, Cao WH, Zhang ZG, Zhang JS, Chen SY (2005) AtNAC2, a transcription factor downstream of ethylene and auxin signaling pathways, is involved in salt stress response and lateral root development. Plant J 44:903. https://doi.org/10.1111/j.1365-313X.2005.02575.x

Heck GR, Perry SE, Nichols KW, Fernandez DE (1995) AGL15, a MADS domain protein expressed in developing embryos. Plant Cell 7(8):1271–1282. https://doi.org/10.1105/tpc.7.8.1271

Hosoda K, Imamura A, Katoh E, Hatta T, Tachiki M, Yamada H, Mizuno T, Yamazaki T (2002) Molecular structure of the GARP family of plant myb-related DNA binding motifs

of the Arabidopsis response regulators. Plant Cell 14(9):2015–2029. https://doi.org/10.1105/tpc.002733

Hu H, Dai M, Yao J, Xiao B, Li X, Zhang Q, Xiong L (2006) Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice. Proc Natl Acad Sci U S A 103:12987. https://doi.org/10.1073/pnas.0604882103

Huh SU, Lee SB, Kim HH, Paek KH (2012) ATAF2, a NAC transcription factor, binds to the promoter and regulates NIT2 gene expression involved in auxin biosynthesis. Mol Cells 34(3):305–313. https://doi.org/10.1007/s10059-012-0122-2

Jakoby M, Weisshaar B, Dröge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F (2002) bZIP transcription factors in Arabidopsis. Trends Plant Sci 7:106. https://doi.org/10.1016/S1360-1385(01)02223-3

Kagaya Y, Ohmiya K, Hattori T (1999) RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants. Nucleic Acids Res 27(2):470–478. https://doi.org/10.1093/nar/27.2.470

Kang IH, Steffen JG, Portereiko MF, Lloyd A, Drews GN (2008) The AGL62 MADS domain protein regulates cellularization during endosperm development in Arabidopsis. Plant Cell 20(3):635–647. https://doi.org/10.1105/tpc.107.055137

Kempin SA, Savidge B, Yanofsky MF (1995) Molecular basis of the cauliflower phenotype in Arabidopsis. Science 267(5197):522–525. https://doi.org/10.1126/science.7824951

Kim S, Koh J, Yoo MJ, Kong H, Hu Y, Ma H, Soltis PS, Soltis DE (2005) Expression of floral MADS-box genes in basal angiosperms: Implications for the evolution of floral regulators. Plant J 43(5):724–744. https://doi.org/10.1111/j.1365-313X.2005.02487.x

Kim SG, Lee AK, Yoon HK, Park CM (2008) A membrane-bound NAC transcription factor NTL8 regulates gibberellic acid-mediated salt signaling in Arabidopsis seed germination. Plant J 55(1):77–88. https://doi.org/10.1111/j.1365-313X.2008.03493.x

Köhler C, Hennig L, Spillane C, Pien S, Gruissem W, Grossniklaus U (2003) The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1. Genes Dev 17(12):1540–1553. https://doi.org/10.1101/gad.257403

Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. Nucleic Acids Res 34:D74. https://doi.org/10.1093/nar/gkj131

Kutter C, Schöb H, Stadler M, Meins F, Si-Ammour A (2007) MicroRNA-mediated regulation of stomatal development in Arabidopsis. Plant Cell 19(8):2417–2429. https://doi.org/10.1105/tpc.107.050377

Lai X, Stigliani A, Vachon G, Carles C, Smaczniak C, Zubieta C, Kaufmann K, Parcy F (2019) Building transcription factor binding site models to understand gene regulation in plants. Mol Plant 12:743. https://doi.org/10.1016/j.molp.2018.10.010

Lee SY, Hwang EY, Seok HY, Tarte VN, Jeong MS, Jang SB, Moon YH (2015) Arabidopsis AtERF71/HRE2 functions as transcriptional activator via cis-acting GCC box or DRE/CRT element and is involved in root development through regulation of root cell expansion. Plant Cell Rep 34(2):223–231. https://doi.org/10.1007/s00299-014-1701-9

Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13:233. https://doi.org/10.1038/nrg3163

Li C, Ng CKY, Fan LM (2015) MYB transcription factors; active players in abiotic stress signaling. Environ Exp Bot 114:80. https://doi.org/10.1016/j.envexpbot.2014.06.014

Liljegren SJ, Ditta GS, Eshed Y, Savidge B, Bowmant JL, Yanofsky MF (2000) SHATTERPROOF MADS-box genes control dispersal in Arabidopsis. Nature 404(6779):766–770. https://doi.org/10.1038/35008089

Lin Z, Yin K, Zhu D, Chen Z, Gu H, Qu LJ (2007) AtCDC5 regulates the G2 to M transition of the cell cycle and is critical for the function of Arabidopsis shoot apical meristem. Cell Res 17(9):815–828. https://doi.org/10.1038/cr.2007.71

Lin RC, Park HJ, Wang HY (2008) Role of Arabidopsis RAP2.4 in regulating lightand ethylene-mediated developmental processes and drought stress tolerance. Mol Plant 1(1):42–57. https://doi.org/10.1093/mp/ssm004

Liu QL, Xu KD, Zhao LJ, Pan YZ, Jiang BB, Zhang HQ, Liu GL (2011) Overexpression of a novel chrysanthemum NAC transcription factor gene enhances salt tolerance in tobacco. Biotechnol Lett 33(10):2073–2082. https://doi.org/10.1007/s10529-011-0659-8

Lu SX, Knowles SM, Andronis C, Ong MS, Tobin EM (2009) Circadian clock associated1 and late elongated hypocotyl function synergistically in the circadian clock of arabidopsis. Plant Physiol 150(2):834–843. https://doi.org/10.1104/pp.108.133272

Michaels SD, Amasino RM (1999) FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell 11(5):949–956. https://doi.org/10.1105/tpc.11.5.949

Michaels SD, Ditta G, Gustafson-Brown C, Pelaz S, Yanofsky M, Amasino RM (2003) AGL24 acts as a promoter of flowering in Arabidopsis and is positively regulated by vernalization. Plant J 33(5):867–874. https://doi.org/10.1046/j.1365-313X.2003.01671.x

Mikhaylichenko O, Bondarenko V, Harnett D, Schor IE, Males M, Viales RR, Furlong EEM (2018) The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev 32:42. https://doi.org/10.1101/gad.308619.117

Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2012) AP2/ERF family transcription factors in plant abiotic stress responses. Biochim Biophys Acta 1819:86. https://doi.org/10.1016/j.bbagrm.2011.08.004

Moreno-Risueno MA, Van Norman JM, Moreno A, Zhang J, Ahnert SE, Benfey PN (2010) Oscillating gene expression determines competence for periodic Arabidopsis root branching. Science 329(5997):1306–1311. https://doi.org/10.1126/science.1191937

Nadeau JA (2009) Stomatal development: new signals and fate determinants. Curr Opin Plant Biol 12(1):29–35. https://doi.org/10.1016/j.pbi.2008.10.006

Nakano T, Suzuki K, Fujimura T, Shinshi H (2006) Genome-wide analysis of the ERF gene family in arabidopsis and rice. Plant Physiol 140:411. https://doi.org/10.1104/pp.105.073783

Nakashima K, Takasaki H, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2012) NAC transcription factors in plant abiotic stress responses. Biochim Biophys Acta 1819:97. https://doi.org/10.1016/j.bbagrm.2011.10.005

Nawy T, Lee JY, Colinas J, Wang JY, Thongrod SC, Malamy JE, Birnbaum K, Benfey PN (2005) Transcriptional profile of the arabidopsis root quiescent center. Plant Cell 17(7):1908–1925. https://doi.org/10.1105/tpc.105.031724

Nesi N, Debeaujon I, Jond C, Pelletier G, Caboche M, Lepiniec L (2000) The TT8 gene encodes a basic helix-loop-helix domain protein required for expression of DFR and BAN genes in Arabidopsis siliques. Plant Cell 12(10):1863–1878. https://doi.org/10.1105/tpc.12.10.1863

Nuruzzaman M, Sharoni AM, Kikuchi S (2013) Roles of NAC transcription factors in the regulation of biotic and abiotic stress responses in plants. Front Microbiol 4:248. https://doi.org/10.3389/fmicb.2013.00248

Pandey SP, Somssich IE (2009) The role of WRKY transcription factors in plant immunity. Plant Physiol 150:1648. https://doi.org/10.1104/pp.109.138990

Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680. https://doi.org/10.1038/nrg2641

Perez-Rodriguez M, Jaffe FW, Butelli E, Glover BJ, Martin C (2005) Development of three different cell types is associated with the activity of a specific MYB transcription factor in the ventral petal of Antirrhinum majus flowers. Development 132(2):359–370. https://doi.org/10.1242/dev.01584

Pinyopich A, Ditta GS, Savidge B, Liljegren SJ, Baumann E, Wisman E, Yanofsky MF (2003) Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature 424(6944):85–88. https://doi.org/10.1038/nature01741

Portereiko MF, Lloyd A, Steffen JG, Punwani JA, Otsuga D, Drews GN (2006) AGL80 is required for central cell and endosperm development in Arabidopsis. Plant Cell 18(8):1862–1872. https://doi.org/10.1105/tpc.106.040824

Prasad K, Zhang X, Tobón E, Ambrose BA (2010) The Arabidopsis B-sister MADS-box protein, GORDITA, represses fruit growth and contributes to integument development. Plant J 62(2):203–214. https://doi.org/10.1111/j.1365-313X.2010.04139.x

Rajani S, Sundaresan V (2001) The Arabidopsis myc/bHLH gene alcatraz enables cell separation in fruit dehiscence. Curr Biol 11(24):1914–1922. https://doi.org/10.1016/S0960-9822(01)00593-0

Ratcliffe OJ, Nadzan GC, Reuber TL, Riechmann JL (2001) Regulation of flowering in Arabidopsis by an FLC homologue. Plant Physiol 126(1):122–132. https://doi.org/10.1104/pp.126.1.122

Ratcliffe OJ, Kumimoto RW, Wong BJ, Riechmann JL (2003) Analysis of the Arabidopsis MADS AFFECTING FLOWERING gene family: MAF2 prevents vernalization by short periods of cold. Plant Cell 15(5):1159–1169. https://doi.org/10.1105/tpc.009506

Riechmann JL, Heard J, Martin G, Reuber L, Jiang CZ, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu GL (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science 290:2105. https://doi.org/10.1126/science.290.5499.2105

Rushton PJ, Somssich IE, Ringler P, Shen QJ (2010) WRKY transcription factors. Trends Plant Sci 15:247. https://doi.org/10.1016/j.tplants.2010.02.006

Sakuma Y, Liu Q, Dubouzet JG, Abe H, Yamaguchi-Shinozaki K, Shinozaki K (2002) DNA-binding specificity of the ERF/AP2 domain of Arabidopsis DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. Biochem Biophys Res Commun 290:998. https://doi.org/10.1006/bbrc.2001.6299

Schönrock N, Bouveret R, Leroy O, Borghi L, Köhler C, Gruissem W, Hennig L (2006) Polycomb-group proteins repress the floral activator AGL19 in the FLC-independent vernalization pathway. Genes Dev 20(12):1667–1678. https://doi.org/10.1101/gad.377206

Serna L, Martin C (2006) Trichomes: different regulatory networks lead to convergent structures. Trends Plant Sci 11(6):274–280. https://doi.org/10.1016/j.tplants.2006.04.008

Smaczniak C, Immink RGH, Angenent GC, Kaufmann K (2012) Developmental and evolutionary diversity of plant MADS-domain factors: Insights from recent studies. Development (Cambridge) 139(17):3081–3098. https://doi.org/10.1242/dev.074674

Song FL, Milliken ON, Pham H, Seyit R, Napoli R, Preston J, Koltunow AM, Parish RW (2009) The Arabidopsis MYB5 transcription factor regulates mucilage synthesis, seed coat development, and trichome morphogenesis. Plant Cell 21(1):72–89. https://doi.org/10.1105/tpc.108.063503

Stracke R, Werber M, Weisshaar B (2001) The R2R3-MYB gene family in Arabidopsis thaliana. Curr Opin Plant Biol 4:447. https://doi.org/10.1016/S1369-5266(00)00199-0

Sun S, Yu JP, Chen F, Zhao TJ, Fang XH, Li YQ, Sui SF (2008) TINY, a dehydration-responsive element (DRE)-binding protein-like transcription factor connecting the DRE- and ethylene-responsive element-mediated signaling pathways in Arabidopsis. J Biol Chem 283(10):6261–6271. https://doi.org/10.1074/jbc.M706800200

Sun X, Wang Y, Sui N (2018) Transcriptional regulation of bHLH during plant response to stress. Biochem Biophys Res Commun 503:397. https://doi.org/10.1016/j.bbrc.2018.07.123

Tapia-López R, García-Ponce B, Dubrovsky JG, Garay-Arroyo A, Pérez-Ruíz RV, Kim SH, Acevedo F, Pelaz S, Alvarez-Buylla ER (2008) An AGAMOUS-related MADS-box gene, XAL1 (AGL12), regulates root meristem cell proliferation and flowering transition in Arabidopsis. Plant Physiol 146(3):1182–1192. https://doi.org/10.1104/pp.107.108647

Toledo-Ortiz G, Huq E, Quail PH (2003) The Arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell 15:1749. https://doi.org/10.1105/tpc.013839

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X et al (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158:1431. https://doi.org/10.1016/j.cell.2014.08.009

Welsch R, Maass D, Voegel T, DellaPenna D, Beyer P (2007) Transcription factor RAP2.2 and its interacting partner SINAT2: stable elements in the carotenogenesis of Arabidopsis leaves. Plant Physiol 145(3):1073–1085. https://doi.org/10.1104/pp.107.104828

Woo HR, Kim JH, Kim J, Kim J, Lee U, Song IJ, Kim JH, Lee HY, Nam HG, Lim PO (2010) The RAV1 transcription factor positively regulates leaf senescence in Arabidopsis. J Exp Bot 61(14):3947–3957. https://doi.org/10.1093/jxb/erq206

Yang SW, Jang IC, Henriques R, Chua NH (2009a) Far-red elongated hypocotyl1 and FHY1-Like associate with the arabidopsis transcription factors LAF1 and HFR1 to transmit phytochrome a signals for inhibition of hypocotyl elongation CW. Plant Cell 21(5):1341–1359. https://doi.org/10.1105/tpc.109.067215

Yang S, Wang S, Liu X, Yu Y, Yue L, Wang X, Hao D (2009b) Four divergent Arabidopsis ethylene-responsive element-binding factor domains bind to a target DNA motif with a universal CG step core recognition and different flanking bases preference. FEBS J 276(23):7177–7186. https://doi.org/10.1111/j.1742-4658.2009.07428.x

Yoo SK, Lee JS, Ahn JH (2006) Overexpression of AGAMOUS-LIKE 28 (AGL28) promotes flowering by upregulating expression of floral promoters within the autonomous pathway. Biochem Biophys Res Commun 348(3):929–936. https://doi.org/10.1016/j.bbrc.2006.07.121

Yoo SY, Kim Y, Kim SY, Lee JS, Ahn JH (2007) Control of flowering time and cold response by a NAC-domain protein in Arabidopsis. PLoS One 2(7):e642. https://doi.org/10.1371/journal.pone.0000642

Yoo SK, Wu X, Lee JS, Ahn JH (2011) AGAMOUS-LIKE 6 is a floral promoter that negatively regulates the FLC/MAF clade genes and positively regulates FT in Arabidopsis. Plant J 65(1):62–76. https://doi.org/10.1111/j.1365-313X.2010.04402.x

Zhang H, Forde BG (1998) An Arabidopsis MADS box gene that controls nutrient-induced changes in root architecture. Science 279(5349):407–409. https://doi.org/10.1126/science.279.5349.407

Zhang F, Gonzalez A, Zhao M, Payne CT, Lloyd A (2003) A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. Development 130(20):4859–4869. https://doi.org/10.1242/dev.00681

Zhao C, Avci U, Grant EH, Haigler CH, Beers EP (2008) XND1, a member of the NAC domain family in Arabidopsis thaliana, negatively regulates lignocellulose synthesis and programmed cell death in xylem. Plant J 53(3):425–436. https://doi.org/10.1111/j.1365-313X.2007.03350.x

Zhu Q, Zhang J, Gao X, Tong J, Xiao L, Li W, Zhang H (2010) The Arabidopsis AP2/ERF transcription factor RAP2.6 participates in ABA, salt and osmotic stress responses. Gene 457(1–2):1–12. https://doi.org/10.1016/j.gene.2010.02.011

Chapter 10
Microbiome and Big-Data Mining

Kang Ning

Abstract Microbiome samples are accumulating at a very fast speed, representing microbial communities from every niche (biome) of our body as well as the environment. The fast-growing amount of microbiome samples, as well as the diversified sources from where the samples are collected, have provided us with an unprecedented scene from where we could obtain a better understanding of the microbial evolution and ecology. While all of these represent profound biological patterns and regulation principles, the understanding of them is heavily dependent on data integration and big-data mining, including the data-driven microbiome marker identification, non-linear relationship mining, dynamic pattern discovery, regulation principle discovery, etc.

In this chapter, we first introduce several terminologies in microbiome research, followed by the introduction of microbiome big-data. Then we emphasize the microbiome databases, as well as mainstream microbiome data mining techniques. We have provided several microbiome applications to showcase the power of microbiome big-data integration and mining for knowledge and clinical applications. Finally, we have summarized the current status of microbiome big-data analysis, pointed out several bottlenecks, and illustrated prospects in this research area.

Keywords Microbiome · Big-data · Evolution · Ecology · Database · Data mining · Knowledge discovery · Clinical applications

K. Ning (✉)
Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-Imaging, Center of AI Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: ningkang@hust.edu.cn

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_10

Microbiome samples are accumulating at a very fast speed, representing microbial communities from every niche (biome) of our body as well as the environment (Mitchell et al. 2020; Integrative HMP (iHMP) Research Network Consortium 2019; Thompson et al. 2017; Sunagawa et al. 2015). The fast-growing amount of microbiome samples, as well as the diversified sources from where the samples are collected, have provided us with an unprecedented scene from where we could obtain a better understanding of the microbial evolution and ecology (Mitchell et al. 2020; Segata et al. 2013; Integrative Human Microbiome Project 2019). While all of these represent profound biological patterns and regulation principles, the understanding of them is heavily dependent on data integration and big-data mining (Knight et al. 2018), including the data-driven microbiome marker identification (Segata et al. 2011), non-linear relationship mining (Surana and Kasper 2017), dynamic pattern discovery (Halfvarson et al. 2017; Ren et al. 2017; Bashan et al. 2016; Backhed et al. 2015; Liu et al. 2019), regulation principle discovery (Han et al. 2020), etc.

In this chapter, we will first introduce several terminologies in microbiome research, followed by the introduction of microbiome big-data. Then we will emphasize the microbiome databases, as well as mainstream microbiome data mining techniques. We will provide several microbiome applications to showcase the power of microbiome big-data integration and mining for knowledge and clinical applications. Finally, we will summarize the current status of microbiome big-data analysis, point out several bottlenecks, and illustrate prospects in this research area.

10.1 Microbial Communities, Metagenome, and Microbiome

As a ubiquitous and important organism in nature, microorganisms usually coexist in the form of a "microbial community" (Thompson et al. 2017; Sunagawa et al. 2015; Segata et al. 2013; Integrative HMP (iHMP) Research Network Consortium 2014). A microbial community usually contains dozens to thousands of different microorganisms, these species cooperate with each other to adapt to the changes in the environment, and their life activities also have a long-term and profound impact on the environment (Thompson et al. 2017; Integrative HMP (iHMP) Research Network Consortium 2014). With the deepening of human understanding of microorganisms, the basic research of microbial community and its application in the fields of health and environment have become increasingly important (Integrative Human Microbiome Project 2019; Biteen et al. 2016). The main research objects of microbiome include all the genetic materials of microbial communities, related environmental parameters and metabolites, as well as their complex relationships and dynamic changes.

In the microbiome research area, several terms need to be explained clearly, including microbiota, metagenome, and microbiome (Whiteside et al. 2015). A microbial community is a mixture of microbial species living, adapting, and evolving in a certain environment. Metagenome refers to the total genetic materials in the

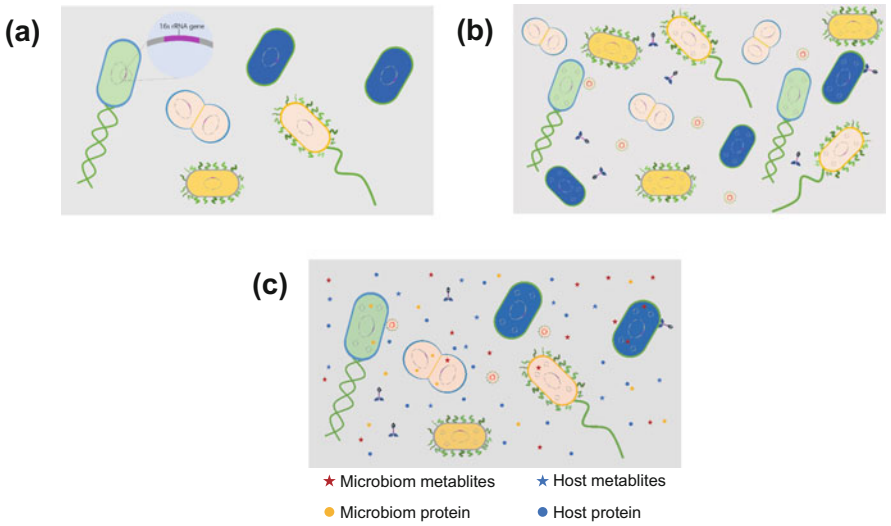

Fig. 10.1 The definitions of microbiota, metagenome, and microbiome. The same shape and color represent the same species, while different symbols represent different entities. (**a**) Microbiota: identification of all species in the microbial community using 16S rRNA sequencing. (**b**) Metagenome: all genetic materials in the microbial community. (**c**) Microbiome: all genetic materials, environmental factors, and metabolites in the microbial community

microbial community, while metagenome could be obtained by shotgun sequencing, many projects are still conducted by 16 s rRNA amplicon sequencing that could only quantitatively profile the species in the community. Microbiome refers to all genetic and non-genetic information contained in the microbial community, including metagenome, as well as all environmental factors and metabolites in the community. A brief illustration of the definitions and relationships of microbial communities, metagenome, and microbiome is provided in Fig. 10.1.

The microbiome research is mostly conducted by the omics approach (Mitchell et al. 2020; Segata et al. 2013). Firstly, samples are collected from niches, stored in a −20°C tube, before DNA extraction and amplification and sequencing. Then high-throughput sequencing is conducted, by means of 16S rRNA sequencing or metagenomic sequencing, and sequencing data are transferred for analysis (Knight et al. 2018).

10.1.1 The Differences Between 16S and Metagenomes

The sequencing principles: 16S rDNA contains nine hypervariable regions and ten conserved regions. A segment of hypervariable region sequence was amplified by PCR and sequenced. Metagenomic sequencing is similar to conventional DNA

library in that it randomly breaks microbial genomic DNA into small fragments and then inserts joints at both ends of the fragments for high-throughput sequencing.

Different fields of study: 16S rRNA sequencing mainly studies the species composition, the evolutionary relationship among species, and the diversity of communities. Besides, metagenomic sequencing can also be used for further research at the genetic and functional levels.

Different degree of species identification: Much of the 16S sequencing results are below species level, while metagenomic sequencing identifies microbes to species level and even to strain level.

The advantages and disadvantages of 16S rRNA and metagenomic sequencing methods for microbial community research have been summarized in (Knight et al. 2018), and we have provided key points in Table 10.1.

Table 10.1 Advantages and disadvantages of 16S rDNA and metagenomic sequencing methods for microbial community research

Method	Advantage	Disadvantage
Marker gene analysis	• Fast, simple, and inexpensive sample preparation and analysis • Closely related to genome content • Suitable for samples with low biomass • Could be compared with existing large public data sets	• Affected by amplification bias • Selection of primers and variable regions will amplify the deviation • Usually need prior knowledge of the microbial community • Resolution is usually only to genus • Need for proper negative control • Limited functional information
Metagenomic analysis	• The relative abundance of microbial functional genes can be directly inferred • For known organisms, microbial classification and phylogenetic identity can be achieved at the species and strain level • It is not assumed to understand the microbial community • No biases associated with PCR • The in situ growth rate of target organisms with sequenced genomes can be estimated • It is possible to assemble a population-average microbial genome • Can be used for new gene families	• Relatively expensive, laborious, and complicated sample preparation and analysis • The default pipeline usually does not annotate viruses and plasmids well • Due to assembly artifacts, population-average microbial genomes are often inaccurate

10.2 The Microbiome Research Is Heavily Dependent on Big-Data

As the number of microbiome samples easily exceeds tens of thousands in a medium-sized data collection (Mitchell et al. 2020), the efficiency and accuracy of sample comparison and search become a critical bottleneck (Knight et al. 2018), not to mention millions of samples from the rapidly diversified biomes from less than a hundred to more than three hundred in public databases (Fig. 10.2). The rapidly increasing number of samples from various niches on the planet has thus created a difficult huddle for knowledge discovery from these samples (Mitchell et al. 2020).

Microbiome research is heavily dependent on big-data, largely due to three reasons: (1) As traditional microbial research strategies could not identify the species in the community, current species identification and quantification is mostly done by sequencing techniques plus data analysis techniques. (2) As heterogeneous microbial community samples are collected from hundreds of different niches around the world, the comparison of these communities could only be performed using big-data mining techniques. (3) The mining of millions to trillions of functional genes from microbial communities is also a data-driven task nowadays.

Big-data technology and machine learning technology are very suitable for the organization, integration, and in-depth analysis of microbiome data (Li et al. 2019; Cheng et al. 2019; Tang et al. 2019; Microbiota meet big data 2014). First of all, microbiome data has all the 4 V characteristics of big data (Volume, Velocity, Variety, Veracity): large **Volume**, a large amount of data, including the amount of collection, storage, and calculation. The starting measurement unit of big data is at least *p* (1000 t), *e* (one million T), or *Z* (1 billion T). There are various types

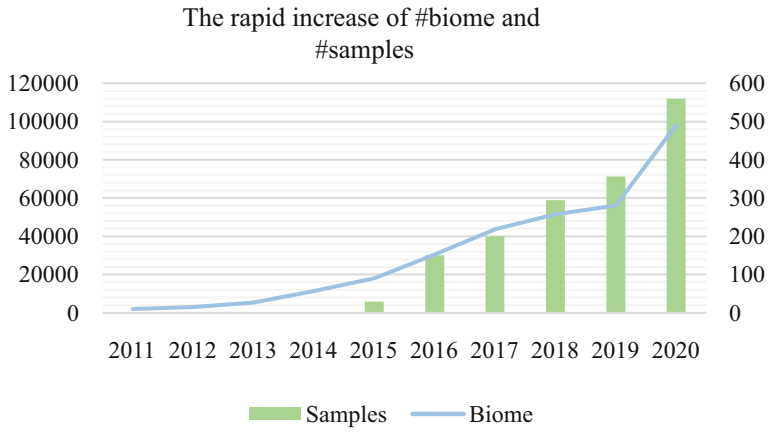

Fig. 10.2 The fast increasing number of microbiome samples, and the rapidly diversified biomes from where they are collected. Results are based on assessment of EBI MGnify database from year 2011 to year 2020

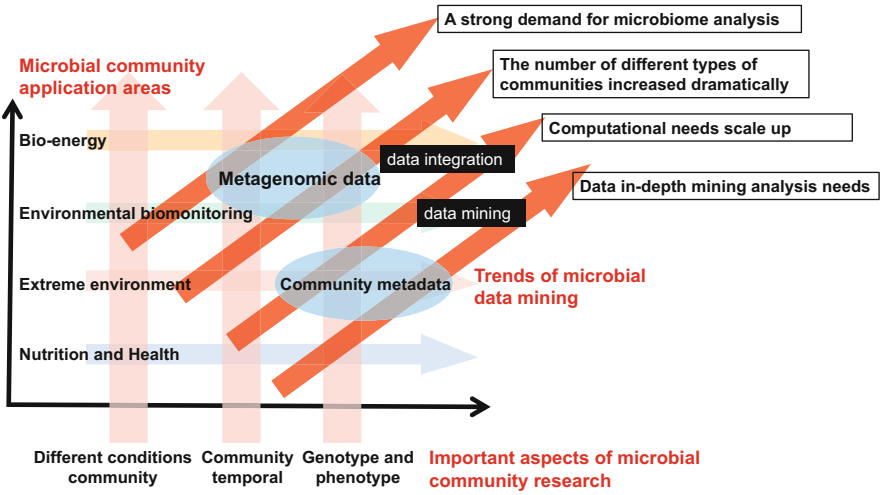

Fig. 10.3 The characteristics and urgent needs in multi-omics researchers

and sources (**Variety**). Including structured, semi-structured, and unstructured data, multi-types of data put forward higher requirements for data-processing ability. The **Value** density is low, and the data value density is relatively low. In other words, it is valuable to wash sand in waves. Information is massive, but the value density is low. How to mine the value of data through powerful machine algorithms is the most important problem to be solved in the era of big data. **Velocity**: this is a significant feature that big data is different from traditional data mining. Secondly, microbiome big data needs to be deeply mined: Data mining (DM) is an emerging interdisciplinary subject that gathers multiple disciplines. It is an extraordinary process, that is, the process of extracting unknown, implied, and potentially valuable information from huge data (Fig. 10.3).

10.3 Microbiome Data Integration and Databases

The development of microbiome research has profoundly boosted the data accumulation as well as the output of the researches. In the past 10 years, an exponential number of publications have been output (Fig. 10.4a), based on more than 100 TB per year of microbiome data accumulated (Fig. 10.4b).

Currently, there are already databases dedicated to microbiome researches (Table 10.2), including MG_RAST ((Meyer et al. 2008), http://metagenomics.anl.gov/), CAMERA ((Seshadri et al. 2007), http://camera.calit2.net/) as specialized databases, and NCBI SRA (http://www.ncbi.nlm.nih.gov/sra) as general databases. Among these databases, NCBI SRA (Kodama et al. 2012), MG-RAST (Meyer

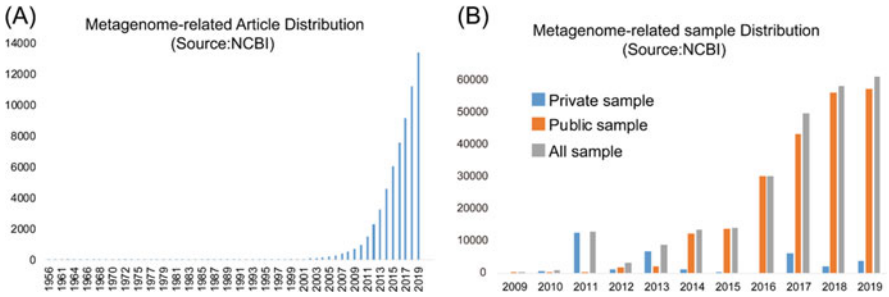

Fig. 10.4 In the past 10 or more years, (**a**) an exponential number of publications have been output, (**b**) based on more than 100 TB per year of microbiome data accumulated

Table 10.2 Commonly used microbiome databases (Zhang et al. 2017)

Database name	Description	Database website	References
NCBI SRA	General database that contains microbiome data of all kinds and formats	http://www.ncbi.nlm.nih.gov/sra	Kodama et al. (2012)
EBI MGnify	Specialized microbiome database with a comprehensive collection of samples, and with a unified analytical pipeline	www.ebi.ac.uk/metagenomics/	Mitchell et al. (2020)
MG-RAST	Specialized microbiome database with a unified analytical pipeline	Metagenomics.anl.gov	Paczian et al. (2019), Meyer et al. (2019)
IMG/M	Specialized microbiome database with a unified analytical pipeline	img.jgi.doe.gov	Markowitz et al. (2008)
Qiita	Specialized microbiome database with a unified analytical pipeline and comprehensive meta-data information	http://qiita.ucsd.edu/	Gonzalez et al. (2018)
CAMERA	Specialized microbiome database, data collection not comprehensive	http://camera.calit2.net/	Seshadri et al. (2007)

et al. 2008), and CAMERA2 (Seshadri et al. 2007) each has more than 10,000 microbiome projects, representing hundreds of thousands of samples and several TB of sequencing data.

However, the microbiome data in several major databases have not been well sorted out, whether in terms of the unification and integration of microbiome data format, or the matching environmental parameters (metadata). One of the key points is that the microbiome data has not been effectively classified and organized, resulting in a bottleneck for sample classification and comparison. Microbial community samples and relevant sequencing data are organized according to the biome ontology organization structure by hierarchical structures. For example: at the end of 2019, EBI MGnify contains sub-millions samples from 491 biomes (https://www.ebi.ac.uk/metagenomics/biomes) (Mitchell et al. 2020), in which the samples from human fecal have the exact biome position at "root > Host-associated > Human > Digestive system > Large intestine > Fecal." This ontology structure is very beneficial to the classification of samples. However, the hierarchical organization structure of the current ontology is not completely tree-like, but has the feature that an ontology belongs to the direct sub-ontology of multiple ontologies. For example, "Fecal" has more than five upper level ontology information. Therefore, the relevant living environment ontology of each microbiome data is likely to have multi-label. On the one hand, the multi-label nature of microbiome data is not conducive to the simple classification of samples, resulting in the bottleneck of sample classification and comparison. On the other hand, the multi-tag attribute of microbiome data conforms to the characteristics of big-data research, and better results are expected to be obtained when processed by machine learning or deep learning.

10.4 Mainstream Microbiome Data Mining Techniques

As regard to microbiome data mining tools, current methods could be categorized according to their purposes (Table 10.3):

1. **Identification of microbial species based on microbiome:** Based on the metagenome sequencing data, the species contained in the metagenome can be assigned to different taxonomic levels, such as phylum, class, order, family, genus, etc. At present, metagenome-based microbial species identification can be categorized into alignment-based and alignment-free sequence classification methods, both of which are based on the assumption that similar sequences originate from similar species. Sequence alignment identifies the species corresponding to the target genome sequence by comparing it with the existing database. Alignment-free sequence classification methods use the characteristics of the sequences themselves, such as GC content, codon usage frequency, etc., to classify them into the species corresponding to the most similar sequences. Typical examples of species identification methods include Megan (Huson et al. 2007), QIIME2 (Bolyen et al. 2019), etc. However, these methods are mostly limited to sequences of known classes and functions in databases (sequences in databases are mostly from model organisms or culturable microorganisms), so

the exact species of the majority of microorganisms in the microbial community remain largely unclear.

2. **Tools for microbial community structure decoding:** Tools for microbial community structure decoding and comparison include those for species composition analysis such as Phyloshop (Shah et al. 2011), Parallel-Meta (Su et al. 2012), MEGAN (Huson et al. 2007), etc., and those for microbial community comparison including UniFrac (Lozupone and Knight 2005) and Fast UniFrac (Hamady et al. 2010). However, these tools still have limitations: MEGAN (Huson et al. 2007) and STAMP (Parks and Beiko 2010) have provided an approach for microbial community sample comparison based on species composition, while such method is largely limited by the ignorance of evolutionary relationships among species (Hamady and Knight 2009). UniFrac (Lozupone and Knight 2005) and Fast UniFrac (Hamady et al. 2010) have taken phylogeny information into consideration, yet they could hardly handle thousands of samples due to large time cost. There is still a lack of efficient and accurate sample comparison and search methods, especially for model-based method.
3. **Microbial-based functional profiling and regulation model generation:** In terms of predicting the main functions of species, the current research is still in its infancy. Methods such as PICRUSt (Langille et al. 2013), based on 16S rRNA data, could analyze differences between samples by inferring the composition of functional genes in the samples. However, this prediction method cannot fully reflect the detailed functional composition and metabolic pathways of different species in a sample. Functional genes in microbial community analysis level, in view of the biosynthesis gene cluster (BGC) and antibiotic resistance gene cluster (ARG) gene functions such as group analysis, in addition to the typical antiSMASH (Medema et al. 2011) and NaPDoS (Ziemert et al. 2012) analysis platform and IMG-ABC (Hadjithomas et al. 2015), DoBISCUIT (Ichikawa et al. 2013), ClusterMine360 (Conway and Boddy 2013) database. Functional annotation and enrichment analysis of microbiome genes can deepen the understanding of microbial community functions and the analysis of key metabolic pathways and microbiome-host metabolic regulation mechanisms. However, the microbiome contains a large number of genes, and the functions of most genes are unknown.
4. **Microbial gene mining from metagenomics data:** At present the main database and the software including DoBISCUIT (Ichikawa et al. 2013) system (http://www.bio.nite.go.jp/pks/) based on manual selection of data, and the databases designed for specific types of metabolites, such as ClusterMine360 (Conway and Boddy 2013) database system, NaPDoS (Ziemert et al. 2012) analysis system (http://napdos.ucsd.edu/) for secondary metabolism genes, COBRA (Becker et al. 2007) for intestinal flora metabolism modeling analysis system, as well as antiSMASH (Medema et al. 2011) biosynthesis gene cluster (BGC) analysis system, etc. Relevant methods, however, largely depend on the reference sequence, known species in the microbial community species reference sequence under the condition of the lack of its completeness is not very ideal. The genes

Table 10.3 Representative analytical platforms for microbiome researches

Name	Description	Website	References
QIIME	Most frequently used package, with comprehensive sets of tools, discontinued in 2018	http://qiime.org	Caporaso et al. (2010)
QIIME 2	QIIME version 2, with a full set of command line and visualized interfaces for interactive and reproducible microbiome analysis	https://qiime2.org	Bolyen et al. (2019)
USEARCH	Fast sequence search and clustering toolset	http://www.drive5.com/usearch	Edgar (2010)
VSEARCH	Fast sequence search and clustering toolset specifically designed for metagenomics sequence analysis	https://github.com/torognes/vsearch	Rognes et al. (2016)
Trimmomatic	Quality control tool for metagenome sequences	http://www.usadellab.org/cms/index.php?page=trimmomatic	Bolger et al. (2014)
Bowtie2	Sequencing data alignment tool	http://bowtie-bio.sourceforge.net/bowtie2	Langmead and Salzberg (2012)
MetaPhlAn2	Microbial community structure profiling tool for	http://segatalab.cibio.unitn.it/tools/metaphlan2	Truong et al. (2015)
Kraken2	k-mer based metagenomic sequence classification tool	https://ccb.jhu.edu/software/kraken2	Wood and Salzberg (2014)
HUMAnN2	Species-level functional profiling for microbial communities	http://www.huttenhower.org/humann2	Franzosa et al. (2018)
MEGAN	Interactive microbial community profiling tool	https://www.wsi.uni-tuebingen.de/lehrstuehle/algorithms-in-bioinformatics/software/megan6/	Huson et al. (2007)
MEGAHIT	Ultrafast metagenome assembly tool	https://github.com/voutcn/megahit	Li et al. (2015)
metaSPAdes	High-quality metagenome assembly tool	http://cab.spbu.ru/software/spades	Nurk et al. (2017)
MetaQUAST	Tool for metagenome sequence assembly quality evaluation	http://quast.sourceforge.net/metaquast	Mikheenko et al. (2016)
MetaGeneMark	Gene prediction from metagenomics sequence	http://exon.gatech.edu/GeneMark/meta_gmhmmp.cgi	Zhu et al. (2010)

Prokka	Fast prokaryotic genome annotation tool	http://www.vicbioinformatics.com/software.prokka.shtml	Seemann (2014)
CD-HIT	Generation of non-redundant gene set	http://weizhongli-lab.org/cd-hit	Fu et al. (2012)
Salmon	k-mer based gene quantification tool	https://combine-lab.github.io/salmon	Patro et al. (2017)
MetaWRAP	Meta tool for metagenomics sequence binning	https://github.com/bxlab/metaWRAP	Uritskiy et al. (2018)
DAS tool	Another tool for metagenomics sequence binning	https://github.com/cmks/DAS_Tool	Sieber et al. (2018)
MOCAT2	A metagenomic assembly, annotation, and profiling framework	https://mocat.embl.de/index.html	Kultima et al. (2016)
ConStrains	Sub-species identification tool for microbial communities	https://bitbucket.org/luo-chengwei/constrains/src	Luo et al. (2015)
MetaPhlAn	A high-resolution microbial community profiling tool based on metagenomics sequences	http://segatalab.cibio.unitn.it/tools/metaphlan/index.html	Truong et al. (2015)
PICRUSt	Functional profile prediction based on species composition profile of the microbial communities	http://picrust.github.io/picrust	Langille et al. (2013)
antiSMASH	Resource and analytical tool on secondary metabolite biosynthetic gene clusters	https://antismash-db.secondarymetabolites.org	Medema et al. (2011)
CARMA	Taxonomic classification of metagenomic sequences	https://wwww.cebitec.uni-bielefeld.de/webcarma.cebitec.uni-bielefeld.de/	Gerlach and Stoye (2011)
SOrt-ITEMS	Metagenomic sequence analysis tool	http://metagenomics.atc.tcs.com/binning/SOrt-ITEMS	Monzoorul Haque et al. (2009)

(continued)

Table 10.3 (continued)

Name	Description	Website	References
PHYLOSHOP	Microbial community profiling tool	https://omics.informatics.indiana.edu/mg/phyloshop/	Shah et al. (2011)
UniFrac	Tool for microbial community species composition comparison	http://bmf.colorado.edu/unifrac	Lozupone and Knight (2005), Hamady et al. (2010)
PhyloPythia	Accurate phylogenetic classification analysis tool		McHardy et al. (2007)
MG-RAST	Analytical platform for microbiome research	https://www.mg-rast.org/	Meyer et al. (2008)
CAMERA	Tool set for metagenomics data analysis	http://camera.calit2.net/	Seshadri et al. (2007)
IBDsite	An integrated package for metagenomic sequence analysis for IBD	https://www.itb.cnr.it/ibd/	Merelli et al. (2012)

around the "environment–microbial community–metabolism" chain are largely unannotated, leaving large room for improvement.

5. **Microbiome data analysis platform:** There are currently several analytics platforms that cover the main steps of microbiome data analysis, such as QIIME (Caporaso et al. 2010), MG-RAST (Glass et al. 2010; Keegan et al. 2016), Camera (Seshadri et al. 2007), and EBI Metagenomics (now known as EBI Mgnify) (Mitchell et al. 2020). These sites often contain large datasets and data-processing platforms. At present, the biggest bottleneck in this regard is that the development of metagenomic data analysis platforms is far behind the rapid accumulation of metagenomic data. In particular, the integration analysis and deep mining of massive metagenomic data and other omics data are in urgent need.

10.5 Integration of Metagenome and Pan-Genome Towards Holistic Analysis of Microbial Communities

The microbiome data is mostly analyzed by the metagenome approach (Fig. 10.5). Metagenomics has been utilized for the studies of changes in community organization and microbial inhabitants, resulting in the discovery of a remarkable amount of genomic diversity and the characterization of new bacterial members (Integrative HMP (iHMP) Research Network Consortium 2014; Riesenfeld et al. 2004). A series of metagenome analysis tools, such as MEGAHIT (Li et al. 2015), MEGAN (Huson et al. 2007), and MetaPhlAn2 (Truong et al. 2015) have been proposed allowing for metagenomics assembly, taxonomy, and functional analysis. The analyses of microbiome composition and function in different sites of human body including skin, oral, and gut show great differences in the microbial structure (Koren et al. 2011; Costello et al. 2009). For example, the taxonomic representation of bacteria on the human skin includes *Staphylococcus*, *Micrococcus*, and *Corynebacterium* (Fredricks 2001; Grice et al. 2009), while the dominant microorganisms in oral are *Streptococci*, *Lactobacillus*, and *Fusobacterium* (Dewhirst et al. 2010; Teng et al. 2015). In addition, the main components of microorganisms in the human gut are *Bacteroides* and *Prevotella* (Costea et al. 2018; Wu et al. 2011). These microbes in human body have coevolved with their hosts, which is also related to human health and disease (Costello et al. 2009; Clemente et al. 2012). The composition of microbes in different hosts varies greatly, and there are dynamic changes under different environmental factors (Costello et al. 2009). For example, Sonnenburg et al. revealed a seasonal cycle of gut microbiota corresponding to the enrichment of functions of the Hadza hunter-gatherers, especially *Bacteroides*, varies with the season, especially between the dry season and the wet season (Smits et al. 2017). Such studies revealed the succession of microbial community that changes with season in human gut. In addition, studies of microbial communities in natural environments such as soil (Daniel 2004), deep-sea (Mason et al. 2014), and wastewater (Guo et al. 2017) have uncovered hundreds of microbes, new genes,

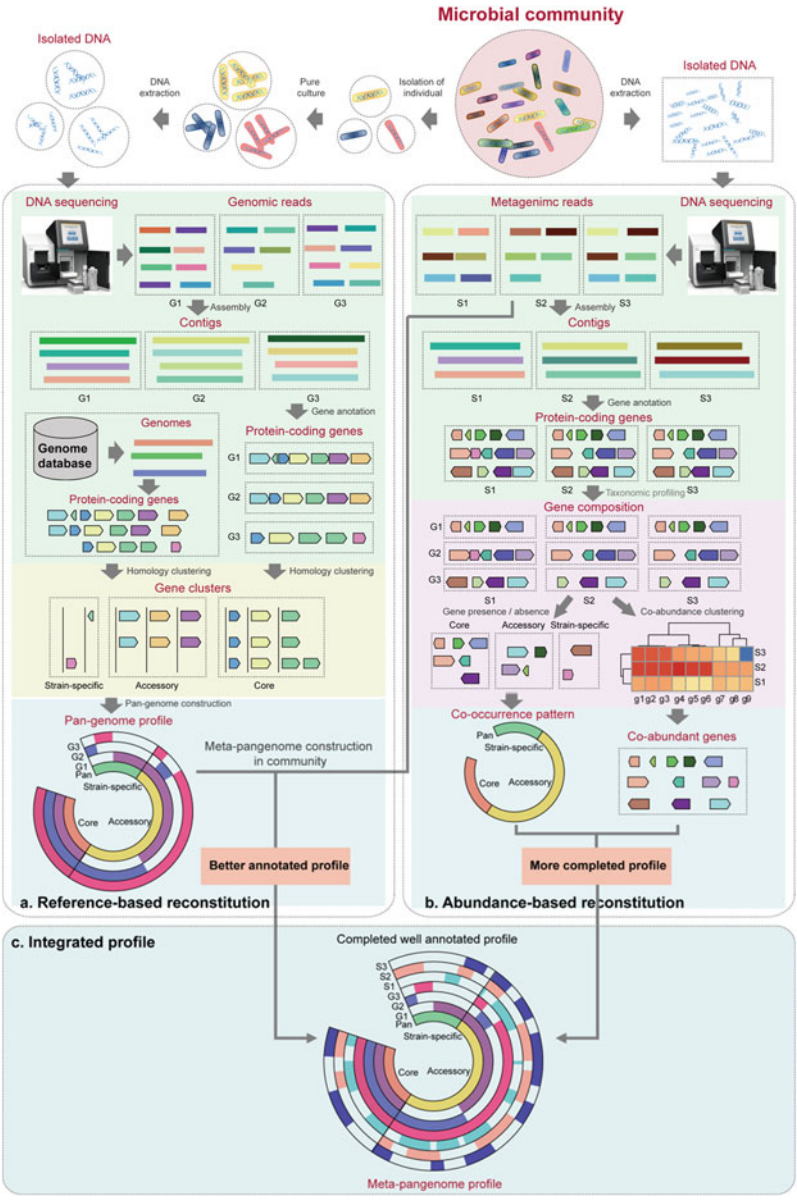

Fig. 10.5 Scheme of integrative pan-genome with metagenome studies on microbial community. (**a**) Using pan-genome of a set of genomes from isolates as a reference to recruit reads from metagenomes to quantify relative frequency of each gene sequence in community. (**b**) Binning co-abundant genes obtained from de novo assembly across metagenomic samples to reconstitute metagenomic species pan-genomes. Co-abundant with core or accessory genes of microbial species co-occurrence in samples and yield co-abundance. This figure was adapted from a previous published work [Integrating pan-genome with metagenome for microbial community profiling. Computational and Structural Biotechnology Journal, 2021, 19:1458–1466] with permission of authors

and uncharacterized metabolism, revealing an incredible microbial diversity and complexity.

10.6 Deep Learning Techniques for Microbiome Research

In recent years, more and more deep learning techniques have been developed for mining microbiome big-data (Li et al. 2019; Tang et al. 2019; Lan et al. 2018; Min et al. 2017; Wang and Gao 2019). These techniques essentially solved the functional gene mining, dynamic pattern discovery, and phenotype prediction problems.

1. **For sample comparison and search:** In microbial community source tracking, the traditional unsupervised learning method SourceTracker (Knights et al. 2011) and FEAST (Shenhav et al. 2019) could achieve very high accuracy when there are hundreds of samples and handful of biomes, while when the number of samples and biomes increase, the running time would increase very rapidly, preventing them from large-scale source tracking. This dilemma could be solved by deep learning solutions: by utilizing model-based methods such as neural network, both speed and accuracy could be achieved for the source tracking problem.
2. **For gene mining:** An example is ARG gene mining, for which traditional BLAST method could find the candidate ARG genes when they could match to those in the database. However, such an approach is limited to known ARG genes, and the search time could be short when faced with millions of candidates to be screened. Again, the deep learning approach has led to the model-based method that could mine novel ARG genes out of millions of candidates in an efficient manner.

All of these limitations have been calling for AI techniques that could discover more knowledge from microbiome dark matters. AI techniques are advantageous in generation of the models from a massive amount of samples, which are representative of the global profile of the context-dependent subjects (Kodama et al. 2012). AI techniques are therefore suitable for accurate and fast search when new samples (either a community, a gene, or a pattern) are searched against the models (Paczian et al. 2019; Markowitz et al. 2008; Daniel 2004). Therefore, AI techniques are especially suitable for microbiome dark matter mining, especially when facing the tradeoff between accuracy and efficiency.

The solutions for eliminating current methods' tradeoffs rely on deep learning approaches (Kodama et al. 2012; Paczian et al. 2019; Meyer et al. 2019; Markowitz et al. 2008; Gonzalez et al. 2018). First of all, model-based methods such as neural networks could be very fast for source tracking: once a rational model has been built, the source tracking could be very fast, and the source tracking accuracy could also be achieved, comparable with or even better than existing distance-based and unsupervised methods. The same approach is suitable for the gene mining problem.

For the spatial-temporal dynamic pattern mining, the deep learning method could also discover the intrinsic patterns out of the cross-section or longitudinal cohorts.

10.7 Representative Microbiome Applications

10.7.1 Case Study 1: Enterotype Analysis (Costea et al. 2018)

In 2011, three sequencing technologies (Illumina, 454, and Sanger) were used to sequence 16S rRNA genes in human fecal samples from three countries (Denmark, Spain, and the USA), and the result was that there were three enterotypes (Costea et al. 2018). The enterotypes were described as "a dense cluster of samples in a multidimensional space composed of communities" and were not affected by age, sex, cultural background, or geographical location. For each enterotype, an indicator/driver group was found at the center of the co-existing microbial network that was most profoundly associated with the enterotype. For example, enterotypes 1 can also be expressed as ET B, and Bacteroides is the best indicator group. Enterotype 2, which can also be expressed as ET P, is driven by Prevotella and its abundance is usually inversely proportional to the abundance of Bacteroides. Enterotype 3, which can also be expressed as ET F, is distinguished by the proportion of Firmicutes, among which the main group is Ruminococcus. All of the above analyses are based on the classification at the genus level, because the genus level can better reflect the ecological niche changes (Costea et al. 2018). Although some genera show functional heterogeneity, such as Streptococci, which contains both common symbiotic and lethal pathogens and groups that can be used for food fermentation, genera level analysis is generally reliable.

10.7.2 Case Study 2: Gene Mining (Qin et al. 2010)

10.7.2.1 Human Intestinal Microbiome Reference Gene Set

The authors describe the assembly and characterization of 3.3 million non-redundant microbial genes from fecal samples of 124 European individuals by Illumina-based metagenomic sequencing. This gene set is 150 times larger than the human gene complement, contains the vast majority of the (more common) microbial genes in the cohort, and probably includes the majority of the human gut microbial genes. These genes are shared to a large extent between individuals in this cohort. More than 99% of the genes were bacterial, suggesting that the entire cohort contained between 1000 and 1150 endemic bacterial species, with each individual containing at least 160 such species, and that they were also largely shared. The authors define and describe the minimum intestinal metagenome and the minimum

intestinal bacterial genome in terms of the functions of all individuals and most bacteria, respectively.

Most of the microbes that live in the gut have profound effects on human physiology and nutrition and are essential to human life. The content, diversity, and function of the gut microbiome are studied in order to understand and utilize the influence of gut microbiome on human health. Methods based on 16S ribosomal RNA gene (rRNA) sequences revealed that two families of bacteria, the Bacteroidaceae and the Antimicrobiaceae, make up more than 90% of the known phylogenetic categories and dominate the distal intestinal flora. Studies have also shown that there is great diversity in the gut microbiome between healthy individuals.

10.7.2.2 Metagenomic Sequencing of the Intestinal Microbiome

As part of the Metahit (Human Intestinal Genomics) project, the authors collected fecal samples from 124 healthy, overweight, and obese adult individuals and patients with inflammatory bowel disease (IBD) in Denmark and Spain. Total DNA was extracted from the fecal samples.

To generate an extensive catalogue of microbial genes from the human gut, the authors first assembled short Illumina readings into longer overlapping clusters, which could then be analyzed and annotated using standard methods. Using SoapDeNovo, the authors assembled all Illumina GA sequence data from scratch. Up to 42.7% of Illumina GA reads were assembled into a total of 6.58 million overlap groups, and nearly 35% of readings from any one sample could map to overlap groups from other samples, indicating the presence of a common sequence core.

To accomplish the overlapping group setup, the authors combined the unassembled reads from all 124 samples and repeated the de novo assembly process. Thus, about 400,000 overlapping groups with a length of 370 Mb and N50 939 bp are generated. Therefore, the total length of the author's final overlap group is 10.7 GB. Approximately 80% of the 576.7 Gb sequences of Illumina GA sequences were able to be compared with the overlap group at a 90% identity threshold to adapt to sequencing errors and strain variability in the gut, almost double the 42.7% of sequences. Soap de novo assembles them into overlapping clusters because the assembly uses more stringent criteria. This indicates that the author's overlap group represents the vast majority of Illumina sequences.

10.7.2.3 Genome Sets of the Human Intestinal Microbiome

To establish a non-redundant human gut microbiome genome, the authors first used the Metagene program to predict ORFs in overlapping populations and found 14,048,045 ORFs longer than 100 bp. They accounted for 86.7% of the overlap, comparable to the 86% found in fully sequenced genomes. Two-thirds of the ORFs appear to be incomplete, possibly due to the size of the author overlap group (N50 is

2.2 KB). Next, the authors remove the excess ORFs by pair comparison using very strict criteria that 95% conformance exceeds 90% of the shorter ORF length, which can be fused with direct homologues but can avoid dataset bloat due to possible sequencing errors.

The authors refer to the genes in the non-redundant set as "epidemic genes" because they are encoded on an overlapping group assembled from the richest read segments. The authors examined the number of prevalent genes found in all individuals, which is a function of the sequencing range and requires at least two gene calls to support reading. An estimate of coverage richness (ICE) based on incidence, determined by 100 people (the maximum number that can be accommodated by the Evaluations21 program), indicates that the authors' catalog captured 85.3% of the prevalence genes. Although this may be an underestimate, it still suggests that the catalogue contains the vast majority of the prevalent genes in this cohort.

Each person carries 536,112 ± 12,167 of the prevalent genes, suggesting that most of the 3.3 million gene pools must be shared. But most of the prevalent genes were found in only a few individuals: 2,375,655 were found in less than 20% of individuals, and 294,110 were found in at least 50% of individuals (these "common" genes, as the authors call them). These values depend on the sampling depth. The sequencing of MH0006 and MH0012 revealed more catalogue genes, which were present in low abundance. Still, even at regular sampling depths, each person still has 204,0566 3603 common genes, suggesting that about 38% of an individual's total gene pool is shared. Interestingly, patients with IBD carried, on average, 25% fewer genes than those without the disease, which is consistent with the observation that the former had less bacterial diversity than the latter.

10.7.3 Case Study 3: Plasticity of Intestinal Flora (Dynamic Pattern) (Liu et al. 2019)

First of all, at the macroscopic research level of the plasticity of the intestinal flora, the project team and the Capital Medical University have been monitoring the dynamics of the intestinal flora of the foreign aid medical team (volunteer team (VT)) for more than a year, aiming to study diet The influence of factor changes on the structure of human intestinal flora.

In this study, we recruited a team of 10 Chinese volunteers who set out from Beijing, stayed in Trinidad and Tobago (TAT) for 6 months and then returned to Beijing. A high-density longitudinal sampling strategy (average of 19 time points for VT members) was used to collect their stool samples (188 samples) and detailed dietary information. We divided the entire longitudinal study into six stages: when VT stays in TAT, T1 represents the pre-travel stage (20 samples), T2 (28 samples), T3 (60 samples), and T4 (21 samples) represents three time slots. After VT returned to Beijing, T5 (35 samples) and T6 (20 samples) sent two time slots, respectively.

At the same time, we also collected samples of Beijing healthy people (BJN, 57 samples), TAT healthy people (TTN, 28 samples), TAT patients (TTP, six samples), and TAT Chinese (TTC, eight samples) as samples control data set. Finally, we sequenced the V4 hypervariable region of the microbial 16S rRNA gene on 287 stool samples from 41 individuals and analyzed high-quality readings using QIIME (Caporaso et al., 2010).

We found that the microbial community in the intestine has two-way plasticity and elasticity during long-term stay and has a variety of dietary changes. First, BJN and TTN show different microbial community patterns (Fig. 10.6a). However, the microbial community of VT members changed from a microbial community similar to BJN to the TTN mode that accompanied them in TAT and returned to the original mode within 1 month after VT returned to Beijing (Fig. 10.6b–f). In addition, although we found that location and population have a great influence on the differentiation of samples (Fig. 10.6g, h), the dynamic changes of each member of VT show a specific trend (Fig. 10.6i, j), indicating that there may be the plasticity mode depending on the intestinal type among VT members. In addition, the relative abundance of Sclerotium and Bacteroides showed strong adaptability on the time axis and was negatively correlated on the time axis (Fig. 10.6k, l). Similarly, the relative abundance of Proteus and Actinomycetes also showed a plasticity pattern (Fig. 10.6m, n). By tracking and comparing at least 10% of the common operational taxonomic units (OTUs) shared by at least 10% of VT members, we found that Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria have unique time dynamics during the long-term stay of VT (Fig. 10.6o–r).

10.7.4 Case Study 4: Athletes' Gut Microbiota (Han et al. 2020)

The gut microbiome of athletes and sedentary individuals differs in diversity and in certain taxa; however, it is unclear to what extent the patterns of the gut microbiome differ between the two and whether athletes' potential can be effectively monitored against the microbiome.

This study recruited a total of 306 fecal samples from 19 Chinese professional female rowers and divided them into three groups according to their daily performance: adult elite athletes (AE), young elite athletes (YE), and young non-elite athletes (YN). The differences of intestinal microbiome in different groups were compared to determine the correlation between intestinal microbiome and diet, physical characteristics and sports performance (Fig. 10.7).

Firstly, the intestinal flora of elite athletes and young non-elite athletes were stratified to find that the intestinal flora of elite athletes and young non-elite athletes had different intestinal types. In terms of taxonomic structure and functional composition, it was found that SCFA-producing bacteria were dominant in the microbial community of elite athletes. Secondly, functional analysis showed that

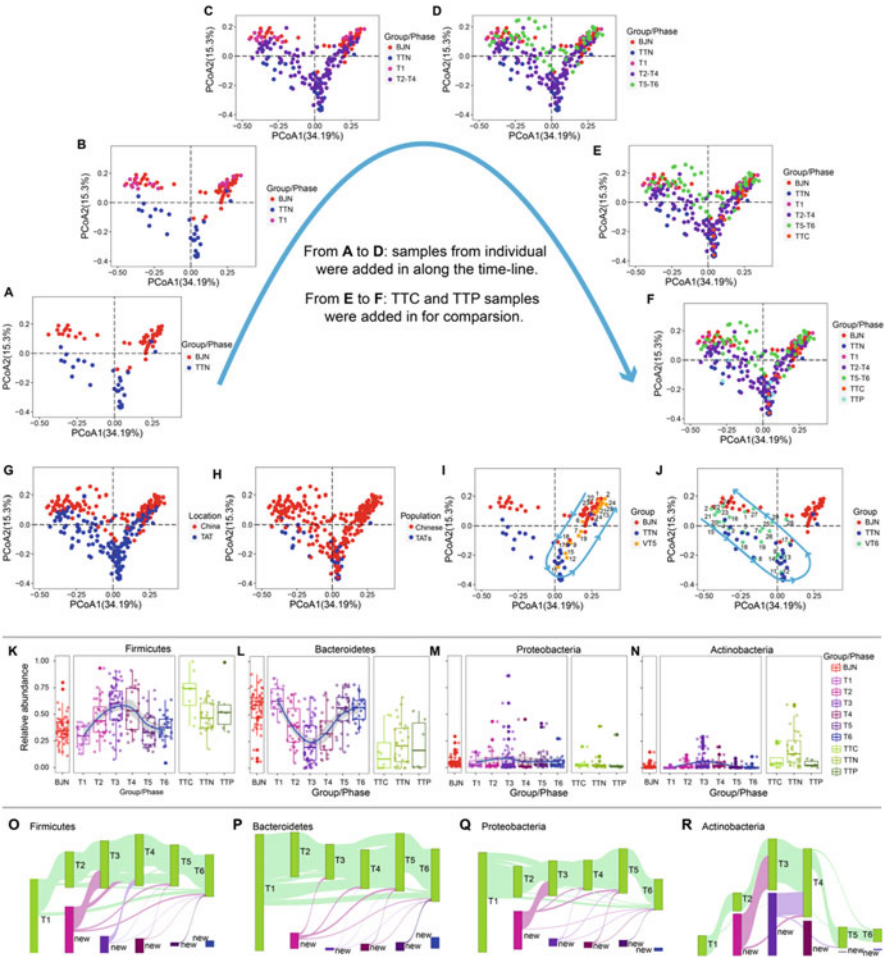

Fig. 10.6 Long-term human gut microbial community pattern and multiple dietary changes (Liu et al. 2019). (Reprinted with permission from authors of Liu et al. (2019))

ATP metabolism, multiple sugar transport systems, and carbohydrate metabolism were enriched in the microbial community of elite athletes. Furthermore, the construction of accurate classifiers based on a combination of taxonomy and functional biomarkers highlights the great potential of monitoring candidate elite athletes from a group of athletes. Finally, it was shown that intestinal flora is closely related to physical characteristics, dietary factors, and exercise-related characteristics. Importantly, the versatility of the athletes' microbiome, which may influence athlete performance by altering the gut microbiome, is associated with dietary factors (29%) and physical characteristics (21%). These findings highlight the complex

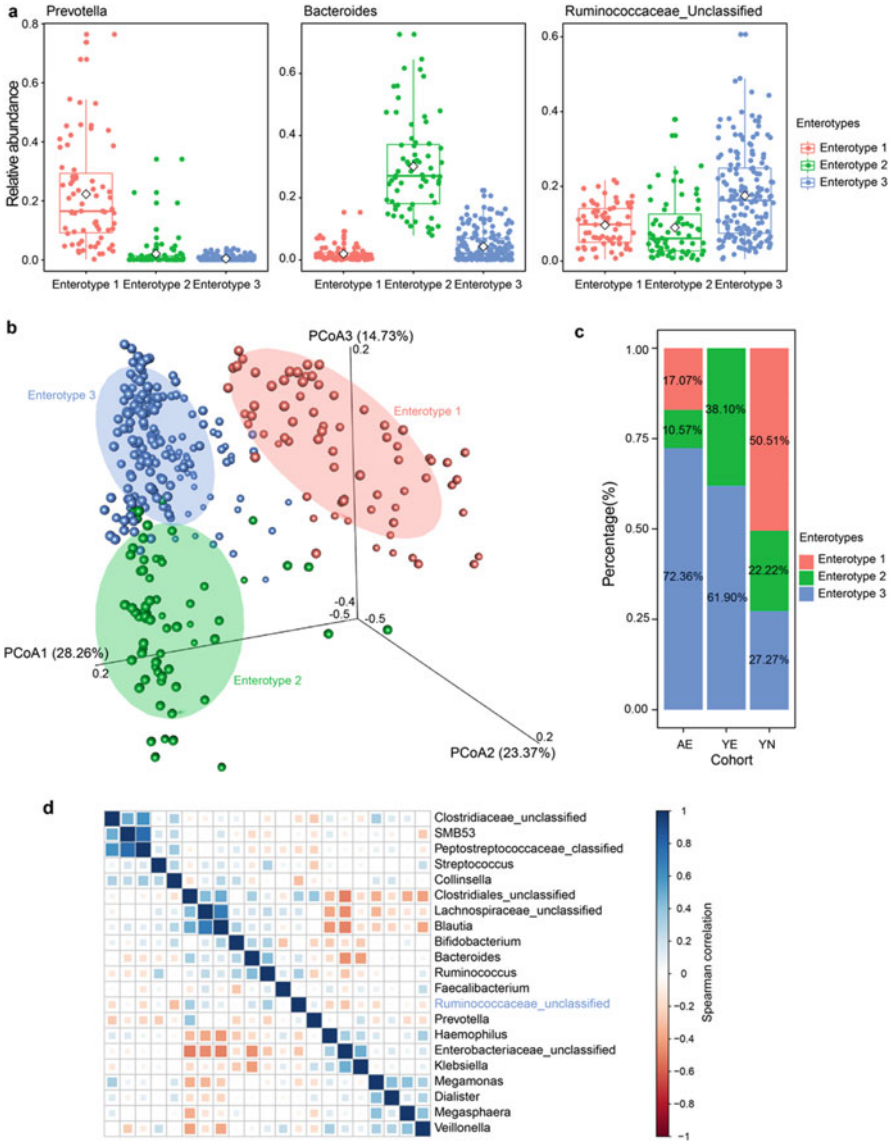

Fig. 10.7 Gut enterotypes in elite and youth non-elite athletes. A total of 306 samples are stratified into three enterotypes. The major contributor in the three enterotypes is *Prevotella, Bacteroides,* and Ruminococcaceae_unclassified, respectively. (**a**) Relative abundances of the top genera (*Prevotella, Bacteroides,* and Ruminococcaceae_unclassified) in each enterotype. (**b**) Three enterotypes were visualized by PCoA of Jensen-Shannon distance at the genus level. (**c**) The proportion of AE, YE, and YN samples distributed in three enterotypes. 72.3% AE, 61.9% YE, and 27.27% YN samples are found in enterotype 3. (**d**) Co-occurrence patterns among the dominant genera (average relative abundance >0.01%) across the samples from enterotype 3, as determined by the Spearman correlation analysis. (Reprinted with permission from authors of Han et al. (2020))

interplay of gut flora, dietary factors, and athletes' physical characteristics and performance, with gut flora as a key factor (Han et al. 2020).

10.8 Microbiome Research: Current Status, Bottlenecks, and Prospects

Today, microbiome research is, from many facets, a data-driven science. Firstly, the sequencing techniques have advanced quickly, thus enabling the fast and batch profiling of millions of microbial community samples. Secondly, data mining techniques have also advanced quickly, thus enabling the batch discovery of functional genes, dynamic patterns, as well as prediction of phenotype with high accuracy and fidelity. Thirdly, although data-driven, many discoveries are later verified by we-lab experiments, such as several probiotics (Whiteside et al. 2015; Routy et al. 2018), verified the power and validity of these data-driven approaches.

However, several bottlenecks remain for the microbiome big-data mining researches. One of the most critical bottlenecks is the big-data integration bottleneck (Integrative Human Microbiome Project 2019), and another is the lack of AI techniques for deep mining of important species, functional genes, and community dynamic patterns from a large amount of microbiome data (Microbiota meet big data 2014).

Despite these bottlenecks, microbiome researches are on the sharp rise, and many problems are on the edge of solution, while many more new frontiers are on the horizon. It is foreseeable that with several millions of samples from thousands of niches that have been collected, sequenced, and analyzed, a much better understanding of the microbial community ecology and evolution patterns would be discovered, together with hundreds of clinical or environmental applications made possible.

10.8.1 Microbiome Research as Part of a Multi-Omics Exploration

The multi-omics studies will continue to grow, in at least two directions: first, from multi-omics for single organisms or single species, to single-cell level omics studies, as well as to population and community level studies; second, the tight integration of multi-omics with data science as well as with clinical applications.

From the aspect of expanding the scope of multi-omics for single organisms or single species, single-cell level omics studies, as well as to population and community level studies, we have already seen rapid progress, largely due to the sequencing technical advances. From the aspect of integration of multi-omics with data science as well as with clinical applications, there are very hard challenges

still lying ahead. For example, it remains to be an open problem to determine the concordance of multi-omics along the timeline.

10.9 Summary

Taken together, it has become clear that microbiome research, which represents a rapidly growing omics research area, has already ensured enough high-quality data, as well as enabled data mining techniques, for large-scale microbiome data mining towards an in-depth understanding of microbial communities. The microbial community niches, species, functional genes and their dynamics, have constituted the microbial dark matter, which has been emerged as a grand challenge for microbiome research. The fast development of microbiome data mining would certainly boost the discovery of much more resources and regulation patterns out of these dark matters. And the integration of microbiome and other omics data would lead to a more complete picture of the dynamic patterns as well as regulation principles in the microbiome world.

References

Backhed F et al (2015) Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 17(6):852

Bashan A et al (2016) Universality of human microbial dynamics. Nature 534(7606):259

Becker SA et al (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. Nat Protoc 2(3):727–738

Biteen JS et al (2016) Tools for the microbiome: nano and beyond. ACS Nano 10(1):6–37

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120

Bolyen E et al (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37(8):852–857

Caporaso JG et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335–336

Cheng M, Cao L, Ning K (2019) Microbiome big-data mining and applications using single-cell technologies and metagenomics approaches toward precision medicine. Front Genet 10:972

Clemente JC et al (2012) The impact of the gut microbiota on human health: an integrative view. Cell 148(6):1258–1270

Conway KR, Boddy CN (2013) ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res 41(Database issue):D402–D407

Costea PI et al (2018) Enterotypes in the landscape of gut microbial community composition. Nat Microbiol 3(1):8–16

Costello EK et al (2009) Bacterial community variation in human body habitats across space and time. Science 326(5960):1694–1697

Daniel R (2004) The soil metagenome – a rich resource for the discovery of novel natural products. Curr Opin Biotechnol 15(3):199–204

Dewhirst FE et al (2010) The human oral microbiome. J Bacteriol 192(19):5002–5017

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460–2461
Franzosa EA et al (2018) Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 15(11):962–968
Fredricks DN (2001) Microbial ecology of human skin in health and disease. J Investig Dermatol Symp Proc 6(3):167–169
Fu L et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152
Gerlach W, Stoye J (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Res 39(14):e91
Glass EM et al (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb Protoc 2010(1):pdb.prot5368
Gonzalez A et al (2018) Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 15(10):796–798
Grice EA et al (2009) Topographical and temporal diversity of the human skin microbiome. Science 324(5931):1190–1192
Guo J et al (2017) Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements. Water Res 123:468–478
Hadjithomas M et al (2015) IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. MBio 6(4):e00932
Halfvarson J et al (2017) Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol 2:17004
Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res 19(7):1141–1152
Hamady M, Lozupone C, Knight R (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J 4(1):17–27
Han M et al (2020) Stratification of athletes' gut microbiota: the multifaceted hubs associated with dietary factors, physical characteristics and performance. Gut Microbes 12(1):1–18
Huson DH et al (2007) MEGAN analysis of metagenomic data. Genome Res 17(3):377–386
Ichikawa N et al (2013) DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. Nucleic Acids Res 41(Database issue):D408–D414
Integrative HMP (iHMP) Research Network Consortium (2014) The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe 16(3):276–289
Integrative HMP (iHMP) Research Network Consortium (2019) The Integrative Human Microbiome Project. Nature 569(7758):641–648
(2019) After the Integrative Human Microbiome Project, what's next for the microbiome community? Nature 569(7758):599
Keegan KP, Glass EM, Meyer F (2016) MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol Biol 1399:207–233
Knight R et al (2018) Best practices for analysing microbiomes. Nat Rev Microbiol 16(7):410–422
Knights D et al (2011) Bayesian community-wide culture-independent microbial source tracking. Nat Methods 8(9):761–763
Kodama Y et al (2012) The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res 40(Database issue):D54–D56
Koren O et al (2011) Human oral, gut, and plaque microbiota in patients with atherosclerosis. Proc Natl Acad Sci U S A 108(suppl 1):4592–4598
Kultima JR et al (2016) MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics 32(16):2520–2523
Lan K et al (2018) A survey of data mining and deep learning in bioinformatics. J Med Syst 42(8):139
Langille MG et al (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31(9):814–821

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359
Li D et al (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31(10):1674–1676
Li Y et al (2019) Deep learning in bioinformatics: introduction, application, and perspective in the big data era. Methods 166:4–21
Liu H et al (2019) Resilience of human gut microbial communities for the long stay with multiple dietary shifts. Gut 68(12):2254–2255
Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71(12):8228–8235
Luo C et al (2015) ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol 33(10):1045–1052
Markowitz VM et al (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36(Database issue):D534–D538
Mason OU et al (2014) Metagenomics reveals sediment microbial community response to deepwater horizon oil spill. ISME J 8(7):1464–1475
McHardy AC et al (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 4(1):63–72
Medema MH et al (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39(Web Server issue):W339–W346
Merelli I, Viti F, Milanesi L (2012) IBDsite: a galaxy-interacting, integrative database for supporting inflammatory bowel disease high throughput data analysis. BMC Bioinformatics 13(suppl 14):S5
Meyer F et al (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386
Meyer F et al (2019) MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. Brief Bioinform 20(4):1151–1159
(2014) Microbiota meet big data. Nat Chem Biol 10(8):605
Mikheenko A, Saveliev V, Gurevich A (2016) MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32(7):1088–1090
Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18(5):851–869
Mitchell AL et al (2020) MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res 48(D1):D570–D578
Monzoorul Haque M et al (2009) SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics 25(14):1722–1730
Nurk S et al (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Res 27(5):824–834
Paczian T et al (2019) The MG-RAST API explorer: an on-ramp for RESTful query composition. BMC Bioinformatics 20(1):561
Parks DH, Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26(6):715–721
Patro R et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14(4):417–419
Qin J et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59–65
Ren T et al (2017) Seasonal, spatial, and maternal effects on gut microbiome in wild red squirrels. Microbiome 5(1):163
Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 38:525–552
Rognes T et al (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584
Routy B et al (2018) Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. Science 359(6371):91–97

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14):2068–2069
Segata N et al (2011) Metagenomic biomarker discovery and explanation. Genome Biol 12(6):R60
Segata N et al (2013) Computational meta'omics for microbial community studies. Mol Syst Biol 9:666
Seshadri R et al (2007) CAMERA: a community resource for metagenomics. PLoS Biol 5(3):e75
Shah N et al (2011) Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. Pac Symp Biocomput:165–176
Shenhav L et al (2019) FEAST: fast expectation-maximization for microbial source tracking. Nat Methods 16(7):627–632
Sieber CMK et al (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 3(7):836–843
Smits SA et al (2017) Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357(6353):802–806
Su X, Xu J, Ning K (2012) Parallel-META: efficient metagenomic data analysis based on high-performance computation. BMC Syst Biol 6(Suppl 1):S16
Sunagawa S et al (2015) Ocean plankton. Structure and function of the global ocean microbiome. Science 348(6237):1261359
Surana NK, Kasper DL (2017) Moving beyond microbiome-wide associations to causal microbe identification. Nature 552(7684):244–247
Tang B et al (2019) Recent advances of deep learning in bioinformatics and computational biology. Front Genet 10:214
Teng F et al (2015) Prediction of early childhood caries via spatial-temporal variations of oral microbiota. Cell Host Microbe 18(3):296–306
Thompson LR et al (2017) A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551(7681):457–463
Truong DT et al (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12(10):902–903
Uritskiy GV, DiRuggiero J, Taylor J (2018) MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6(1):158
Wang W, Gao X (2019) Deep learning in bioinformatics. Methods 166:1–3
Whiteside SA et al (2015) The microbiome of the urinary tract–a role beyond infection. Nat Rev Urol 12(2):81–90
Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15(3):R46
Wu GD et al (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334(6052):105–108
Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38(12):e132
Ziemert N et al (2012) The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One 7(3):e34064
Zhang G et al (2017) Development of Comprehensive Microbiome Big Data Warehouse/Center for Long-term Scientific Impact[J]. Bulletin of Chinese Academy of Sciences 32(3):280–289

Chapter 11
Data Integration Applications in Medical Information Systems

Marcel Friedrichs

Abstract Medical information systems play a vital role in the everyday successful treatment of patients in hospitals, general practitioners' offices, and beyond. From storing patient information in electronic health records to the recommendation of treatment options, and the warning on wrong prescriptions or dosages, the information systems provide a multitude of different features. These can be utilized in prospective, direct, and retrospective patient care. One especially important task is the prevention of drug interactions and their potential adverse drug reactions in polypharmacy patients. All of these tasks require a solid data basis and data integration processes to provide the latest recommendations and information to healthcare professionals. Where historically, single large databases such as ABDAmed on the German market provided all required information, newer systems use a multitude of different data sources of high quality. This chapter analyzes different examples of medical information systems, the underlying data integration, and how a solid integration workflow can elevate the potential of old and new healthcare information. The examples range from drug therapy safety systems using official healthcare database, over potentially inadequate medication systems, to molecular biological analysis tools. Finally, the chapter outlines an approach how new data integration efforts may bring all of these systems together for the prospect of patient treatment in a personalized manner.

11.1 Introduction

As whole economic sectors adopt new digital solutions under terms such as "industry 4.0" and new technological paradigms like IoT (internet of things), the healthcare sector is changing as well. Medical imaging was one of the first key areas to adopt digital solutions for storing, processing, and distribution of patient data

M. Friedrichs (✉)
Faculty of Technology, Bioinformatics/Medical Informatics Department, Bielefeld University, Bielefeld, Germany
e-mail: mfriedrichs@techfak.uni-bielefeld.de

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_11

such as MRI (magnetic resonance imaging) or PET (positron emission tomography) scans. The "picture archiving and communication system" (PACS) replaced the need for physically stored images and allowed healthcare professionals remote access to all present or archive scans of their patients (Duerinckx and Pisa, 1982). This greatly reduced cost for long-term storage and time for images to be transferred from one station to another or in the worst case from an external storage facility to the hospital. Other medical areas followed, from medical samples being automatically processed and analyzed by laboratory information systems (LIS), electronic health records (EHR), to decision support systems ensuring drug therapy safety.

For the German healthcare sector, all of these individual efforts now culminate in the construction of the teleinformatics infrastructure ("Telematikinfrastruktur"). This infrastructure will provide a complete, fast, and safe exchange of information between patients and healthcare professionals. This digitization may result in faster adoption of new results and tools from research projects further improving drug therapy safety and reducing adverse drug events (ADR).

All of these systems require data of some sort. Be it patient information for electronic health records or medical knowledge databases for decision support systems. Therefore, this chapter describes the need for drug therapy safety tools and the data integration efforts of several medical information systems. The chapter concludes with an outlook on combining all of these efforts to further improve drug therapy safety.

11.2 Drug Therapy Safety

The safety and appropriateness of pharmacotherapy is an important topic in the field of medicine and under extensive research. Where younger patients rarely require more than one medication, the number of drugs taken in the elderly increases. In aging populations, multimorbidity is increasing with a corresponding increase in polypharmacy, which in turn is the prime risk factor for inappropriate prescribing. The evidence is well-known by several studies that the use of certain groups of medications in elderly and vulnerable patients is associated with falls (Fiss et al., 2010) and an increase in mortality (Chrischilles et al., 2009). With an increasingly older population in Germany, prognosticated to be 22% of the population aged 65 and older in 2022 (German Federal Statistical Office, 2021), the prevalence of multimorbidity is growing. Furthermore, inappropriate medications can impair cognitive properties (Boustani et al., 2010), reduce the quality of life, and cause additional costs for the healthcare system (Fick, 2001). The major challenges in gerontopharmacology are both over-treatment and undertreatment associated with polypharmacy.

Approximately 2.7 million BARMER insured people in Germany are suffering from five or more chronic diseases (Grandt et al., 2018). In addition, every fourth BARMER insured person aged 65 and older received at least one potentially inadequate medication (PIM) based on the PRISCUS list (Grandt et al., 2018; Holt

et al., 2010). As more PIM lists have been published and new ones are emerging, like FORTA (Pazan et al., 2019) or EU(7)-PIM (Renom-Guiteras et al., 2015) this result would likely be even higher.

Further increasing the complexity of the prescription process is the growing number of available medications. The German Federal Institute for Drugs and Medical Devices (BfArM) reported for April 2021 approximately 103,975 medications on the German market. From these medications, 34,911 are freely available and 52,478 without a prescription (BfArM, 2021). Furthermore, polypharmacy increases the risk of drug-related problems such as medication errors and adverse drug reactions. Without the help of medical decision support systems, healthcare professionals are likely unable to review all potential issues for every patient case.

The increased interest in molecular analyses, not only by researchers but also by healthcare professionals, may finally lead the way toward personalized medicine. The adoption of sequencing technologies and others in hospitals is positive, but staff needs to be properly trained and new safety measures implemented to prevent errors in data interpretation. Subsequent changes in a patient's drug therapy on an individual molecular basis need to be thoroughly tested and regulated to increase and not reduce drug therapy safety.

11.3 Medical Information System Examples

Following, different medical information systems are analyzed for their specific use-cases and needs in terms of data integration.

11.3.1 KALIS

KALIS is a web-based information system for checking drug-related problems in the medication process (Alban et al., 2017). It is comprised of multiple components, each tailored to a specific use-case. The main component is the pharmacological risk check. Here, medications and compounds can be checked with indications, side effects, and intolerances for interactions and other risks. Other modules help in finding potentially inappropriate medication for elderly patients, pharmacogenetic interactions in light of CYP enzyme defects, and guideline compliant analyses for hypertension.

Figure 11.1 visualizes the data integration architecture of KALIS. It is divided into integration, conception, and merging. The resulting KALIS-DWH has a uniform data structure and provides comprehensive information for the aforementioned

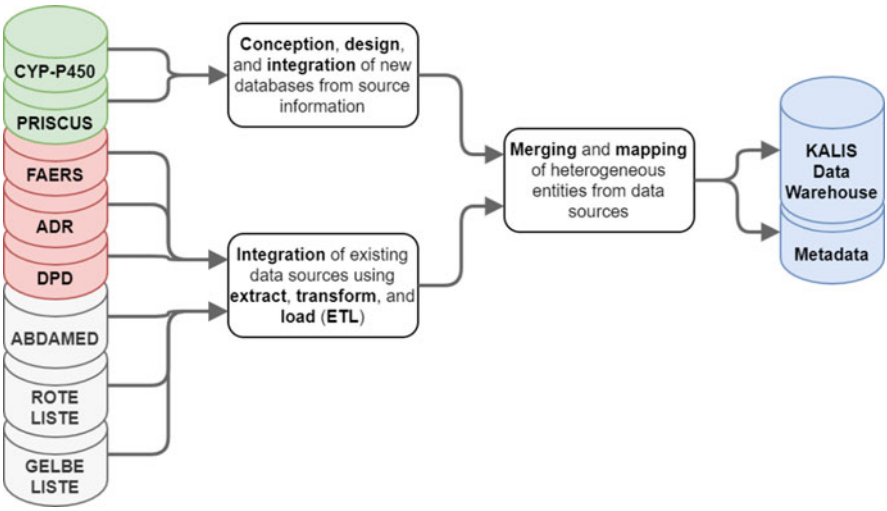

Fig. 11.1 Overview of the KALIS data warehouse integration pipeline (Alban et al., 2017)

risk-check components. Eight different data sources are integrated into the data warehouse:

- Pharmacological databases (gray): ABDAmed (ABDATA Pharma-Daten-Service, 2021), ROTE LISTE® (Rote Liste® Service GmbH, 2021), and GELBE LISTE® (Vidal MMI Germany GmbH, 2021)
- International databases with patient-related case reports of adverse drug events (red): FAERS (FDA, 2021), ARD (Health Canada, 2021), and DPD (Health Canada, 2021)
- Newly developed databases (green): CYP-P450 and PRISCUS-Liste (Holt et al., 2010)

The newly developed databases are based on information sources from scientific literature. Aggregating this knowledge into databases and merging it with pharmacological data enriches the risk analysis with new components.

The family of Cytochrome P450 enzymes (CYP) plays a crucial role in the metabolism of many substances. Variabilities between patients in the metabolism of medications by enzyme induction or inhibition and other genetic factors indicate a significant issue of pharmacotherapy. A new database CYP-P450 was designed, which contains information on substance-CYP enzyme interactions in the liver and kidney. This data is primarily based on the results of the literature research of Dippl (2011).

The PRISCUS list was created as a part of the joint project "PRISCUS" (Holt et al., 2010), which was funded by the German Federal Ministry of Education and Research (BMBF). The PRISCUS list includes 83 drugs available on the German drug market. The risk of these drugs for any side effects or age-related complications

prevails the medical benefits. A new database was derived from the published list. For these 83 potentially inadequate drugs information such as reason, therapy alternatives, and more were integrated into a suitable tabular data format.

Due to different exchange formats (XML, ASCII, CSV, MDB) and license models, specific parsers were implemented in Java for each data source. These parsers were used to extract the datasets, transform the data into the respective MySQL database, and load it efficiently into KALIS-DWH. Additional metadata for data analysis is stored in a separate database.

DAWIS-M.D. (Hippe et al., 2010) is a data warehouse for molecular information including data sources such as DrugBank (Knox et al., 2010), SIDER (Kuhn et al., 2010), and KEGG (Kanehisa, 2000). The pharmacological databases of KALIS-DWH were fused with the biomolecular databases of DAWIS-M.D. This data can be used for knowledge discovery of the underlying mechanism of drug action or the potential impact on the disease. The data integration of biomolecular data sources was performed by implementing XML parsers in Java and using the software kit BioDWH (Töpel et al., 2008). National and international identification standards were used for coding, mapping, and assignment of medical information such as drugs, therapeutic indications, diseases, and side effects. These include the ATC index (Anatomical Therapeutical Chemical classification), ICD-10 (International Statistical Classification of Diseases and Related Health Problems), and MedDRA (Medical Dictionary for Regulatory Activities) (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use ICH, 2021). In this way, the homogeneous data warehouses KALIS-DWH and DAWIS-M.D. provide pharmacological and biomolecular information for efficient and goal-oriented risk analysis of drugs. The standardized codes support the accuracy of data inputs and processing as well as a simple data exchange and uniform communication between KALIS and the end-user.

11.3.2 GraphSAW

GraphSAW is a web-based medical information system on drug interactions and side effects from pharmaceutical and molecular databases (Shoshi et al., 2015). Where KALIS focused mainly on vetted and official pharmaceutical databases such as ABDAmed (ABDATA Pharma-Daten-Service, 2021), GraphSAW provides a visual analysis and comparison with molecular databases such as DrugBank (Knox et al., 2010). The analyses are split into different components including drug–drug, drug–side effects, drug–molecule, drug–disease, drug–pathway, and pathway–disease interactions. A screenshot of the GraphSAW website is shown in Fig. 11.2.

The data integration utilized the two commercial databases ABDAmed (ABDATA Pharma-Daten-Service, 2021) and KEGG (Kanehisa, 2000) and the two freely available databases SIDER (Kuhn et al., 2010) and DrugBank (Knox et al., 2010) as visualized in Fig. 11.3.

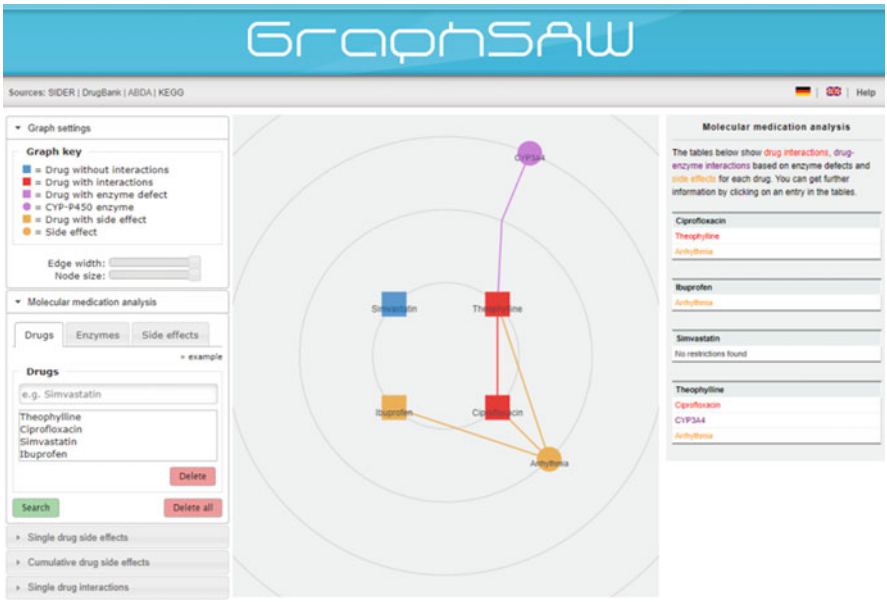

Fig. 11.2 Screenshot of the GraphSAW website. The analysis modules are shown on the left. Results are listed on the right and the main visualization in the middle. Here the molecular medication analysis is shown

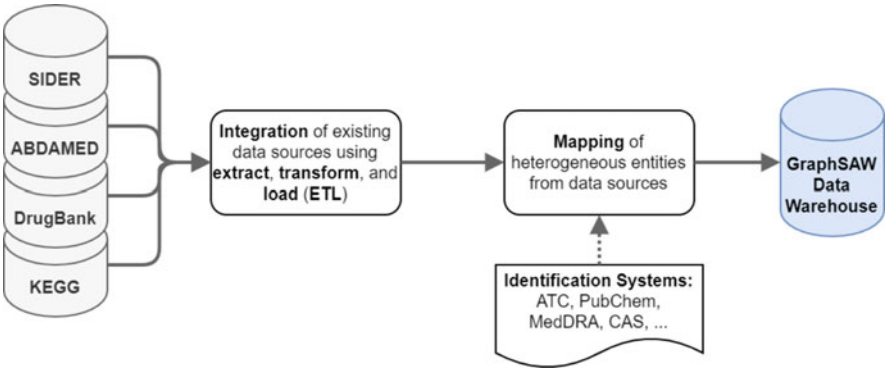

Fig. 11.3 Overview of the GraphSAW data warehouse integration pipeline (Shoshi et al., 2015)

DrugBank is the largest resource that collects binding data on small molecules, in particular those of drugs and proteins. At the time of creation 6711 approved and experimental drugs were integrated from DrugBank. As of April 2021, DrugBank contains more than double the number of drugs (14,460). DrugBank provides information on drug–drug as well as drug–target interactions, including CYP enzymes as mentioned in the KALIS section.

Further drug interactions were obtained from the commercial database ABDAmed that is based on approved and validated drug-related data in comparison to DrugBank. ABDAmed contains comprehensive facts for dealing with more than 47,000 drugs such as information about application and composition, risks, and drug interactions. The ABDAmed database includes also the side effects of drugs. More than 4500 side effects (3135 different; 1381 synonyms) were extracted automatically from full-text information in German and translated manually into English.

An additional 4192 different drug side effects were obtained from SIDER. Information on metabolic pathways was obtained from KEGG, which already integrates compounds from DrugBank (Knox et al., 2010), PubChem (Kim et al., 2020), CAS (American Chemical Society, 2021), and more. Therefore, DrugBank identifiers were used for mapping drugs between the data sources.

To map drugs between DrugBank and ABDAmed, the ATC classification system was used. MedDRA terms (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use ICH, 2021) were used for coding drug side effects of both SIDER and ABDAmed. The mapping between drugs of DrugBank and SIDER was realized by drug names because these databases did not have corresponding identifiers for compounds. By introducing mappings between the heterogeneous databases, interaction and side effect information were assigned to all drugs.

The data integration was implemented as parsers written in Java for the bio data warehouse BioDWH (Töpel et al., 2008). Using the data warehouse architecture ensures both the availability and the relevance of the data sources. Additional metadata for data analysis is stored in a separate database such as extracted and translated side effects from ABDAmed.

11.3.3 PIMBase

In recent years, lists, criteria, and classification systems for assessing potentially inappropriate medication for geriatric patients were developed and published. Besides these PIM lists of medication with a negative risk–benefit balance (i.e. PRISCUS (Holt et al., 2010), AUSTRIAN PIM (Mann et al., 2011)), lists with a positive balance (i.e. FORTA (Pazan et al., 2019), EU(7)-PIM (Renom-Guiteras et al., 2015)) are also becoming the focus of interest.

However, those PIM lists are spread across scientific journals and difficult to access for patients or health professionals in the context of treatment. The integration of the various lists into a uniform database and subsequent merging as well as an implementation of a unique rating scale are essential for the qualitative improvement of the drug therapy in the elderly and offer opportunities for practical application to identify and reduce inappropriate prescribing.

The data integration for the PIMBase database is divided into multiple steps. First, the original PIM lists were collected and the format analyzed. Most lists are only accessible as tables in PDF files either as supplementary information or directly embedded in their respective publications. Freely available tools for automatic extraction of tables from PDF files are often unsuitable and error-prone. As a correct data transformation could not be guaranteed, most PIM lists were transferred by hand into a machine-readable tabular format. This step was thoroughly checked to prevent any copy errors or loss of context.

With all lists in a machine-readable state common information such as drug names, ratings, reasons, or alternatives were compared. A collection of different list entries is listed below.

- Magnesium hydroxide.
- Docusate sodium (oral).
- Spironolactone >25 mg/d.
- Concomitant use of theophylline together with ciprofloxacin may increase the risk of theophylline toxicity in the elderly.
- In the elderly, avoid doses of acetylsalicylic acid greater than 325 mg per day due to increased risk of gastrointestinal bleeding and peptic ulcer disease.

The EU(7)-PIM list already provides annotations with the ATC index for the different drugs. The entries of the other PIM lists were manually annotated with ATC codes. For the first iteration of the PIMBase database, a simple relational database schema was developed mainly consisting of textual, numerical, and listing information. A simple python integration pipeline uses the created machine-readable lists and generates an SQL script readily usable in MySQL database installations.

Using the generated database, the first iteration of the PIMBase website[1] allows users to search for names and ATC codes and to see detailed information for each PIM entry. A screenshot of the website is shown in Fig. 11.4. With the addition of more PIM lists, multiple issues become apparent. For example, when searching for acetylsalicylic acid (ATC B01AC06 and N02BA01), four entries exist in the FORTA, one in the AUSTRIAN, and one in the EU(7)-PIM lists. All six different entries are annotated with matching ATC codes but still disconnected. The problem becomes more complex, where PIM entries are not only relevant for a single, but multiple drugs or even whole therapeutic categories. An example is the FORTA entry “Class I-III antiarrhythmic agents: All except Amiodarone and Dronedarone.” Not only do different lists may use slightly different synonyms for drugs but also use names, synonyms, or abbreviations of therapeutic categories which are not standardized. An example for these synonyms is {“Antimuscarinics,” “Antimuscarinic drugs,” “Muscarinic antagonists,” “Muscarinic-blocking agents,” “Muscarinic-blocking drug”}. When a user now searches for a certain drug or drug class, all relevant entries should be found. If a specific drug is searched for, but an

[1] https://pimbase.kalis-amts.de.

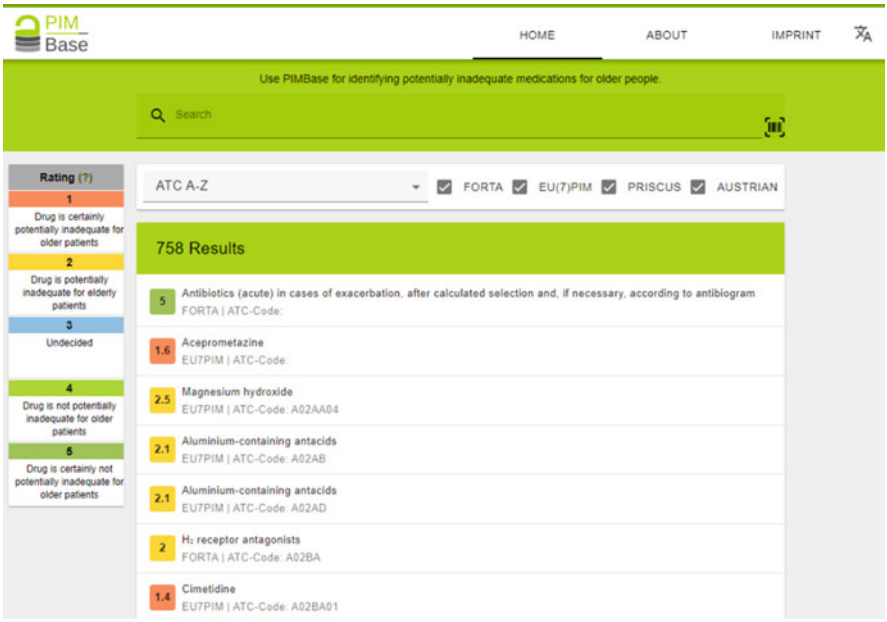

Fig. 11.4 Screenshot of the PIMBase website with the rating scale on the left and the entries of potentially inappropriate medication in the center

entry only provides a drug class that includes the specified drug, the entry should still be found.

This challenge necessitates the integration of therapeutic class hierarchies. Multiple databases provide their own categories and hierarchies such as NDF-RT, KEGG, and DrugBank. Independent hierarchies such as the ATC index and USP drug classification exist as well. However, each of these hierarchies has different intentions, number of hierarchy levels and drugs listed. An excerpt comparison of databases and hierarchies is visualized in Fig. 11.5. These hierarchies are used to implement a better search strategy in finding entries by drugs and drug classes. Drug entries in the leaf nodes need to be mapped to the ATC codes used for the PIMBase entries. Additionally, mappings between drug class hierarchies improve the number of entries found under category terms and reduce the number of duplicates in search suggestions.

In addition to drugs and drug classes, PIM entries are in most cases specific to certain patient conditions such as indications, age, gender, and laboratory measurements. Providing only relevant entries for a specific patient is therefore even more complex. Diseases, side effects, and other keywords need to be annotated for each PIM entry. Furthermore, the logical relationship between them needs to be encoded in a suitable data structure such as decision trees. If a user only searches for a specific drug, but the entry is only relevant in combination with a condition, the entry should still be shown. Vice versa, if only a condition is entered, the matching

USP Drug Classification (Extracted from KEGG)	ATC (Extracted from KEGG)	ABDAMED	DrugBank
Antiparkinson Agents	N04 ANTI-PARKINSON DRUGS	Antiparkinsonika	Anticholinergic Agents*
Anticholinergics	N04A ANTICHOLINERGIC AGENTS	Trihexyphenidyl	Aclidinium
Trihexyphenidyl	N04AA Tertiary amines	Biperiden	Agmatine
Diphenhydramine	N04AA01 Trihexyphenidyl	Procyclidin	Alcuronium
Benztropine	N04AA02 Biperiden	Bornaprin	Amantadine
Dopamine Precursors and/or L-Amino Acid Decarboxylase Inhibitors	N04AA03 Metixene	Levodopa	Amitriptyline
Levodopa	N04AA04 Procyclidine	Carbidopa	Amobarbital
Carbidopa	...	Amantadin	Amoxapine
Carbidopa/ Levodopa	N04AB Ethers chemically close to antihistamines	Bromocriptin	Anisotropine methylbromide
Dopamine Agonists	N04AB01 Etanautine	Pergolid	Aprobarbital
Bromocriptine	N04AB02 Orphenadrine (chloride)	α-Dihydroergocryptin	Aripiprazole
Ropinirole	N04AC Ethers of tropine or tropine derivatives	α-Dihydroergocryptin mesilat	Atracurium
Pramipexole	N04AC01 Benzatropine	Ropinirol	Atracurium besylate
Apomorphine	N04AC30 Etybenzatropine	Pramipexol	Atropine
Rotigotine	N04B DOPAMINERGIC AGENTS	Cabergolin	Barbital
Monoamine Oxidase B (MAO-B) Inhibitors	N04BA Dopa and dopa derivatives	Apomorphin	Batefenterol
Selegiline	N04BA01 Levodopa	Piribedil	Benactyzine
Rasagiline	N04BA02 Levodopa and decarboxylase inhibitor	Rotigotin	Benzatropine
Safinamide	N04BA03 Levodopa, decarboxylase inhibitor and COMT inhibitor	Selegilin	Benzilone
Antiparkinson Agents, Other	...	Rasagilin	Benzquinamide
Istradefylline	N04BB Adamantane derivatives	Safinamid	Bevonium
Carbidopa/ Levodopa/ Entacapone	N04BB01 Amantadine	Tolcapon	Biperiden
Entacapone	N04BC Dopamine agonists	Entacapon	Bornaprine
Tolcapone	N04BC01 Bromocriptine	Budipin	Brompheniramine
Amantadine	N04BC02 Pergolide	Opicapon	Buclizine
	N04BC03 Dihydroergocryptine mesylate	Benserazid	Butabarbital
	...		Butalbital
	N04BD Monoamine oxidase B inhibitors		Butobarbital
	N04BD01 Selegiline		Butylscopolamine
	N04BD02 Rasagiline		Carnylofin
	N04BD03 Safinamide		Chloroprocaine
	N04BX Other dopaminergic agents		Chlorphenoxamine
	N04BX01 Tolcapone		Chlorpromazine
	N04BX02 Entacapone		Chlorprothixene
	N04BX03 Budipine		Cisatracurium
	N04BX04 Opicapone		Clidinium
	N04C OTHER ANTIPARKINSON DRUGS		Clozapine
	N04CX Other antiparkinson drugs		Cocaine
	N04CX01 Istradefylline		...

Fig. 11.5 Excerpt of different drug classes and therapeutic groups for antiparkinson agents from KEGG, ABDAmed, and DrugBank. NDF-RT therapeutic categories had no categories matching either antiparkinson or anticholinergic agents. * DrugBank has no category for antiparkinson agents and was substituted with the sub-category of anticholinergic agents

drugs should be shown as well. This requires even more databases for disease information and suitable ontologies for measurements such as creatinine clearance. These challenges are currently under development.

Encoding each entry of all PIM lists with appropriate logical rulesets will result in a powerful toolset for healthcare professionals and patients. The quick access to relevant information for the specific patient situation will increase drug therapy safety and hopefully reduce inappropriate prescribing without the need to manually scan all PIM lists and a step further toward personalized medicine.

11.4 Outlook

Medical decision support and drug therapy safety are important but complex challenges. This chapter introduced several medical information systems and presented their data integration needs. Each of these systems represents a specific area of medical decision support and provides a piece to drug therapy safety as a whole.

The primary issue is the adoption by healthcare professionals. While being easily accessible and intuitive, none of the presented systems can communicate with other software such as hospital information systems. Communication standards like Health Level Seven (HL7) and FHIR and exchange standards for electronic health records need to be implemented. This will allow populating information system inputs directly from patient records and reduce the time and effort it takes to use the systems. Implementing these standards requires an extension of the data integration pipelines. This includes mapping relevant entities to the identification systems used in these standards.

In a future project, the concepts of all presented systems are planned to be merged into a single decision support system. Aside from the data integration needs, the entities and information from all systems need to be mapped. In most cases, this should be trivial where suitable identification systems are already present such as ICD-10 codes for diseases and ATC codes for medications. The development of an interaction check between KALIS and PIMBase with KATIS requires new information on the molecular composition and mechanics of remedies. This molecular data could then be used in the context of GraphSAW finding interactions between drugs and remedies.

Personalized medicine needs to analyze a patient as a whole, not only what medication he uses or which side effects are present. Allergies, diet, physical activity, and potential use of remedies all need to be considered to provide the best and safest treatment possible and to reduce adverse drug reactions. Therefore, the combination of all presented tools could provide a basis for personalized medicine in the future.

References

ABDATA Pharma-Daten-Service: Abdamed. https://abdata.de/datenangebot/abdamed/. Accessed: 2021-04-29

Alban S, Ulrich M, Arben S, Venus O, Ralf H (2017) Kalis—an ehealth system for biomedical risk analysis of drugs. Stud Health Technol Inform 236(Health Informatics Meets eHealth):128–135

American Chemical Society: Chemical abstracts service (CAS). https://www.cas.org. Accessed: 2021-04-29

BfArM: Verkehrsfähige arzneimittel im zuständigkeitsbereich des bfarm. https://www.bfarm.de/DE/Service/Statistiken/AM_statistik/statistik-verkf-am-zustBfArM.html. Updated: 2021-04-16

Boustani M, Baker MS, Campbell N, Munger S, Hui SL, Castelluccio P, Farber M, Guzman O, Ademuyiwa A, Miller D, Callahan C (2010) Impact and recognition of cognitive impairment among hospitalized elders. J Hosp Med 5(2):69–75

Chrischilles EA, VanGilder R, Wright K, Kelly M, Wallace RB (2009) Inappropriate medication use as a risk factor for self-reported adverse drug effects in older adults. J Am Geriatr Soc 57(6):1000–1006

Dippl H (2011) Hepatische cytochrom-wechselwirkungen von pharmakologischen substanzen—eine literaturrecherche für den zeitraum 2000–2008

Duerinckx AJ, Pisa EJ (1982) Filmless picture archiving and communication in diagnostic radiology. In: Duerinckx AJ (ed) 1st International conference and workshop on picture archiving and communication systems, vol 0318. International Society for Optics and Photonics, SPIE, New York, pp 9–18

FDA: FDA adverse event reporting system (FAERS). https://www.fda.gov. Accessed: 2021-04-29

Fick D (2001) Potentially inappropriate medication use in a medicare managed care population: association with higher costs and utilization. J Manag Care Pharm 7(5):407–413

Fiss T, Dreier A, Meinke C, van den Berg N, Ritter CA, Hoffmann W (2010) Frequency of inappropriate drugs in primary care: analysis of a sample of immobile patients who received periodic home visits. Age Ageing 40(1):66–73

German Federal Statistical Office: 14. koordinierte bevölkerungsvorausberechnung. https://service.destatis.de/bevoelkerungspyramide. Accessed: 2021-04-29

Grandt D, Lappe V, Schubert I (2018) BARMER Arzneimittelreport 2018 Schriftenreihe zur Gesundheitsanalyse. BARMER

Health Canada: Canada vigilance adverse reaction online database. https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada/adverse-reaction-database.html. Accessed: 2021-04-29

Health Canada: Drug product database (DPD). https://health-products.canada.ca/dpd-bdpp/. Accessed: 2021-04-29

Hippe K, Kormeier B, Töpel T, Janowski SJ, Hofestädt R (2010) DAWIS-M.D.—A Data Warehouse System for Metabolic Data, pp 720–725. Ges. für Informatik

Holt S, Schmiedl S, Thürmann PA (2010) Potentially inappropriate medications in the elderly. Deutsches Aerzteblatt Online

International council for harmonisation of technical requirements for pharmaceuticals for human use (ICH): medical dictionary for regulatory activities (MEDDRA). https://www.meddra.org. Accessed: 2021-04-29

Kanehisa M (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2020) PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49(D1):D1388–D1395

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2010) DrugBank 3.0: a comprehensive resource for omics' research on drugs. Nucleic Acids Res 39(Database):D1035–D1041

Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010) A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 6(1):343

Mann E., Böhmdorfer B, Frühwald T, Roller-Wirnsberger RE, Dovjak P, Dückelmann-Hofer C, Fischer P, Rabady S, Iglseder B (2011) Potentially inappropriate medication in geriatric patients: the Austrian consensus panel list. Wien Klin Wochenschr 124(5-6):160–169

Pazan F, Weiss C, Wehling M (2019) The FORTA (fit fOR the aged) list 2018: third version of a validated clinical tool for improved drug treatment in older people. Drugs Aging 36(5):481–484

Renom-Guiteras A, Meyer G, Thürmann PA (2015) The EU(7)-PIM list: a list of potentially inappropriate medications for older people consented by experts from seven European countries. Eur J Clin Pharmacol 71(7):861–875

Rote Liste® Service GmbH: Rote liste®. https://www.rote-liste.de. Accessed: 2021-04-29

Shoshi A, Hoppe T, Kormeier B, Ogultarhan V, Hofestädt R (2015) GraphSAW: a web-based system for graphical analysis of drug interactions and side effects using pharmaceutical and molecular data. BMC Med Inform Decis Mak 15(1):1–10

Töpel T, Kormeier B, Klassen A, Hofestädt R (2008) BioDWH: A data warehouse kit for life science data integration. J Integr Bioinform 5(2):49–57

Vidal MMI Germany GmbH: Gelbe liste®. https://www.gelbe-liste.de. Accessed: 2021-04-29

Part IV
Network Modeling and Simulation

Chapter 12
Visualising Metabolic Pathways and Networks: Past, Present, Future

Falk Schreiber, Eva Grafahrend-Belau, Oliver Kohlbacher, and Huaiyu Mi

Abstract Visualisations of metabolites and metabolic pathways have been used since the early years of research in biology, and pathway maps have become very popular in biochemistry textbooks, on posters, as well as in electronic resources and web pages about metabolism. Visualisations help to present knowledge and support browsing through chemical structures, enzymes, reactions and pathways. In addition, visual and immersive analytics of metabolism connects network analysis algorithms and interactive visualisation methods to investigate structures in the network such as centralities, motifs and paths, or to compare pathways for finding differences between species or conditions. The graphical depiction of networks supports the mapping and investigation of additional data such as metabolomics, enzyme activity, flux and transcriptomics data, and the exploration of the data in the network context. It builds a foundation for investigating the dynamics of metabolic processes obtained either experimentally or via modelling and simulation. Here we discuss past, present and future of the visualisation of metabolic networks and pathways and provide links to several resources.

Keywords Visualisation · Metabolism · Metabolic networks · Visual analytics · Immersive analytics · Biovis

F. Schreiber (✉)
University of Konstanz, Konstanz, Germany
Monash University, Clayton, VIC, Australia
e-mail: falk.schreiber@uni-konstanz.de

E. Grafahrend-Belau
Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
e-mail: eva.grafahrend-belau@pharmazie.uni-halle.de

O. Kohlbacher
University of Tübingen, Tübingen, Germany
e-mail: oliver.kohlbacher@uni-tuebingen.de

H. Mi
University of Southern California, Los Angeles, CA, USA
e-mail: huaiyumi@usc.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_12

12.1 Introduction

Visualisations of metabolic pathways and pathway components such as enzymes and compounds have been used since the early years of research in biology, and metabolic pathway maps have become very popular in biochemistry textbooks, on posters, as well as in electronic resources and web pages. One example is Gerhard Michal's famous poster Biochemical Pathways (Michal, 1968, 1998) which has been printed over a million times. The first example of computational representation of pathways was the EcoCyc database (Karp and Mavrovouniotis, 1994; Keseler et al., 2016). Another well-known example is the KEGG pathway database (Kanehisa and Goto, 2000; Kanehisa et al., 2012), the largest collection of manually curated pathway maps and related metabolic information.

Visualisations are commonly used in biology (Gehlenborg et al., 2010; Kerren et al., 2017). Metabolic pathway visualisations help to present knowledge and to support browsing through chemical structures, enzymes, reactions and pathways. Visual and immersive analytics of metabolism connects network analysis algorithms and (interactive and/or immersive) visualisation methods to investigate hubs, motifs, paths and so on in the network, or to compare pathways for finding differences between species or conditions. In addition, network visualisation also supports the mapping and investigation of further data such as metabolomics, proteomics, transcriptomics, enzyme activity and flux data, and the exploration of the data in the network context. It builds a foundation for exploring and navigating the dynamics of metabolic processes obtained either experimentally or via modelling and simulation. In conclusion, visualising and visually exploring metabolic pathways and networks helps in understanding them, is important in making sense of the complex biological data and knowledge that is being produced these days, and is an important research area. A simple visualisation example is shown in Fig. 12.1.

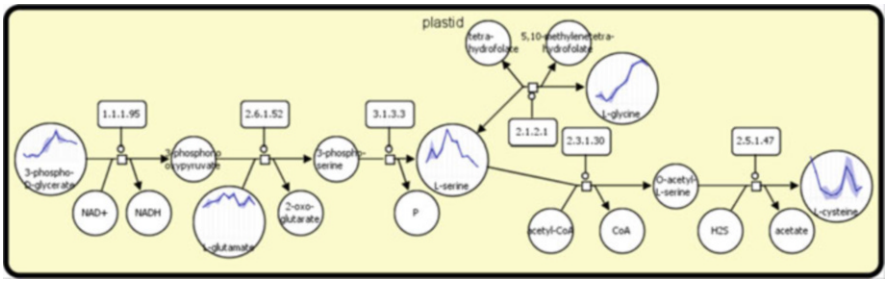

Fig. 12.1 A metabolic pathway example with some time series data of metabolite concentrations shown within the vertices representing metabolites (excerpt from a MetaCrop pathway (Schreiber et al., 2012) rendered by the Vanted tool (Junker et al., 2006))

12.1.1 Network Representation

A network representing metabolic processes consists of a set of elements (called vertices or nodes) and their connections (called edges or arcs) which have a defined appearance (e. g. size of vertices) and are placed in a specific layout (e. g. coordinates of vertices). Typical representations of metabolic reactions as graphs with different interpretations of vertices and edges are shown in Fig. 12.2. Although initiatives for a uniform representation of metabolic pathways have been presented earlier (Kitano, 2003; Kitano et al., 2005; Michal, 1998), no graphical representation became a standard to represent metabolic processes. In 2006 an international consortium started developing a standard for the graphical representation of cellular processes and biological networks including metabolism called the Systems Biology Graphical Notation (SBGN) (Le Novère et al., 2009). SBGN allows the visualisation of complex biological knowledge, including metabolic networks (see also the information in Fig. 12.3). Within this chapter, we will use SBGN for representing metabolic pathways and networks where possible.

12.1.2 Network Layout

Metabolic pathway maps have been produced manually for a long time. These drawings are manually created (usually with help of computer programs) long before their actual use and provide a static view of the data defined by the creator. They show the knowledge at the time of the map's generation and an end-user cannot change the visualisation. Some navigation may be supported in electronic systems using such pre-drawn pictures, but the result of an action (the new picture) either

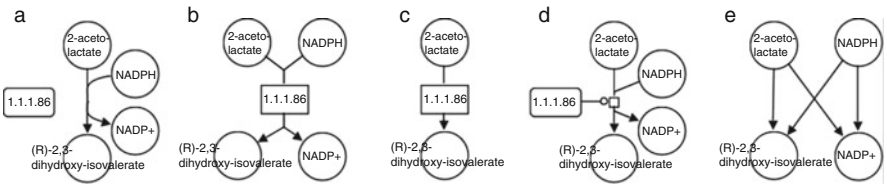

Fig. 12.2 Different representations of biochemical reactions: (**a**) hypergraph (vertices denote metabolites and enzymes, edges denote reactions); (**b**) bipartite graph with enzymes represented within reactions (vertices denote metabolites and reactions including enzymes, edges connect metabolites with reactions); (**c**) simplified representation of (**b**) without co-substances (as used in KEGG), (**d**) bipartite graph with enzymes represented as separate entities (SBGN notation, vertices denote metabolites, enzymes and reactions, edges connect metabolites with reactions (consumption, production) and enzymes with reactions (catalysis)); (**e**) simplified metabolite network (vertices denote metabolites, edges connect metabolites transformed by reactions). Note that the classical representations such as the ones in Michal's poster (Michal, 1998) and in Stryer's biochemistry textbook (Stryer, 1988) are similar to (**a**)

The **Systems Biology Graphical Notation** aims at an unambiguous graphical representation of biological networks and cellular processes. To encode levels of detail and different aspects of the interactions, SBGN provides three orthogonal and complementary languages:
Process description (PD) focuses on temporal dependencies of transformations and interactions in a network and describes them in a mechanistic way. It is used to represent networks of events which convert biological entities into each other, change their state or transport them to another location. Entity pool nodes represent pools of simple chemicals, macromolecules, etc.; process nodes encode their transformation or transport. For details see Rougny et al. (2019). SBGN PD is commonly used to graphically represent metabolic pathways and networks.
Entity relationship (ER) focuses on the relationship and influences in which the entities are involved or which they have onto each other. It does not consider temporal aspects but describes relationships in a mechanistic manner and shows all possible relationships at once. Entity nodes in ER represent entities that exist; relationships are rules that decide whether an entity node exists. Details can be found in the technical specification (Sorokin et al., 2015).
Activity flow (AF) focuses on the biological activity. It shows the sequential influence of activities and can be ambiguous when it comes to the underlying mechanism. Activity nodes represent the biological activities; modulation arcs represent the influence of activities onto other activities. For details see Mi et al. (2015).
Detailed information on SBGN is available at www.sbgn.org, as part of the yearly special issue on standards in systems and synthetic biology (Schreiber et al., 2020) and here (Junker et al., 2012). Software support for SBGN maps is described by van Iersel et al. (2012) and on www.sbgn.org.

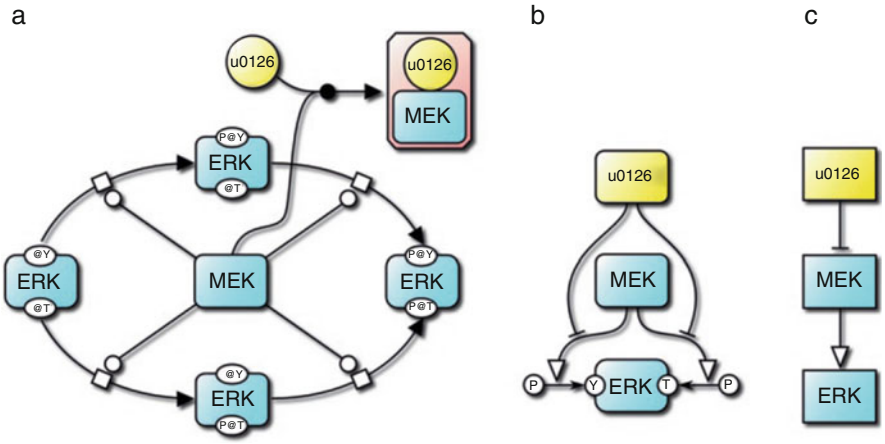

Fig. 12.3 Box SBGN: Explanation of SBGN and SBGN languages, the image shows as example protein phosphorylation catalysed by an enzyme and modulated by an inhibitor in all three SBGN languages: (**a**) PD, (**b**) ER, (**c**) AF (image from Le Novère et al. (2009))

replaces the current image or it is visualised in a new view, and the visualisations are not interactive.

However, the automatic computation of visualisations and interactive exploration methods are highly desirable, due to size and complexity of biological networks, a

steady growth of knowledge and the derivation of user-specific parts of networks from databases. The computer-based generation (layout) of a network map on demand at the time it is needed is called dynamic visualisation. These visualisations are created by the end-user from up-to-date data with help of a layout algorithm. They can be modified to provide specific views of the data, and several navigation methods such as the extension of an existing drawing or map with new additional parts are supported.

The automatic layout of networks, that is the computation of maps from a given network, is called graph drawing. Graph drawing methods take a network (or graph) and compute a layout consisting of coordinates for the vertices and routings of the edges. See the books by Di Battista et al. (1999) and Kaufmann and Wagner (2001) for general graph drawing algorithms. Although standard graph drawing algorithms can be used for laying out metabolic networks, domain specific network visualisations that conform to biological representational conventions are advantageous (Bourqui et al., 2011; Schreiber, 2002).

In the following sections we will discuss major resources for metabolites, reactions and pathways (Sect. 12.2); the visualisation of metabolites and enzymes, which are the building blocks of metabolic pathways (Sect. 12.3); and the visualisation of the pathways themselves (Sect. 12.4). We will focus on key questions which can be addressed using visualisation, present important graph drawing algorithms in brief (both standard and domain specific algorithms) and discuss a selection of useful tools. Next we will discuss the exploration and analytics of pathways and data, in particular visual analytics and immersive analytics of metabolic pathways and related data (Sect. 12.5). We conclude with perspectives and research questions in this field. Boxes contain additional background information regarding, for example, standards for metabolic network representation and layout algorithms.

12.2 Resources for Metabolites, Reactions and Pathways

Large amounts of knowledge about metabolites, enzymes, metabolic reactions, pathways and networks have been accumulated and are derived at increasing speed. Several databases and information systems have been developed to provide a comprehensive way to manage, explore and export this knowledge in meaningful ways. We will concentrate on the most important primary databases and briefly discuss their typical content and important features.

Databases in this area can be divided into *metabolite/compound databases* providing information about the chemical compounds used or produced in biochemical reactions; *reaction/enzyme databases* containing information about enzymes and the reactions catalysed by them; and *pathway databases* providing information about metabolic pathways. See also Table 12.1 for more information and a comparison of relevant databases.

Table 12.1 Databases. Abbreviations: W - Web, S - Web services, F - FTP

Database	Ease of use	Level of detail	Data access	Support of exchange formats	URL
Metabolite/compound databases					
ChEBI	++	+	W, S, F	TSV, XML, OBO, OWL, MOL, SDF	bit.ly/zWCpYl
KEGG COMPOUND	+	+	W, S, F	MDL/MOL, KCF, Jmol, KegDraw, RDF	bit.ly/wxHkVi
PubChem	++	++	W, S, F	ASNT, XML, CSV, SDF, JSON, PNG	bit.ly/2JsSyOm
Reaction/enzyme databases					
BRENDA	++	++	W, S	SBML, Fasta, CSV	bit.ly/xGVzg3
ExPASy-ENZYME	+		W, F		bit.ly/33yii31
KEGG ENZYME			W, S, F	RDF	bit.ly/3fSXiZE
Rhea			W	TSV, BioPAX, RDF, RXN	bit.ly/AfG1d4
Sabio-RK	+	++	W, S	SBML, BioPAX, SBPAX, XLS, MatLab, TSV	bit.ly/zZ7Ax0
Metabolic pathway databases					
MetaCyc	+	++	W, S	SBML, BioPAX, DB, TSV, ...	bit.ly/2JfJigV
KEGG PATHWAY	++	+	W, S, F	KGML	bit.ly/w2urRG
PANTHER Pathway	+	++	W, S, F	SBML, SBGN BioPAX	bit.ly/x6aQ9n
Reactome	++	++	W, S	SBML, BioPAX, DB, SBGN, PSI PPTX, PNG, ...	bit.ly/9RtvaZ

12.2.1 Metabolite/Compound Databases (Chemical Databases)

The typical content is information about metabolites (compounds) and their properties such as name, synonyms, molecular weight, molecular formula and structure. Often associated information such as chemical reactions, metabolic pathways, publications and various links to other databases can also be found.

Important resources are PubChem (Kim et al., 2015, 2020; Wang et al., 2009) is a comprehensive source of compound and substance information (consisting of the three primary databases: Compounds, Substances and BioAssay). KEGG COMPOUND (Goto et al., 2002) is a database of small molecules, biopolymers and other chemical substances of biological interest. ChEBI (Hastings et al., 2015) is a database of small molecules with detailed information about nomenclature, molecular structure, formula and mass.

The visualisation and visual analysis of data from these databases is discussed in Sect. 12.3.

12.2.2 Reaction/Enzyme Databases

The typical content is information about enzymes and their properties such as nomenclature, enzyme structure, functional parameters and specificity. Often additional information about the reactions catalysed by the given enzyme, metabolic pathways, references and links to other databases can also be found.

Important resources are: BRENDA (Chang et al., 2020; Scheer et al., 2011) is a comprehensive enzyme information system providing detailed molecular and biochemical information on enzymes based on primary literature. ExPASy-ENZYME (Gasteiger et al., 2003) is an enzyme database which covers information related to the nomenclature of enzymes. Rhea (Lombardot et al., 2018) is an expert-curated reaction database with information about biochemical reactions and reaction participants. The KEGG databases ENZYME and REACTION (Kanehisa and Goto, 2000; Kanehisa et al., 2004) provide enzyme- and reaction-specific information about chemical reactions in the KEGG metabolic pathway database. Sabio-RK (Wittig et al., 2012, 2017) is an expert-curated biochemical reaction kinetics database with detailed kinetic information.

The visualisation and visual analysis of data from these databases is discussed in Sect. 12.3.

12.2.3 Metabolic Pathway Databases

The typical content is information about metabolic pathways and their single reactions, involved enzymes and reactants and associated information such as organism-specific information about genes, their related gene products, protein functions and expression data. Often several types of information are provided in the context of the graphical representation of pathways.

Major databases are: KEGG PATHWAY (Kanehisa et al., 2002, 2020), a multi-organism pathway database which contains metabolic pathways, represented as curated, manually drawn pathway maps consisting of links to information about compounds, enzymes, reactions and genes. BioCyc/MetaCyc (Caspi et al., 2012, 2019; Krieger et al., 2004) is a collection of organism-specific pathway databases including MetaCyc, a curated multiorganism pathway database, which contains metabolic pathways curated from the literature, lists of compounds, enzymes, reactions, genes and proteins associated with the pathways. Reactome (Croft et al., 2014; Matthews et al., 2009) is a curated multi-organism pathway database initially focussing on human biology. PANTHER pathway (Mi and Thomas, 2009; Mi et al., 2020) is an expert-curated multi-organism pathway database.

In addition to the mentioned primary databases, there are secondary pathway databases and collaborative databases. Secondary metabolic pathway database systems are collecting and presenting information from various sources. Examples are NCBI BioSystems (Geer et al., 2010) and Pathway Commons (Cerami et al., 2011; Rodchenkov et al., 2019). The former is a centralised repository for metabolic pathway information containing biological pathways from multiple databases (e. g. KEGG, Human Reactome, BioCyc and the National Cancer Institute's Pathway Interaction database). The latter provides access to various public metabolic pathway databases, such as Reactome, HumanCyc and IMID. A well-known community-driven collaborative platform dedicated to the curation and representation of biological pathways is WikiPathways (Martens et al., 2021).

There is also BioModelsDB (Chelliah et al., 2013; Malik-Sheriff et al., 2019), a database of mathematical models representing biological processes including metabolism, and the BioModelsDB part Path2Models (Büchel et al., 2013), an automatic translation of metabolism from databases such as KEGG into biological models using the SBML and SBGN standards. In addition, there are also special metabolic pathway databases covering specific species or groups of species, for example, for plants PlantCyc (Schläpfer et al., 2017; Zhang et al., 2010), MetaCrop (Grafahrend-Belau et al., 2008; Weise et al., 2006) and Plant Reactome (Naithani et al., 2016, 2019).

The visualisation of data from these databases is discussed in Sect. 12.4. The above mentioned databases also provide static (e. g. KEGG) or dynamic (e. g. BioCyc) visualisations of pathways and networks. Furthermore, they often come with integrated analysis tools to support high-throughput experimental data analysis. For example, Reactome visualises pathways and maps expression data using colour-coding onto pathway maps. A Cytoscape plugin enables to generate new pathways based on database queries and to perform some graph analysis on these networks. KegArray is a light-weight data mapping utility, good for easily mapping expression data (csv) onto KEGG pathways to colour-code vertices and provides also some scatter plots of the data. And PANTHER Pathways allow users to view results in both SBGN process view and an automatically converted activity flow view. However, these tools are specifically developed for specific databases and often provide less functionality than the best general purpose tools presented in Sect. 12.4.3; therefore we will not present them in detail here.

12.2.4 Exchange Formats

To represent metabolic pathway information in a unified way and to support the exchange of pathway models between software tools, exchange formats have been proposed. Two exchange formats which focus on the exchange of information between software tools and databases are SBML (Hucka et al., 2003) and BioPAX (Demir et al., 2010), see the box in Fig. 12.4. Although they also partly support the exchange of graphical information, they are mainly relevant for software

SBML (Systems Biology Markup Language) is a machine-readable format for representing pathway models [Keating et al., 2020]. SBML has been developed by aninternational community of software developers and systems biologists to provide a common format for data sharing between various computer-modelling software applications. It is neutral with respect to software encoding and programming languages, and oriented towards enabling XML-encoded models. Software tools which use SBML as a format for writing and reading models can exchange the same computable representation of those models. Today, around 230 software packages support SBML. Detailed documentation on the SBML format is available online at `www.sbml.org`.

BioPAX is a collaborative effort to create a data-exchange format for biological pathway data [Demir et al., 2010]. The aim is to support accessing, sharing and integrating data from multiple pathway databases. BioPAX supports the representation of metabolic and signalling pathways, molecular and genetic interactions as well as gene regulation. It describes relationships between genes, small molecules, complexes and their states, including the results of events. Detailed documentation on the BioPAX format is available online at `www.biopax.org`. BioPAX is complementary to other standard pathway information-exchange languages, such as SBML, focusing on qualitative, large networks and their integration rather than on mathematical modelling of quantitative, small models.

SBML, BioPAX as well as other standards such as SBGN are part of the COMBINE initiative [Waltemath et al., 2020] and are used from single reaction to small pathway to larger and whole cell models [Waltemath et al., 2016].

Fig. 12.4 Box BioPAX and SBML

developers and modellers. The exchange format relevant for transferring graphical information and most relevant for users only interested in the visual representation of pathways is SBGN, the Systems Biology Graphical Notation (Le Novère et al., 2009), see the box in Fig. 12.3. Exchange of metabolite structures relies on several formats popularised in cheminformatics and drug design. MDL Molfiles are suitable for storing individual structures; collections of these files can then be assembled into SD Files (Dalby et al., 1992). While some pathway databases provide small molecule structures as MOL or SD files (e. g. KEGG, ChEBI), others provide the structures as SMILES (Weininger, 1988). SMILES, a so-called line notation, encodes the structure as a string and thus does not provide atom coordinates, but provides a more compact representation. A multitude of other file formats exists; most of these formats can be easily accessed and interconverted by cheminformatics toolkits and libraries (e. g. CDK (Steinbeck et al., 2003)) or conversion utilities (e. g. OpenBabel (O'Boyle et al., 2011)).

12.3 Visualising Metabolites and Enzymes

Textbook views of metabolic pathways often illustrate the underlying biochemical mechanisms. To this end it is essential not just to provide the name of the metabolites. Structural drawings are much better suited to illustrate the molecular details of an enzymatic reaction. The example in Fig. 12.5 shows the reactions catalysed by triosephosphate isomerase and glyceraldehyde 3-phosphate dehydrogenase, the isomerisation of dihydroxyacetone phosphate to D-glyceraldehyde-3-phosphate to D-glycerate 1,3-bisphosphate. Visualisation of the metabolites by their names, IDs or abbreviation makes it hard to understand the mechanism, while the structural drawings immediately reveal the conversion of the hydroxyl group to an aldehyde and of the keto group to a hydroxyl group and subsequent introduction of a phosphate group. The layout of the structural formulas has been designed to highlight the fact that the larger part of the structure remains unchanged during the two reactions. Only parts of the structure (highlighted by the boxes) are modified in the reaction. Manual layouts of metabolic pathways typically found in biochemistry textbooks are thus careful with the layout of both the structures and the pathway to maximise the mental map preservation between adjacent structures.

While drawing structural formulas comes natural to chemists and biochemists, the automated generation of structural formulas is a difficult task. The drawings have to adhere to numerous conventions developed since their initial conception by Kekulé towards the end of the 1800s. While many of these conventions have been standardised by the International Union of Pure and Applied Chemistry (IUPAC), there is no unique way for drawing a chemical structure; it can be adapted depending on the context, the level of detail required, and the information that needs to be conveyed.

Most small molecule chemical structures can be represented by planar graphs and thus can be laid out in 2D without major issues (Rücker and Meringer, 2002). Specific conventions have to be followed with respect to angles, representation of stereochemistry or bond orders, to name just a few. Ring systems pose particular

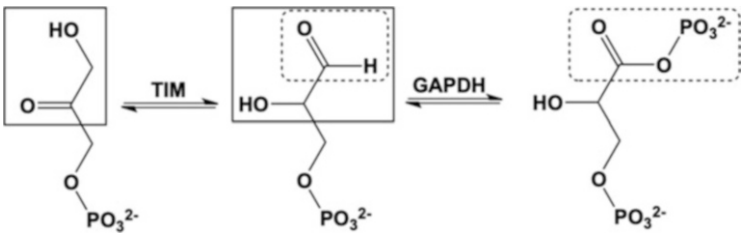

Fig. 12.5 Triosephosphate isomerase (TIM) catalyses the conversion of dihydroxyacetone phosphate to D-glyceraldehyde-3-phosphate, which in turn can be converted to D-glycerate 1,3-bisphosphate by glyceraldehyde 3-phosphate dehydrogenase (GAPDH). A consistent layout of the three metabolites involved makes it easier to grasp the structural changes entailed by each metabolic reaction (highlighted by the boxes)

challenges, since they are typically drawn in very specific ways and more often than not projection of the three-dimensional shape is preferred over a non-crossing planar embedding of the final structure. It could be shown that already the drawing of planar graphs with fixed edge lengths (as is the case for structural diagrams) is NP-hard (Eades and Wormald, 1990), most of these algorithms have to resort to heuristics to generate good layouts.

Several algorithms have been proposed over the years to layout structures in an aesthetic manner (Clark et al., 2006; Helson, 1990). In addition, a number of algorithms have been implemented in commercial tools for structure editing and cheminformatics, for example, in the ChemDraw suite.[1] in Accelrys Draw,[2] or in MOE.[3] Also academic cheminformatics projects such as CACTVS (Ihlenfeld et al., 1994) or the more recent Chemistry Development Kit, CDK, (Steinbeck et al., 2003) permit the layout of molecular structures. Based on the structure stored in pathway databases (see Sect. 12.2) these tools permit the rendering of the structure into a 2D image. Another option for the retrieval of structure drawings is PubChem,[4] which contains pre-computed structural formulas. These can be downloaded in PNG format.

A challenge in the visualisation of metabolic networks is currently the joint layout of the metabolic network and its constituent metabolites. While it is in principle possible to layout metabolic networks and simultaneously display the structural formulas of its metabolites, current pathway visualisation tools do not consider this problem (see, for example, Fig. 12.6). Not all tools are able to display structural formulas at all. Those that do, resort to pre-rendered images of the structures. If the structures are drawn individually, their orientation depends mostly on the algorithm used—there is no canonical orientation of a molecular structure. The orientation, size and general layout of any two structures adjacent in a metabolic network are thus mostly random, and it becomes very difficult to match the conserved common substructure between the two structures and thus to comprehend the underlying mechanism. Mental map preservation between any two adjacent structures would of course be preferable and clearly enhance readability of the pathway. Hand-made pathway diagrams found in textbooks are thus so far vastly superior to automatically drawn pathways with structural formulas. The simultaneous constrained drawing of metabolite structures and metabolic pathways is one of the more difficult problems in this area. Some algorithms for the constrained drawing of structures that should be suitable to solve this problem have been suggested in the literature in different contexts (Boissonnat et al., 2000; Fricker et al., 2004).

For small molecules (metabolites) 2D visualisation is the method of choice, because the structures are easier to comprehend and—to the schooled eye—the three-dimensional aspects of the structures are typically obvious. The same does

[1] CambridgeSoft Corp., Cambridge, MA, USA.

[2] Accelrys, Inc., San Diego, CA, USA.

[3] Chemical Computing Group, Montreal, Canada.

[4] pubchem.ncbi.nlm.nih.gov.

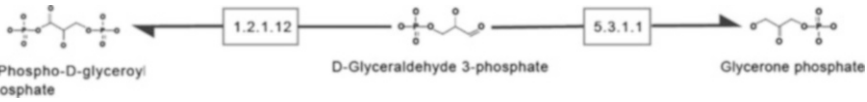

Fig. 12.6 Visualisation of metabolic pathways usually relies on pre-computed metabolite structures. As a consequence, the resulting pathway layout does not match the (usually random) orientation of the structural formulas embedded in the pathway and hampers understanding of the pathway mechanisms; excerpt from a KEGG (Kanehisa et al., 2014) pathway rendered by BiNA (Gerasch et al., 2014)

not apply to proteins, however. Representing proteins as structural formulas is not only impractical, but the function of proteins can only be understood from their three-dimensional structure.

12.4 Visualising Reactions and Pathways

12.4.1 Visualising the Structure of Metabolic Reactions and Pathways

Visual representations of metabolic pathways are widely used in the life sciences. They help in understanding the interconnections between metabolites, analysing the flow of substances through the network, and identifying main and alternative paths. Important visualisation requirements are (Schreiber, 2002):

- For *parts of reactions*: The level of detail shown concerning specific substances and enzymes is very much dependent on the goal of the visualisation, see also Sect. 12.3. Often for main substances their name and/or structural formula should be shown, for co-substances the name or abbreviation, and for enzymes the name or EC-number.
- For *reactions*: The reaction arrows should be shown from the reactants to the products with enzymes placed on one side of the arrow and co-substances on the opposite side. Both sides of a reaction as well as their reversibility should be visible.
- For *pathways*: The main direction of reactions should be visible to show their temporal order. Few exceptions to the main direction are used to visualise specific pathways such as the fatty acid biosynthesis and the citric acid cycle. The arrangement of these cyclic reaction chains should be emphasised: a repetition of a reaction sequence in which the product of the sequence re-enters as reactant in the next loop, either as cycle (the reactant and the product of the reaction sequence are identical from loop to loop, e. g. citric acid cycle) or as spiral (the reactant of the reaction sequence varies slightly from the product, e. g. fatty acid biosynthesis).

Besides specific visualisation requirements, reaction and pathway visualisations should meet the usual quality criteria of network layouts such as low number of edge crossings and a good usage of the overall area. See Fig. 12.7 for an example which meets these requirements.

12.4.2 Layout Algorithms for Visualising Metabolic Pathways and Networks

Metabolic networks are usually represented as directed graphs. Common approaches to automatically layout these networks are force-directed and hierarchical (or layered) layout methods. Although quite common as visualisation principal, for example, in the manual KEGG maps layout, automatic orthogonal (or grid) methods are less often used. See the box in Fig. 12.8 as well as the images in Fig. 12.9 for these layout methods. Force-directed methods are widely used, and several network analysis tools support such layouts. However, these approaches do not meet common visualisation requirements. Different vertex sizes, the special placement of co-substances and enzymes, the partitioning of substances into reactants and products and the general direction of pathways are not considered. A few approaches extend the force-directed layout method to deal with application specific requirements. An example is implemented in the PATIKA system (Demir et al., 2002; Dogrusöz et al., 2006) where the layout algorithm considers directional and rectangular regional constraints which can be used to enforce layout directions and sub-cellular locations.

Layered layout methods are often used as they emphasis the main direction within a network. Tools which support such layered layout methods are often based on existing layout libraries. These approaches show the main direction of reactions and are sometimes able to deal with different vertex sizes. However, there is no special placement of co-substances or specific pathways (e. g. cycles). Some improved approaches consider cyclic structures or depict pathways of different topology with different layouts, e. g. the algorithm by Becker and Rojas (2001) which emphasises cyclic structures, and PathDB (Mendes, 2000; Mendes et al., 2000) which visualises metabolic networks based on hierarchical layout allowing co-substances to be represented in a smaller font on the side of the reaction arrow.

There are some advanced methods for the automatic layout of metabolic pathways and networks such as the mixed, the extended layered and the constraint layout. The mixed layout approach (Karp and Mavrovouniotis, 1994) depicts (sub-)pathways of different topology with suitable layout algorithms such as linear, circular, tree and hierarchical layout, and places co-substances and enzymes beside reaction arrows. It is used in the MetaCyc/BioCyc database system. The extended layered approach (Schreiber, 2002) extends the hierarchical layout for different vertex sizes, consideration of co-substances and enzymes, and special layout of open and closed cycles; it is implemented in BioPath system (Brandenburg et al.,

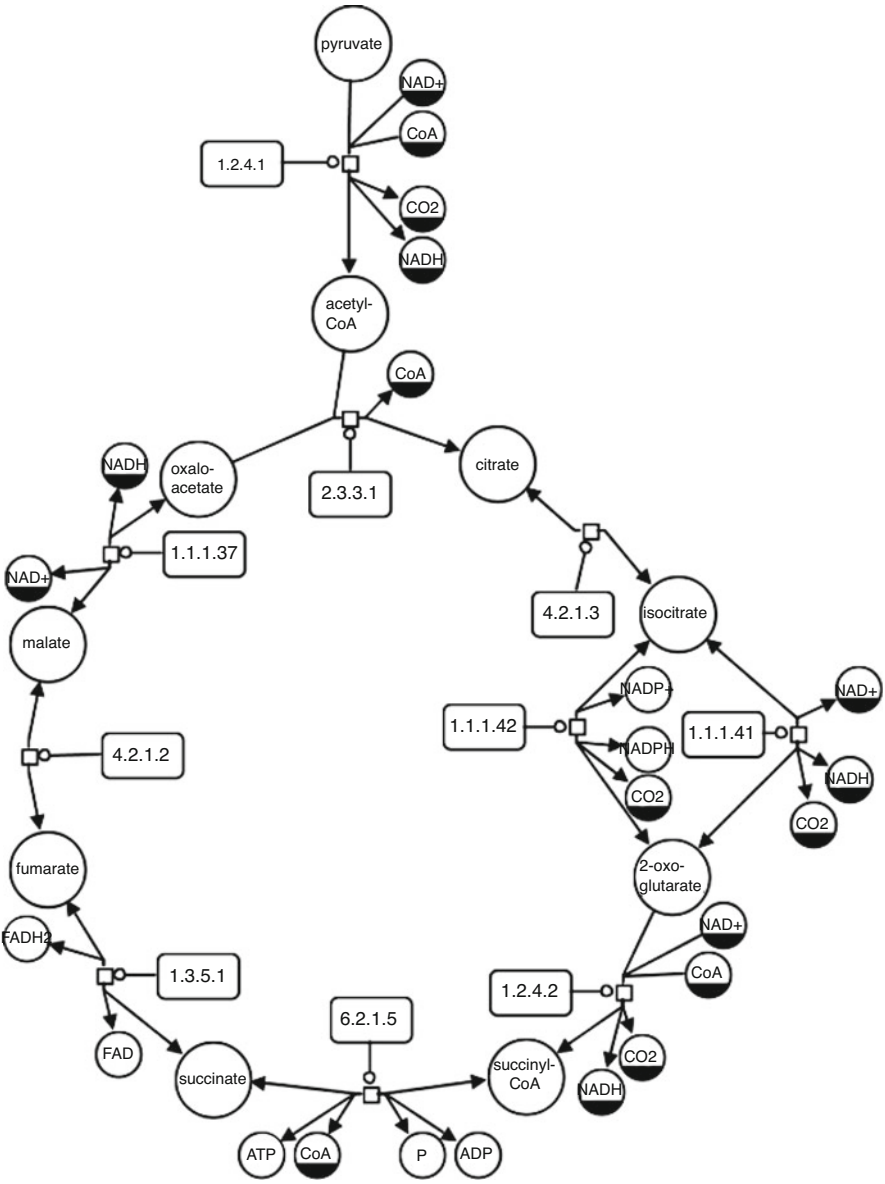

Fig. 12.7 Example of metabolic pathway visualisation which meets the requirements outlined in Section 4.1 (citric acid cycle, including reversible and irreversible reactions and circular shape of the pathway; excerpt from a MetaCrop (Hippe et al., 2010) pathway rendered by Vanted (Colmsee et al., 2013))

Force-directed methods

The idea of force-directed layout methods is to simulate a system of physical forces (e. g. spring or magnetic forces) on the graph and compute a distribution of the vertices with low overall force or energy. Standard force-directed graph drawing algorithms look like:

1. Compute the initial layout d_0 of graph G
   $d_0(G) = \text{init}(G)$
2. While not finished
   a) for all vertices v from vertex set V
      Compute the current force to this vertex, e. g. energy, magnetic, distribution, penalties
      $F(v) = \text{force}(v, d_i(G))$
   b) for vertex / vertices from vertex set V (chosen one vertex, vertex set depending of strategy s, or all vertices)
      Compute a new layout $d_{i+1}(G)$
      $d_{i+1}(G) =$ move v according to $F(v)$

For specific algorithms the parameters of this general method have to be specified; several methods could be used to compute the initial layout (e. g. random placement or pre-defined from an application), the current force, the strategy for choosing a vertex and how the vertex should be moved, the criteria to finish the computation and so on. Force-directed methods can be applied to undirected and directed graphs.

Layered methods

Layered graph drawing algorithms (also called hierarchical layouts) consist of the following four phases:

1. The removal of cycles in the graph as a result of changing temporarily the direction of some edges (decycling)
2. The assignment of vertices to layers in a way that all edges obtain the same direction, e. g. from top to bottom (layering)
3. The permutation of vertices within each layer with the aim to reduce edge crossings (crossing reduction)
4. The assignment of coordinates to vertices and of bend points (coordinates) to edges (coordinate assignment)

Layered methods work on directed graphs. To apply them to undirected graphs these graphs have to be changed into directed graphs, e. g. by choosing a start vertex v, applying breath first search (BFS) starting with v and directing each edge depending on the order given by the BFS algorithm.

Fig. 12.8 Box layout algorithms

2004). Finally, the constraint layout approach (Schreiber et al., 2009) allows the expression of visualisation requirements including positions of co-substances and specific pathways as constraints and produces a layout by solving these constraints. This approach is particularly well suited in cases when parts of the layout are predefined as shown in Czauderna et al. (2013).

Figure 12.10 shows examples of visualisations computed by these layout algorithms.

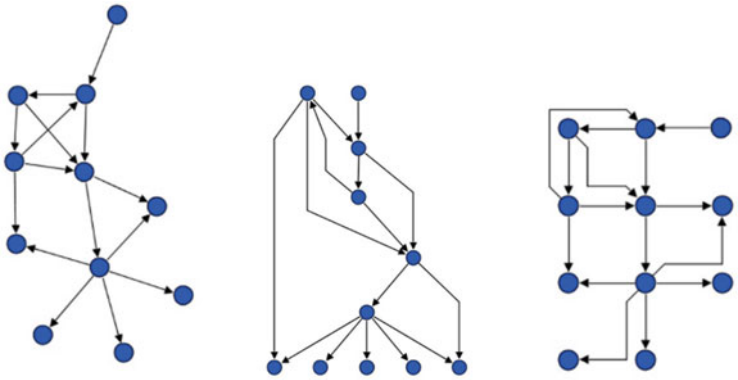

Fig. 12.9 The same network with three different layouts: (from left to right) force-directed, layered (top to bottom) and orthogonal layout

12.4.3 Tools

There are more than 170 tools available, and previous reviews have already compared a number of them. Kono et al. focus in their comparison on pathway representation, data access, data export and exchange, mapping, editing and availability (Kono et al., 2009). Suderman and Hallett compare more than 35 tools relevant in 2007 regarding several aspects of network and data visualisation (Suderman and Hallett, 2007). Rohn et al. present a comparison of 11 non-commercial tools for the network-centred visualisation and analysis of biological data (Rohn et al., 2012). And Gehlenborg et al. present visualisation tools for interaction networks and biological pathways including tools for multivariant omics data visualisation (Gehlenborg et al., 2010). It should be noted that progress in this field is fast, many new tools appeared and old tools obtained new features since then. Well-known tools supporting network visualisation and analysis are:

- *BiNa* (Gerasch et al., 2014; Küntzer et al., 2007) (http://bit.ly/y6ix9i)
- *BioUML* (Kolpakov, 2002; Kolpakov et al., 2006) (http://bit.ly/yIETIt)
- *CellDesigner* (Funahashi et al., 2003, 2006) (http://bit.ly/A0FQiF)
- *CellMicrocosmos* (Sommer and Schreiber, 2017a; Sommer et al., 2010) (http://bit.ly/WJ8cnE)
- *Cytoscape* (Shannon et al., 2003; Smoot et al., 2011) (http://bit.ly/wY2sbG)
- *MapMan* (Thimm et al., 2004; Usadel et al., 2005) (http://bit.ly/3yaa6UE)
- *OMIX* (Droste et al., 2011) (http://bit.ly/wY2sbG)
- *Ondex* (Köhler et al., 2006) (http://bit.ly/AetZjz)
- *Pathway Projector* (Kono et al., 2009) (http://bit.ly/zo5x2M)
- *PathVisio* (van Iersel et al., 2008) (http://bit.ly/zunwxW)
- *SBGN-ED* (Czauderna et al., 2010) (http://bit.ly/17m7KfW)
- *Vanted* (Junker et al., 2006; Rohn et al., 2012) (http://bit.ly/Aigr0T)
- *VisAnt* (Hu et al., 2004, 2009) (http://bit.ly/agZBni)

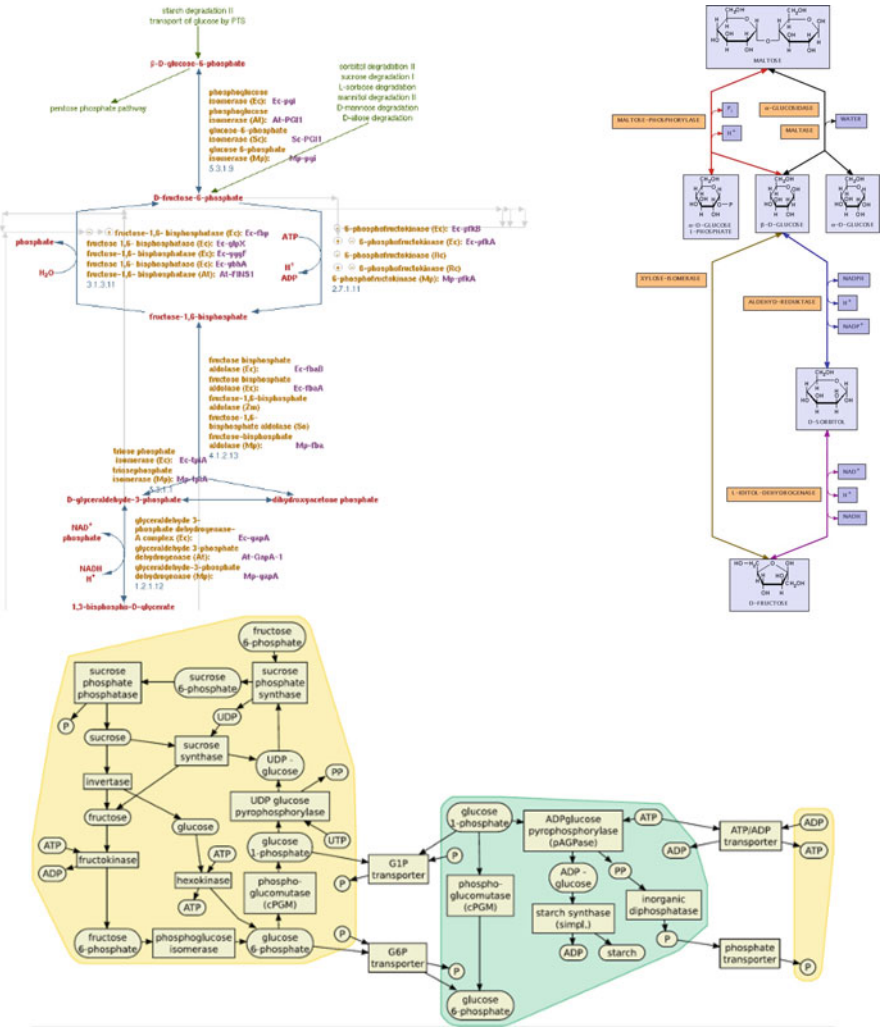

Fig. 12.10 Example visualisations computed by layout algorithms specifically tailored to metabolic networks: (top left) mixed layout (from the MetaCyc webpage), (top right) extended layered layout (from BioPath) and (bottom) constraint layout (from a prototype implementing of the constraint layout approach; note that these networks are not in SBGN notation)

These tools often provide a selection of standard and partly specific layout algorithms for metabolic pathways, the possibility to map additional data onto pathways as well as analysis algorithms.

Note that for a specific metabolic database or pathway collection often several different visualisation methods exists. For example, the visualisation of KEGG pathways can be done with tools and layout methods such as implemented in Pathway projector (Kono et al., 2009), KEGGgraph (Zhang and Wiemann, 2009)

and Vanted (Rohn et al., 2012), can be rebuilt and visualised as in Gerasch et al. (2014), or can be even translated into SBML or SBGN and then layouted and visualised (Czauderna et al., 2013; Wrzodek et al., 2011).

12.5 Visual and Immersive Analytics of Metabolic Pathways and Related Data

For a fast and automatic production of pictures or maps of metabolic networks layout algorithms are very useful. However, a layout is just the first step, and in interactive systems many additional requirements exist, for example, for interactive exploration, structural analysis of the networks, visualisation of experimental data (transcriptomics, metabolomics, fluxes, etc.) in the network context, studying networks in their spatial (3D) context and so on. Here we discuss some typical examples.

12.5.1 Multiscale Representation of Metabolism and Navigation Through Metabolic Networks

Metabolic networks can be huge, and a visualisation may become unreadable due to the large number of objects and connections. Several abstraction and exploration techniques have been transferred from the field of information visualisation to navigate in metabolic networks. As metabolic pathways are hierarchically structured (e. g. carbohydrate metabolism includes a number of sub-pathways such as TCA cycle and glycolysis) this information can be used to help navigating through the network. Often used navigation techniques include clickable overview maps (in many databases and tools, e. g. KEGG (Kanehisa et al., 2002) and iPath (Letunic et al., 2008)), maps showing increasing levels of detail (e. g. the MetaCyc website (Caspi et al., 2012)), interconnected maps (e. g. in GLIEP (Jusufi et al., 2012)), overview and detail diagrams (e. g. method by Garkov et al. (2019)) and interactive extension of pathways within a map (e. g. the method in KGML-ED (Klukas and Schreiber, 2007)).

It should be noted that there is a major obstacle for simple interactive visualisation methods including automatic layout: the mental map of the user (Misue et al., 1995). When browsing through pathways the user builds a mental representation of the objects, their relative position and connections. Basically the user's mental map is its understanding of the network based on the current view. However, sudden or large changes between the current and the next view destroy the user's mental map and therefore hinder interactive understanding of networks. So far there are only few approaches which address this problem.

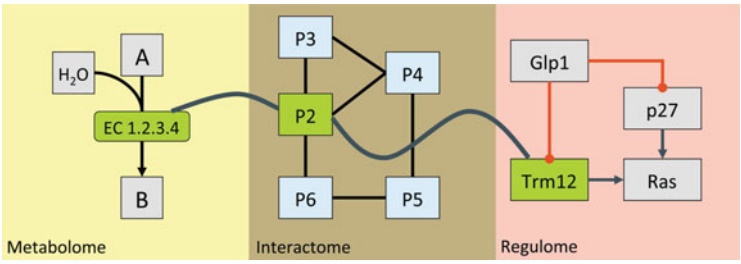

Fig. 12.11 Multivariate networks: Different networks are connected through shared entities (from Kohlbacher et al. (2014))

Metabolic networks can be part of multivariate networks (Kohlbacher et al., 2014) (see Fig. 12.11) and heterogeneous networks (Schreiber et al., 2014), both increase the complexity for representation and navigation. The development of interactive layout algorithms for these structures is still an open research problem, and so far only some initial approaches exists such as the previously mentioned constraint layout approach (Schreiber et al., 2009).

12.5.2 Visual Analytics of the Structure of Metabolic Networks

Analysing structural properties in biological networks can help in gaining new insights, and there are several structural properties of interest in metabolic networks: shortest paths between metabolites which may represent preferred routes, network motifs within the network which may indicate functional properties, different centralities of metabolites and reactions which may correspond to their importance, and clusters or communities within the metabolic network which may structure the network into functional modules. Many network analysis algorithms which can be used for the investigation of structural properties in networks have been developed; overviews can be found in the book of Brandes and Erlebach (2005) and Junker and Schreiber (2008).

"Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces" (Thomas and Cook, 2006). An important aspect of this field is that data analysis is combined with interactive visualisation methods. Here, analytics includes structural analysis of networks as well as investigating additional data as discussed in the following Sect. 12.5.3. Also for metabolic networks interaction plays an important role in visual analytics (Kerren and Schreiber, 2012).

Several tools implement visual analytics methods, for example, Cytoscape, Ondex, Vanted, and VisAnt, often provided via additional Plugins/Add-ons (see also Sect. 12.4.3 for details and references). Some tools also allow the integration of a wide range of other data into the analysis (Rohn et al., 2012). To better understand the analysis results, visualisation algorithms can help by highlighting the relevant

structures such as straightening the shortest path in the map, putting central elements in the centre of the image or laying out the same motifs in the same way. A few specialised layout algorithms have been developed for a better visualisation and graphical investigation of structures and connections in networks such as coordinated perspectives for the analysis of network motifs (Klukas et al., 2006) or visually comparing pathways, for example, to understand metabolic pathways across organisms using two and a half dimensional layout (Brandes et al., 2004).

12.5.3 *Integration and Visualisation of Omics Data in Metabolic Networks*

Data mapping deals with the integration of additional data into metabolic networks. Examples are metabolomics, transcriptomics and fluxomics measurements, which can be mapped on different network elements (such as metabolites, enzymes and reaction edges), see also Figs. 12.1 and 12.12. A common problem for data mapping and subsequent analysis such as correlation analysis and clustering is the usage of correct identifiers, that is having the correct name in both the data and the network. To help the user several tools support mapping tables which translate

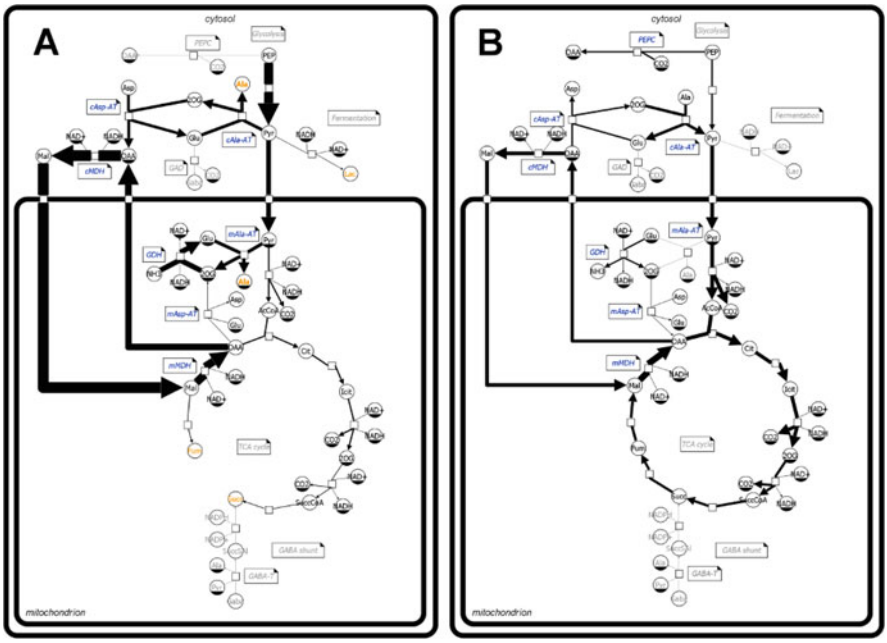

Fig. 12.12 An example of flux visualisation showing the flux distribution in a metabolic network under two scenarios (rendered by Vanted, data from Rolletschek et al. (2011))

identifiers in the data into identifiers in the network, and translation services such as BridgeDB (van Iersel et al., 2010) exist. Depending on the data different diagram types are desired in the vertices or at the edges of the metabolic pathway. Examples are bar charts, pie charts, line charts, box plots and heat maps.

Whereas most tools support the visualisation of data connected to vertices of the network, only few tools provide mapping of data onto edges. Metabolomics data, in particular the results of stable isotope tracer experiments, yield important details on the dynamics of networks, and flux visualisation is important as it provides insights on the integrated response of a biochemical reaction network to environmental changes or genetic modifications. Thus, such representations are also important tools in metabolic engineering (Wiechert, 2001). To support the analysis and understanding of simulated or experimentally measured flux distributions, the visualisation of flux information in the network context is essential and is mainly performed by scaling the width of the reaction edges according to the flux data or by displaying the flux values in the corresponding reaction vertices, see Fig. 12.12. Tools such as FBASimViz (Grafahrend-Belau et al., 2009), MetaFluxNet (Lee et al., 2003), Omix (Droste et al., 2011) and OptFlux (Rocha et al., 2010) support such visualisations.

12.5.4 Immersive Analytics of Metabolic Networks

The visualisation of structures and pathways in 3D has advantages and disadvantages. A good 2D visualisation may be easier to understand and is directly printable on paper. For small molecules 2D visualisation is the method of choice, because the structures are easier to comprehend and the three-dimensional aspects of the structures are typically obvious to an expert. However, the same does not apply to proteins. Visualising proteins as structural formulas is not only impractical, but the function of proteins can only be understood from their three-dimensional structure. This provides arguments for an integration of 2D (mainly Information Visualisation) and 3D (mainly Scientific Visualisation) techniques (Kerren and Schreiber, 2014).

Early work of representing metabolic pathways in 3D by Qeli et al. (2004) and Rojdestvenski et al. (2003; 2002) goes back to the early 2000. In the last years the novel research field of immersive analytics (Chandler et al., 2015) is developed which is concerned with "the use of engaging, embodied analysis tools to support data understanding and decision making" with a focus on immersive (3D) environments (Dwyer et al., 2018). It builds on and combines ideas from the fields of data visualisation, visual analytics, virtual reality, computer graphics and human–computer interaction. The key idea is to get immersed into the data and employ all senses, not only vision. This area has many potential applications in the life and health sciences (Czauderna et al., 2018). Some initial applications for the visualisation and exploration of metabolism in immersive environments include MinOmics, an immersive tool for multi-omics analysis (Maes et al., 2018) and

Fig. 12.13 Exploration of metabolic pathways within the spatial context using an immersive environment based on CAVE2 (stereoscopic 3D) and zSpace (stereoscopic fishtank 3D) (from Sommer and Schreiber (2017b))

the integration and exploration of pathways in a cell environment (Sommer and Schreiber, 2017b) as shown in Fig. 12.13.

12.6 Perspectives

The visual exploration and analytics of metabolic networks is a fast developing field. Although there are already several methods and tools that help in understanding metabolic networks, continuous development is imminent. Here we outline some current directions of research and tool developments in this area:

- Connection to other networks: Metabolism is strongly linked to other biological processes represented, for example, by protein interaction or gene regulatory networks (see also Sect. 12.5). The combined visualisation and easy visual travelling from one network to the next may help in better understanding effects such as regulation of metabolism.
- Context for combined omics data: Although several tools support integration and visualisation of omics data within metabolic networks, the visualisation of complex data sets covering several domains (networks, images, sequences, omics data, etc.) is not yet sufficiently solved. Initial solutions have been presented (e. g. Rohn et al. (2011)), but as more and more such data sets are produced in experiments, there is an increasing need for better analysis and visualisation approaches.

– Mental map preserving layouts: A mental map of a layout is a mental picture of the structure of the layout and helps understanding changing maps (Misue et al., 1995), see also Sect. 12.5.1. It is often used to measure the quality of a dynamic network layout (Archambault et al., 2011), and has been shown to be important in dynamic layouts (Purchase and Samra, 2008; Purchase et al., 2007). Most existing layout algorithms are not mental map preserving and often the same algorithm would produce different visualisations when applied to the same network. Also, there are only a few studies regarding mental map preserving network layouts in visual and immersive analytics (e. g. Kotlarek et al. 2020). However, acceptance of visualisation and exploration methods also depends on better support of the user's mental map and this is an important area for future research.

Biological network visualisation and the layout of metabolic networks is an interesting area in graph drawing (Binucci et al., 2019). More open questions and major problems arising in biological network visualisation are also discussed in Albrecht et al. (2009). Metabolic network and pathway visualisation is only a small aspect of biological data visualisation. As biology aims to provide insights into the overall system, that is into processes on cellular, tissue, organ and even organism levels, visualisation of metabolism has to be embedded into broader visualisation frameworks. Beside networks and related data, other data modalities are also important, for example, imaging data and phenotypical data.

This chapter presented history and state-of-the-art of visualisation and visual analysis of metabolic pathways and networks, provided descriptions of important metabolic network databases and exchange formats, gave a brief overview of often used tools and discussed future research directions including immersive analytics. Methods and tools presented here are a building block of such a broader visualisation framework for biological data.

References

Albrecht M, Kerren A, Klein K, Kohlbacher O, Mutzel P, Paul W, Schreiber F, Wybrow M (2010) On open problems in biological network visualization. In: Eppstein D, Gansner ER (eds) Graph Drawing 17th International Symposium GD 2009. LNCS, vol 5849. Springer, Berlin, pp 256–267

Archambault D, Purchase H, Pinaud B (2011) Animation, small multiples, and the effect of mental map preservation in dynamic graphs. IEEE Trans Visual Comput Graphics 17(4):539–552

Becker MY, Rojas I (2001) A graph layout algorithm for drawing metabolic pathways. Bioinformatics 17(5):461–467

Binucci C, Brandes U, Dwyer T, Gronemann M, von Hanxleden R, van Kreveld M, Mutzel P, Schaefer M, Schreiber F, Speckmann B (2019) 10 reasons to get interested in Graph Drawing. In: Computing and software science—state of the art and perspectives. LNCS, vol 10000, pp 85–104

Boissonnat JD, Cazals F, Flötotto J (2000) 2D-structure drawings of similar molecules. In: Proceedings of the Graph Drawing. Lecture Notes in Computer Sciences, vol 1984. Springer, Berlin, pp 115–126

Bourqui R, Purchase HC, Jourdan F (2011) Domain specific vs generic network visualization: an evaluation with metabolic networks. In: Proceedings of the Australasian conference on user interface (AUIC '11). Australian Computer Society Inc, New York, pp. 9–18

Brandenburg FJ, Forster M, Pick A, Raitner M, Schreiber F (2004) BioPath—exploration and visualization of biochemical pathways. In: Jünger M, Mutzel P (eds) Graph Drawing Software. Springer, Berlin, pp 215–236

Brandes U, Erlebach T (2005) Network analysis: methodological foundations. In: Springer book series lecture notes in computer science tutorial

Brandes U, Dwyer T, Schreiber F (2004) Visual understanding of metabolic pathways across organisms using layout in two and a half dimensions. J Integr Bioinf 1:e2

Büchel, F, Rodriguez N, Swainston N, Wrzodek C, Czauderna T, Keller R, Mittag F, Schubert M, Glont M, Golebiewski M, van Iersel MP, Keating, SM, Rall, M, Wybrow M, Hermjakob H, Hucka M, Kell DB, Müller W, Mendes P, Zell A, Chaouiya C, Saez-Rodriguez J, Schreiber F, Laibe C, Dräger A, Le Novère N (2013) Path2models: large-scale generation of computational models from biochemical pathway maps. BMC Syst Biol 7:116

Caspi R, Altman T, Dreher K, Fulcher C, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 40(1):D742–753

Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD (2019) The MetaCyc database of metabolic pathways and enzymes—a 2019 update. Nucleic Acids Res 48(1):D445–D453

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C (2011) Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res 39(1):D685–690

Chandler T, Cordeil M, Czauderna T, Dwyer T, Glowacki J, Goncu C, Klapperstück M, Klein K, Marriott K, Schreiber F, et al (2015) Immersive analytics. In: Big data visual analytics (BDVA), vol 2015. IEEE, New York, pp 1–8

Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D (2020) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 49(1):D498–D508

Chelliah V, Laibe C, Le Novère, N (2013) BioModels Database: A repository of mathematical models of biological processes. Methods Mol Biol 1021:189–199

Clark AM, Labute P, Santavy M (2006) 2D structure depiction. J Chem Inf Model 46(3):1107–1123

Colmsee C, Czauderna T, Grafahrend-Belau E, Hartmann A, Lange M, Mascher M, Weise S, Scholz U, Schreiber F (2013) Optimas-DW, MetaCrop and Vanted: a case study for data integration, curation and visualisation in life sciences. In: Proceedings of ontologies and data in life sciences, vol LNI P-220, pp 1834–1840

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio, P (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42(1):D472–D477

Czauderna T, Klukas C, Schreiber F (2010) Editing, validating, and translating of SBGN maps. Bioinformatics 26(18):2340–2341

Czauderna T, Wybrow M, Marriott K, Schreiber F (2013) Conversion of KEGG metabolic pathways to SBGN maps including automatic layout. BMC Bioinf 14:250

Czauderna T, Haga J, Kim J, Klapperstück M, Klein K, Kuhlen T, Oeltze-Jafra S, Sommer B, Schreiber F (2018) Immersive analytics applications in life and health sciences. In: Marriott K, Schreiber F, Dwyer T, Klein K, Riche NH, Itoh T, Stuerzlinger W, Thomas BH (eds.) Immersive Analytics. Springer, Berlin, pp 289–330

Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J Chem Inf Comput Sci 32(3):244–255

Demir E, Babur O, Dogrusöz U, Gürsoy A, Nisanci G, Çetin Atalay R, Ozturk M (2002) PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. Bioinformatics 18(7):996–1003

Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Reubenacker O, Samwald M, van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le Novère NL, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD (2010) The BioPAX community standard for pathway data sharing. Nat Biotechnol 28(9):935–942

Di Battista G, Eades P, Tamassia R, Tollis IG (1999) Graph drawing: algorithms for the visualization of graphs. Prentice Hall, New Jersey

Dogrusöz U, Erson EZ, Giral E, Demir E, Babur O, Cetintas A, Colak R (2006) Patika*web*: a web interface for analyzing biological pathways through advanced querying and visualization. Bioinformatics 22(3):374–375

Droste P, Miebach S, Niedenführ S, Wiechert W, Nöh K (2011) Visualizing multi-omics data in metabolic networks with the software Omix: a case study. Biosystems 105(2):154–161

Dwyer T, Marriott K, Isenberg T, Klein K, Riche N, Schreiber F, Stuerzlinger W, Thomas BH (2018) Immersive analytics: An introduction. In: Marriott K, Schreiber F, Dwyer T, Klein K, Riche NH, Itoh T, Stuerzlinger W, Thomas BH (eds) Immersive analytics. Springer, Berlin, pp 1–23

Eades P, Wormald N (1990) Fixed edge length graph drawing is NP-hard. Discrete Appl Math 28:111–134

Fricker PC, Gastreich M, Rarey M (2004) Automated drawing of structural molecular formulas under constraints. J Chem Inf Comput Sci 44(3):1065–1078

Funahashi A, Morohashi M, Kitano H (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. Biosilico 1(5):159–162

Funahashi A, Matsuoka Y, Jouraku A, Kitano H, Kikuchi N (2006) CellDesigner: a modeling tool for biochemical networks. In: Proceedings of the 38th conference on Winter simulation. Winter Simulation Conference, pp 1707–1712

Garkov D, Klein K, Klukas C, Schreiber F (2019) Mental-map preserving visualisation of partitioned networks in Vanted. J Integr Bioinf 16(3):e0026

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31(13):3784–3788

Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH (2010) The NCBI BioSystems database. Nucleic Acids Res 38(1):D492–D496

Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin AC (2010) Visualization of omics data for systems biology. Nat Methods 7:S56–S68

Gerasch A, Faber D, Küntzer J, Niermann P, Kohlbacher O, Lenhof HP, Kaufmann M (2014) BiNA: a visual analytics tool for biological network data. PLoS One 9(2):e87397

Gerasch A, Kaufmann M, Kohlbacher O (2014) Rebuilding KEGG maps: algorithms and benefits. In: 2014 IEEE Pacific Visualization Symposium (PacificVis), pp 97–104

Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M (2002) LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res 30(1):402–404

Grafahrend-Belau E, Weise S, Koschützki D, Scholz U, Junker BH, Schreiber F (2008) MetaCrop—a detailed database of crop plant metabolism. Nucleic Acids Res 36(1):D954–D958

Grafahrend-Belau E, Klukas C, Junker BH, Schreiber F (2009) FBASimViz: interactive visualization of constraint-based metabolic models. Bioinformatics 25(20):2755–2757

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C (2015) ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res 44(1):D1214–D1219

Helson HE (1990) Structure diagram generation. In: Lipkowitz, B, Boyd DB (eds) Reviews in computational chemistry. Wiley-VCH, New York, pp 313–398

Hippe K, Colmsee C, Czauderna T, Grafahrend-Belau E, Junker BH, Klukas C, Scholz U, Schreiber F, Weise S (2010) Novel developments of the metacrop information system for facilitating systems biological approaches. J Integr Bioinf 7(3):125

Hu Z, Mellor J, Wu J, DeLisi C (2004) VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinf 5(1):e17

Hu Z, Hung JH, Wang Y, Chang YC, Huang CL, Huyck M, DeLisi C (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res 37:W115–W121

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin I, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novère N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531

Ihlenfeld WD, Takahashi Y, Abe H, Sasaki S (1994) Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and flexibility. J Chem Inf Comput Sci 34:109–116

Junker BH, Schreiber F (2008) Analysis of Biological Networks. In: Wiley Series on Bioinformatics, Computational Techniques and Engineering. Wiley, New York

Junker BH, Klukas C, Schreiber F (2006) Vanted: A system for advanced data analysis and visualization in the context of biological networks. BMC Bioinf 7(109):1–13

Junker A, Rohn H, Czauderna T, Klukas C, Hartmann A, Schreiber F (2012) Creating interactive, web-based and data-enriched maps using the systems biology graphical notation. Nat Protoc 7:579–593

Jusufi I, Klukas C, Kerren A, Schreiber F (2012) Guiding the interactive exploration of metabolic pathway interconnections. Inf. Visualization 11(2):136–150

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30(1):42–46

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32(1):D277–280

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40(1):D109–D114

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42(1):D199–D205

Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M (2020) KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 49(D1):D545–D551

Karp PD, Mavrovouniotis ML (1994) Representing, analyzing, and synthesizing biochemical pathways. IEEE Expert 9(2):11–21

Kaufmann M, Wagner D (2001) Drawing graphs: methods and models. In Lecture Notes in Computer Science Tutorial, vol 2025. Springer, Berlin

Keating SM, Waltemath D, König M, Zhang F, Dräger A, Chaouiya C, Bergmann FT, Finney A, Gillespie CS, Helikar T, Hoops S, Malik-Sheriff RS, Moodie SL, Moraru II, Myers CJ, Naldi A, Olivier BG, Sahle S, Schaff JC, Smith LP, Swat MJ, Thieffry D, Watanabe L, Wilkinson DJ, Blinov ML, Begley K, Faeder JR, Gomez HF, Hamm TM, Inagaki Y, Liebermeister W, Lister AL, Lucio D, Mjolsness E, Proctor CJ, et al (2020) SBML Level 3: an extensible format for the exchange and reuse of biological models. Mol Syst Biol 8(16):e9110

Kerren, A, Schreiber F (2012) Toward the role of interaction in visual analytics. In: Rose O, Uhrmacher AM (eds) Prodeedings winter simulation conference, p 420

Kerren, A, Schreiber F (2014) Why integrate infovis and scivis? an example from systems biology. IEEE Comput Graphics Appl 34(6):69–73

Kerren A, Kucher K, Li YF, Schreiber F (2017) BioVis Explorer: a visual guide for biological data visualization techniques. PLoS One 12(11):1–14

Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, Latendresse M, Muñiz-Rascado L, Ong Q, Paley S, Peralta-Gil M, Subhraveti P, Velázquez-Ramírez DA, Weaver D, Collado-Vides J, Paulsen I, Karp PD (2016) The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res 45(1):D543–D550

Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH (2015) PubChem Substance and Compound databases. Nucleic Acids Res 44(1):D1202–D1213

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2020) PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49(1):D1388–D1395

Kitano H (2003) A graphical notation for biochemical networks. Biosilico 1(5):169–176

Kitano H, Funahashi A, Matsuoka Y, Oda K (2005) Using process diagrams for the graphical representation of biological networks. Nat Biotechnol 23:961–966

Klukas C, Schreiber F (2007) Dynamic exploration and editing of KEGG pathway diagrams. Bioinformatics 23(3):344–350

Klukas C, Schreiber F, Schwöbbermeyer H (2006) Coordinated perspectives and enhanced force-directed layout for the analysis of network motifs. In: Misue K, Sugiyama K, Tanaka J (eds) Proceedings Asia-Pacific symposium on information visualization (APVis'06). CRPIT, vol. 60. ACS, New York, pp. 39–48

Köhler J, Baumbach J, Taubert J, Specht M, Skusa A, Rüegg A, Rawlings C, Verrier P, Philippi S (2006) Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics 22(11):1383–1390

Kohlbacher O, Schreiber F, Ward MO (2014) Multivariate networks in the life sciences. In: Multivariate network visualization, Springer, Berlin, pp 61–73

Kolpakov FA (2002) BioUML—framework for visual modeling and simulation of biological systems. In: Proceedings of the international conference on bioinformatics of genome regulation and structure, pp 130–133

Kolpakov F, Puzanov M, Koshukov A (2006) BioUML: Visual modeling, automated code generation and simulation of biological systems. In: Proceedings of the fifth international conference on bioinformatics of genome regulation and structure (BGRS), pp 281–284

Kono N, Arakawa K, Ogawa R, Kido N, Oshita K, Ikegami K, Tamaki S, Tomita M (2009) Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. PLoS One 4(11):e7710

Kono N, Arakawa K, Ogawa R, Kido N, Oshita K, Ikegami K, Tamaki S, Tomita M (2009) Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google maps API. PLoS One 4(11):e7710

Kotlarek J, Kwon OH, Ma KL, Eades P, Kerren A, Klein K, Schreiber F (2020) A study of mental maps in immersive network visualization. In: Proceedings of the IEEE Pacific visualization. IEEE, New York, pp 1–10

Krieger CJ, Zhang P, Müller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 32:438–442

Küntzer J, Backes C, Blum T, Gerasch A, Kaufmann M, Kohlbacher O, Lenhof HP (2007) BNDB—the biochemical network database. BMC Bioinf 8:367.1–9

Lee SY, Lee DY, Hong SH, Kim TY, Yun H, Oh YO, Park S (2003) MetaFluxNet, a program package for metabolic pathway construction and analysis, and its use in large-scale metabolic flux analysis of *escherichia coli*. Genome Inform 14:23–33

Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem M, 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 T, 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 27:735–741

Letunic I, Yamada T, Kanehisa M, Bork P (2008) iPath: interactive exploration of biochemical pathways and networks. Trends Biochem Sci 33:101–103

Lombardot T, Morgat A, Axelsen KB, Aimo L, Hyka-Nouspikel N, Niknejad A, Ignatchenko A, Xenarios I, Coudert E, Redaschi N, Bridge A (2018) Updates in Rhea: SPARQLing biochemical reaction data. Nucleic Acids Res 47(1):D596–D600

Maes A, Martinez X, Druart K, Laurent B, Guégan S, Marchand CH, Lemaire SD, Baaden M (2018) MinOmics, an integrative and immersive tool for multi-omics analysis. J Integr Bioinf 15(2):20180006

Malik-Sheriff RS, Glont M, Nguyen TVN, Tiwari K, Roberts MG, Xavier A, Vu MT, Men J, Maire M, Kananathan S, Fairbanks EL, Meyer JP, Arankalle C, Varusai TM, Knight-Schrijver V, Li L, Dueñas-Roca C, Dass G, Keating SM, Park YM, Buso N, Rodriguez N, Hucka M, Hermjakob H (2019) BioModels—15 years of sharing computational models in life science. Nucleic Acids Res 48(1):D407–D415

Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, Millerand RA, Digles D, Lopes EN, Ehrhart F, Dupuis LJ, Winckers LA, Coort SL, Willighagen EL, Evelo CT, Pico AR, Kutmon M (2021) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res 49(1):D613–D621

Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P (2009) Reactome knowledge-base of human biological pathways and processes. Nucleic Acids Res 37(1):D619–622

Mendes P (2000) Advanced visualization of metabolic pathways in PathDB. In: Proceedings of the conference on plant and animal genome

Mendes, P, Bulmore DL, Farmer AD, Steadman PA, Waugh ME, Wlodek ST (2000) PathDB: a second generation metabolic database. In: Hofmeyr JHS, Rohwer JM, Snoep JL (eds) Proceedings of the International BioThermoKinetics Meeting. Stellenbosch University Press, Stellenbosch, pp 207–212

Mi H, Thomas P (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol 563:123–140

Mi H, Schreiber F, Moodie SL, Czauderna T, Demir E, Haw R, Luna A, Le Novère N, Sorokin AA, Villéger A (2015) Systems biology graphical notation: activity flow language level 1 version 1.2. J Integr Bioinf 12(2):265

Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, Thomas PD (2020) PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res 49(1):D394–D403

Michal G (1968) Biochemical pathways (Poster). Boehringer Mannheim (latest version 2005)

Michal, G (1998) On representation of metabolic pathways. BioSystems 47:1–7

Misue K, Eades P, Lai W, Sugiyama K (1995) Layout adjustment and the mental map. J Vis Lang Comput 6:183–210

Naithani S, Preece J, D'Eustachio P, Gupta P, Amarasinghe V, Dharmawardhana PD, Wu G, Fabregat A, Elser JL, Weiser J, Keays M, Fuentes AMP, Petryszak R, Stein LD, Ware D, Jaiswal P (2016) Plant Reactome: a resource for plant pathways and comparative analysis. Nucleic Acids Res 45(1):D1029–D1039

Naithani S, Gupta P, Preece J, D'Eustachio P, Elser JL, Garg P, Dikeman DA, Kiff J, Cook J, Olson A, Wei S, Tello-Ruiz MK, Mundo AF, Munoz-Pomer A, Mohammed S, Cheng T, Bolton E, Papatheodorou I, Stein L, Ware D, Jaiswal P (2019) Plant Reactome: a knowledgebase and resource for comparative pathway analysis. Nucleic Acids Res 48(1):D1093–D1103

O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. J Cheminf 3:33

Purchase HC, Samra A (2008) Extremes are better: Investigating mental map preservation in dynamic graphs. In: Diagrammatic representation and inference. LNCS, vol 5223. Springer, Heidelberg, pp. 60–73

Purchase HC, Hoggan E, Görg C (2007) How important is the "Mental Map"?—an empirical investigation of a dynamic graph layout algorithm. In: Graph Drawing, LNCS, vol 4372. Springer, Heidelberg, pp. 184–195

Qeli E, Wiechert W, Freisleben B (2004) 3D visualization and animation of metabolic networks. In: Proceedings of the 18th European simulation multiconference, pp 1–4

Rocha I, Maia P, Evangelista P, Vilaca P, Soares S, Pinto JP, Nielsen J, Patil KR, Ferreira EC, Rocha M (2010) OptFlux: an open-source software platform for in silico metabolic engineering. BMC Syst Biol 4:45

Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, Mistry H, Mosier L, Dlin J, Wen Q, O'Callaghan C, Li W, Elder G, Smith PT, Dallago C, Cerami E, Gross B, Dogrusoz U, Demir E, Bader GD, Sander C (2019) Pathway commons 2019 update: integration, analysis and exploration of pathway data. Nucleic Acids Res 48(1):D489–D497

Rohn H, Klukas C, Schreiber F (2011) Creating views on integrated multidomain data. Bioinformatics 27(13):1839–1845

Rohn H, Hartmann A, Junker A, Junker BH, Schreiber F (2012) Fluxmap: a Vanted add-on for the visual exploration of flux distributions in biological networks. BMC Syst Biol 6:33.1–9

Rohn H, Junker A, Hartmann A, Grafahrend-Belau E, Treutler H, Klapperstück M, Czauderna T, Klukas C, Schreiber F (2012) Vanted v2: a framework for systems biology applications. BMC Syst Biol 6(139):1–13

Rojdestvenski I (2003) Metabolic pathways in three dimensions. Bioinformatics 19(18):2436–2441

Rojdestvenski I, Cottam M (2002) Visualizing metabolic networks in VRML. In: Proceedings of the international conference on information visualisation (IV'02), pp 175–180

Rolletschek H, Melkus G, Grafahrend-Belau E, Fuchs J, Heinzel N, Schreiber F, Jakob PM, Borisjuk L (2011) Combined non-invasive imaging and modeling approaches reveal metabolic compartmentation in the barley endosperm. Plant Cell 23(8):3041–3054

Rougny A, Toure V, Moodie S, Balaur I, Czauderna T, Borlinghaus H, Dogrusoz U, Mazein A, Dräger A, Blinov ML, Villeger A, Haw R, Demir E, Mi H, Sorokin A, Schreiber F, Luna A (2019) Systems biology graphical notation: Process description language level 1 version 2. J Integr Bioinf 16(3):22

Rücker C, Meringer M (2002) How many organic compounds are graph-theoretically nonplanar? MATCH Commun Math Comput Chem 45:153–172

Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Söhngen C, Stelzer M, Thiele J, Schomburg D (2011) BRENDA, the enzyme information system in 2011. Nucleic Acids Res 39(1):D670–D676

Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T, Kahn D, Rhee SY (2017) Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. Plant Physiol 173(4):2041–2059

Schreiber F (2002) High quality visualization of biochemical pathways in BioPath. In Silico Biol 2(2):59–73

Schreiber F, Dwyer T, Marriott K, Wybrow M (2009) A generic algorithm for layout of biological networks. BMC Bioinf 10:375

Schreiber F, Colmsee C, Czauderna T, Grafahrend-Belau E, Hartmann A, Junker A, Junker BH, Klapperstück M, Scholz U, Weise S (2012) MetaCrop 2.0: managing and exploring information about crop plant metabolism. Nucleic Acids Res 40(1):D1173–D1177

Schreiber F, Kerren A, Börner K, Hagen H, Zeckzer D (2014) Heterogeneous networks on multiple levels. In: Multivariate network visualization. Springer, Berlin, pp 175–206

Schreiber F, Sommer B, Czauderna T, Golebiewski M, Gorochowski TE, Hucka M, Keating SM, König M, Myers C, Nickerson D, Waltemath D (2020) Specifications of standards in systems and synthetic biology: status and developments in 2020. J Integr Bioinf 17(2–3):22

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27(3):431–432

Sommer, B, Schreiber F (2017a) Integration and virtual reality exploration of biomedical data with CmPI and VANTED. Inf Technol 59(4):181–190

Sommer B, Schreiber F (2017b) Integration and virtual reality exploration of biomedical data with CmPI and Vanted. Inf Technol 59(4):181–190

Sommer B, Künsemöller J, Sand N, Husemann A, Rumming M, Kormeier B (2010) CELLmicrocosmos 4.1—an interactive approach to integrating spatially localized metabolic networks into a virtual 3D cell environment. In: Fred ALN, Filipe J, Gamboa H (eds) Bioinformatics 2010—Proceedings of the International Conference on Bioinformatics, pp 90–95

Sorokin AA, Le Novère N, Luna A, Czauderna T, Demir E, Haw R, Mi H, Moodie SL, Schreiber F, Villéger A (2015) Systems biology graphical notation: Entity relationship language level 1 version 2. J Integr Bioinf 12(2):264

Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen EL (2003) The Chemistry Development Kit (CDK): an Open-Source Java Library for Chemo- and Bioinformatics. J Chem Inf Comput Sci 43(2):493–500

Stryer L (1988) Biochemistry. W H Freeman, New York

Suderman, M, Hallett M (2007) Tools for visually exploring biological networks. Bioinformatics 23(20):2651–2659

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914–939

Thomas JJ, Cook KA (2006) A visual analytics agenda. IEEE Comput Graph Appl 26(1):10–13

Usadel B, Nagel A, Thimm O, Redestig H, Blaesing OE, Palacios-Rojas N, Selbig J, Hannemann J, Piques MC, Steinhauser D, Scheible WR, Gibon Y, Morcuende R, Weicht D, Meyer S, Stitt M (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. Plant Physiol 138:1195–1204

van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C (2008) Presenting and exploring biological pathways with PathVisio. BMC Bioinf 9:399.1–399.9

van Iersel M, Pico A, Kelder T, Gao J, Ho I, Hanspers K, Conklin B, Evelo C (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinf 11:5

Van Iersel MP, Villéger AC, Czauderna T, Boyd SE, Bergmann FT, Luna A, Demir E, Sorokin A, Dogrusoz U, Matsuoka Y, Funahashi A, Aladjem M, Mi H, Moodie S, Kitano H, Le Novère N, Schreiber F (2012) Software support for SBGN maps: SBGN-ML and LibSBGN. Bioinformatics 28(15):2016–2021

Waltemath D, Karr J, Bergmann F, Chelliah V, Hucka M, Krantz M, Liebermeister W, Mendes P, Myers C, Pir P, Alaybeyoglu B, Aranganathan N, Baghalian K, Bittig A, Burke P, Cantarelli M, Chew YH, Costa R, Cursons J, Schreiber F (2016) Toward community standards and software for whole-cell modeling. IEEE Trans Biomed Eng 63(10):2007–2014

Waltemath D, Golebiewski M, Blinov ML, Gleeson P, Hermjakob H, Hucka M, Inau ET, Keating SM, König M, Krebs O, Malik-Sheriff RS, Nickerson D, Oberortner E, Sauro HM, Schreiber F, Smith L, Stefan MI, Wittig U, Myers CJ (2020) The first 10 years of the international coordination network for standards in systems and synthetic biology (COMBINE). J Integr Bioinf 17(2–3):20200005

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37(7):W623–W633

Weininger D (1988) SMILES, a chemical language and information system 1. Introduction to methodology and encoding rules. J Chem Inf Model 28:31

Weise S, Grosse I, Klukas C, Koschützki D, Scholz U, Schreiber F, Junker BH (2006) Meta-All: a system for managing metabolic pathway information. BMC Bioinf 7:e465

Wiechert W (2001) 13C metabolic flux analysis. Metab Eng 3(3):195–206

Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, Algaa E, Weidemann A, Sauer-Danzwith H, Mir S, Krebs O, Bittkowski M, Wetsch E, Rojas I, Müller W (2012) SABIO-RK—database for biochemical reaction kinetics. Nucleic Acids Res 40(1):D790–D796

Wittig U, Rey M, Weidemann A, Kania R, Müller W (2017) SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. Nucleic Acids Res 46(1):D656–D660

Wrzodek C, Dräger A, Zell A (2011) KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. Bioinformatics 27(16):2314–2315

Zhang JD, Wiemann S (2009) KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. Bioinformatics 25:1470–1471

Zhang P, Dreher K, Karthikeyan A, Chi A, Pujar A, Caspi R, Karp P, Kirkup V, Latendresse M, Lee C, Mueller LA, Muller R, Rhee SY (2010) Creation of a genome-wide metabolic pathway database for *populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. Plant Physiol 153(4):1479–1491

Chapter 13
Comprehensive Open-Source Petri Net Toolchain for Modeling and Simulation in Systems Biology

Christoph Brinkrolf and Lennart Ochel

Abstract In systems biology, the process of modeling and the process of simulating the biological system of interest are essential. Implementing these processes in research marks a major difference between traditional biology and systems biology. There are several approaches to model a system e.g., discrete, continuous, stochastic, and hybrid modeling. Depending on the systems's properties, a matching modeling approach needs to be selected as well as a method or tool which offers access to this approach. Such methods or tools could for instance be rule based, system of ODEs, and Petri nets. In this chapter we will focus on a Petri net formalism that covers discrete, continuous, and stochastic models among other features. The open-source implementation of the editor used for modeling and visualization of simulation results, as well as the open-source implementation of the Petri net library and simulation engine makes this toolchain with all its implemented features and supported Petri net formalism unique. The Petri net library PNlib is written in Modelica, an equation-based modeling language for cyber-physical systems. VANESA, the editor, is written in Java and exports the Petri net model of the biological system for simulation to Modelica. The exported model and the PNlib are compiled by the open-source OpenModelica Compiler (OMC), executed, and simulation results are made available in VANESA. VANESA can be downloaded at: http://agbi.techfak.uni-bielefeld.de/vanesa.

Keywords Petri net · Modeling · Simulation · xHPN · PNlib · Modelica · OpenModelica · Systems biology · Hybrid functional timed simulation

C. Brinkrolf (✉)
Bioinformatics Department, Bielefeld University, Bielefeld, Germany
e-mail: cbrinkro@cebitec.uni-bielefeld.de

L. Ochel
RISE AB, Linköping, Sweden
e-mail: lennart.ochel@ri.se

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_13

13.1 Introduction to Systems Biology

Systems biology is a branch of life sciences which aims to understand a biological system of interest on system-level in its entirety. This differs from e.g., biochemistry where for example one particular enzyme and its kinetics is investigated solely. For representation and further investigation of the biological system, a model is used. Usually, different data types (kinetics, chemical and physiological parameters, sequencing data) from literature, experiments, databases, and further data sources are integrated. Thus, the choice of an appropriate modeling approach regarding the integrated data and abstraction level of the model is crucial. Only the data which can be represented by the chosen modeling approach is able to get integrated into the model. Once the model is created, it can be analyzed (e.g., dependencies and connectivity of components of the model) and missing parameters could be estimated which results in a refinement of the model. The simulation of the model and the analysis of the simulation result with e.g., existing experimental data from wet lab could also lead to an improvement of the model. But the major advantage of simulation is the possibility to manipulate the model, test and formulate hypotheses, and predict future behavior of the biological system. These processes (modeling, simulation, formulating, and testing of hypotheses) lead to a better understanding of the biological system of interest and might reduce costly (chemicals, manpower, time, other resources) wet lab experiments.

Choice of Modeling Approach
The choice of modeling approach depends mainly on the abstraction level of the model, the data and its data types to describe the model, the included or excluded process of simulation, and the availability of tools which supports the chosen modeling approach as well as offering its simulation. There are several classifications of modeling approaches, and some data types can be described with several approaches. A kinetic of an enzyme for example can be modeled with a set of rules or with a set of ordinary differential equations (ODE). Time can be omitted, modeled as discrete time intervals, or be treated continuously. The amount of a certain component (concentration) can be represented also as a discrete number or continuously. It is also possible to combine discrete and continuous values as a hybrid model. Further, the model and its simulation can be either deterministic or influenced by probabilistic factors.

General modeling approaches could be categorized as: graphs (such as Boolean networks, state charts, Bayesian networks, Petri nets), rule-based systems, system of mathematical equations (set of ODE, differential-algebraic system of equations), grammars and corresponding automata, among others. Hybrid models consist of a combination of more than one modeling approach. A broader overview of modeling techniques in systems biology is presented in Bartocci and Lió (2016).

In this chapter, we will focus on the modeling and simulation using Petri nets. The basic Petri net formalism is a discrete approach omitting time to describe parallel behavior. By time, it got extended by several modeling aspects and as of today, different kinds of Petri net formalism exist which may combine one or more

extensions of the basic Petri net approach. They differ in the treatment of time (no time, discrete time, sophisticated modeling of continuous time), representation and change of concentrations (discrete and continuous amounts), concepts of inhibition, concepts of fuzzy logic, concept of stochastic events, and many more. Thus, Petri nets are a versatile and powerful approach which are not limited to systems biology.

13.2 Petri Nets

In this section, an informal overview of important Petri net concepts and extensions is given. Formal and more detailed definitions are given in David and Alla (2010). Petri nets were introduced by Carl Adam Petri in 1962.

A discrete Petri net is a bipartite graph with two disjoint and finite sets of two types of nodes: places and transitions. Directed arcs always connect two nodes of different types. Arcs connecting two places or two transitions are not allowed. Each place holds a non-negative number (integer) of tokens. The vector which assigns each place its number of tokens is called marking. A positive integer arc weight is assigned to each arc. The default arc weight is 1.

The change of marking is performed by the firing events of transitions. A transition may only fire if it is enabled. Each transition has a set of pre-places with arcs from each pre-place to the transition and a set of post-places with arcs from the transition to each post-place. If for all pre-places the number of token in each place is greater or equal to the arc weight of this specific place to the transition, the transition is enabled. Transitions with an empty set of pre-places are always enabled. When a transition fires, the number of tokens defined by the arc weight are removed from each pre-place. For each place of the set of post-places, the number of tokens defined by the arc weight is added. Firing of a transition often changes the marking of the Petri net and also the sum of all tokens in the Petri net. The time when a transition fires is not defined and thus depends on the concrete implementation.

In a graphical representation, places are shown as circles while transitions are shown as rectangles or bars. Tokens are usually drawn inside of each place or the number of tokens for each place is shown inside of each place. The default arc weight of 1 is often omitted in the graphical representation. Some implementations show which transitions are enabled and some implementations even animate the token flow from and to places.

In Fig. 13.1 a discrete Petri net of the abstract model of the photosynthesis as a single reaction is shown. It represents the chemical reaction of $6\,CO_2 + 12\,H_2O \rightarrow 6O_2 + 6\,H_2O$ + glucose. Initial number of tokens is shown in the places. Running the simulation for 10 time steps results in an increase of glucose and a decrease of H_2O as shown in Fig. 13.2. After firing 5 times, the transition *Photosynthesis* is not enabled anymore since the place H_2O holds 6 tokens but 12 are required. In the following subsections, this example will be improved and extended as new features and concepts are introduced.

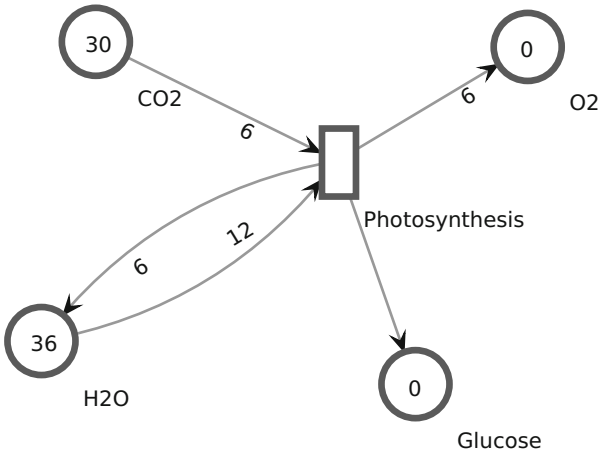

Fig. 13.1 Discrete Petri net of the photosynthesis modeled as a single reaction

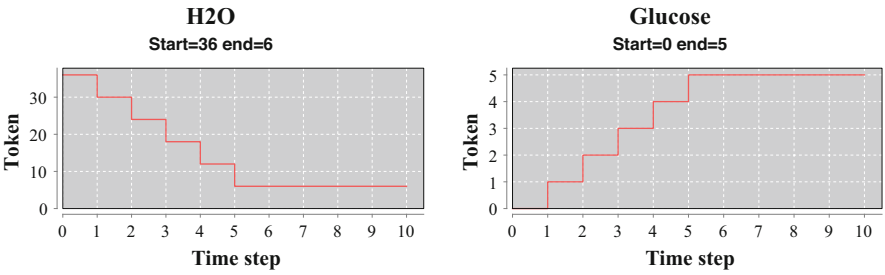

Fig. 13.2 Selected simulation results of discrete Petri net shown in Fig. 13.1, simulated for 10 time steps

13.2.1 Continuous Petri Net

A continuous Petri net is a bipartite graph with two disjoint and finite, not empty, sets of continuous places and continuous transitions. Each place holds a non-negative real number of tokens. In some literature, these continuous tokens are called marks. The nodes are connected by directed arcs, but similar to discrete Petri nets, two nodes of the same type are not allowed to be connected. The arc weight assigned to each arc is a positive rational number. If for each place of pre-places of a specific transition the number of tokens is greater or equal to the arc weight, the transition is enabled. By firing, the number of tokens indicated by the corresponding arc weight from pre-places is subtracted and for post-places tokens are added. Continuous transitions are able to fire a real number of times. The tokens added and deleted are multiplied by this factor.

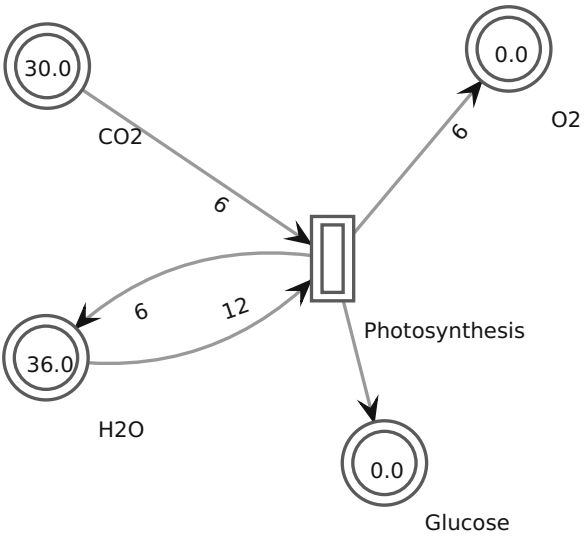

Fig. 13.3 Continuous Petri net of the photosynthesis modeled as a single reaction. Firing speed of transition *Photosynthesis* is set to $v_{max} = 1.5$

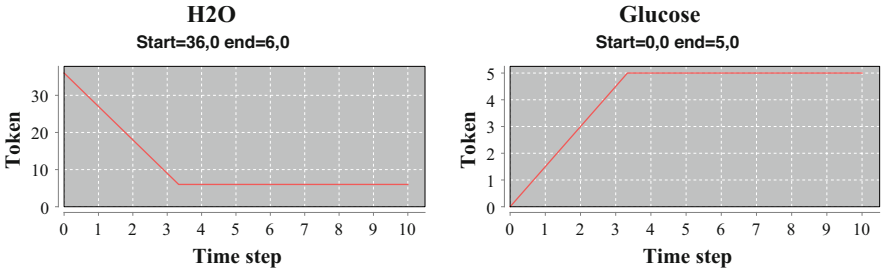

Fig. 13.4 Selected simulation results of continuous Petri net shown in Fig. 13.3, simulated for 10 time steps

As graphical representation, continuous places and transitions are shown as two nested circles and rectangles. The extension of continuous Petri nets by a concept of time are defined as timed continuous Petri nets. The only difference is that a maximal speed function as a rational number is associated with each transition which is treated as an additional factor during firing (David and Alla, 2010).

In Fig. 13.3 the photosynthesis reaction from Sect. 13.2 is modeled as a timed continuous Petri net. Initial number of tokens as well as arc weights remains the same, but the maximal speed function of the continuous transition is set to $v_{max} = 1.5$. As selected simulation results show in Fig. 13.4, there is an increase of glucose and a decrease of H_2O. The number of tokens in each place after 10 time steps of simulation are the same as for the discrete Petri net.

13.2.2 Hybrid Petri Net, Hybrid Dynamic Net, and Functional Net

A hybrid Petri (HPN) (David and Alla, 2010) net is a combination of a discrete and a continuous Petri net. Thus, it consists of a set of discrete places, continuous places, discrete transitions, and continuous transitions. The sets of places and transitions are disjointed. Discrete places hold an integer number of tokens and continuous places hold a real number of tokens. Directed arcs connect places with transitions and vice versa. In contrast to continuous and discrete Petri nets, there are further rules for arcs: It must be assured that only an integer number of tokens are taken from or added to discrete places. Thus, in general, arcs from continuous transitions to discrete places and arcs from discrete places to continuous transitions are not allowed. There are two more arcs: inhibitory and test arcs, which only connect a pre-place with a transition. Enabling and firing of transitions are very similarly defined as for discrete and continuous Petri nets. Besides, an inhibitory arc enables a transition only if the number of tokens in the place is 0. If a transition fires, no tokens are subtracted from pre-places which are connected to the transition with a test arc. For timed hybrid Petri nets, a positive or zero rational number is assigned to each discrete transition as its timing and to each continuous transition as its flow rate.

A hybrid dynamic net (HDN) (Drath et al., 1998) is very similar to HPN. The two major differences are that HDN does not allow different number of tokens subtracted and added by firing of a transition. Thus, the arc weights for all incoming and outgoing arcs of a transition have to have the same value. The second difference is the firing speed of continuous transitions. Beside a constant, the speed can be given as a function of values in the places. The concept of functions of values in the places is also used in functional nets (Hofestädt and Thelen, 1998) which are an extension of discrete Petri nets. In functional nets, the arc weight is either a positive integer or a function depending on the places and its values.

Figure 13.5 shows the photosynthesis introduced in Sect. 13.2 reaction as a hybrid Petri net. In addition to the timed continuous Petri net shown in Fig. 13.3, the sunlight intensity and one inhibitor are taken into account and modeled as discrete elements. The sunlight intensity has to be greater than 100 to enable the transition *Photosynthesis*. As long as the inhibitor is present, the transition *Photosynthesis* is not enabled. The transition *Inhibitor_decay* fires with a delay of 2. Selected simulation results are shown in Fig. 13.6. Similar to the simulation result of the continuous Petri net, there is an increase of glucose and a decrease of H_2O. The amount of glucose and H_2O after 10 time steps are the same as in the simulation result of the continuous Petri net, but the reaction is 4 time steps delayed due to the presence of the inhibitor. After 4 time steps, the place *Inhibitor* has zero tokens and the transition *Photosynthesis* is enabled. Tokens from the place *sunlight_intensity* are not consumed and thus modeled with a test arc. The number of tokens does not change during the simulation and remain 500.

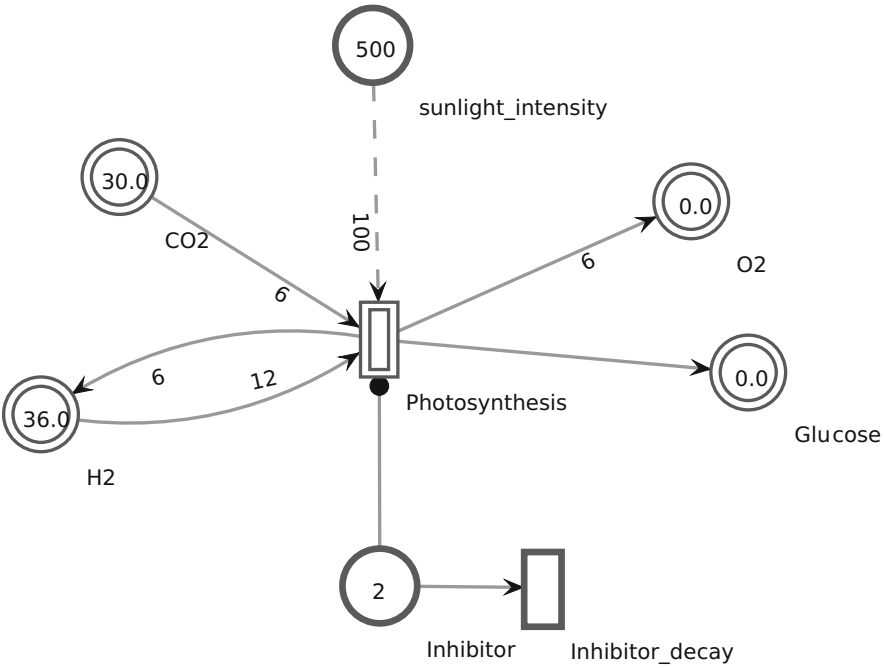

Fig. 13.5 Hybrid Petri net of the photosynthesis modeled as a single reaction. An inhibitor and its decay are modeled as discrete entities as well as the sunlight intensity

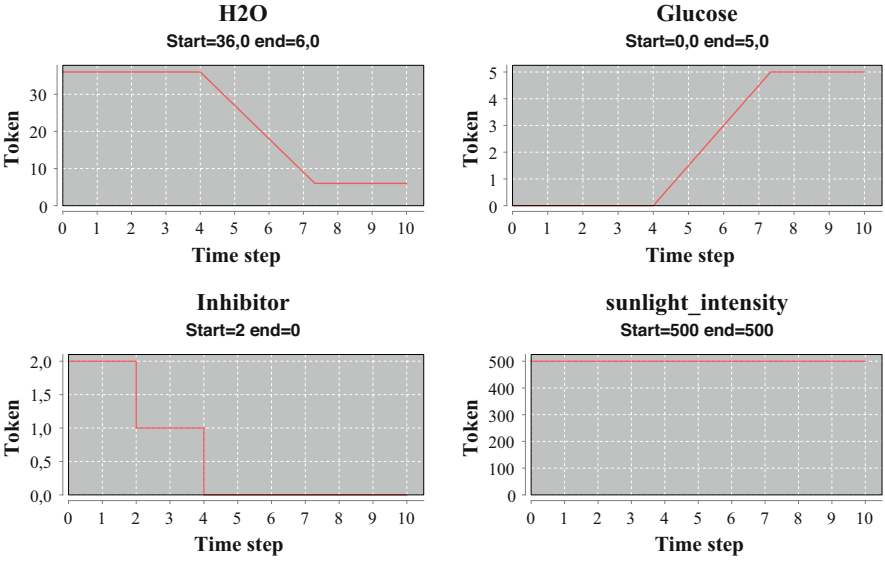

Fig. 13.6 Selected simulation results for HPN shown in Fig. 13.5, simulated for 10 time steps

13.2.3 Hybrid Functional Petri Net

A hybrid functional Petri net (HFPN) (Matsuno et al., 2003) shares the same elements as a hybrid Petri net and extends it by the features of hybrid dynamic nets and functional nets. Broadly speaking, a HFPN is a HPN allowing maximum speed functions of values in the places for continuous transitions. Functions of values in the places are also allowed as arc weights, as it was proposed for functional nets. A non-negative integer-valued delay function can be assigned to discrete transitions. A firing condition is a property of all transitions which allows further control of firing, since a transition may only fire if its firing condition is true.

The Petri net elements of the HFPN version of the photosynthesis model are the same as for the HPN shown in Fig. 13.5. Only the maximal speed function of the transition *Photosynthesis* is set to $v_{max} = v \cdot \min(H_2O, CO_2) \cdot rate$ with parameters $v = 1.5$ and $rate = 0.25$ as a non-linear function. Similar to the previous examples, there is an increase of glucose and a decrease of H_2O, but the gradient is not constant due to the parameterized maximal speed function. The simulation results of the inhibitor and sunlight intensity remain the same (Fig. 13.7).

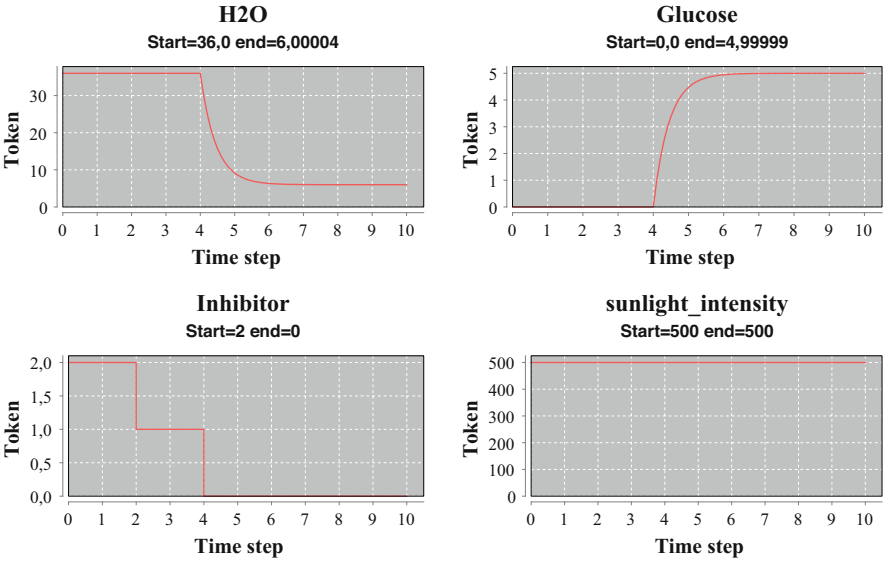

Fig. 13.7 Selected simulation results of HFPN, simulated for 10 time steps

13.2.4 Further Petri Net Concepts

Capacities (Genrich and Lautenbach, 1981) can be used to define lower and an upper limit of tokens in each place. Lower capacities of pre-places and upper capacities of post-places need to be taken into account for the determination if a transition is enabled. Thus, capacities have a direct influence on the definition of firing.

A conflict (David and Alla, 2010) can occur if a place is connected to more than one transition and more than one transition are enabled, but the place does not hold enough tokens that all enabled transitions can fire. Thus, it has to be determined which transitions will fire. Such a conflict can be solved, for example, by assigning priorities or probabilities to the concurrent transitions. Especially in hybrid Petri nets and its extensions other types of conflicts need to be solved which requires conflict solving strategies.

A stochastic transition (David and Alla, 2010) is a discrete transition with a delay determined by a non-negative random variable, which allows modeling probabilistic behavior.

Colored Petri nets (Jensen, 1987) extend discrete Petri nets by distinguishable types (colored) of tokens. Each colored Petri net can be unfolded into a (much) larger discrete Petri net. Thus, colored Petri nets can represent large discrete Petri nets in a small and condensed way.

Uncertainty can be represented using Fuzzy logic, which is implemented in Fuzzy Petri nets (Cardoso et al., 1996).

One concept of hierarchical Petri nets is introduced in Fehling (1993). It allows the refinement of places and transitions. For each hierarchical node, there is a sub-Petri net which represents the complexity of the node. Usually, the hierarchical node is shown instead of its sub-Petri net. Nesting of hierarchical and regular Petri net nodes is not limited to a certain level. Thus, hierarchical Petri nets allow modeling huge systems in a structured way and give visual aid to hide and show only parts of interest of the Petri net at a time.

13.2.5 Advantages of Petri Nets

Petri nets are widely used to model, simulate, and analyze systems, especially in the field of biology for more than 30 years (Fuss, 2013). Recently, HFPN and its simulation results were used for kinetic parameter estimation (Li et al., 2021), continuous fuzzy Petri net were applied to model uncertainty and lack of information (e.g. exact kinetic parameter) of a system (Liu et al., 2019), and a discrete Petri net and its analysis were used to discover unknown properties and dependencies within the modeled system (Gutowska et al., 2020). Their basic structure that states are represented as places and actions are represented as transitions as well as the firing rule are easy to understand. Given a graphical user interface, Petri nets are intuitive to use even for users who do not have a strong

background in computer science or mathematics, since all mathematical definitions are hidden.

Sometimes, when starting the process of modeling, only little information is available. In this case, it is already possible to start with a simple and qualitative model using discrete Petri nets. Once more information is present, the Petri net can be easily extended. Using hybrid (functional) Petri nets, some parts remain discrete and other parts of the system can be modeled quantitatively using continuous elements along with functions. Thus, the model will grow and improve with each iteration without the need to start from scratch again.

From the scientific point of view, a lot of work and effort has been put already into the theory of Petri nets and the definition of their extensions, which makes their usage very reliable and its simulation result comprehensible. There are also a lot of algorithms available to analyze (mostly discrete) Petri nets. For reachability analysis, all possible markings of a Petri net are calculated, and it can be decided if a certain marking is reachable. Since this algorithm requires exponential time and space (Lipton, 1976), it should only be applied for small or bounded Petri nets (e.g., Petri nets with upper capacity for all places). A similar analysis is the computation of coverability of a certain marking. It is less strict than reachability, because a marking does not need to be matched exactly, but all elements of the marking need to be matched by an equal or higher value. The internally computed coverability graph still needs exponential resources in time and space depending on the size of the Petri net, but its computation is always finite (Finkel, 1993).

There are other characteristics which describe a Petri net, for example its liveness and boundedness. The degree of liveness describes how often each transition is at least able to fire (never, sometimes, arbitrary times, infinite times, always) (Murata, 1989). The degree of boundedness of a Petri net gives the maximum number of tokens of all places in all reachable markings (Murata, 1989). If all places of a Petri net have an upper capacity, the degree of boundedness is simply the maximum number of all upper capacities.

The property that a Petri net is a graph makes it possible to apply general graph analysis algorithms. Beside ordinary graph algorithms, there are also algorithms to analyze the structure of a Petri net for reduction (Ackermann et al., 2012) or decomposition (Sackmann et al., 2006). The decomposition tries to find (mostly) independent subnets, which can then be analyzed with fewer resources.

13.3 Requirements on Petri Nets in Systems Biology and Available Tools

Requirements for Modeling in Systems Biology

One limitation of modeling is the selected modeling approach. If the approach only allows a qualitative way of modeling, it will be impossible to transfer quantitative behavior of the system of interest to the model. In systems biology, metabolic

reactions and pathways are a great field of interest. Modeling kinetic behavior of enzymes is not possible with discrete Petri net elements, but with continuous places and transitions continuous enzymatic reaction can be modeled. Stoichiometry of reactions can be modeled by arc weights of a continuous Petri net, but inhibitors cannot be considered. With the inhibitory arc of hybrid Petri nets, there is an easy and intuitive way to model inhibitions, but kinetic functions occurring for example in mass action law or Michaelis–Menten equation in combination with not identical stoichiometry of all reactants of each reaction require at least the combined concepts of HPN, HDN, and functional nets, as they are defined in HFPN. Further offered concepts might be very useful, but in general not as essential. With lower and upper capacities, a minimum or maximum concentration can be easily defined. Since in the real world an infinite concentration of a chemical does not exist, it is often useful to set a limit of concentration for all components of the model.

In order to model and simulate sophisticated models in systems biology, only Petri net extensions and tools are considered which offer at least the functionality as HFPN does.

Available Tools

In the last 25 years, many tools for Petri net modeling, simulation, and analysis were developed and published. The majority of modeling and simulation tools focuses either on discrete, hybrid, stochastic, colored Petri nets, or other Petri net concepts. Currently, there are two tools available which match the requirements: Cell Illustrator (Nagasaki et al., 2010) and Snoopy (Heiner et al., 2012).

Cell Illustrator is a commercial tool to model and simulate complex biological systems using HFPN. Its graphical user interface and representation of the Petri net are strongly tailored to use cases from molecular and systems biology. Beside the Petri net symbols, for each node of the Petri net, an icon of a biological entity or relation can be chosen from a library of biological elements. Thus, the visualization of the model is similar to other systems biology modeling tools (e.g. CellDesigner (Funahashi et al., 2003) or tools supporting SBGN (Le Novère et al., 2009)) which increases usability. Properties of all Petri net elements can either be modified by selecting a specific element or by editing a table containing all model parameters. Simulation results are visualized real time and can be exported, as well as the model, in several file formats including images.

Snoopy is a Petri net modeling and simulation tool offering many Petri net classes, such as discrete, continuous, hybrid, colored, and fuzzy Petri nets. Its general purpose graphical user interface is not oriented for a specific scientific field. Before modeling, a Petri net class for the model has to be chosen. Since a Petri net model can be transformed to a different Petri net class, an initial discrete model can be then extended by continuous elements after transferring it to the hybrid Petri net class. If a firing speed function depends on the value of a specific place, this place needs to be connected to the transition by a modifier arc. This special arc does not have any impact on the Petri net but makes this specific variable available for the transition. After computation of simulation finished, the simulation results are visualized and can be exported.

13.4 Open-Source Components for a Petri Net Toolchain

13.4.1 Motivation

Our aim is to provide a tool which assists in modeling and simulation of sophisticated application cases in systems biology, using hybrid Petri nets with support of functions. There are only two competitive tools with similar functionality available: Cell Illustrator and Snoopy. Both tools have their strengths, but also their weaknesses. While modeling kinetic speed functions, both tools support a syntax check for functions avoiding structural mistakes, e.g., missing closing bracket or function argument. Further structural mistakes and inconsistencies could be revealed by applying physical units to each parameter, e.g., two values that are added together must have matching units. Cell Illustrator takes physical units into account, while Snoopy only supports the extraction of variables, which makes it easier to change a value of a variable. Unfortunately, neither of both tools offers an intuitive rendering of mathematical expression. Such a rendering supports a lot to comprehend the general structure of a mathematical expression. Both tools support knock out experiments by disabling a transition. Snoopy realizes it user-friendly by check boxes. If the dynamics of one biological entity, e.g., a metabolite, should not be simulated, Snoopy offers the ability to set the tokens in a place to a constant value.

For larger biological systems, some entities, e.g., ATP or ADP, are reactants within several reactions. If for example ADP is modeled as one single place, Snoopy improves the visualization of the overall Petri net by logical duplicates of this place. Thus, for visualization, this single place may occur multiple times in the network to avoid arcs crossing large parts of the model.

Cell Illustrator supports lower and upper capacities, which for instance is useful to describe saturation processes of biological entities. It also provides user-defined conflict handling, which means that the user may directly influence the internal conflict resolution strategies by setting priorities to concurrent transitions.

Both tools support the visualization of simulation results for places (change rate of tokens) and transitions (firing speed over time), but none offers the ability to visualize the flow of tokens on the arcs. This feature would be very useful to investigate continuous parts of a Petri net. A constant token value in a continuous place does not imply that there is no token flow. A rather typical case for continuous places in biological applications is an equilibrium of in and out flows, which obscures the dynamics in this particular part of the model.

A major concern of Snoopy is that in general for continuous and hybrid Petri nets non-negative markings and non-negative firing speed of transitions are not assured. If a transition fires with negative speed, its pre-places and post-places are reversed, so tokens flow in the opposite direction of the arcs. This violates one of the fundamental properties of Petri nets. For continuous Petri nets, the use of adaptive semantics ensures non-negative marking of pre-places, but that does not apply to continuous parts of a hybrid Petri net.

Since both tools are closed-source, the validation of simulation can only be checked by comparing the simulation results of example Petri net models. The simulation itself is a black box implementation which lacks transparency. Additionally, due to the negative number of tokens and firing speed, the simulation results of Snoopy are difficult to compare and comprehend. Further comparison and discussion can be found in Brinkrolf et al. (2018).

Both tools do not match the requirements, and they cannot be adapted or improved by plug-ins to fulfill our needs, since some design choices are fundamental. This resulted in the development of a new open-source modeling and simulation toolchain.

13.4.2 The Extended Hybrid Petri Net Formalism

The extended hybrid Petri net formalism (xHPN) (Proß, 2013; Proßet al., 2012) combines several Petri net concepts including all of those which are supported by HFPN. By definition, xHPN is an extension of HPN, offering hybrid modeling with inhibitory and test arcs. It also supports functions depending not only on values in the places as maximum firing speed of continuous transitions and as arc weights. This includes also the arc weight of an inhibitory arc, defining its threshold to inhibit the transition. A delay can be assigned to discrete transitions, lower and upper capacities can be assigned to places. Stochastic transitions with a variable hazard function are supported as well. Transitions have an additional firing condition given as a Boolean expression. The four different types of conflicts defined in David and Alla (2010) are defined in xHPN accordingly. For certain conflicts, resolution strategies based on probabilities and priorities are also supported.

13.4.3 Modelica and OpenModelica

Modelica (Modelica Association, 2021) is a free equation-based modeling language for cyber-physical systems. There are several compilers (Modelica Association, 2021) available on the market, such as the commercial tools Dymola, Wolfram SystemModeler, JModelica, and SimulationX as well as the open-source implementation OpenModelica Compiler (OMC) (Fritzson et al., 2005) which is part of the OpenModelica project. Development of OpenModelica started more than 20 years ago and is actively and financially supported by the Open Source Modelica Consortium (OSMC). Its aim is to be a free implementation for academic and industrial usage.

In a nutshell, the OMC reads a Modelica model and compiles it to an executable. Running the executable will then compute simulation results, which are either written to a file or directly sent via socket communication to other tools.

13.4.4 PNlib

The PNlib (Proßand Bachmann, 2012) is an open-source implementation of the entire xHPN formalism in Modelica. It is available as stand-alone Modelica library and also as a part of the OpenModelica environment. The implementation comprises all Petri net elements as well as additional definitions and algorithms including conflict solving strategies. Thus, the PNlib is independent of further software tools, such as the compiler. The simulation results computed by the executable are supposed to be the same, independent of the chosen compiler. Because all aspects of the Petri net models are implemented in this open-source library, the simulation results are comparable, transparent, and comprehensible, which would not be the case with (partial) black box implementation like Snoopy or Cell Illustrator. This is a huge advantage for academic and industrial use cases.

13.4.5 VANESA

VANESA (Brinkrolf et al., 2014) is an open-source software tool written in Java which aims to assist scientists in the processes of modeling, simulating, and analyzing biological systems. For the process of modeling, two graph-based approaches are supported: biological networks and Petri nets, which in contrast to biological networks can be simulated. A biological network is a graph which nodes represent biological entities (e.g., metabolites, enzymes, genes) and edges represent biological relations (e.g., reaction, inhibition). Several different node and edge types are supported, and its visualization can be manipulated by the user.

Modeling a biological network can be either done from scratch based on lab data and data from literature, or the online data warehouse DAWIS-MD (Hippe et al., 2010) can be requested. It provides access to KEGG pathways (Kanehisa et al., 2012) and molecular biological databases (e.g., Mint (Licata et al., 2012), IntAct (Kerrien et al., 2012), Brenda (Scheer et al., 2011)) which query result leads to an initial network based on depth search. The retrieved network is fully editable.

For all graphs (biological networks and Petri nets), basic graph algorithms can be applied (comparison, intersection, merging of graphs, shortest path calculation). Hierarchical modeling is supported, which is convenient if models get more complex and grow in their number of nodes.

Models are saved as a SBML Level 3 Version 1 (Chaouiya et al., 2015) file, which was extended by VANESA-specific attributes as SBML annotations. Thus, the model excluding VANESA-specific attributes can be opened by any other tool supporting SBML Level 3 Version 1. This SBML export is performed by JSBML (Rodriguez et al., 2015) library, which ensures that the saved model is a valid SBML model.

13.5 VANESA: Hub of the Open-Source Petri Net Toolchain

In this section, modeling and simulation of Petri nets with VANESA and its interplay with PNlib and OpenModelica Compiler as well as unique features of this toolchain are presented.

13.5.1 Modeling

The graphical user interface of VANESA provides an intuitive way to create a Petri net. While creating nodes and arcs with the mouse, it is ensured that only arcs defined in xHPN are created, which keeps the Petri net structure valid. If a non-defined arc is created (e.g., an arc connecting two places or transitions, inhibitory arc from transition to place), a warning is given, and the arc is deleted. Adding a place with a name that already exists for a place in the Petri net results in adding a logical place of the already existing place instead (Fig. 13.8).

Places, transitions, and arcs are created with default values. Default number of tokens in places is zero, as well as their lower capacity. Upper capacity is set to infinity. Delay of discrete transitions, maximum firing speed of continuous transitions, and default arc weight are set to 1. After selecting an element, all properties of this selected element can be modified easily. This includes to set the number of tokens for places to a constant value, knock out transitions and set their firing conditions, define conflict solution strategies for discrete places if there is a structural conflict, and attach parameters to continuous transitions.

Since kinetic speed functions might get large and complex, VANESA supports the extraction of parameters. Thus, the value of a parameter can be changed easily for all its occurrences in the function. These parameters do have a physical unit which, in combination with the syntax check, avoids and reveals structural mistakes. While entering the function, it gets syntax checked and rendered to a Latex image in real time. If the syntax check detects a mistake, the position of the mistake is indicated. These features support the user significantly while dealing with complex functions.

13.5.2 Simulation

For the simulation of the Petri net, the user can specify the duration of simulation and the size of the result, given by the number of returned time steps of evaluation. The time unit for the duration depends on the time unit used for modeling (e.g., seconds or minutes). The number of returned time steps does not influence the accuracy of numerical calculation of simulation, but the accuracy of visualization of simulation results.

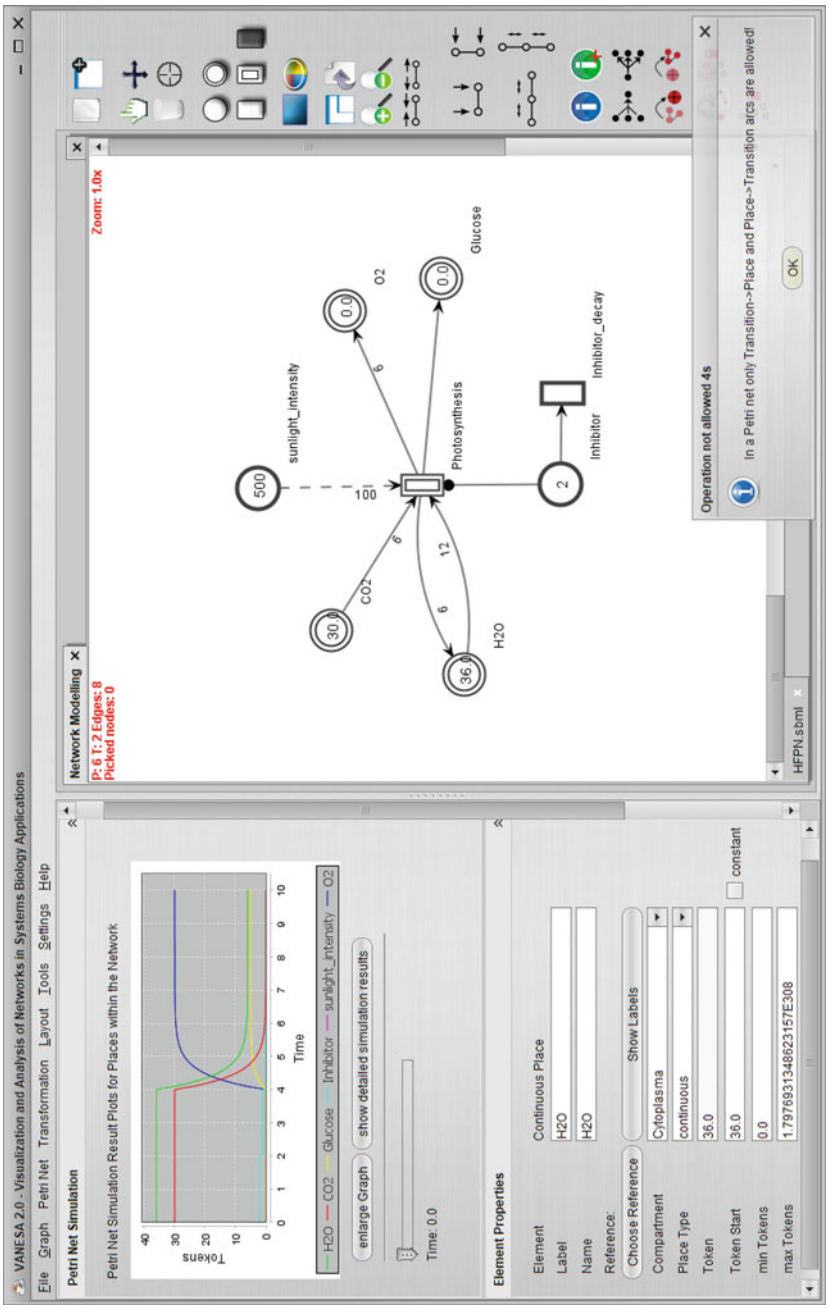

Fig. 13.8 Graphical user interface of VANESA. In the center, the Petri net is shown. On the left, the simulation results and detailed information about previously selected place H2O are displayed. On the right-hand side is the toolbar for Petri net manipulation. Simulation results are zoomed in to show only results between 0 and 40 tokens. A warning is displayed in the lower right corner, indicating that an illegal connection was created

In the background, the Petri net is exported to Modelica in a way that it is compliant with the elements defined in PNlib. When the export is finished, the OpenModelica Compiler is called with simulation specific parameters and compiles the Petri net model with the help of PNlib to an executable. After compilation, the executable is run and the actual simulation is computed.

The default option is saving the simulation results in a file, and once simulation finished, the results can be further processed and visualized. Some simulations take several minutes to be computed, which leads to the drawback that the user cannot stop the simulation based on visualization of the first computed steps of the simulation result. Thus, the communication between the executable and VANESA is realized by a TCP/IP client–server model. VANESA acts as a server and as soon as the executable connects to VANESA, calculation of simulation starts and simulation results of evaluated time steps are sent as byte stream to VANESA and get processed and visualized. The communication between VANESA and OpenModelica, as well as between VANESA and executable, forwards warnings or errors that might occur which are shown to the user. That includes problems with physical units within mathematical expressions, as well as numerical problems of the integrator or inconsistencies of the Petri net formalism itself. This feedback can be very useful to avoid mistakes and improve the quality of the model.

13.5.3 Visualization of Simulation Results

The visualization of simulation results covers places, transitions, and arcs. The lines which are drawn in the chart based on the connected values for each time step are interactive so that the user can zoom in and request specific values for specific time steps by hovering the drawn value. For places, the chart shows the number of tokens for each time step, and for transitions the actual firing speed for each time step is shown. For arcs, two lines are drawn. One line shows the actual token flow for each time step, and the second line shows the cumulative token flow. If multiple places are selected, the chart shows the number of tokens for the selected places combined in a single chart. VANESA is able to manage multiple runs of simulation of the same Petri net which were obtained for example by varying one kinetic parameter. Each single simulation result can be enabled or disabled for visualization. If more than one simulation result is enabled for visualization and a single place is selected, the chart shows all simulation results of this specific place combined with multiple lines.

The slider allows the user to investigate the simulation result of a specific time step. When the slider is set to a specific time step, for each place, the number of tokens is drawn into the place and enabled transitions are colored red.

There is also a detailed view of simulation results provided. It shows all the tokens for all places for all time steps in a table and plots the tokens of each place in separate charts. These charts are either scaled individually or have the same scaling.

Individual scaling focuses on the individual properties and dynamics of each place, while a common scaling for all charts makes it possible to represent the global behavior of change of tokens in the places.

13.5.4 Exports and Documentation

All graphs can be exported in PNG and SVG file format. Charts can be exported in PNG, SVG, and PDF file format. Excerpts of the graphs can also be created by zooming in on the desired area.

Storing simulation results in the SBML file would increase its file size drastically. Thus, each single simulation result can be exported in CSV format for saving, sharing with colleagues, and for further analysis of the raw simulation data using external tools. This export includes all attributes of the simulation result (for all places, transitions, and arcs) and can be also imported and mapped to the graph.

Sharing the entire model with all its elements and parameters is important for collaborative work and for transparency. This is ensured by a Latex export for a Petri net, which generates a Latex file which can be compiled as a PDF file. The Latex file contains an image of the Petri net as it is visualized in VANESA at the time of generation, all initial values of the places, all equations with their preconditions, post-conditions, speed functions, and all parameters of each equation. Physical units for all initial values and parameters are taken into account as well. The table of initial values also indicates if a place is set to a constant number of tokens. If an equation is disabled (knocked out), its speed function equals 0. Thus, this automatically generated Latex document provides all information necessary to model or adapt the Petri net.

13.6 Conclusion and Discussion

Petri nets are a well established method of choice for modeling, simulating, and analyzing systems in systems biology. Since the introduction of Petri nets in 1962, the formalism got extended by many concepts to fulfill the need to represent more sophisticated models. Since then, lots of research focused on the mathematical foundation of Petri nets including proofs, theorems, and Petri net properties.

Modeling sophisticated systems requires a composition of concepts such as discrete and continuous elements, inhibitor arcs, and speed maximal functions including parameters. There are only a very limited number of tools which offer modeling and simulation of Petri nets combining these concepts. Due to limitations of the existing tools, the combination of the xHPN formalism implemented in the Modelica library PNlib, the OpenModelica Compiler, and VANESA is developed to offer a transparent open-source environment for modeling, simulation, visualization, and analysis of simulation results.

References

Ackermann J, Einloft J, Nöthen J, Koch I (2012) Reduction techniques for network validation in systems biology. J Theor Biol 315:71–80

Bartocci E, Lió P (2016) Computational modeling, formal analysis, and tools for systems biology. PLoS Comput Biol 12(1):e1004591

Brinkrolf C, Janowski SJ, Kormeier B, Lewinski M, Hippe K, Borck D, Hofestädt R (2014) VANESA—a software application for the visualization and analysis of networks in system biology applications. J Integr Bioinform 11(2):239. https://doi.org/10.2390/biecoll-jib-2014-239

Brinkrolf C, Henke NA, Ochel L, Pucker B, Kruse O, Lutter P (2018) Modeling and simulating the aerobic carbon metabolism of a green microalga using Petri nets and new concepts of VANESA. J Integr Bioinform 15(3). https://doi.org/10.1515/jib-2018-0018

Cardoso J, Valette R, Dubois D (1996) Fuzzy petri nets: an overview. IFAC Proc Vol 29(1):4866–4871. https://doi.org/10.1016/S1474-6670(17)58451-7. https://www.sciencedirect.com/science/article/pii/S1474667017584517. 13th World Congress of IFAC, 1996, San Francisco USA, 30 June–5 July

Chaouiya C, Keating SM, Berenguier D, Naldi A, Thieffry D, van Iersel MP, Le Novere N, Helikar T. (2015) The systems biology markup language (SBML) level 3 package: qualitative models, Version 1, Release 1. J Integr Bioinform 12(2):270

Computer Science YUD, Lipton R (1976) The reachability problem requires exponential space. Research report (Yale University, Department of Computer Science). Department of Computer Science, Yale University. https://books.google.ca/books?id=7iSbGwAACAAJ

David R, Alla H (2010) Discrete, continuous, and hybrid petri nets, 2nd edn. Springer, Heidelberg

Drath R, Engmann U, Schwuchow S (1998) Hybrid aspects of modelling manufacturing systems using modified petri nets. IFAC Proc Vol 31(31):145–151. https://doi.org/10.1016/S1474-6670(17)41019-6. https://www.sciencedirect.com/science/article/pii/S1474667017410196. 5th IFAC Workshop on Intelligent Manufacturing Systems 1998 (IMS'98), Gramado, Brazil, 9–11 November

Fehling R (1993) A concept of hierarchical petri nets with building blocks. In: Rozenberg, G. (ed) Advances in Petri Nets 1993. Springer, Berlin, pp 148–168

Finkel A (1993) The minimal coverability graph for petri nets. In: Rozenberg G (ed) Advances in Petri Nets 1993. Springer, Berlin, pp 210–243

Fritzson P, Aronsson P, Lundvall H, Nyström K, Pop A, Saldamli L, Broman D (2005) The openmodelica modeling, simulation, and software development environment. Simulation News Europe 15(44/45):8–16

Funahashi A, Morohashi M, Kitano H, Tanimura N (2003) Celldesigner: a process diagram editor for gene-regulatory and biochemical networks. BIOSILICO 1(5):159–162. https://doi.org/10.1016/S1478-5382(03)02370-9. https://www.sciencedirect.com/science/article/pii/S1478538203023709

Fuss H (2013) Simulation of biological systems with petri nets—introduction to modelling of distributed systems. In Möller DPF (ed) Erwin-Riesch workshop: system analysis of biological processes. Vieweg+Teubner, Wiesbaden, pp 3–12. https://doi.org/10.1007/978-3-663-19445-3_1

Genrich H, Lautenbach K (1981) System modelling with high-level petri nets. Theor Comput Sci 13(1):109–135. https://doi.org/10.1016/0304-3975(81)90113-4. https://www.sciencedirect.com/science/article/pii/0304397581901134. Special Issue Semantics of Concurrent Computation

Gutowska K, Formanowicz D, Formanowicz P (2020) Systems approach based on petri nets as a method for modeling and analysis of complex biological systems presented on the example of atherosclerosis development process. In: Bartoszewicz A, Kabziński J, Kacprzyk J (eds) Advanced, Contemporary Control. Springer, Cham, pp 579–586

Heiner M, Herajy M, Liu F, Rohr C, Schwarick M (2012) Snoopy—a unifying petri net tool. In: Haddad S, Pomello L (eds) Application and theory of petri nets. Springer, Heidelberg, pp. 398–407

Hippe K, Kormeier B, Töpel T, Janowski S, Hofestädt R (2010) DAWIS-M.D.—A data warehouse system for metabolic data. In Fähnrich KP, Franczyk B (eds) Informatik 2010: Service science—Neue perspektiven für die Informatik, Beiträge der 40. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Band 2, 27.09. - 1.10.2010, Leipzig, Deutschland, *LNI*, vol 175, pp 720–725. GI

Hofestädt R, Thelen S (1998) Quantitative modeling of biochemical networks. In Silico Biol 1(1):39–53

Jensen K (1987) Coloured petri nets. In: Brauer W, Reisig W, Rozenberg G (eds.) Petri Nets: central models and their properties. Springer, Heidelberg, pp 248–299

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40(Database Issue):109–114

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40(Database Issue):841–846

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. https://doi.org/10.1038/nbt.1558

Li C, Qin J, Kuroyanagi K, Lu L, Nagasaki M, Satoru M (2021) High-speed parameter search of dynamic biological pathways from time-course transcriptomic profiles using high-level petri net. Biosystems 201:104332. https://doi.org/10.1016/j.biosystems.2020.104332. https://www.sciencedirect.com/science/article/pii/S0303264720302033

Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40(Database Issue):857–861

Liu F, Sun W, Heiner M, Gilbert D (2019) Hybrid modelling of biological systems using fuzzy continuous Petri nets. Brief. Bioinform. 22(1):438–450. https://doi.org/10.1093/bib/bbz114

Matsuno H, Tanaka Y, Aoshima H, Doi A, Matsui M, Miyano S (2003) Biopathways representation and simulation on hybrid functional Petri net. In Silico Biol 3(3):389–404

Modelica Association: Modelica Tools webpage. https://www.modelica.org/tools/

Modelica Association: Modelica webpage. https://www.modelica.org/

Murata T (1989) Petri nets: properties, analysis and applications. Proc IEEE 77(4):541–580. https://doi.org/10.1109/5.24143

Nagasaki M, Saito A, Jeong E, Li C, Kojima K, Ikeda E, Miyano S (2010) Cell Illustrator 4.0: a computational platform for systems biology. In Silico Biol 10(1):5–26

Petri CA (1962) Kommunikation mit automaten. Ph.D. thesis, Universität Hamburg, Hamburg

Proß S (2013) Hybrid modeling and optimization of biological processes. Ph.D. thesis, Bielefeld University, Bielefeld

Proß S, Bachmann B (2012) PNlib—An advanced Petri Net library for hybrid process modeling. In: Otter M, Zimmer D (eds.) Proceedings of the 9th International Modelica Conference. Linköping University Electronic Press, Linköping, pp 47–56. https://doi.org/10.3384/ecp1207647

Proß S, Janowski S, Bachmann B, Kaltschmidt C, Kaltschmidt B (2012) PNlib—A modelica library for simulation of biological systems based on extended hybrid petri nets. In Heiner M, Hofestädt R (eds.) Proceedings of the 3rd International Workshop on Biological Processes & Petri Nets (BioPPN 2012), satellite event of Petri Nets 2012, Hamburg, Germany, June 25, 2012, *CEUR Workshop Proceedings*, vol 852, pp 47–61. CEUR-WS.org. http://CEUR-WS.org/Vol-852/

Rodriguez N, Thomas A, Watanabe L, Vazirabad IY, Kofia V, Gómez HF, Mittag F, Matthes J, Rudolph J, Wrzodek F, Netz E, Diamantikos A, Eichner J, Keller R, Wrzodek C, Fröhlich S, Lewis NE, Myers CJ, Le Novère N, Palsson BØ, Hucka M, Dräger A (2015) JSBML 1.0: providing a smorgasbord of options to encode systems biology models. Bioinformatics 31(20):3383–3386. https://doi.org/10.1093/bioinformatics/btv341

Sackmann A, Heiner M, Koch I (2006) Application of Petri net based analysis techniques to signal transduction pathways. BMC Bioinf 7:482

Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Söhngen C, Stelzer M, Thiele J, Schomburg D (2011) BRENDA, the enzyme information system in 2011. Nucleic Acids Res 39(Database Issue):670–676

Chapter 14
Immersive Exploration of Cell Localization Scenarios Using VR, Spatialized Video Communication, and Integrative Bioinformatics

Bjorn Sommer, Ayn Sayuti, Chang Hee Lee, Zidong Lin, Jenny Hu, and Ashley Hall

Abstract Integrating spatially localized molecular networks into virtual cell environments is an approach which is only provided by a very small number of tools. As this task requires the combination of a set of Biotechnology/Bioinformatics-related information sources, it can be seen as an appropriate example for Integrative Bioinformatics research. Here, we want to show new immersive perspectives for cytological pathway integration by combining recent explorative technologies with the software CELLmicrocosmos 4 PathwayIntegration. A mesoscopic-localized metabolic pathway—i.e. the citrate cycle and the glycolysis—is localized based on database entries onto an abstract cell environment of *Arabidopsis thaliana*. The created cell model is used in three different contexts providing different degrees of immersion: (1) Web-based 2D exploration of 3D Scenarios (using Gather.town), (2) Exploration and Annotation in a VR Design Application (using Gravity Sketch), and (3) Large-Scale VR Visualization and Navigation (using the CAVE2 and zSpace). All these examples promise to be very useful in the context of Integrative Bioinformatics-related education as well as communication.

Keywords CELLmicrocosmos · Design · Visualization · Virtual reality · Immersive analytics · Gravity sketch · Gather.town

B. Sommer (✉) · A. Sayuti · C. H. Lee · Z. Lin · J. Hu · A. Hall
School of Design, Royal College of Art, London, UK
e-mail: bjoern@cellmicrocosmos.org

A. Sayuti
Faculty of Art and Design, Universiti Teknologi MARA, Shah Alam, Malaysia

C. H. Lee
School of Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_14

14.1 Introduction

The visualization of biochemical pathways has a long tradition. Since Michal started to illustrate an overview of all metabolic reactions by publishing his Biochemical Pathways maps in 1968 (Michal 1998, 2012), a variety of new approaches have been developed to visualize and explore the structure and connectivity of metabolic pathways (Becker and Rojas 2001; Genc and Dogrusoz 2003; Karp and Paley 1994; Schreiber 2002; Schreiber et al. 2009). These approaches are relevant in terms of biochemical education as well as to provide an overview of the current status quo in research as well as to identify gaps in current biological knowledge.

Since Michal's visualization attempt, a wide variety of databases providing metabolic data, such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) and Reactome, were established (Croft et al. 2011; Kanehisa et al. 2012). Maps in those databases are visualized in two dimensions (2D). Whereas the Reactome maps illustrate enzyme localizations inside abstract 2D cell compartments, KEGG in most visualizations does not depict subcellular localizations of network components.

Whereas 2D visualization of biochemical networks is around since a couple of decades, there are not many approaches which combine those networks with 3D models depicting cell environments whereby providing spatial context. The *CELLmicrocosmos 4 Pathway-Integration (CmPI)* is a software which combines 2D network visualization with 3D cell environments by using protein or gene localization data (Sommer et al. 2010a). At the mesoscopic level, it provides various cell component models which are based on a number of microscopic visualization techniques. The provided cell component models are ranging in size from a few thousand nanometres down to a few nanometres (Sommer 2012). In this way, CmPI is a good example for Integrative Bioinformatics, as it uses Bioinformatics-related data sources to map localization information onto 3D cell models resulting from microscopic images.

Prior to visualize genes/proteins in the context of a spatial cell environment, they have to be assigned to cell component location(s). Doing so for 2D cell visualization, related localization approaches are, for example, COMPARTMENTS and CellWhere (Binder et al. 2014; Zhu et al. 2015). Both approaches provide specific methods to predict the subcellular localization of proteins—partly only for a single specific tissue. But they do not aim to take the spatial structure of biological cells into account.

To the best of our knowledge, there are currently no related tools available enabling the network mapping into virtual cell environments. Outdated projects which were following a similar idea were, for example, MetNetVR, or The Interactorium which visualizes a single localization scenario (Wurtele et al. 2003; Widjaja et al. 2009). Also, there are Cytoscape plugins available, such as the 2D graph localization visualization Cerebral or the fragmentary subcellular visualization tool 3DScape which has never been fully functional (early prototype status) (Barsky et al. 2007; Wang 2011). None of those tools is able to semi-automatically localize and visualize genes/proteins in the context of a spatial cellular environment in 3D.

14.1.1 New Ways for Interactively Exploring Cellular Structures

Since the arrival of the Oculus Rift DK1™ in 2013, *Virtual Reality (VR)* headset technologies drastically matured and the recent sale numbers of the Oculus Quest 2™ show that VR is about to become mainstream with a quickly growing number of users (Hamish 2021; Bol 2021). In Bioinformatics, the use of VR technologies is not completely new. Tools like the previously mentioned MetNetVR used CAVE technology already 15 years ago (Yang et al. 2006). The prospects of recent VR technologies for Integrative Bioinformatics were discussed in a journal special issue (Sommer et al. 2018): For example, the HoloLens was used to explore molecular models in *Augmented Reality (AR)* (Müller et al. 2018), a large tiled stereoscopic screen was used for multi-omics analysis (Maes et al. 2018), or *Head-mounted Displays (HMDs)* were used to create a game-inspired visualization of extracellular matrix elements (Belloy et al. 2018). New tools like VRdeo can be used to explore molecular models with HMDs and create educational narrated videos while doing so (Brøuža et al. 2021).

During the COVID pandemic, another important development was starting: the major daily face-to-face communication platforms quickly became video chats like Zoom and Skype—Zoom fatigue was a common problem, especially in teaching environments (Wiederhold 2020). Therefore, new tools were combining video chats with dynamic group building and interactions—usually by making use of web technologies, we call them here Spatialized Video Communication platforms. One of these approaches is Gather.town. This web platform provides video chats in customizable 2D spaces, enabling serendipitous interactions by enabling users to mostly freely move between different spaces, rooms, and quickly change communication partners. Gather.town early attracted a decent amount of venture capital and was already used at a couple of conferences and educational institutions bringing back some of the joy of physical meetings (Mascarenhas 2021; Gather - Crunchbase Company Profile and Funding 2021).

Therefore, we are presenting here three scenarios which are based on either VR technology to provide immersive exploration of created localization scenarios, or web technology enabling dynamic video communication between multiple participants. These three scenarios are based on a combination of an abstract plant cell and a specific localization scenario based on *Arabidopsis thaliana* which was created by using CmPI. Over the years, CmPI was already used in a couple of scenarios. Originally intended as a tool to generate *and* explore cytological localization scenarios in 3D, we started to use it also as an authoring tool to create cell models which can be used in the next step for advanced visualization or educational approaches in third-party tools such as Blender (Biere et al. 2018).

In particular, the following three new data exploration and visualization scenarios will be discussed in the following sections: (1) Web-based 2D exploration of 3D Scenarios, (2) Exploration and Annotation in a VR Design Application, and (3) Large-Scale VR Visualization and Navigation.

14.2 Methods

As the base for this work, the CELLmicrocosmos 4.2.1 PathwayIntegration (CmPI) is used. This software has been developed over a number of years (Sommer et al. 2010a, 2015). The Java standalone version (requires Java 7 or 8) is accessible via the website http://Cm4.CELLmicrocosmos.org. Those readers who are familiar with CmPI can skip the following sections and continue with the chapter *Application Cases*.

14.2.1 Mesoscopic Modelling

As the base for a cell environment, first a cell model is needed as the starting structure which can be correlated in the next steps with biochemical pathways. For this purpose, CmPI provides a number of preconfigured cell models which can be used to create eukaryotic as well as prokaryotic cell models. The cell models can be configured by using a number of different cell component models, such as Chloroplast, Mitochondria, Nucleus, etc. They are available on three different *Cytological Abstraction Levels (CAL),* see Fig. 14.1 (Sommer 2012; Spevacek 2000):

- CAL1: 3D-microscopy/−spectroscopy-based (Image),
- CAL2: interpretative (Allegory), and.
- CAL3: abstract cell visualization (Abstraction).

CAL1 and CAL2 models are based on different image resources (Flickr 2014; The Cell 2014; Cell Press 2014). For models of CAL1, online databases such as the Cell-Centered Database (CCDB) were used (Martone et al. 2002). Figure 14.1 CAL1 shows, for example, the tip of a mitochondrion based on a dataset containing 256 electron-microscopic images acquired from the CCDB. Amira® was used for semi-automatically segmentation (FEI Visualization Sciences Group 2014). Then, the result was optimized by using Autodesk® 3ds max® (Autodesk 2012). The

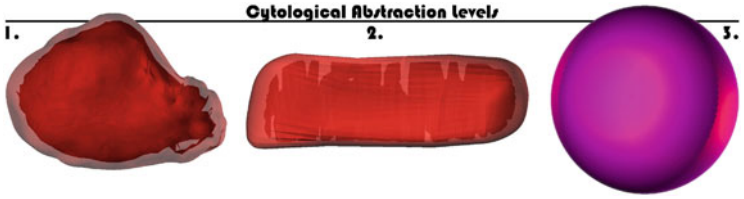

Fig. 14.1 Three cytological abstraction levels (Sommer 2012). (1) 3D-microcsopy/-spectroscopy-based cell visualization, (2) interpretative one, and (3) abstract one. (Reprinted with permission from Sommer (2012))

final CAL1 model was exported in the VRML 2.0 format and imported into CmPI. Alternatively to commercial software like Amira®, Open Source software like Fiji (IsJustImageJ) with TrakEM may be used (Cardona et al. 2012; Schindelin et al. 2012).

Using the interpretative approach (CAL2), cell models featuring a reduced visual complexity can be created—which are more similar to those models often seen in traditional educational textbooks. For this purpose 3D modelling programs, such as Autodesk 3ds max® or Blender can be used (Autodesk 2012; Blender 2014). Different microscopic images can be used as inspiration, e.g. light-microscopic images to acquire the overall structure of the cell at a few thousand nanometres as well as the colour staining, as well as electron-microscopic images which are able to depict the granular structure of the cell.

If the internal structure is not of relevance for the intended visualization, CAL3 can be used, where cell components are substituted by simple geometrical objects, like cubes or spheres (Wurtele et al. 2003; Widjaja et al. 2009).

Import and export of cell models is supported by using the VRML97/2.0 format, making it compatible with many modelling packages, such as Blender or Autodesk 3ds max®. The CELLmicrocosmos 3.2 CellEditor (CmCE) can be used to prepare cell models for CmPI. CmCE is able to work with different VRML97 formats and save same in a CmCX-compatible format incl. the different cell component layers: http://Cm3.CELLmicrocosmos.org.

14.2.2 Functional Modelling

Now that the initial cell model is created, the structure has to be combined in the next step with a biochemical pathway. CmPI can combine structural data at the mesoscopic level with the functional level based on gene−/protein-related localization data.

The previously described process of generating a cell model is followed by importing gene- or protein-related data. Figure 14.3 **Centre** shows the citrate cycle which was imported from the KEGG database into CmPI—the layout is based on the standard KGML layout. The 2D visualization in CmPI—based on the JUNG library—depicts compounds as blue nodes and enzymes using their localization colour, whereas the arrows depict the direction of the corresponding reaction (O'Madadhain et al. 2003; Sommer et al. 2013).

To combine biological networks with e spatial structure of the cell model, localization data is required. For this purpose, CmPI was connected to a number of databases UNIPROT, BRENDA, Gene Ontology (GO), and ANDCell (Chan et al. 2012; UniPort Consortium 2013; Chang et al. 2014; Podkolodnaya et al. 2011; Ivanisenko et al. 2020). All these databases were integrated in the materialized database structure DAWIS-M.D. with the purpose to enable immediate access (Kormeier 2014; Töpel et al. 2008; Sommer et al. 2010b). A new version of

BioDWH—the data warehouse system which is the base of DAWIS-M.D.—was just recently published (Friedrichs 2021).

Moreover, CmPI provides very simple network modelling capabilities. However, it is not intended as an advanced Systems Biology network modelling tool. For this purpose, readers are referred to projects like VANTED, VANESA, or Cytoscape (Brinkrolf et al. 2014; Rohn et al. 2012; Smoot et al. 2011). Alternatively, in a previous project an early prototype was built as a bridge between VANTED and CmPI, enabling to integrate 2D networks created in VANTED into CmPI and localize and visualize the corresponding networks (Sommer and Schreiber 2017).

For our example here, only the UNIPROT database integration in DAWIS-M.D. was used for localization purposes which also integrates Gene Ontology. For the localization of proteins, the UniProt Knowledgebase (UniProtKB) is relevant with the following categories:

- General Annotation (Comments).
  - Subcellular location.
- Ontologies.
  - Keywords
- Cellular component
  - Gene ontology.
  - Cellular component.

Figure 14.2 shows the localization layers. Database terms, such as "mitochondrial chromosome" or "mitochondrial pyruvate dehydrogenase complex" which can be found in the UNIPROT database, are mapped onto the Matrix of the mitochondrion model.

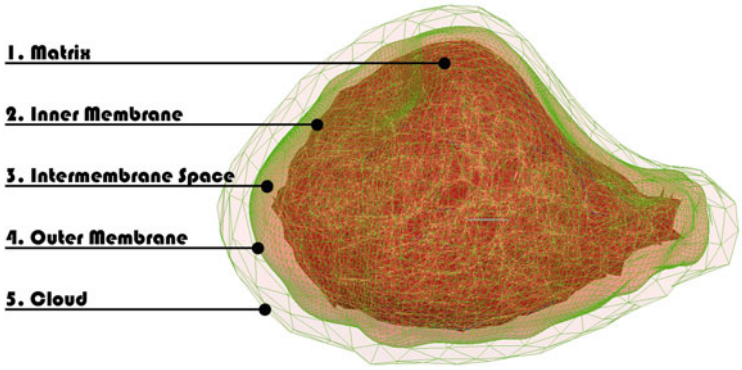

Fig. 14.2 Localization layers using the mitochondrion modelled in Fig. 14.1 CAL1. (Reprinted with permission from Sommer (2012))

To select the appropriate localization for a gene or protein, the Subcellular Localization Charts are used which provide a number of different visualization categories (Sommer et al. 2013; Mueller et al. 2016). Different selection categories supported by the localization visualization are the amount of localization entries, the protein-specific localization, or the protein co-localizations. These chart categories can be used to assign specific localizations directly to all associated data entries.

Now, the mesoscopic/spatial data of the cell components can be combined with the functional/biological network data. The network is laid out based on the interconnections by selecting one of the available algorithms, such as the Fruchterman–Reingold algorithm (Fruchterman and Reingold 1991), the ISOM layout (Meyer 1998), or a 2D mapping layout. In the latter case, the corresponding 2D layout is shown in Fig. 14.3 centre (GUI right top) and will be mapped onto a unit sphere using polar coordinates. Afterwards, the layout is mapped onto the surface of the corresponding cell component and finally, the result can be visualized and explored in 3D space.

14.2.3 Cell Exploration

The created cell model can already be explored with CmPI. For this purpose, three different modes providing a 6DOF (six degrees of freedom) navigation are integrated into CmPI, enabling the user to move around cell component models or travelling across them (Sommer et al. 2010a): Floating, Flight, and Object-bound mode. In Fig. 14.3 the change between different perspectives was achieved by navigating in Floating mode.

The following basic windows are part of the application:

- CellEditor (CELLmicrocosmos 3.1): a simple interface to add new cell component models to the cell environment which can be used to add new cell components and manipulate their position with mouse and keyboard (Fig. 14.3 centre: GUI left bottom),
- PathwayIntegration (CELLmicrocosmos 4.2): the pathway (Fig. 14.3 centre: GUI top) and protein localization table (Fig. 14.3 centre: GUI right bottom) which can be used to download KEGG pathways and protein localization from DAWIS-M.D.,
- 2DViewer: 2D network visualization using the JUNG library (Fig. 14.3 centre: GUI right top),
- CellUniverse: 3D visualization of the cell (Fig. 14.3 centre: GUI centre-top),
- NodeDetails: information about the currently selected protein (Fig. 14.3 centre: GUI left top).

CmPIweb is a simplified web-based version of CmPI based on three.js and D3.js can be used as an online viewer of localization scenarios of original files from Kovanci et al. (2016): http://Cm4web.CELLmicrocosmos.org.

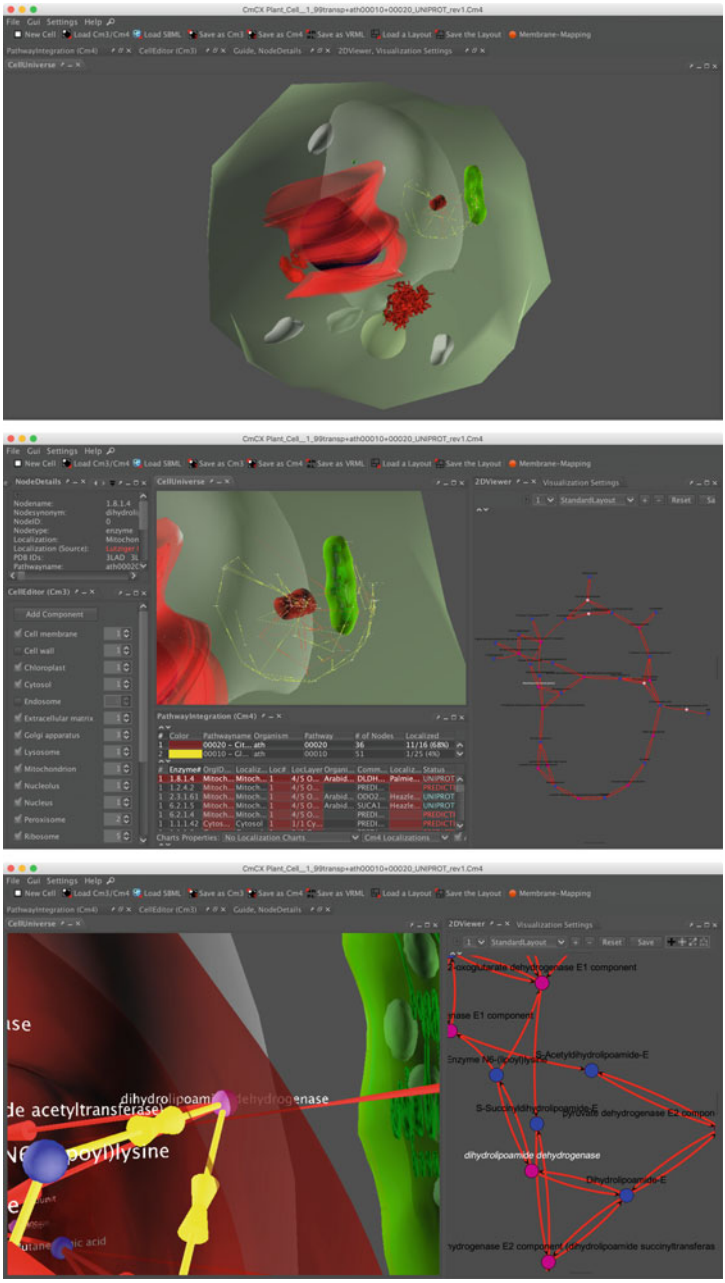

Fig. 14.3 Creating the cell model in CmPI. *Top:* plant cell model in the cell explorer in 3D; *Centre:* intracellular view of the cell components in 3D (left, mitochondrion, chloroplast, and endoplasmic reticulum), as well as the overview of the citrate cycle in 2D (right); *Bottom:* 3D close-up view of dihydrolipoamide dehydrogenase (EC 1.8.1.4) which is part of the citrate cycle (red) as well as the glycolysis (green) (left) and citrate cycle in 2D (right)

14.3 Application Cases

Now, we will present three application cases to explore cytological localization scenarios using web technology and Virtual Reality (VR): (1) Web-based 2D exploration of 3D Scenarios, (2) Exploration and Annotation in a VR Design Application, and (3) Large-Scale VR Visualization and Navigation.

But prior to discussing these application scenarios, the generation of the cell model and its corresponding network localization will be discussed in the following section.

14.3.1 Creation of the Cytological Localization Scenario

An abstract plant cell model based on *Arabidopsis thaliana* was created using standard 3D cell components of the CellExplorer *(Plant_Cell__1_99transp.Cm3)*. The KEGG pathways for the citrate cycle (ath00020) and glycolysis (ath00010) were downloaded from the KEGG database integration of DAWIS-M.D. The localization was done using the UniProt integration of DAWIS-M.D. Where localization information were fragmentary or had to be predicted, most-recent localization information from the online resources UniProt 2021 and BRENDA 2021 were acquired and applied to CmPI (Chang et al. 2021; Nucleic Acids Res 2021).

To both pathways the GEM layout was applied in 2D (Frick et al. 1994). Both pathways were then mapped onto the 3D cell components using the Sphere Mapping layout with the parameters half-sphere and same-enzyme-same-place. The latter parameter guarantees that the same enzyme with the same localization in both pathways are placed onto the same position—which applies to enzymes EC 1.8.1.4, 4.1.1.49, 1.2.4.1, and 2.3.1.12.

Figure 14.3 **top** shows the complete plant cell from the front perspective, and Fig. 14.3 **bottom** shows a detail view of the cell model with associated pathway on the left side, as well as the detail of the 2D (GEM) layout of the metabolic pathway/citrate cycle on the right side.

Compounds which are connected to enzymes with two different localizations (e.g. mitochondrion and chloroplast) are mapped onto a cytosol localization. In this cell model, the cytosol is a transparent abstract sphere surrounding the mitochondrion.

14.3.2 Web-Based 2D Exploration of 3D Scenarios

In the beginning of this chapter, a 2D map of metabolic pathways based on the GEM layout was mapped onto the spatial structure of an abstract plant cell. Now, we are using the 3D model to create renderings which can be integrated into Gather.town,

providing a new approach to present and explore cytological location scenarios (Gather 2021).

Gather.town is a relatively new web-based Spatialized Video Communication platform which enables dynamic communication providing a 2D game-like environment. The big advantage is that multiple people can easily use this platform in parallel (the authors claim that up to 2000 people can use this platform), enabling serendipitous interactions as well as being accessible via a simple URL.

Gather.town is highly customizable by providing the option to upload a background image and defining then abstract boundaries, portals, placing furniture, and integration of Google documents, external websites, etc. The simple square-based navigation is possible by moving user's avatar with the arrow keys or the AWDS keys. Most importantly, it provides video communication for those people who are in close proximity to each other.

A number of screenshots were created by using the previously generated CmPI plant cell model associated with the two metabolic pathways. The screenshots were uploaded into Gather.town by using its proprietary map editor—run in a web browser—which is shown in Fig. 14.4 **top**: the screenshot shows the mitochondrion, chloroplast, and cytosol associated with the corresponding pathways. The blue rectangles are portals which can be used to change between different rooms. Arrows are placed in front of the portals to indicate the entrances. Each room here represents another perspective, usually zooming in or out: the transparent cell revealing its internal cell components and depicted pathways. In Fig. 14.4 **centre-top** shows the external opaque view of the cell. Figure 14.4 **centre-bottom** shows the transparent cell revealing its internal cell components and depicted pathways. Figure 14.4 **bottom** shows a detail of the chloroplast with its associated glycolysis pathway.

This approach provides, e.g. university teachers, with a new way to present cell models and its underlying functionality to students. Teachers could prepare corresponding lectures covering different biochemical topics by using CmPI and enable multiple students to explore and discuss these environments on their own or in a guided fashion by accompanying the exploration using the video communication functionality to narrate the journey.

This early prototype can be currently tested by using the link: https://gather.town/app/pX2Id9BY1Y4COAJP/PlantCell1-99transp.

14.3.3 Exploration and Annotation in a VR Design Application

Whereas Virtual Reality exploration required investing into expensive computers as well as headset hardware in the past, the advent of devices like the Oculus Quest™ enables users to enter VR by using affordable standalone devices.

We decided to use the Gravity Sketch software to explore the previously created model of a cytological localization scenario (Gravity Sketch 2021). Early studies using Gravity Sketch in the context of product design showed that promising results can be achieved with this software from an educational perspective (Joundi et al.

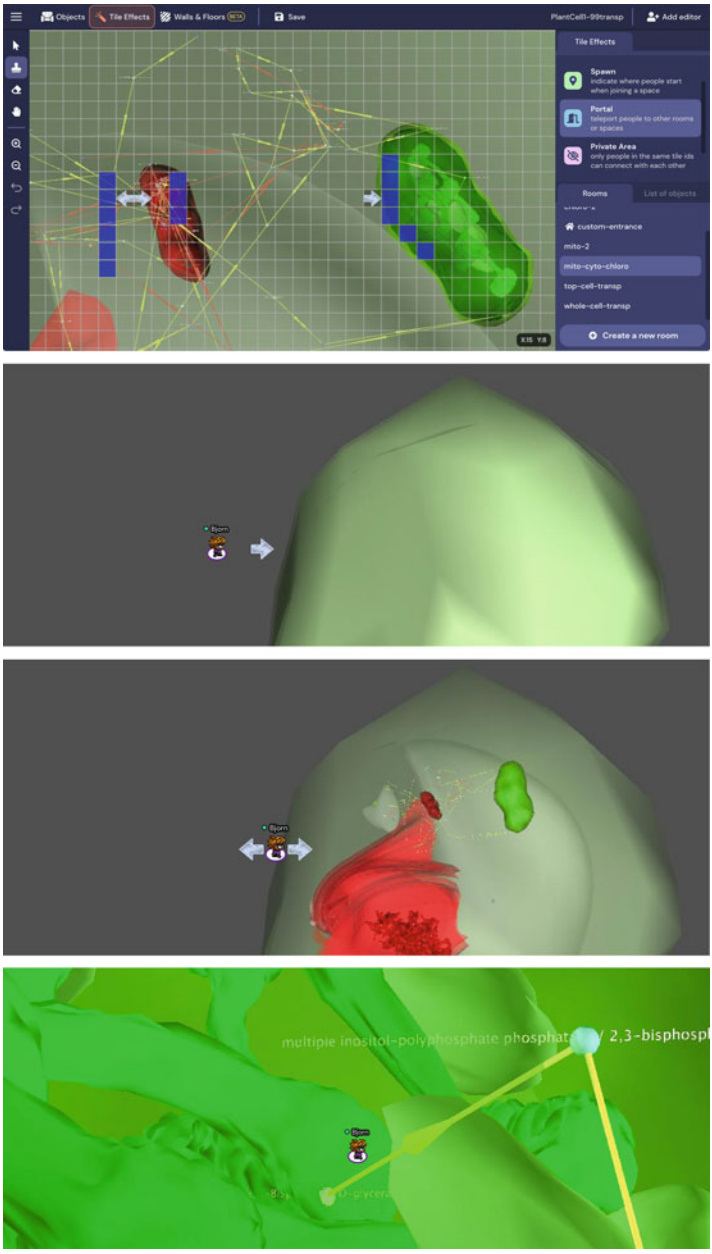

Fig. 14.4 Exploring the 3D cell model in Gather.town. *Top:* the map editor of Gather.town, whereas blue rectangles indicate portals between different rooms, e.g. the left-most portals lead towards the centre images; *Centre-top:* the cell view from outside, going forward, leads to; *Centre-bottom:* the transparent cell view showing all cell components; *Bottom:* the close up of the glycolysis pathway localized at the chloroplast and at its internal thylakoid. The arrows indicate portals to other rooms—each of the four scenes here represents a different Gather.town room

2020; Van Goethem et al. 2020). As compatible VR device, we used the Oculus Quest 2 ™. The huge advantage of this device is that it can be used standalone without connecting it to an external computer and a high-spec graphics card. Gravity Sketch supports to explore models by zooming in and out using the Oculus Quest controllers. Being a tool for virtual design and modelling—which is, e.g. used in the context of architecture, car, or shoe design—Gravity Sketch provides various sketching modes and tools which can be used to annotate the cell. In addition, the professional version of Gravity Sketch enables a VR co-working space which allows multiple users to explore and annotate a model in parallel. Whereas in Gather.town, it is easily possible to have multiple rooms and to change between these rooms, Gravity Sketch currently provides one basic large room without direct connections to other rooms, but in parallel the user can be visually fully immersed into this space (in contrast to Gather.town). And this is especially very useful in case 3D models have to be explored or created.

The cell model created in CmPI was exported to a VRML 2.0 model and then converted via Blender 2.91.2 into an OBJ file using a MacBook Pro 2015. This OBJ file was imported to Gravity Sketch software version 5.1.58-qc by using the Landing Pad which can be launched on the computer's web browser. Launching Gravity Sketch on the Oculus Quest 2 enables the user to choose the uploaded OBJ file and place it in 3D space. Figure 14.5 shows the model imported to Gravity Sketch which was annotated for the mitochondrion (blue circle), and two enzymes located at the cytosol (green) and chloroplast (yellow). The annotation feature is an important functionality of Gravity Sketch which can be used here to highlight specific regions of interested or to indicated connections between different cell components or proteins.

Currently, Gravity Sketch only imports OBJ files without materials. Therefore, the imported cell model does not show colours and is depicted in grayscale, but the transparency levels are correctly shown. However, as the intention is here to highlight certain regions of the cell by using colours and most cells components in reality are anyway semi-transparent, this is not a big disadvantage. Moreover, the exploration in VR supported by 3D-stereoscopic visualization enables improved differentiation of semi-transparent cell components.

14.3.4 Large-Scale VR Visualization and Navigation

CmPI cannot only be used to create the cell models, it also supports stereoscopic 3D exploration of cells in case the appropriate hardware is available (professional 3D monitor or 3D TVs, with, e.g. an NVIDIA Quadro graphics) including position-based adjustment of the stereo vision (Sommer et al. 2014).

In 2015, we have introduced the SpaceMap approach which was based on a semi-immersive 3D monitor, the zSpace 200™ (Fig. 14.6 bottom): For navigation inside the 3D environment, (a) head-tracking is used to change the perspective/navigate

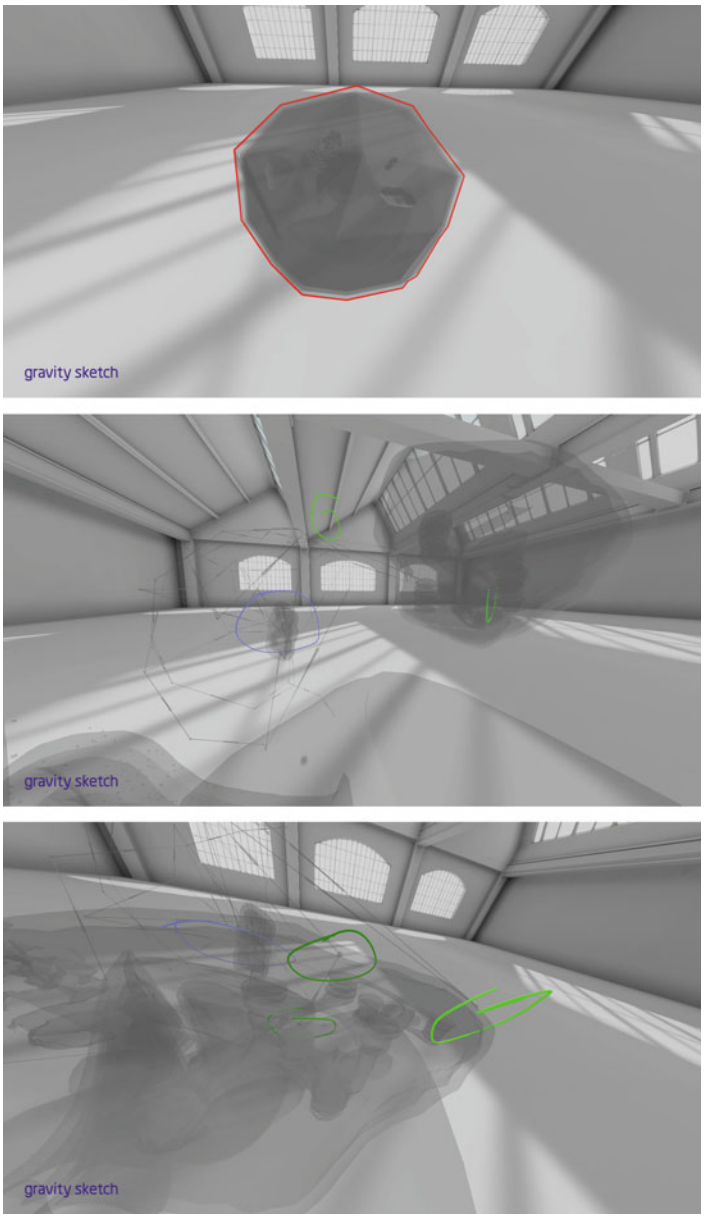

Fig. 14.5 Exploring the Plant Cell with an Oculus Quest 2 using Gravity Sketch which enables annotation in 3D space using the Oculus Quest 2 controllers. Blue circle highlights the mitochondrion, the green an enzyme localized at the cytosol, the yellow an enzyme localized at the Chloroplast. *Top:* External view of the Plant Cell; *Centre:* intracellular view of the cell components (mitochondrion, chloroplast, and endoplasmic reticulum); *Bottom:* close-up view of the chloroplast. The background shows the factory environment of Gravity Sketch which supports the user's orientation in 3D space

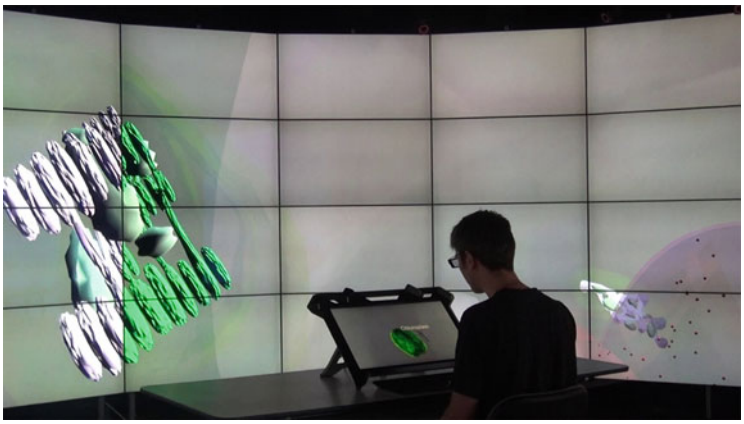

Fig. 14.6 Exploring the Chloroplast with the SpaceMap approach, integrating CAVE2 ™ with the zSpace ™. The operate can use the zSpace to prepare and navigate the journey in the virtual environment which is shown on the CAVE2. The audience can observe the journey while being immersed using 3D glasses

around cell components and (b) a 3D stylus pen is used to grab cell components and rotate around them (Sommer et al. 2015).

In Fig. 14.6 the zSpace is used in combination with CmPI to navigate inside the CAVE2™—a cylindrical large-scale Virtual reality environment consisting of 80 monitors (Sommer et al. 2016; Febretti et al. 2013). For this purpose, CmPI connects to a special CmPI implementation based on Omegalib running on the CAVE2 which enables large-scale visualization of cytological data (Febretti et al. 2014).

The zSpace system provides the SpaceMap which is used to precisely navigate to a specific place in 3D space. For example, it is possible to point with the 3D cursor at a specific location where the camera position inside the CAVE2 environment smoothly moves to. Figure 14.6 for example shows the internal structures of a chloroplast inside a plant cell. The zSpace can also be used to find an appropriate perspective in 3D space which is then transferred from the SpaceMap to the large and immersive CAVE2.

Providing a large presentation space, SpaceMap enables a new lecture format, where, e.g. a university teacher could prepare a cytological localization scenario and present it to students by operating the zSpace shown here. By wearing 3D glasses, the students and the teacher would be fully immersed in the virtual environment which could be operated similarly to a spaceship. Obviously, this is the most complex approach, as the expensive CAVE2 environment is needed which only exist at a few locations in the world, and also the university teacher will have to be prepared for operating the SpaceMap.

14.4 Discussion and Outlook

Whereas CmPI is used already since a couple of years for exploring localization scenarios in 3D, it was not often used as an authoring tool, except CmPIweb (Kovanci et al. 2016). Here we explored three external frameworks which can be used to explore those cell models for educative purposes:

Firstly, we presented how CmPI screenshots can be used as a simple base to create Gather.town rooms. Different rooms can present different magnification levels (outside and inside cell down to cell component-internal areas). The rooms can be interconnected, so that visitors can freely explore those scenarios. Obviously, the 3D properties are getting lost, but the big advantage is that Gather.town is run in a web browser and it provides a very simple and intuitive navigation and video communication interface for visitors in close proximity to each other. Moreover, the authors claim that currently up to 2000 users can enter a space. The free version of Gather.town enables at the moment around 25 people to explore the space (Gather 2021). Teacher could accompany the exploration process and provide a narrated journey or they could prepare scenarios which can be explored by the students on their own.

Secondly, Gravity Sketch was used to explore and annotate cell models in 3D space using an Oculus Quest 2. In this way, users can explore the spatial structure of the cell, naturally navigate in 3D space and can draw, e.g. circles around specific positions they want to highlight and discuss. By providing an optional collaborative environment, this exploration process can also be accompanied by a teacher. Obviously, the big advantage over Gather is the option to explore 3D objects using intuitive 3D navigation. Conversely, the communication is currently not supported via video chat and there is the technological overhead requiring HMDs and dedicated experience.

Thirdly, a large-scale cell environment was setup using the CAVE2 in conjunction with the zSpace. A plant cell was explored using the 3D stylus pen of the zSpace in order to navigate the space. This complex approach enables a central operator to explore a cell environment with a cohort of students in an immersive way. While with Gravity Sketch every user explores the scenarios from a first-person perspective, in this approach the virtual environment is explored like operating a central spaceship—therefore, all students are in the same location. However, as the CAVE2 provides nearly a narrow 360° perspective, students can walk inside the CAVE2 and explore different perspectives.

In summary, potential target groups for these approaches are researchers discussing certain localization scenarios, educators who want to present metabolic pathways to school or university students in, e.g. Biology or Bioinformatics, or for conference/workshop organizers who want to use biology-related themes in, e.g. Gather.town.

Obviously, the rate of accessibility is decreasing from the first to the third approach. Whereas the first one can be accessed from basically everywhere using web technologies, the second approach is compatible to all VR devices which are

supported by Gravity Sketch. Apart from Gravity Sketch, CmPI could as well be used to create virtual cell environments in game engines such as Unity and Unreal. The last example is only compatible to CAVE2 environments and obviously is only addressing a very specific audience. The big advantage of the CAVE2 is that 20–40 people could experience the virtual journey at the same time while a central navigator is guiding the virtual experience.

For those readers interested in related frameworks to those discussed in this chapter, we would recommend our review publication on *Immersive Design Engineering* (Sommer et al. 2020). For readers interested in comparing the usage of a CAVE2 vs. HMDs we recommend the publication from Cordeil et al. (2017).

For future scenarios, we are looking into combining cell visualization-based approaches with bio-inspired design. Based on previous studies it was found that there is an interest in integrating biological materials in everyday products (Sayuti and Ahmed-Kristensen 2020). In the context of this research project we are exploring the ownership regarding biological products, e.g. research and educational approaches (Sayuti et al. 2021, 2020).

References

Autodesk: 3ds Max - 3D modeling, animation, and rendering software – Autodesk. http://usa.autodesk.com/3ds-max/. Accessed 7 Jan 2012

Barsky A, Gardy JL, Hancock REW, Munzner T (2007) Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. Bioinformatics 23:1040–1042

Becker MY, Rojas I (2001) A graph layout algorithm for drawing metabolic pathways. Bioinformatics 17:461–467

Belloy N, Wong H, Révoteau-Jonquet J, Baud S, Dauchez M (2018) Mesoscopic rigid body modelling of the extracellular matrix self-assembly. J Integr Bioinform 15:20180009. https://doi.org/10.1515/jib-2018-0009

Biere N, Ghaffar M, Doebbe A, Jäger D, Rothe N, Friedrich BM, Hofestädt R, Schreiber F, Kruse O, Sommer B (2018) Heuristic modeling and 3D stereoscopic visualization of a Chlamydomonas reinhardtii cell. J Integr Bioinform 15:20180003. https://doi.org/10.1515/jib-2018-0003

Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, Jensen LJ (2014) COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database 2014:bau012

Blender.: blender.org - Home of the Blender project - free and open 3D Creation Software. http://www.blender.org/. Accessed 22 July 2014

Bol M (2021) Has oculus quest sold one-million lifetime units? https://arinsider.co/2020/09/21/has-oculus-quest-sold-one-million-units/. Accessed 16 May 2021

Brinkrolf C, Janowski SJ, Kormeier B, Lewinski M, Hippe K, Borck D, Hofestädt R (2014) VANESA-A software application for the visualization and analysis of networks in systems biology applications. J Integr Bioinform 11:239. https://doi.org/10.2390/biecoll-jib-2014-239

Brøuža V, Byška J, Mičan J, Kozlíková B (2021) VRdeo: creating engaging educational material for asynchronous student-teacher exchange using virtual reality. Comput Graph 98:280–292

Cardona A, Saalfeld S, Schindelin J, Arganda-Carreras I, Preibisch S, Longair M, Tomancak P, Hartenstein V, Douglas RJ (2012) TrakEM2 software for neural circuit reconstruction. PLoS One 7:e38011

Cell Press: Picture Show: Cell Press. http://www.cell.com/pictureshow. Accessed 8 Dec 2014
Chan J, Kishore R, Sternberg P, Van Auken K (2012) The gene ontology: enhancements for 2011. Nucleic Acids Res 40:D559–D564
Chang A, Schomburg I, Placzek S, Jeske L, Ulbrich M, Xiao M, Sensen CW, Schomburg D (2014) BRENDA in 2015: exciting developments in its 25th year of existence. Nucleic Acids Res 43:D439
Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D (2021) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 49:D498–D508
Cordeil M, Dwyer T, Klein K, Laha B, Marriott K, Thomas BH (2017) Immersive collaborative analysis of network connectivity: CAVE-style or head-mounted display? IEEE Trans Vis Comput Graph 23:441–450
Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39:D691–D697
Febretti A, Nishimoto A, Thigpen T, Talandis J, Long L, Pirtle JD, Peterka T, Verlo A, Brown M, Plepys D (2013) CAVE2: a hybrid reality environment for immersive simulation and information analysis. In: IS&T/SPIE electronic imaging. International Society for Optics and Photonics, Bellingham, pp 864903–864903–12
Febretti A, Nishimoto A, Mateevitsi V, Renambot L, Johnson A, Leigh J (2014) Omegalib: a multi-view application framework for hybrid reality display environments. In: Virtual reality (VR). IEEE, pp 9–14
FEI Visualization Sciences Group: Amira | FEI Visualization Sciences Group. http://www.vsg3d.com/amira/overview. Accessed 22 July 2014
Flickr: Welcome to Flickr – Fotosharing. https://www.flickr.com/. Accessed 22 July 2014
Frick A, Ludwig A, Mehldau H (1994) A fast adaptive layout algorithm for undirected graphs (extended abstract and system demonstration). In: International symposium on graph drawing. Springer, Berlin, pp 388–403
Friedrichs M (2021) BioDWH2: an automated graph-based data warehouse and mapping tool. J Integr Bioinform 18:167
Fruchterman TMJ, Reingold EM (1991) Graph drawing by force-directed placement. Softw Pract Exp 21:1129–1164
Gather. https://gather.town/. Accessed 11 May 2021
Gather - Crunchbase Company Profile & Funding. https://www.crunchbase.com/organization/gather-4189. Accessed 16 May 2021
Genc B, Dogrusoz U (2003) A constrained, force-directed layout algorithm for biological pathways. In: Graph drawing. Springer, Berlin, pp 314–319
Gravity Sketch | 3D design and modelling software. https://www.gravitysketch.com/. Accessed 11 May 2021
Hamish H (2021) Oculus Quest 2 sales figures prove VR has finally gone mainstream. https://www.techradar.com/news/oculus-quest-2-sales-figures-prove-vr-has-finally-gone-mainstream. Accessed 16 May 2021
Ivanisenko TV, Saik OV, Demenkov PS, Ivanisenko NV, Savostianov AN, Ivanisenko VA (2020) ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. BMC Bioinformatics 21:1–21
Joundi J, Christiaens Y, Saldien J, Conradie P, De Marez L (2020) An explorative study towards using VR sketching as a tool for ideation and prototyping in product design. In: Proceedings of the design society: DESIGN conference. Cambridge University Press, Cambridge, pp 225–234
Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40:D109–D114
Karp PD, Paley S (1994) Automated drawing of metabolic pathways. In: Third international conference on bioinformatics and genome research, pp 225–238
Kormeier B (2014) Data warehouses in bioinformatics. In: Approaches in integrative bioinformatics. Springer, Berlin, pp 111–130

Kovanci G, Ghaffar M, Sommer B (2016) Web-based hybrid-dimensional visualization and exploration of cytological localization scenarios. J Integr Bioinform 13:298

Maes A, Martinez X, Druart K, Laurent B, Guégan S, Marchand CH, Lemaire SD (2018) MinOmics, an integrative and immersive tool for multi-omics analysis. J Integr Bioinform 15:20180006. https://doi.org/10.1515/jib-2018-0006

Martone ME, Gupta A, Wong M, Qian X, Sosinsky G, Ludäscher B, Ellisman MH (2002) A cell-centered database for electron tomographic data. J Struct Biol 138:145–155

Mascarenhas N (2021) Sequoia Capital puts millions of dollars into Gather, a virtual HQ platform | TechCrunch. https://techcrunch.com/2021/03/11/sequoia-capital-puts-millions-of-dollars-into-gather-a-virtual-hq-platform/. Accessed 16 May 2021

Meyer B (1998) Self-organizing graphs—a neural network perspective of graph layout. In: Lecture notes in computer science (graph drawing). Springer, Berlin, pp 246–262

Michal G (1998) Biochemical pathways: an atlas of biochemistry and molecular biology. Wiley-Spektrum, Heidelberg

Michal G (2012) ExPASy - biochemical pathways. https://www.webcitation.org/6qd4XvLnc. Accessed 18 March 2012

Mueller, S.C., Sommer, B., Backes, C., Haas, J., Meder, B., Meese, E., Keller, A.: From single variants to protein cascades: multi-scale modeling of SNV sets in genetic disorders. J Biol Chem 291, 1582–1590 (2016). https://doi.org/https://doi.org/10.1074/jbc.M115.695247

Müller C, Huber M, Biener V, Herr D, Koch S, Reina G, Weiskopf D, Ertl T (2018) Interactive molecular graphics for augmented reality using HoloLens. J Integr Bioinform 15:20180005. https://doi.org/10.1515/jib-2018-0005

(2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480–D489

O'Madadhain J, Fisher D, White S, Boey Y (2003) The JUNG (Java Universal Network/Graph) framework. University of California, Irvine

Podkolodnaya OA, Yarkova EE, Demenkov PS, Konovalova OS, Ivanisenko VA, Kolchanov NA (2011) Application of the ANDCell computer system to reconstruction and analysis of associative networks describing potential relationships between myopia and glaucoma. Russ J Genet Appl Res 1:21–28

Rohn H, Junker A, Hartmann A, Grafahrend-Belau E, Treutler H, Klapperstück M, Czauderna T, Klukas C, Schreiber F (2012) VANTED v2: a framework for systems biology applications. BMC Syst Biol 6:139

Sayuti A, Ahmed-Kristensen S (2020) Understanding emotional responses and perception within new creative practices of biological materials. In: Proceedings of the sixth international conference on design creativity (ICDC 2020), pp 144–151

Sayuti NA, Sommer B, Ahmed-Kristensen S (2020) Identifying the purposes of biological materials in everyday designs. Environ-Behav Proc J 5:29–37

Sayuti NA, Sommer B, Ahmed-Kristensen S (2021) Bio-related design genres: a survey on familiarity and potential applications. In: Interactivity and game creation: 9th EAI international conference, ArtsIT 2020, Aalborg, Denmark, 10–11 Dec 2020. Springer, Berlin, p 379

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B (2012) Fiji: an open-source platform for biological-image analysis. Nat Methods 9:676–682

Schreiber F (2002) High quality visualization of biochemical pathways in BioPath. In Silico Biol 2:59–73

Schreiber F, Dwyer T, Marriott K, Wybrow M (2009) A generic algorithm for layout of biological networks. BMC Bioinformatics. 10:1

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27:431–432

Sommer B (2012) CELLmicrocosmos - integrative cell modeling at the molecular, mesoscopic and functional level. http://pub.uni-bielefeld.de/download/2557380/2557411

Sommer B, Schreiber F (2017) Integration and virtual reality exploration of biomedical data with CmPI and VANTED. It-Inf Technol 59:181

Sommer B, Künsemöller J, Sand N, Husemann A, Rumming M, Kormeier B (2010a) CELLmicrocosmos 4.1: an interactive approach to integrating spatially localized metabolic networks into a virtual 3D cell environment. In: Fred A, Filipe J, Gamboa H (eds) BIOINFORMATICS 2010 - proceedings of the 1st international conference on bioinformatics, part of the 3rd international joint conference on biomedical engineering systems and technologies (BIOSTEC 2010), pp 90–95

Sommer B, Tiys ES, Kormeier B, Hippe K, Janowski SJ, Ivanisenko TV, Bragin AO, Arrigo P, Demenkov PS, Kochetov AV, Ivanisenko VA, Kolchanov NA, Hofestädt R (2010b) Visualization and analysis of a cardio vascular disease-and MUPP1-related biological network combining text mining and data warehouse approaches. J Integr Bioinform 7:148

Sommer B, Kormeier B, Demenkov PS, Arrigo P, Hippe K, Ates Ö, Kochetov AV, Ivanisenko VA, Kolchanov NA, Hofestädt R (2013) Subcellular localization charts: a new visual methodology for the semi-automatic localization of protein-related data sets. J Bioinforma Comput Biol 11:1340005

Sommer B, Bender C, Hoppe T, Gamroth C, Jelonek L (2014) Stereoscopic cell visualization: from mesoscopic to molecular scale. J Electron Imaging 23:011007-1–011007-10. https://doi.org/10.1117/1.JEI.23.1.011007

Sommer B, Wang SJ, Xu L, Chen M, Schreiber F (2015) Hybrid-dimensional visualization and interaction - integrating 2D and 3D visualization with semi-immersive navigation techniques. In: Big data visual analytics (BDVA). IEEE, pp 1–8

Sommer B, Hamacher A, Kaluza O, Czauderna T, Klapperstück M, Biere N, Civico M, Thomas B, Barnes DG, Schreiber F (2016) Stereoscopic space map – semi-immersive configuration of 3D-stereoscopic tours in multi-display Environments. In: Proceedings of stereoscopic displays and applications XXVII

Sommer B, Baaden M, Krone M, Woods A (2018) From virtual reality to immersive analytics in bioinformatics. J Integr Bioinform 15:20180043

Sommer B, Lee CH, Martin N, Torrisi VS (2020) Immersive design engineering. In: Electronic imaging 2020 Proceedings of stereoscopic display and applications. XXXI, pp 265-1–265-16. https://doi.org/10.2352/ISSN.2470-1173.2020.2.SDA-265

Spevacek G (2000) Lernen mit Bildern in Texten: in Abhangigkeit von bereichsspezifischem Wissen, Gert Spevacek

The Cell: An Image Library. http://www.cellimagelibrary.org. Accessed 8 Dec 2014

Töpel T, Kormeier B, Klassen A, Hofestädt R (2008) BioDWH: a data warehouse kit for life science data integration. J Integr Bioinform 5:93

UniPort Consortium (2013) Update on activities at the universal protein resource (UniProt) in 2013. Nucleic Acids Res 41:D43–D47

Van Goethem S, Watts R, Dethoor A, Van Boxem R, van Zegveld K, Verlinden J, Verwulgen S (2020) The use of immersive technologies for concept design. In: International conference on applied human factors and ergonomics. Springer, Berlin, pp 698–704

Wang Q (2011) 3DScape: three dimensional visualization plug-in for Cytoscape. Nat Prec. https://doi.org/10.1038/npre.2011.6094.1

Widjaja YY, Pang CNI, Li SS, Wilkins MR, Lambert TD (2009) The interactorium: visualising proteins, complexes and interaction networks in a virtual 3D cell. Proteomics 9:5309–5315

Wiederhold BK (2020) Connecting through technology during the coronavirus disease 2019 pandemic: avoiding “Zoom Fatigue”. Mary Ann Liebert, Inc, Larchmont

Wurtele ES, Li J, Diao L, Zhang H, Foster CM, Fatland B, Dickerson J, Brown A, Cox Z, Cook D (2003) MetNet: software to build and model the biogenetic lattice of Arabidopsis. Comp Funct Genomics 4:239–245

Yang Y, Wurtele ES, Cruz-Neira C, Dickerson JA (2006) Hierarchical visualization of metabolic networks using virtual reality. In: Proceedings of the 2006 ACM international conference on virtual reality continuum and its applications. ACM, pp 377–381

Zhu L, Malatras A, Thorley M, Aghoghogbe I, Mer A, Duguez S, Butler-Browne G, Voit T, Duddy W (2015) CellWhere: graphical display of interaction networks organized on subcellular localizations. Nucleic Acids Res 43:W571–W575

Part V
Integrative Tools and Workflow

Chapter 15
IoS: A Needed Platform for Scientific Workflow Management

Savas Takan, Visam Gültekin, and Jens Allmer

Abstract Data analytics, machine learning, and artificial intelligence have found widespread application in science. They are usually employed as part of more extensive data analysis pipelines starting with raw data processing. Unfortunately, many tools that are not tested yet are used to support critical decision-making. The intended internet of science platform aims to overcome this issue and lead to sustainable, interoperable, reusable, and correct scientific workflow development. This paper calls for action to develop the internet of science to facilitate a future focus on collaborative knowledge discovery.

Keywords Scientific workflows · Workflow management · Data analytics · Data integration · Internet of science · IoS

15.1 Introduction

Today science has become more and more data-driven. While physics has needed to deal with humongous amounts of data for quite some time, biomedical sciences started facing big data stemming from sequencing and other measurement initiatives only in the last few decades. Unlike filtering strategies available for physics, the biomedical community did not agree on means to immediately discard the majority of measurements but needs to hold on to the data at this point.

S. Takan
Faculty of Engineering, Computer Engineering, Uludağ University Görükle Campus, Bursa, Turkey

V. Gültekin
Bielefeld University, Bioinformatik und Medizinische Informatik Bielefeld, Deutschland

Hochschule Ruhr West, University of Applied Sciences, Medical Informatics and Bioinformatics, Mülheim an der Ruhr, Germany

J. Allmer (✉)
Hochschule Ruhr West, University of Applied Sciences, Medical Informatics and Bioinformatics, Mülheim an der Ruhr, Germany
e-mail: jens@allmer.de

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_15

With more and more data becoming available, it is now vital to share the data effectively so that it can be explored from different perspectives and under different scenarios. The FAIR data initiative aims to make data findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). Findable and accessible are the more easily achievable goals. Interoperability and reusability are, however, harder to realize. The latter two are tied to the development and implementation of standards for the underlying fields producing the data, such as exemplified in proteomics via mzML (Martens et al. 2011) and MIAPE (Taylor et al. 2007). With agreed-upon standards and minimum information criteria in place, interoperability becomes achievable.

Data analytics workflows are becoming more and more complex. This complexity is accompanied by an increase in available workflow management systems and data analytics platforms. Popular platforms are KNIME (Berthold et al. 2008), RapidMiner (Mierswa et al. 2006), and Galaxy (Goecks et al. 2010), to name just a few. These platforms provide a means of creating data flows, including arbitrary data transformations, to develop a reproducible data analytics workflow ranging from simple formatting via statistical and advanced mathematical transformations to machine learning. One caveat is that workflows are not easily recreated among platforms (Beukers and Allmer n.d.). Any of these platforms are also extensible by the user so that additional functions become available over time. Such functionality remains largely untested but is often quickly embraced by the community. On the workflow level, data analytics platforms also fall short when workflow testing is considered. Thus, today, untested modules are put together into untested workflows. The expectable result is obvious. While, therefore, the confidence in the results should be low, this aspect remains largely unexplored.

With the internet of science (IoS), a platform to overcome such issues was proposed (Allmer 2019). The IoS aims to create a collaborative community solution to solve the underlying problems. Any scientist, engineer, or developer is welcome to join the initiative (https://bitbucket.org/allmer/ios/). Here we will expand on how scientific workflows shall be developed using the IoS in the future.

15.2 The Internet of Science

15.2.1 Platform Development

The IoS has been introduced earlier (Allmer 2019), and more information will become available on the associated bitbucket repository as the community forms around it. The defining purpose of the IoS is to ensure quality and correctness. Therefore, any new tool and any significant changes to existing tools in the platform need to pass a peer review process before making it available to the research community. Otherwise, they are only available in the development community for testing and development. This strict separation will ensure quality, as workflows build on comprehensively tested modules are more trustworthy than those built with untested modules in the development community. The tools which pass the review process and become part of the IoS can then be used to build scientific workflows.

15.2.2 Building Scientific Data Analysis Workflows

One of the principal aims of the IoS is to facilitate collaboration. Collaboration naturally includes workflow development which is modeled as a collaborative process within the IoS. Similar to collaborative coding integrated development environments such as Codeanywhere and Cloud9, workflow development within the IoS will allow simultaneous workflow design and testing for multiple users (Fig. 15.1). Here, FAIR data from the sequence read archive (Leinonen et al. 2011) on NCBI and the gene ontology (The Gene Ontology Consortium 2009) provide the workflow development data source. Alternatively, private or proprietary data could be used as a data source. To the best of our knowledge, no workflow management platform allows the simultaneous development of data analytics workflows and combines it with a versioning system. Another novelty envisioned for IoS workflows is a unit and integration test system (Fig. 15.1; blue user). Any part of the workflow with at least one processing step can be tested. Tests are not part of the production workflow but ensure correctness. In Fig. 15.1, the gray and green users collaboratively build a workflow, while the blue user develops tests. All changes are visible to all users immediately but will only persist if the user introducing the changes commits them to the versioning system.

Recently, we compared three workflow management systems and implemented RNA-seq analysis workflows. Workflow reproducibility among workflow management systems was a major issue (Beukers and Allmer n.d.). We also identified that creating functional sub workflows is crucial with increasing data analysis complexity, but only one of the workflow management tools supports this (Beukers

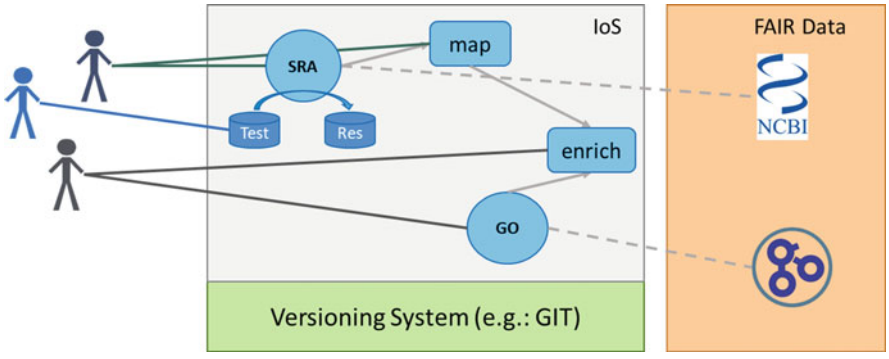

Fig. 15.1 Overview of the workflow development process using the IoS. Based on FAIR data (orange box), processing workflows can be developed within the IoS (gray box). Here three users (gray, blue, and green) simultaneously access the workflow development platform. The gray user uses the GO connector (blue circle) and a data transformation tool (enrich; blue rounded rectangle). The green user uses the SRA connector and the map function and connects it to the enriching process from the gray user. The blue user ensures correctness by implementing tests. All users simultaneously see all changes. However, only changes committed via the versioning system will remain when the respective users log out

and Allmer n.d.). Therefore, IoS workflows will support subworkflows and nested subworkflows. Since the IoS system will allow testing of any subpart of the workflow, other functionality can also be offered. One essential criterion is parameter optimization over a subworkflow. For example, an optimization algorithm (e.g., genetic algorithm) can extract parameters from the subworkflow modules. The algorithm can then determine practical ranges and optimize the parameter values. For example, support vector machine learning depends on proper parameter settings, and these parameters are often manually optimized. This optimization could be done automatically with the IoS system.

Current workflow systems have large amounts of tools included directly. For KNIME, there are more than 3000 tools available, and for Galaxy, the number is unknown since any installation can add tools. Tool documentation is generally good and online help is available, but with the sheer amount of tools, it is still hard to develop workflows without proper training and experience. The IoS system will take several measures to overcome this issue. First, many platforms have alternative algorithms for the same purpose or even alternative algorithm implementations as separate tools. The IoS system will never have multiple implementations of the same algorithm. However, slight changes to algorithms may adapt them to be most effective for a given input space. To shield the user from this, processors will be available to encapsulate alternative algorithms for the same task and decide by the input which particular algorithm to use. The same will be possible on a higher level. For example, optimization is a general procedure and can be performed using many algorithms such as the genetic algorithm or ant colony optimization. Although these algorithms work differently, their aim is the same so that they will be bundled in an optimization processor. The user may choose a particular approach from the processor or ignore the choice and leave the automatic selection. This strategy effectively reduces the number of visible tools for workflow development. Nonetheless, many tools will be available, and it may still be hard to choose which tools to connect to each other in the workflow system. To reduce the challenge, the IoS provides suggestions for which nodes to connect next on three levels: (1) any tool that can handle the input of the previous tool is listed for selection, (2) the tools that most often follow the selected tool are highlighted, and (3) an intelligent selection which highlights the tools which most often follow the given tool concerning the other tools already used in the workflow. The smart system can be based on a market basket analysis strategy. Note that smart downstream tool suggestions depend on a larger user group that develops workflows for many different purposes and makes their workflows publicly available.

15.2.3 Workflow Repository and Publishing Workflows

Sharing of workflows is of significant importance to avoid duplication of efforts and provide a basis for further developing existing workflows. To effectively achieve this, workflows need to be comprehensively tested and well documented (Fig. 15.2).

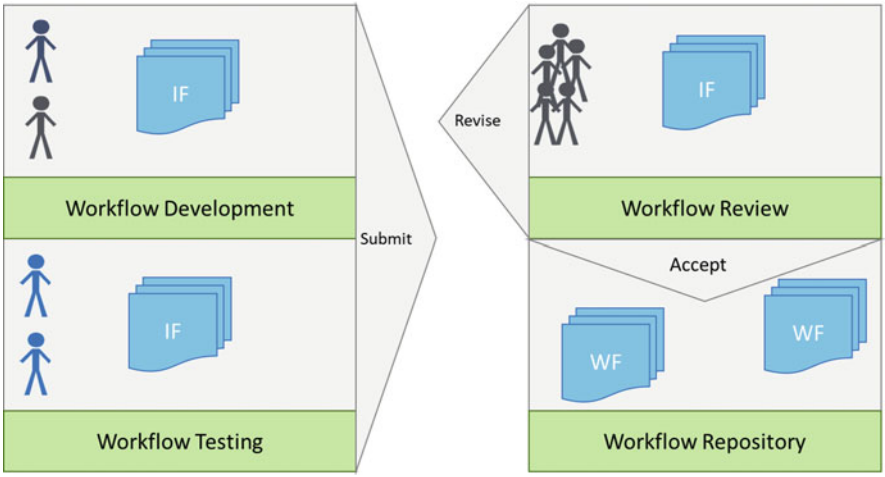

Fig. 15.2 The review process for including workflows into the public repository. As in Fig. 15.1, green and gray users developed a workflow (IF). Similarly, the blue user developed integration tests for the workflow. For inclusion into the repository, other users (here a second blue one) need to develop integration tests. The well-documented and comprehensively tested workflow can be submitted for inclusion into the public workflow repository. A review committee consisting of experts in the field and development experts reviews the workflow, the integration tests, and the documentation. Either the workflow is then included in the repository, or the committee asks for revisions

15.2.3.1 Workflow Documentation

Workflow documentation is needed on three levels: (1) for understanding the workflow and its intent, (2) for further development, and (3) for having a clear and complete output of what happened during workflow execution. The latter should be directly usable in the materials and methods section of a scientific manuscript. Workflow documentation for documenting the design's intent needs to enable grouping of nodes, highlighting of data streams with arbitrarily associated annotations. These annotations are different from, for example, Javadoc or similar code documentation tools. Like everything else, workflow documentation will be developed collaboratively and will make use of the versioning system. Tests developed for the workflows can also be documented. Thus the workflow and all its tests can be well documented for intent.

Documentation of the actual workflow execution will be on the tool level and may only be framed from the workflow level to ensure proper organization of information for publication.

15.2.3.2 Workflow Testing

Workflows like tool development for the IoS need proper unit testing and integration testing to ensure quality and correctness. Comprehensive workflow testing is also

required for inclusion into the public repository. Workflows, just like tools and about anything else within the IoS, will be assigned digital object identifiers (DOIs) so that they become citable, which can help honor the effort put into development, testing, and documentation. Comprehensive workflow testing needs to provide synthetic data to test correctness for various scenarios and experimental data to show functionality (DOI for both). For subworkflows, unit tests need to be provided, and for the overall workflow, integration tests are mandatory. Tests need to be documented for their intent in order to simplify the review process.

15.2.3.3 Workflow Sharing

Workflow sharing or publishing is an essential process because it can reduce the duplication of efforts. Additionally, it enables others to perform data analysis using those workflows. Currently, workflows are, for example, published alongside manuscripts and are only peer-reviewed by a few reviewers who are typically domain experts and not workflow experts. Additionally, workflows usually do not include automated testing and generally are only applied to the data they were developed for analyzing. These complications entail that it is mandatory to comprehensively test workflows from others before applying them to one's data. This step is generally neglected due to time constraints. It is self-evident that this is a dangerous strategy. Therefore, the IoS will put into effect a strict workflow testing routine before they can enter the public repository (Fig. 15.2). Consequently, the workflows in the repository can be used by domain experts without the need for further scrutiny. Upon sharing the workflow in the repository, others can develop new tests and include them into the repository following the review process. Thus, heavily used workflows will likely have large amounts of tests that are run nightly. Should errors be detected at one point, all users of the workflow will immediately be notified. This level of transparency and security affords the confident application of IoS workflows. Apart from complete workflows, subworkflows can be shared to be put together into larger workflows. Collaborative workflow development, testing, and documentation coupled with publishing in the public repository will lead to community accepted workflows for different data analysis questions.

15.2.3.4 Review Committee

It is evident that comprehensively tested and effectively peer-reviewed workflows can be used with confidence. The review process needs to be adequate for workflow reviewing within a given scientific domain to achieve such confidence. Therefore, the review committee will consist of workflow experts who ensure that the workflows are adequately assembled, documented, and tested. Additionally, experts for the specific scientific domain and statisticians will be part of each review committee. When in doubt, the review committee can invite external peer reviewers. In dialogue with the workflow's submitters, an agreement will be reached

when to resubmit a revision if the workflow was rejected inclusion in the public repository. The assembly of the review committee is automatic. Workflow experts can be determined according to their tool usage statistics. Domain experts can be determined via many routes, for example, the data provider for the workflow, application statistics of similar workflows, and contribution to manuscripts in the research field. Additionally, workflow developers may suggest reviewers. The review process is entirely open and transparent, and all reviews, comments, and decisions are public and citable via DOIs. Each workflow is associated with an audit trail for public assessment, including the review committee's output, the developers' rebuttals, and nightly workflow tests.

15.3 Contribution to the IoS

At this point, it may be important to ascertain that any contribution is appreciated and measured in the IoS. The core designers and developers of the IoS framework, the developers and testers of tools, developers and testers of workflows, users applying workflows to data, the review committee memberships, etc., all together will enable the IoS to bring science back to a collaborative endeavor striving for knowledge.

The internet of science depends on a community effort involving all scientists and engineers willing to change to put science at the forefront again and reinstate trust in scientific findings. The IoS is by no means targeting a particular research domain but all domains involving data analytics. While current approaches to workflow management are manifold, an equal amount of issues is apparent. The same is true for tool development which suffers from overworked reviewers and increasing numbers of paper submissions. The IoS will eliminate duplications of effort and ensure that any contribution is appreciated. Currently, workflow development can be performed with many workflow management systems, and cross-platform reproducibility efforts have been excerpted. An alternative approach is given by the IoS, representing a monolithic platform not allowing the automation of third-party tools. This approach, coupled with testing and review processes on multiple levels, seems safer than other workflow systems. Many good workflow development strategies have been devised, and it is time to consolidate the current knowledge into a new development, the IoS.

15.4 Outlook

The idea of the IoS has first been published in (Allmer 2019) and is still very fresh. Its development platform has not much to show yet (https://bitbucket.org/allmer/ios). However, a group of interested collaborators is forming. The next step is to agree upon aims and procedures and put them into place. In parallel, the first coding

efforts will commence shortly. By the end of 2021, a working prototype can be expected.

Acknowledgments We want to thank all colleagues who already indicated their willingness to contribute to the IoS.

References

Allmer J (2019) Towards an internet of science. J Integr Bioinform 16:20190024. https://doi.org/10.1515/jib-2019-0024

Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME: the Konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thime L, Decker R (eds) Data analysis, machine learning and applications. Springer, Berlin, pp 319–326

Beukers M, Allmer J (n.d.) Challenges for the development of automated RNA-seq analyses pipelines. BMC Bioinformatics under review

Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86. https://doi.org/10.1186/gb-2010-11-8-r86

Leinonen R, Sugawara H, Shumway M (2011) The sequence read archive. Nucleic Acids Res 39:D19–D21. https://doi.org/10.1093/nar/gkq1019

Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Römpp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz P-A, Deutsch EW (2011) mzML–a community standard for mass spectrometry data. Mol Cell Proteomics 10(R110):000133. https://doi.org/10.1074/mcp.R110.000133

Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T (2006) YALE. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '06. ACM Press, New York, p 935

Taylor CF, Paton NW, Lilley KS, Binz P-A, Julian RKJ, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJR, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR 3rd, Hermjakob H Jr, Andrew R (2007) The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 25:887–893. https://doi.org/10.1038/nbt1329

The Gene Ontology Consortium (2009) The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res 38:D331. https://doi.org/10.1093/nar/gkp1018

Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Chapter 16
Revealing Genotype–Phenotype Interactions: The AgroLD Experience and Challenges

Pierre Larmande and Konstantin Todorov

Abstract Understanding genotype–phenotype relationships is one of the most important areas of research in agronomy. The new challenges aim at understanding these relationships on the level of the different molecular entities responsible for the expression of complex phenotypic traits. Recent advances in high-throughput technologies have resulted in tremendous increase in the amount of data in the agronomic domain. Unfortunately, they can only partially capture these dynamics. It is important to effectively integrate additional information and extract knowledge to understand the biological system as a whole. To this end, the Semantic Web offers a stack of powerful technologies for the integration of information from diverse sources, making knowledge explicit by the help of ontologies and explicit semantic relations between entities. In particular, knowledge graphs have gained popularity as means to structure and semantically represent the data and knowledge in a particular field, opening up new and enhanced ways of information retrieval and knowledge discovery. We have developed AgroLD, a knowledge graph that exploits the Semantic Web technology and some of the relevant standard domain ontologies, to integrate knowledge on plant crop species and in this way facilitate the formulation of new scientific hypotheses. This chapter provides an overview of the AgroLD project focusing on the data integration and semantic annotation processes which initially focused on genomics, proteomics, and phenomics. Likewise, we present the different data exploration strategies developed to make the platform available to a large audience. Our objective is to offer a domain specific knowledge platform to solve complex biological and agronomical questions related to the implication of genes in, for instance, plant disease resistance or high yield traits.

P. Larmande (✉)
DIADE, IRD, University of Montpellier, Montpellier, France

French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, Montpellier, France

P. Larmande · K. Todorov
LIRMM, University of Montpellier, CNRS, Montpellier, France
e-mail: pierre.larmande@ird.fr

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_16

Finally, we will present several current challenges in knowledge extraction from heterogeneous biological data sources.

Keywords Data integration · FAIR · Knowledge graphs · Bioinformatics · Plant science

16.1 Introduction

The demand for food is expected to grow substantially in the next decades (FAO n.d.). To meet the challenges of this global growth in a context of climate change, a better understanding of genotype–phenotype relationships is crucial to improve production capacities. Agronomic research is witnessing an unprecedented revolution in the acquisition of various data such as phenotypic and genomic data, as well as data related to the study of specific genes. The 3000 Rice Genomes project (Wang et al. 2018) is an excellent illustration of this problem since it generates terabytes of genomic data associated with the results of large-scale phenotypic experiments carried out in environments with different conditions.

A better understanding of genotype–phenotype relationships requires the integration of biological information of various kinds. However, this information is often dispersed in several databases on the Internet each with different data models, scales, or distinct means of access. For biologists, it is difficult to search relevant information in these databases as the mass of information can be incomplete and hard to manage. These problems are particularly relevant in the context of genetic association analyses or GWAS (Genome Wide Association Studies), which allow to associate large regions of the genome (locus) with a phenotypic trait (trait). GWAS loci often include several hundred genes that need to be analysed in order to identify only a fraction of the genes associated with the trait under study. At some point, each scientist will have to choose which genes to investigate further in the laboratory. Often, this choice is subjective, as it is based on inferences from partial data. Today's major challenges are related to the development of methods to integrate these heterogeneous data and to enrich biological knowledge. The scientists also need methods to dig into this mass of data and to highlight relevant information that identifies key genes. In order to overcome the limitations posed by the heterogeneity of data, international scientific communities encourage the dissemination of information according to the FAIR principles (Findable, Accessible, Interoperable, Re-usable) (Wilkinson et al. 2016), increasing the interoperability of data.

Semantic Web technologies, a concept coined by Tim Berners-Lee et al. (2001) and standardized by the World Wide Web Consortium (W3C), offer a FAIR solution to facilitate this integration and enable data interoperability. Among these technologies, the Resource Description Framework (RDF) (W3C n.d.) is widely used to publish data on the Web and interconnect it to form what we call the Web of Data. RDF allows a resource and its relationships to other resources to be described in the form of triples of the kind *Subject-Predicate-Object*. These

triples can be combined to build large data networks (also known as RDF, or knowledge graphs), integrated from different data sources. In recent years, many initiatives emerged in the biomedical and bioinformatics fields aiming at providing integrated environments to formulate scientific hypotheses about the role of genes in the expression of phenotypes or the emergence of diseases. Among them, we cite Bio2RDF (Belleau et al. 2008), EBI RDF (Jupp et al. 2014), or Uniprot RDF (Redaschi and Consortium 2009). However, to the best of our knowledge, there was no equivalent in the agronomic field before the AgroLD platform (Venkatesan et al. 2018) was launched.

16.2 Overview of the AgroLD Platform

We have developed AgroLD, a knowledge graph powered by Semantic Web technologies as a structure to integrate data, to enable knowledge sharing and to allow information retrieval at scale. It is designed to integrate available information on various plant species in the agronomic domain such as rice (genus Oryza), Arabidopsis thaliana, wheat (genus Triticum), to name a few. Table 16.1 shows the complete list of species with the total number of related protein entities. In the following, we describe the components of the knowledge graph and the process of its construction.

16.2.1 Integrated Data Sources

The conceptual framework of AgroLD is based on well-established ontologies in the field such as Gene Ontology (The Gene Ontology Consortium 2019), Plant Ontology (The Plant Ontology Consortium 2002), Plant Trait Ontology (Cooper et al. 2018), or Plant Environment Ontology (Buttigieg et al. 2013). Table 16.2 shows the complete list of the used ontologies. The majority of these ontologies are hosted by the OBO Foundry project (Smith et al. 2007). Furthermore, we decided to build AgroLD in several phases. The current phase (second phase) covers information on genes, proteins, predictions of homologous genes, metabolic pathways, plant phenotypes, and genetic studies. At this stage, we have integrated data from several resources such as Ensembl plants (Bolser et al. 2016), UniProtKB (The UniProt Consortium 2018), Gene Ontology Annotation (Huntley et al. 2015). The choice of these sources has been guided by the biological community. They are indeed widely used and have a strong impact on user's confidence. We have also integrated resources developed by the local SouthGreen platform (South Green, Collaborators 2016) such as TropGeneDB (Hamelin et al. 2013), a tropical plant genetics database, OryGenesDB (Droc et al. 2009) a rice genomics database, GreenPhylDB (Valentin et al. 2021), a comparative genomics database for tropical plants, OryzaTagLine (P Larmande et al. 2008) a rice phenotype database, and SniPlay (Dereeper et al. 2015)

Table 16.1 List of available species in AgroLD

Species	Proteins
Arabidopsis thaliana	91,917
Coffea canephora	23,615
Manihot esculenta	20,437
Musa acuminata	47,304
Oryza barthii	36,673
Oryza brachyantha	40,639
Oryza glaberrima	47,570
Oryza glumipatula	72,546
Oryza longistaminata	11,548
Oryza meridionalis	24,651
Oryza punctata	8110
Oryza rufipogon	89,831
Oryza sativa	4519
Oryza sativa f. spontanea	11,545
Oryza sativa Indica group	191,871
Oryza sativa Japonica group	151,069
Setaria italica	16,775
Sorghum bicolor	24,226
Theobroma cacao	2273
Triticum aestivum	26,705
Triticum urartu	64,588
Vitis vinifera	6971
Zea mays	87,433

The table summarizes protein entities per species available in AgroLD

a rice genomic variation database. These resources bring together experimental data produced by local researchers and their partners. Table 16.3 provides an overview of the integrated data sources.

16.2.2 Towards Automation of RDF Transformations

Our contributions focus on the development of various RDF conversion workflows for large agronomic datasets. Although several generic tools exist within the Semantic Web community, including Datalift (Scharffe et al. 2012), Tarql,[1] RML.io (Dimou et al. 2014), none of them was adapted to take into account the complexity of data formats in the biological domain (e.g. VCF format) or even the complexity of the information they could contain. A simple example illustrates this complexity

[1] http://tarql.github.io

Table 16.2 List of available ontologies used to link datasets in AgroLD

Ontology	Website	Example(s)
Gene Ontology (GO)	http://geneontology.org/	http://purl.obolibrary.org/obo/GO_0008150
Plant Ontology (PO)	http://planteome.org/	http://purl.obolibrary.org/obo/PO_0025131
Plant Trait Ontology (TO)		http://purl.obolibrary.org/obo/TO_0000387
Plant Environment Ontology (EO)		http://purl.obolibrary.org/obo/EO_0007359
Sequence Ontology (SO)	http://www.berkeleybop.org/ontologies/	http://purl.obolibrary.org/obo/SO_0000104
Phenotype and Attribute Ontology (PATO)		http://purl.obolibrary.org/obo/PATO_0000462
NCBI Taxonomy		http://purl.obolibrary.org/obo/NCBITaxon_4565
Evidence code Ontology		http://purl.obolibrary.org/obo/ECO_0000033

through the GFF (Generic Feature Format) (Sequence Ontology Consortium n.d.), which represents genomic data in a TSV type format (file with tabs as separators). It contains a column with *key* = *value* type information, of variable length and having different information depending on the data source. In this case, the transformation needs to be adapted according to the data source. Furthermore, the large volume of data was a limiting factor for the above-mentioned tools.

In this context, we developed RDF conversion tools adapted to a large range of genomics data standards such as GFF, Gene Ontology Annotation File (GAF) (The Gene Ontology Consortium n.d.), Variant Call Format (VCF) (1000 Genome project Consortium n.d.) and are currently working on packaging these tools in an API (SouthGreenPlatform/AgroLD_ETL 2018 2020). These data standards represent a first step, as they are indeed the most widely used in the community. We plan to develop new models for other standards, especially for phenotypic data. For more details, refer to Venkatesan et al. (2018).

16.2.3 Semantic Annotation with Bio-Ontologies

For this phase, each dataset was downloaded from selected sources and semantically annotated with URIs of ontological terms. By the end of 2020, AgroLD included around 100 million RDF triples created by transforming more than 50 datasets from 10 data sources. In Addition, when possible, we used semantic annotations already found in datasets, such as genes or traits annotated, respectively, with GO or TO identifiers (i.e. GO:0005524 is transformed in URI). In this case, we generated

Table 16.3 Overview of the Integrated Sources

Data source	Information	Website	Species	Ontologies used
Ensembl plants (Bolser et al. 2016)	Genes, annotations	https://plants.ensembl.org	All	GO
Oryzabase (Kurata and Yamazaki 2006)	Genes, ontology associations, publications	https://shigen.nig.ac.jp/rice/oryzabase	R	GO,PO,TO
GOA (Huntley et al. 2015)	Gene ontology associations	https://www.ebi.ac.uk/GOA	A,R, W	GO
Rice Genome Hub	Genes, annotations, ontology associations	https://rice-genome-hub.southgreen.fr	R, S, A	GO, SO
Gramene (Tello-Ruiz et al. 2018)	QTL, pathways, and ontology associations	https://www.gramene.org/	R,A, W, S, M	GO, PO, TO, EO
Interpro (Mitchell et al. 2015)	Classification of protein families	https://www.ebi.ac.uk/interpro	All	GO
RAPDB (Sakai et al. 2013)	The rice annotation project	https://rapdb.dna.affrc.go.jp	R	GO
MSU RGAP	MSU rice genome annotation project	http://rice.plantbiology.msu.edu	R	GO
UniprotKB (The UniProt Consortium 2018)	Protein information	https://www.uniprot.org	All	GO
Oryza Tag Line (Larmande et al. 2008)	Rice mutant database	https://oryzatagline.cirad.fr	R	PO, TO
TropGeneDB (Hamelin et al. 2013)	Genetic, genomic, and phenotypic database	https://tropgenedb.cirad.fr	R	GO, TO, PO
GreenPhylDB (Valentin et al. 2021)	Comparative genomics	https://www.greenphyl.org	R,A	GO
RiceNetDB (Lee et al. 2015)	Gene networks database	http://bis.zju.edu.cn/ricenetdb	R	GO
StringDB (Szklarczyk et al. 2019)	Protein–protein interactions network database	https://string-db.org	R,A	GO

Species and Ontologies are referenced as follows: *R* rice, *W* wheat, *A* Arabidopsis, *S* Sorghum, *M* maize, *All* all species listed in Table 16.1. *GO* gene ontology, *PO* plant ontology, *TO* plant trait ontology, *EO* plant environment ontology, *SO* sequence ontology

additional properties with the corresponding ontologies, adding 22% more triples (see details in Table 16.3). The OWL versions of the used ontologies have been directly loaded in the knowledge graph, but not counted in total.

In addition, we used the AgroPortal Web Service API (Jonquet et al. 2018) to enrich data with semantic annotations, for instance, to extract URIs corresponding to taxons in the GFF files, but also to identify ontological concepts in the data such as a plant organ (e.g. leaf is annotated with PO:0025034) or a phenotypic trait (e.g. plant height is annotated with TO:0000207). Moreover, we developed a dedicated application (Larmande and Jibril 2020) to handle semi-structured file formats (tsv, csv, excel) and to better control semantic annotations made by AgroPortal and manage the different annotation exceptions for an optimal result.

16.2.4 Data Linking Methods

In our knowledge graph construction pipeline, RDF graphs share a common namespace and are named according to the corresponding data sources. Entities in RDF graphs are linked by the common URI principle. In general, we build URIs by referring to Identifiers.org (Laibe et al. 2014) which provides design patterns for each registered source. For instance, genes integrated from Ensembl Plants are identified by the base URI [http://identifiers.org/ensembl.plant/Os12g010180]. When they are not provided by Identifiers.org, new URIs are constructed and in this case URIs take the form [http://www.southgreen.fr/agrold/resource/Entity_ID]. In addition, the properties linking the entities are constructed as from [http://www.southgreen.fr/agrold/vocabulary/property].

In order to link identical entities from different data sources, we used the approach based on key identification, which is the most common one. Its principle is to scan the URIs in order to look for similar patterns in the terminal part of the URI (i.e. {Entity_ID}). In addition, we also followed the common URI approach which recommends to use the same URI pattern for two similar entities. Therefore, for the same entity, this allowed us to aggregate information from different RDF graphs. In addition, we used cross-reference links by transforming them to URIs and linking the resource to the rdfs predicate *seeAlso*. This significantly increases the number of outbound links, making AgroLD better integrated with other data sources. In the future, we plan to implement a *similarity entity profile* approach to identify matches between entities with different URIs (see Sect. 16.3 of the chapter).

In order to match the different data types and properties, we developed a schema that associates the classes and properties identified in AgroLD with corresponding ontologies. For instance, the *Protein class* [http://www.southgreen.fr/agrold/vocabulary/Protein] is associated with the SO *polypeptide class* [http://purl.obolibrary.org/obo/SO_000010] with the OWL property *equivalentClass*. Similar mappings have been done for the properties. For example, the has_function property is linked with the RO class [http://purl.obolibrary.org/obo/RO_0000085], with the owl property: equivalentProperty. When an equivalent property did not exist,

we associated it with the higher level property with rdfs: subPropertyOf. For example, the property *has_trait* [http://www.southgreen.fr/agrold/vocabulary/has_trait], linking entities with TO terms is associated with a more generic property from RO causally related to [http://purl.obolibrary.org/obo/RO_0002410]. So far, 55 mappings have been identified.

16.2.5 Facilitating Access to Linked Data

Regarding access to RDF data, although the SPARQL language is efficient to build queries, it remains difficult to handle for our main users, which are bioinformaticians and biologists with little or no background in formal query languages. Therefore, we propose a web application implementing various elements of semantic search systems (i.e. pattern-based querying, graphical visualization, information retrieval tools—http://agrold.org). Thus the AgroLD platform provides four entry points, as described in Venkatesan et al. (2018):

– **Quick Search**, a faceted search plugin provided by Virtuoso, which allows users to perform keyword searches and to browse easily through the results (Fig. 16.1a).

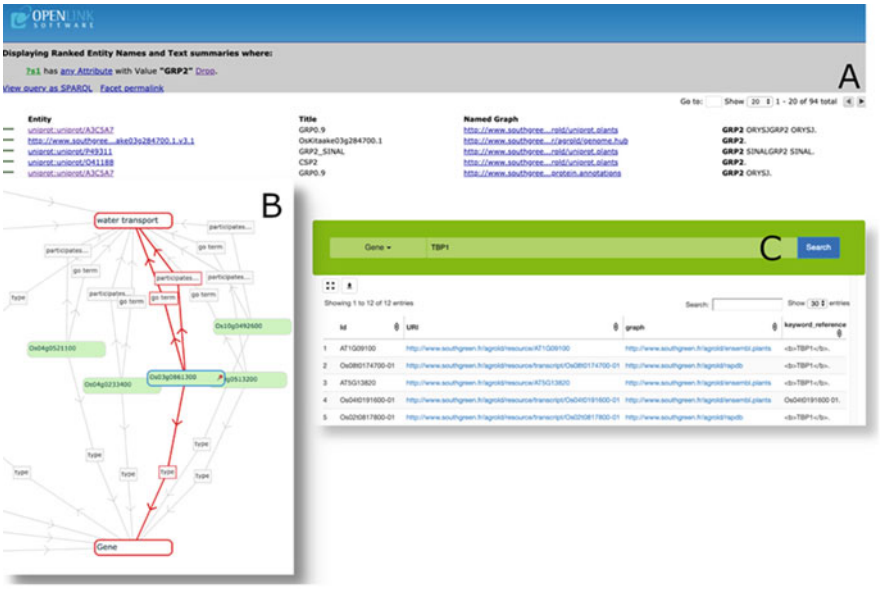

Fig. 16.1 Overview of AgroLD Web interfaces. (**a**) displays the Faceted search interface. (**b**) Displays results from the Relfinder tool. (**c**) Displays results from the advanced search interface

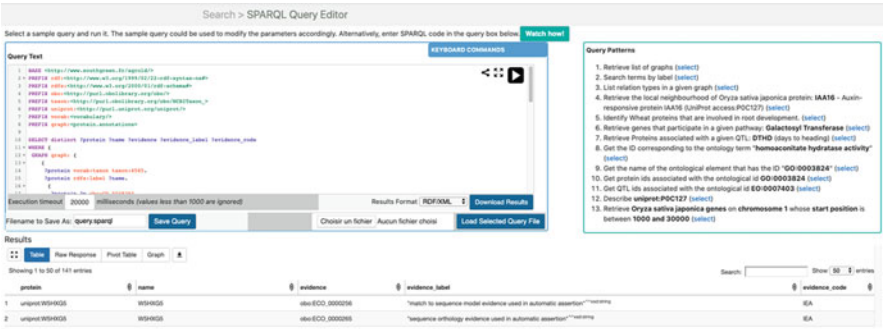

Fig. 16.2 The SPARQL query editor. The Query patterns frame allows to select a query from a natural language question. The Query text frame allows to visualize and modify the SPARQL query. The results frame displays results returned from the query

- **SPARQL Editor**, a SPARQL query editor that provides an interactive environment for formulating SPARQL queries. We developed the editor based on the YASQE and YASR (Rietveld and Hoekstra 2015) tools and adapted them for our system. In addition, we provided several SPARQL query patterns corresponding to search questions in order to help the users to dive into SPARQL syntax (Fig. 16.2).
- **Explore Relationships** is an adapted version of RelFinder (Heim et al. 2009) that allows users to explore and visualize the relationships between entities (Fig. 16.1b).
- **Advanced Search**, a search interface offering specific filters such as filtering by Gene, Protein, PathWay classes and having an aggregation engine of external web resources such as retrieving publication summary from PubMed (Fig. 16.1c).

16.3 Challenges and Future Work

The main observation made during our first phase of AgroLD development is that database resources remain limited to produce sufficient and relevant knowledge in order to formulate research questions based on molecular genes functions. Moreover, very few interactions between genes and phenotypes are explicitly mentioned. However, improving crop production requires a better understanding of these interaction mechanisms. More generally, we observed that biological resources are rich and powerful, but their potential is not currently fully unlocked due to limitations that need to be addressed and shape a set of future challenges. We believe that these limitations can be circumvented through the development of new methods. We propose in the following section to describe these approaches.

16.3.1 Extraction of Biological Entities and Their Relationships

16.3.1.1 Challenges

One of these challenges will be to enrich AgroLD with unstructured data contained in scientific publications as well as in text fields of databases (so far we have focused on structured data in these databases). Many of these text fields contain molecular mechanisms and phenotypes of interest that are often described by complex expressions associating biological entities linked by specialized semantic relationships (e.g. "*Ehd1 and Hd3a can also be downregulated by the photoperiodic flowering genes Ghd7 and Hd1*" source PMID: 20566706). In this case, the objective will be to develop computational tools to extract biological entities and their relationships in order to extract relevant information, here the entities *Ehd1, Hd3a, Ghd7, and Hd1* and the *downregulated* relationship.

16.3.1.2 State of the Art

Recently, word embedding methods have been used to improve text mining approaches. In general, they allow representation of words as vectors in *n*-dimensional space. For example, the word *man* could be represented by the vector here in two dimensions [0.33 0.98]. By representing all the words in a dictionary by the same method, it is easy to imagine that words having a semantic similarity such as *man* and *woman* will have close values in the same vector space. Moreover, these vector representations can be used to perform operations based on analogical relations, such as $king - man + woman = queen$ (Mikolov et al. 2013b).

Therefore, many representation models have been developed in order to create word embeddings. Among them we can mention the most popular Word2Vec (Mikolov et al. 2013a), Glove (Pennington et al. 2014), ELMo (Peters et al. 2018), and BERT (Devlin et al. 2018). The main difference between these models is that Word2vec and Glove do not take into account the word order in the sentence (i.e. they are independent of the context of the sentence; *in prison cell and blood cell*, the word *cell* will have the same vector), whereas ELMo and BERT take into account the word order—but with two different approaches—(i.e. they will generate different vectors for the same words depending on their context; in the previous example the word *cell* will have two different vectors).

Recent advances in biological text mining tools became possible, thanks to advances in deep learning techniques used in Natural Language Processing (NLP). For instance, to identify named entities, recent use of neural networks approaches showed better results than previous approaches. Among them, Bi-LSTM-CRF (Bidirectional Long Short Term Memory model combined with Conditional Random Fields) offered encouraging results. Habibi et al. (2017) adopted the Lample et al. (2016) model and used word vectors from word embedding (i.e. Word2Vec) as

input vectors in a bidirectional LSTM-CRF (Bi-LSTM-CRF) model. However, these methods require a large amount of data in order to optimize the training phases. More recently, improvements have been made to this method by using transfer learning (Corbett and Boyle 2018), multi-layer learning (Yoon et al. 2019) and multi-task learning (Wang et al. 2019). Finally, more recent approaches used more context-sensitive representation models (i.e. ELMo and BERT). The DTranNER application (Hong and Lee 2020) used ELMo in a Bi-LSTM-CRF architecture by enhancing the CRF labelling step with a deep learning architecture. The BioBERT application (Lee et al. 2020) used the BERT model to create contextualized word vectors. In addition, BioBERT has been trained on biomedical corpora and has been trained for relation extraction.

16.3.1.3 Planned Action

In order to compare these different approaches, we developed a corpus of rice data that can be used as a training model to detect entities and their relationships in the text. This corpus, OryzaGP (Larmande et al. 2019) consists of more than 15,000 titles and abstracts of published scientific papers on rice downloaded from PubMed. In addition, we extracted and annotated 123,146 gene mentions along with RapDB and MSU identifiers. We will use the OryzaGP as training and validation data for both approaches to feature extraction.

16.3.2 Semantic Annotation

16.3.2.1 Challenges

Semantic annotation refers to the automatic creation of a link between an entity and an ontological term. For example in the following sentence: *the protein IAA16 is expressed in the coleoptile,* Coleoptile is a biological entity referring to the same class identified by the URI obo:PO_0020033. Thus, it is possible to link entities from the same RDF graph or from different graphs as soon as they share the same semantic annotations. Ontologies allow us to create semantic links between biological entities. In our field, the conceptual framework for knowledge management is based on well-established ontologies (see Sect. 16.2). Identifying semantic links within data is an important part of building knowledge networks in AgroLD. It is also an active discipline in the computer science community (Faria et al. 2013; Otero-Cerdeira et al. 2015).

16.3.2.2 State of the Art

Several methods have been proposed to match terms from text to concept labels from ontologies in order to augment knowledge. However, few studies provide efficient methods for complex phenotypic traits or phenotypes (Harrow et al. 2017). We provide some examples describing this complexity below.

- A natural language term referring a biological entity can be represented by its symbol or acronym: for example, the MOC1 gene refers MONOCULM 1, the APO1 protein refers ABERRANT PANICLE ORGANIZATION 1;
- A natural language term referring a biological entity can be polysemic and ambiguous, therefore difficult to annotate;
- A term corresponding to a phenotype can implicitly refer to several concept labels from different ontologies. For example, the *Dwarfism* phenotype can be annotated with the *dwarf-like* concept from the PATO (Phenotype And Trait Ontology) ontology, but it also matches the *Tillering* concept from the PO (Plant Ontology) ontology and the *Tiller angle* concept from the TO (Trait Ontology) ontology;
- A term corresponding to a phenotype can be annotated using two ontologies. For example, the *wrinkled seed* phenotype is composed of the *wrinkled* concept label from the PATO ontology and the *seed* concept label from the PO ontology.

16.3.2.3 Planned Action

To meet these challenges, powerful semantic annotation tools often rely on a combination of word processing, knowledge bases, semantic similarity measures, and machine learning techniques (Jovanović and Bagheri 2017). Agroportal (Jonquet et al. 2018) aims to develop a portal of reference ontologies for agronomy. It also aims to provide several search and semantic annotation tools. As indicated in Jonquet et al. (2018), we plan to develop an annotation workflow between AgroPortal and AgroLD which will include similarity measures, word processing, and use AgroPortal features to annotate data with ontological concepts.

16.3.3 Reasoning over Linked Data

16.3.3.1 Challenges and State of the Art

The RDF model used in AgroLD is also complemented by other structuring languages to describe data (RDFS (RDF Schema 1.1 n.d.), OWL (OWL Web Ontology Language Overview n.d.), and SKOS (SKOS Simple Knowledge Organization System Namespace Document 30 July 2008 'Last Call' Edition n.d.)) or describe their constraints (ShEx (Shape Expression Vocabulary n.d.), SHACL

(Shapes Constraint Language (SHACL) n.d.)). The use of schemas—also called ontologies when their structure is more complex—on data allows to classify them in the form of classes of entities, relationships, and instances. Thus, it is possible to implement reasoning mechanisms, thanks to ontologies. For example, generalization/specialization relationships are frequently used in reasoning to propagate information. In this case, if we define *Class B subclass of Class A*, if the entity E1 is an instance of B, then it will be also classified as an instance of A. It is also possible to use reasoning to enrich the links existing between data. For example, when using symmetric or transitive relations. For example, in the case of protein–protein interaction networks, defining *interact_with* as reflexive or for gene co-expression networks, defining *coexpress_with* as transitive will allow the reasoner to enrich the information if the data is incomplete.

Another advantage offered by Semantic Web technologies is the use of rule-based languages to validate constraints on data. This includes the emerging languages ShEx and SHACL. In general, few reasoning methods and tools have been developed on real data and specifically in the agronomic field. This will open a large field of exploration in the future.

16.3.4 Data Linking

16.3.4.1 Challenges

The Data linking process aims at establishing semantic links of equivalence or other type between entities from different RDF graphs. Data linking is an important part of the integration process because it allows the aggregation of various properties to the same entity, thus enriching its overall description. For example, as shown in Fig. 16.3, let us consider a biological entity identified by two distinct URI *ensembl:OS01G0675800* and *oryzabase:11464* in two different datasets. It is therefore not possible to determine whether they are identical. However, a

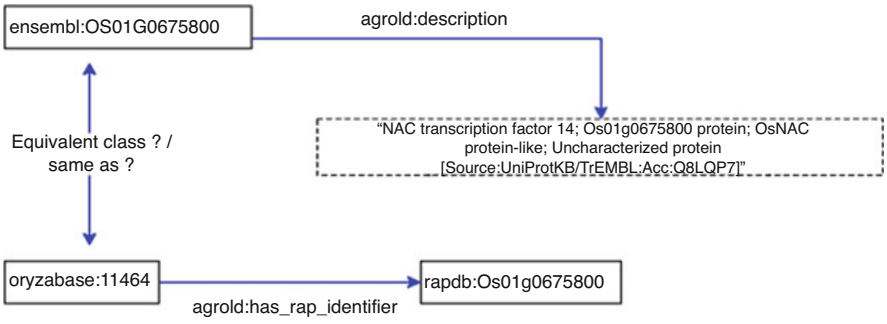

Fig. 16.3 Data Linking issue example in AgroLD

biologist can confirm their similarity based on their properties *agrold:description* and *agrold:has_rapdb_identifier*. In fact, we find the presence of the identifier "*OS01G0675800*" from the rapdb resource, associated with the second entity, in the description and URI of the first entity. Example in Fig. 16.4 shows two biological entities. These entities correspond to the APO1 protein but are considered distinct because they have different URIs. Moreover, the linking task is even more difficult when the properties describing them are heterogeneous. One question is to determine the properties to be used as a basis for comparison. Another question is to determine how the attributes are valued or structured in order to avoid creating erroneous links or missing links. As shown in Fig. 16.4, descriptions can be expressed in different natural languages, with different vocabularies or different values.

16.3.4.2 State of the Art

These limits can be classified into three dimensions listed in Fig. 16.4: value dimension, ontological dimension, and logical dimension.

The value dimension refers to properties containing literal (text) values expressed in natural language or numerical values that can lead to binding errors. The authors (Achichi et al. 2019) identify four levels of heterogeneity in this dimension, also indicated in Fig. 16.4: value type, terminology, linguistics, best practices.

- ***Value type heterogeneity.*** This heterogeneity concerns the way literal values are encoded (e.g. string, integer, etc.). In this case, the challenge lies in the harmonization of the value types, for example, standardizing the formats of dates, numerical measurements, etc.
- ***Terminology heterogeneity.*** In this case the differences will concern a term corresponding to a word or a group of words. This variation can be expressed in different ways: (1) synonymy when different terms represent the same concept; (2) polysemy when similar terms have different meanings; (3) acronyms and abbreviations. As can be seen in Fig. 16.4, a name of the entities corresponds to an abbreviation (i.e. APO1). To overcome this problem, it is possible to expand acronyms and abbreviations.
- ***Linguistic heterogeneity.*** The terms involved come from different languages. This is a frequent issue when working with experimental data from diverse sources that reflect the diversity of information that can be found on the Web. In this case, with English and Japanese, the similarity search tools are not efficient. It is necessary to go through an automatic translation step in prior.
- ***Best practices heterogeneity.*** Knowledge representation is subject to design pattern practices. Their transgression is a barrier in the discovery of correspondences.

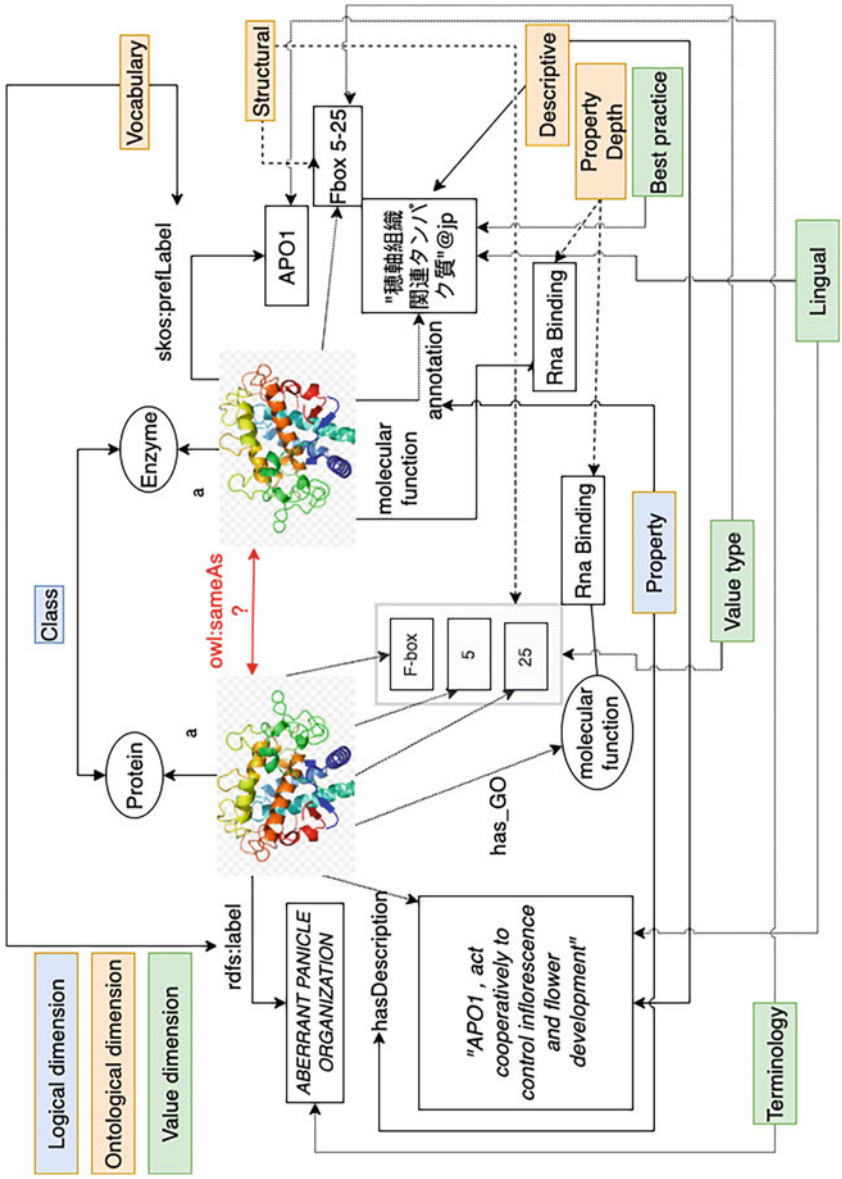

Fig. 16.4 Data Linking challenges in biology. (Inspired from Achichi et al. 2019)

The ontological dimension refers to the class or property variations associated with the compared instances. Four levels of heterogeneity are identified: vocabulary, structure, property depth, and descriptions.

- ***Vocabulary heterogeneity*.** Classes and properties are often described, by different data producers, using different vocabularies. This problem is even more complicated in the context of the Web of Data where not all resources are described in the same way. The use of mapping between vocabularies, such as Agroportal in our case, can help overcome this problem.
- ***Structural heterogeneity***. The description of an entity can be made at different levels of granularity. In this example the term Fbox 5-25 is structured differently: the information is embedded in the data structure for the first entity while it is included in a literal for the second. The use of NLP methods to extract the information can help in the linking process.
- ***Property depth heterogeneity.*** This heterogeneity is located at the resource schema level and corresponds to property modelling variations. Here, the literal "DNA Binding", which is a molecular function, is modelled from a GO class for the first entity and a property for the second. So the distance between the entity and the elements is greater for the first one. Possible methods to solve this type of problem could be to index the literals with their context in order to be able to compare them.
- ***Descriptive heterogeneity.*** Entities can be described with a larger set of properties in a dataset compared to another. In Fig. 16.4, we can see that entities contain more descriptive information (text literal fields) than the set of properties describing them. It is obvious that comparing these resources only by their properties will be less efficient than approaches taking into account the whole set of information.

The logical dimension refers to the fact that the equivalence between two entities is implicit, but can be deduced using reasoning methods. Two main heterogeneity problems are identified:

- ***Class heterogeneity.*** This type of heterogeneity refers to the level of the class hierarchy. This is generally the case when two resources belonging to different classes have explicit or implicit hierarchical relationships (the concepts "Protein" and "Enzyme", in Fig. 16.4, illustrate this problem because the Enzyme class is a subclass of Protein).
- ***Property heterogeneity.*** At this level, the equivalence between two values is deduced after the completion of a property reasoning task. For example, two resources referring to the same entity can have two properties that are semantically inverted (i.e. the has Description and is Annotated By properties). In this case, these two properties contain the same information, as shown in the example in Fig. 16.4.

16.3.4.3 Planned Action

Data linking is an active research field that developed a plethora of approaches (Manel et al. 2016). Regarding the state of the art, we listed various software that implement linking methods, the following are the most cited or recent ones (Silk (Jentzsch et al. 2010), Limes (Ngomo and Auer 2011), Legato (Achichi et al. 2017)). One of the major challenges is to manage datasets with limited overlap in terms of the properties used to describe their resources, what we call complementary datasets. This missing information makes it difficult for recent systems based solely on property analysis to assess relationships between instances. The datasets integrated in AgroLD present largely this problem.

Few methods have been developed with real life data and none in the field of plant genomics or phenomics. We will propose to develop a method adapted to the context of AgroLD taking into consideration the challenges outlined above.

- **Text mining:** We plan to exploit the textual content of RDF graphs using natural language processing techniques to identify named entities and reconstruct their relationships, allowing the discovery of relevant links between related RDF graphs.
- The **knowledge graph augmentation** techniques add structured information to existing RDF graphs by exploring relevant external data on the Web (e.g. markup data, scientific articles, (social) media, other knowledge graphs). We will apply these methods to enrich our datasets and reconstruct missing information.
- **Machine learning for complementary datasets.** We will explore the relevant criteria that effectively represent inter-graph resources and classify them as identical (or not) by machine learning. We will use vector models for pairs of instances and train on the input–output relationships from the training data. A training dataset on AgroLD data is currently under construction.

16.4 Discussion

We have seen that the use of a common representation format is important to integrate many heterogeneous and distributed biological data sources. Semantic Web technologies are well adapted to enable this integration. Indeed, they provide a FAIR structure for sharing biological knowledge and benefit from the support of a large computer science community. However, in order to fully unlock the potential of these biological resources, new methods need to be developed.

- First, we need to develop methods to extract meaningful information embedded in unstructured data such as text fields from databases or even web documents and scientific publications. Because molecular mechanisms and phenotypes associating biological entities are often described in natural language by human experts. Therefore it is important to be able to process such data and create links with database entries. In addition, related information can be extracted from

images. We refer to Ubbens and Stavness (2017), Pound et al. (2017), Choudhury et al. (2019) for a review of these methods.
– Second, we need to improve semantic annotation methods in order to cover a large domain of plant science. Indeed, currently semantic methods are well adapted to the genomic domain because they are shared with the biomedical domain. But these methods are less advanced for the plant phenomics or other related plant field studies. Since plant high-throughput phenotyping technologies are gaining in popularity, consequently we are witnessing the development of several related ontology projects such as the Agronomy ontology, the planteome project, and Agroportal (Cooper et al. 2018; Jonquet et al. 2018; Devare et al. n.d.). This will help to strengthen semantic annotation methods in our domain.
– Third, we need to apply more symbolic approaches such as reasoning or rule-based constraint checking than we used to do formerly. Indeed, biology data is often incomplete and contains implicit knowledge, reasoning approaches can improve the pre-processing step by enriching the data. Frequently, data scientists methods are used to clean and normalize data as a pre-processing step but such approaches are rarely used. Combining these approaches with traditional machine learning could be a powerful way to achieve the ultimate goal of revealing genotype–phenotype interactions (van Harmelen and ten Teije 2019; Marcus 2020).
– Finally, because biology data is complex, incomplete, and with low coverage complementarity between datasets, we need to develop new data linking methods. These methods should combine both state-of-the-art techniques from the computer science community and the specificity of the biological domain in order to overcome these barriers. Research direction should combine (1) natural language processing techniques to extract embedded information in unstructured text fields, (2) knowledge augmentation by exploiting external resources, and (3) machine learning techniques to infer new relationships.

As a perspective of this experience, applying a candidate gene prioritization approach makes it possible to identify and classify among a large number of genes those that are strongly associated with the phenotype. There are many approaches to identifying candidate genes (Moreau and Tranchevent 2012). The recent success of graph models and deep learning in bioinformatics suggests the possibility of systematically incorporating multiple sources of information into a heterogeneous network and learning the non-linear relationship between phenotype and candidate genes (Alshahrani and Hoehndorf 2018). Graphs are powerful tools to represent the interactions between entities. Thus, they are well-fitted to represent each type of interaction that occurs in biological networks. Because AgroLD is based on RDF, a label-oriented multi-graph representation model, the platform will be suitable for evaluating this type of approach. Information retrieval among knowledge graphs requires the development of methods to sort the results in a meaningful way. In the future, we will seek to develop an approach adapted to the context of AgroLD that takes into consideration the challenges outlined above.

References

1000 Genome project Consortium (n.d.) Variant Call Format (VCF). Accessed 20 March 2018

Achichi M, Bellahsene Z, Todorov K (2017) Legato results for OAEI 2017. In: Proceedings of the 12th international workshop on ontology matching co-located with the 16th international semantic web conference (ISWC 2017), Vienna, Austria, 21 Oct 2017, pp 146–152. http://ceur-ws.org/Vol-2032/oaei17_paper6.pdf

Achichi M, Bellahsene Z, Ellefi MB, Todorov K (2019) Linking and disambiguating entities across heterogeneous RDF graphs. J Web Semant 55:108–121. https://doi.org/10.1016/j.websem.2018.12.003

Alshahrani M, Hoehndorf R (2018) Semantic disease gene embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes. Bioinformatics 34(17):i901–i907. https://doi.org/10.1093/bioinformatics/bty559

Belleau F, Tourigny N, Good B, Morissette J (2008) Bio2RDF: a semantic web atlas of post genomic knowledge about human and mouse. In: Data integration in the life ..., pp 153–160

Berners-Lee T, Hendler J, Lasilla O (2001) The Semantic Web. Scientific American 284(5):34–43

Bolser D, Staines DM, Pritchard E, Kersey P (2016) Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. Methods Mol Biol 1374:115–140. https://doi.org/10.1007/978-1-4939-3167-5_6

Buttigieg, Luigi P, Morrison N, Smith B, Mungall CJ, Lewis SE, ENVO Consortium (2013) The environment ontology: contextualising biological and biomedical entities. J Biomed Semant 4(1):43. https://doi.org/10.1186/2041-1480-4-43

Choudhury D, Sruti AS, Awada T (2019) Leveraging image analysis for high-throughput plant phenotyping. Front Plant Sci 10:508. https://doi.org/10.3389/fpls.2019.00508

Cooper L, Meier A, Laporte MA, Elser JL, Mungall C, Sinn BT, Cavaliere D et al (2018) The planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res 46(D1):D1168. https://doi.org/10.1093/nar/gkx1152

Corbett P, Boyle J (2018) Improving the learning of chemical-protein interactions from literature using transfer learning and specialized word embeddings. Database 2018:bay066. https://doi.org/10.1093/database/bay066

Dereeper A, Homa F, Andres G, Sempere G, Sarah G, Hueber Y, Dufayard J-F, Ruiz M (2015) SNiPlay3: a web-based application for exploration and large scale analyses of genomic variations. Nucleic Acids Res 43(W1):W295–W300. https://doi.org/10.1093/nar/gkv351

Devare M, Aubert C, Laporte M-A, Valette L, Arnaud E, Buttigieg PL (2016) Data-driven agricultural research for development: a need for data harmonization via semantics. International Conference on Biomedical Ontologies (ICBO). pp 3–5

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint ArXiv:1810.04805

Dimou A, Sande MV, Colpaert P, Verborgh R, Mannens E, Van De Walle R (2014) RML: a generic language for integrated RDF mappings of heterogeneous data. In: CEUR workshop proceedings, vol 1184

Droc G, Périn C, Fromentin S, Larmande P (2009) OryGenesDB 2008 update: database interoperability for functional genomics of rice. Nucleic Acids Res 37(Database issue):D992–D995. https://doi.org/10.1093/nar/gkn821

FAO (n.d.) How to feed the world in 2050. UN. http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf. Accessed 19 Feb 2021

Faria D, Pesquita C, Santos E, Palmonari M, Cruz IF, Couto FM (2013) The AgreementMakerLight ontology matching system. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 8185 LNCS. https://doi.org/10.1007/978-3-642-41030-7_38

Habibi M, Weber L, Neves M, Wiegandt DL, Leser U (2017) Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33(14):i37–i48. https://doi.org/10.1093/bioinformatics/btx228

Hamelin C, Sempere G, Jouffe V, Ruiz M (2013) TropGeneDB, the multi-tropical crop information system updated and extended. Nucleic Acids Res 41(D1):D1172. https://doi.org/10.1093/nar/gks1105

van Harmelen F, ten Teije A (2019) A boxology of design patterns for hybrid learning and reasoning systems. ArXiv:1905.12389 [Cs]. https://doi.org/10.13052/jwe1540-9589.18133

Harrow I, Jiménez-Ruiz E, Splendiani A, Romacker M, Woollard P, Markel S, Alam-Faruque Y, Koch M, Malone J, Waaler A (2017) Matching disease and phenotype ontologies in the ontology alignment evaluation initiative. J Biomed Semant 8(1):1–13. https://doi.org/10.1186/s13326-017-0162-9

Heim P, Hellmann S, Lehmann J, Lohmann S, Stegemann T (2009) RelFinder: revealing relationships in RDF knowledge bases. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 5887 LNCS, pp 182–187. https://doi.org/10.1007/978-3-642-10543-2_21

Hong SK, Lee J-G (2020) DTranNER: biomedical named entity recognition with deep learning-based label-label transition model. BMC Bioinform 21(1):53. https://doi.org/10.1186/s12859-020-3393-1

Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C (2015) The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res 43(Database issue):D1057–D1063. https://doi.org/10.1093/nar/gku1113

Jentzsch A, Jentzsch A, Isele R, Bizer C (2010) Silk – generating RDF links while publishing or consuming linked data. In: Proceedings of ISWC

Jonquet C, Toulet A, Arnaud E, Aubin S, Yeumo ED, Emonet V, Graybeal J et al (2018) AgroPortal: a vocabulary and ontology repository for agronomy. Comput Electron Agric 144(October 2016):126–143. https://doi.org/10.1016/j.compag.2017.10.012

Jovanović J, Bagheri E (2017) Semantic annotation in biomedicine: the current landscape. J Biomed Semant 8(1):44. https://doi.org/10.1186/s13326-017-0153-x

Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, Gaulton A et al (2014) The EBI RDF platform: linked open data for the life sciences. Bioinformatics 30:1338. https://doi.org/10.1093/bioinformatics/btt765

Kurata N, Yamazaki Y (2006) Oryzabase. An Integrated Biological and Genome Information Database for Rice. Plant Physiology 140(1):12–17. https://doi.org/10.1104/pp.105.063008

Laibe C, Wimalaratne S, Juty N, Le Novère N, Hermjakob H (2014) Identifiers. Org: integration tool for heterogeneous datasets. Dils 2014:14. https://doi.org/10.6084/m9.figshare.1232122.v1

Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C (2016) Neural architectures for named entity recognition. http://arxiv.org/abs/1603.01360

Larmande P, Jibril KM (2020) Enabling a fast annotation process with the Table2Annotation tool. Genomics Informat 18:e19. https://doi.org/10.5808/GI.2020.18.2.e19

Larmande P, Gay C, Lorieux M, Périn C, Bouniol M, Droc G, Sallaud C et al (2008) Oryza tag line, a phenotypic mutant database for the genoplante rice insertion line library. Nucleic Acids Res 36(Database issue):D1022–D1027. https://doi.org/10.1093/nar/gkm762

Larmande P, Do H, Wang Y (2019) OryzaGP: rice gene and protein dataset for named-entity recognition. Genomics Informat 17(2):e17. https://doi.org/10.5808/GI.2019.17.2.e17

Lee T, Oh T, Yang S, Shin J, Hwang S, Kim CY, Kim H et al (2015) RiceNet v2: an improved network prioritization server for rice genes. Nucleic Acids Res 43(W1):W122–W127. https://doi.org/10.1093/nar/gkv253

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Manel A, Zohra B, Konstantin T (2016) A survey on web data linking. Ingénierie Des Systèmes d'information 21(5–6):11–29. https://doi.org/10.3166/isi.21.5-6.11-29

Marcus G (2020) The next decade in AI: four steps towards robust artificial intelligence. ArXiv:2002.06177 [Cs]. http://arxiv.org/abs/2002.06177

Mikolov T, Chen K, Corrado G, Dean J, Sutskever L, Zweig G (2013a) Word2vec. https://Code.Google.Com/p/Word2vec

Mikolov T, Yih W-t, Zweig G (2013b) Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 746–751

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C et al (2015) The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43(Database issue):D213–D221. https://doi.org/10.1093/nar/gku1243

Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet 13(8):523–536. https://doi.org/10.1038/nrg3253

Ngomo ACN, Auer S (2011) Limes-a time-efficient approach for large-scale link discovery on the web of data. In: Proceedings of IJCAI, 2312–17. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-385

Otero-Cerdeira L, Rodríguez-Martínez FJ, Gómez-Rodríguez A (2015) Ontology matching: a literature review. Expert Syst Appl 42(2):949–971. https://doi.org/10.1016/j.eswa.2014.08.032

OWL Web Ontology Language Overview (n.d.). https://www.w3.org/TR/owl-features/. Accessed 11 Feb 2021

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. ArXiv Preprint ArXiv 1802:05365

Pound MP, Atkinson JA, Townsend AJ, Wilson MH, Griffiths M, Jackson AS, Bulat A et al (2017) Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. GigaScience 6(gix083):1–10. https://doi.org/10.1093/gigascience/gix083

RDF Schema 1.1 (n.d.). https://www.w3.org/TR/rdf-schema/. Accessed 11 Feb 2021

Redaschi, Nicole, and UniProt Consortium (2009) UniProt in RDF: tackling data integration and distributed annotation with the semantic web. Nat Proc. https://doi.org/10.1038/npre.2009.3193.1

Rietveld L, Hoekstra R (2015) The YASGUI family of SPARQL clients. Semantic Web J 8:373

Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, Wakimoto H et al (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant & Cell Physiology 54(2):e6. https://doi.org/10.1093/pcp/pcs183

Scharffe F, Atemezing G, Troncy R, Gandon F, Villata S, Bucher B, Hamdi F et al (2012) Enabling linked data publication with the Datalift Platform. http://www.eurecom.fr/en/publication/3707/detail

Sequence Ontology Consortium (n.d.) GFF3 specification. https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

Shape Expression Vocabulary (n.d.). https://www.w3.org/ns/shex. Accessed 11 Feb 2021

Shapes Constraint Language (SHACL) (n.d.). https://www.w3.org/TR/shacl/. Accessed 11 Feb 2021

SKOS Simple Knowledge Organization System Namespace Document 30 July 2008 'Last Call' Edition (n.d.). https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html. Accessed 11 Feb 2021

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ et al (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech 25(11):1251–1255. https://doi.org/10.1038/nbt1346

South Green, Collaborators (2016) The South Green Portal: a comprehensive resource for tropical and Mediterranean Crop Genomics South Green Collaborators. Curr Plant Biol 78:6–9. https://doi.org/10.1016/j.cpb.2016.12.002

SouthGreenPlatform/AgroLD_ETL (2018) 2020. Python. South Green Bioinformatics platform. https://github.com/SouthGreenPlatform/AgroLD_ETL.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M et al (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47(D1):D607–D613. https://doi.org/10.1093/nar/gky1131

Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, Wei S et al (2018) Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucleic Acids Res 46(D1):D1181–D1189. https://doi.org/10.1093/nar/gkx1111

The Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still going strong. Nucleic Acids Res 47(D1):D330–D338. https://doi.org/10.1093/nar/gky1055

The Gene Ontology Consortium (n.d.) Gene annotation file (GAF) specification. http://geneontology.org/page/go-annotation-file-format-20. Accessed 20 March 2018

The Plant Ontology Consortium (2002) The Plant Ontology Consortium and plant ontologies. Comp Funct Genomics 3(2):137–142. https://doi.org/10.1002/cfg.154

The UniProt Consortium (2018) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47(D1):D506–D515. https://doi.org/10.1093/nar/gky1049

Ubbens JR, Stavness I (2017) Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. Front Plant Sci 8:1190. https://doi.org/10.3389/fpls.2017.01190

Valentin G, Abdel T, Gaëtan D, Jean-François D, Matthieu C, Mathieu R (2021) GreenPhylDB v5: a comparative pangenomic database for plant genomes. Nucleic Acids Res 49(D1):D1464–D1471. https://doi.org/10.1093/nar/gkaa1068

Venkatesan A, Ngompe GT, El Hassouni N, Chentli I, Guignon V, Jonquet C, Ruiz M, Larmande P (2018) Agronomic Linked Data (AgroLD): a knowledge-based system to enable integrative biology in agronomy. PLoS One 13:17. https://doi.org/10.1371/journal.pone.0198270

W3C (n.d.) Resource Description Framework (RDF): concepts and abstract syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. Accessed 3 April 2010

Wang W, Mauleon R, Zhiqiang H, Chebotarov D, Tai S, Zhichao W, Li M et al (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557(7703):43–49. https://doi.org/10.1038/s41586-018-0063-9

Wang X, Zhang Y, Ren X, Zhang Y, Zitnik M, Shang J, Langlotz C, Han J (2019) Cross-type biomedical named entity recognition with deep multi-task learning. Bioinformatics (Oxford, England) 35(10):1745–1752. https://doi.org/10.1093/bioinformatics/bty869

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Yoon W, So CH, Lee J, Kang J (2019) CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. BMC Bioinformatics 20(10):249

Chapter 17
Interactive Data Analyses Using TBtools

Chengjie Chen and Rui Xia

Abstract Increasing biological data provide us unprecedented ability to uncover the mystery of life. To leverage oncoming large biological data, efficient and effective data analysis is indispensable for biological research. However, data analysis has become a major challenge to biologists, most of who are not skillful in computer science, thereby limiting the utilization efficiency of biological data. Although a lot of bioinformatics software have been developed in the community, the majority of them require users to work under command-line environments or even be familiar with programming languages, with few of them focusing on freeing users from elaborate command-line-based tasks.

Here, we present TBtools, an out-of-box solution to routine biological data analyses. The toolkit integrates ~150 practical functions for data analyses and visualization, with a user-friendly graphical interface. In this chapter, we describe the design philosophy, development objectives, and main characteristics of TBtools. We also provide a comprehensive introduction of its main functions, especially those included in the "Sequence Toolkits" and "Graphics" catalogs, and advanced features, like R plugins that contributed by senior users. A few practical tutorials are presented to demonstrate the superb functionalities and outstanding interactive nature of TBtools.

Keywords TBtools · Bioinformatics · Function integration · Data analysis · Data visualization

17.1 Design Philosophy and Development Objectives

In recent years, high-throughput sequencing technologies have been developing rapidly in the field of life sciences, and big data analyses have become an indispensable part of biological research. For the vast majority of biologists, the effective

C. Chen · R. Xia (✉)
College of Horticulture, South China Agricultural University, Guangzhou, Guangdong, China
e-mail: rxia@scau.edu.cn

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_17

use of these data is not only an opportunity, but also a challenge. For instance, through genome-wide association analysis with population resequencing and large-scale trait surveys, we can identify key genes associated with certain important traits efficiently. Constructing gene co-expression networks from transcriptional expression profiles, we can screen out potential hub regulatory factors of key traits and steer the direction of further gene function studies. Most of these analyses often involve two stages, which we simplified as upstream and downstream data analyses, and for both of them, researchers are required to have two major aspects of knowledge, computer science and biology. Upstream data analyses often require large-scale data operations, which run on high-performance servers and consume a lot of computation resources, such as profiling whole-genome SNP sites from resequencing data with data size of >10 TB, or calculating gene expression levels from >100 GB transcriptome sequencing data. These analyses are common procedure for different projects, mainly involving large data calculations and little relevance to specific biological questions. Many powerful bioinformatics software or tools have been developed to meet this type of common demand. Often commercial service providers can be relied on this part of analyses, alternatively, researchers can learn to use these tools in a project and reuse them repeatedly for different projects. In contrast, downstream data analyses are much more complicated and need "personalized recipe" of analyses to solve various biological questions. Normally there is no routine way to follow for these analyses, including various small tasks of different purposes, such as conversion of all kinds of file formats, sorting and extraction of certain text information, representation of distinct results, etc. To handle these small and specific analyses, researchers are often required to search, test, and learn to use a large number of tools (commands or tools composed in different programming languages) for different functions or to program them into different pipelines to achieve certain analysis goals. And in most cases, this process of searching, testing, learning is not repeatable among different projects, therefore greatly increases the cost of simple data analyses and reduces the efficiency of scientific research. Based on these observations, since 2015, we have been developing a bioinformatics combo toolkit, TBtools, to integrate hundreds of data analysis functions routinely needed from biological laboratories, for streamlined and simplified usage. This chapter gives an overview and introduction on the development and usage of TBtools.

17.1.1 Development Logic

"Know the tricks of the trade," this idiom fits to the process of learning and mastering any bioinformatics software.

17.1.1.1 Practical First, Concise Utmost

The development of TBtools starts from the demand of daily data analyses of the authors. The realization of bioinformatics functions is designed to fit the needs of daily data analysis in the most practical and simplest way, for example, the routine bulk sequence extraction from fasta sequence files and the simple local BLAST sequence search (Altschul et al. 1997). The design of the software interface follows the rule of "the simpler, the better," only retaining concise prompt message and necessary parameter settings. All the rest, including intermediate files generated automatically are hidden without showing. And many options of certain functions are simplified or automated, such as the automatic recognition of the format of input data (e.g., the automatic selection of BLAST sub-programs based on the query and subject sequence types), automatic file format conversion, and programmed adjustments of file names (such as file names containing spaces or special characters) (Fig. 17.1).

17.1.1.2 The Simple "IOS" Logic

The implementation of each function in TBtools strictly follows the most basic programming logic:

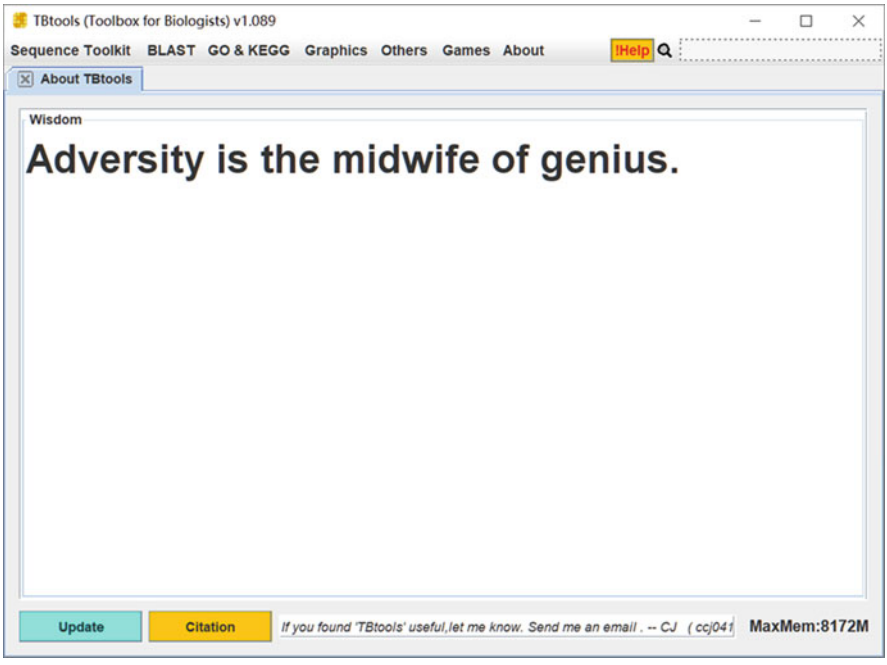

Fig. 17.1 TBtools Main Interface

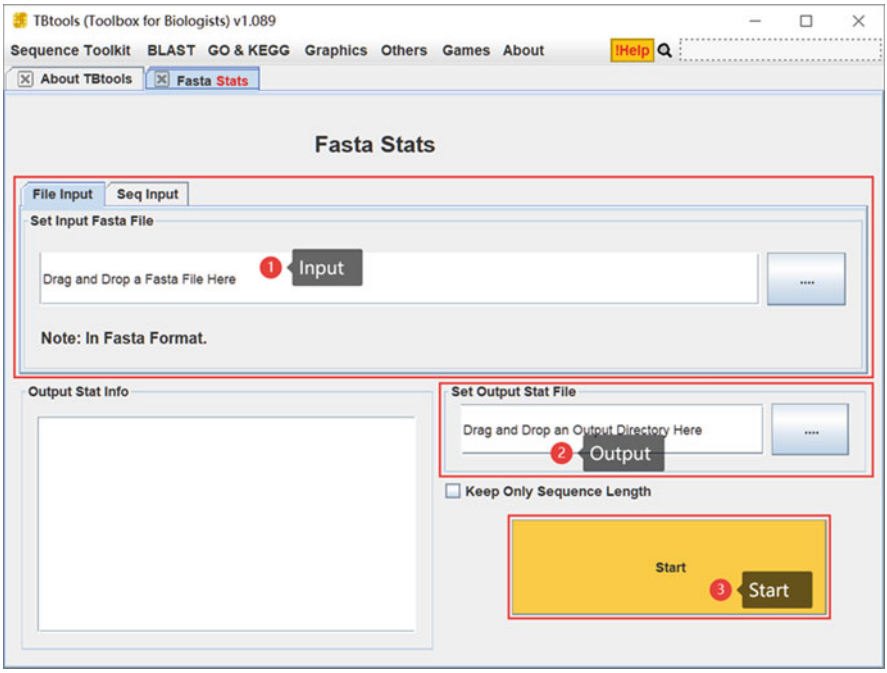

Fig. 17.2 The "IOS" Logic of TBtools Interface

(a) set Input data
(b) set Output data
(c) press "Start" button to proceed

that is, "Input, Output, and Start" logic (Fig. 17.2).

When using any function of TBtools, users only need to set the input and output files, and then click the "Start" button to start the analysis. Worthy of note is that the "Start" button is always bright yellow, which is easy for a quick notice.

17.1.1.3 User-Driven Development

The main driving force for the development of TBtools is real data analysis demands from hundreds of thousands of scientific researchers. Since the first release of TBtools in 2016, there are >50,000 stable users. Users' feedbacks such as suggestions, comments, and new demands are the motives of our continuous development and innovation of the software. The software has a public program repository at Github, https://github.com/CJ-Chen/TBtools/releases, where installers of TBtools can be downloaded and users can report usage issues. We have also established real-time user communication communities in Tencent and Telegram (Fig. 17.3; http://118.24.17.128/TBtoolsUserGroup.png and https://t.me/

Fig. 17.3 TBtools Real-Time User Community

joinchat/Q6WHOhWOlHHOkMgRGoN7nw), and a TBtools user forum (http://www.tbtools.icu:1234). These communities ensure that the development team to communicate with software users in real time, collecting user feedbacks in time for troubleshooting and further development of new features.

17.1.2 Development Objectives

17.1.2.1 One-Click Environment Configuration

In projects dominated by biological questions, bioinformatics is more of a powerful technology. In actual data analyses, users are required to install different software for different needs. For example, when plotting a genome circle map, we need to configure the Perl language environment before installing the Circos software package (Connors et al. 2009). During this period, a lot of source code compilation work is involved, and different software installation problems are often encountered, such as incompatible system platform versions. In fact, installing and configuring software on servers is already a critical step in bioinformatics data analysis that consumes lots of time and energy. Therefore, we have implemented almost all functions of TBtools using pure Java code, ensuring cross-platform characteristics. For a small number of mature software, we have gathered their cross-platform binaries and packaged them into single software installers.

17.1.2.2 Graphics User Interface for Data Analysis

At present, common bioinformatics data analysis tools often require users to be familiar with programming languages such as R language, or familiar with command-line working environments such as DOS or Shell. The learning curve is steep. For most researchers most of whose work is not done on computers, learning costs are too high. It is difficult to master and easy to forget. The initial goal of TBtools development is to enable all users to quickly master the use of the software and start data analyses right out of the box. Although TBtools can also be run through the command line, we focus on creating specific user-friendly interfaces for each useful function.

17.1.2.3 Function Integration

In daily data analyses, users need to use different software in combination to complete a simple analysis task. For instance, to "assess the similarity of two protein sequences in a certain species," we may need to use a text editor or write a script to extract two protein sequences, then use BLAST (local or web) to compare the two sequences, and finally use other software or tools to visualize the results of the comparison. Frequent switching of software takes much time of scientific researchers, and besides, it is easy to interrupt users' thoughts. One of the development goals of TBtools is to fully integrate simple analysis functions. Users can quickly complete sequence extraction, BLAST and direct visualization in TBtools. Covering more than 150 functions, users can integrate them according to their needs, fully meeting other analysis scenarios.

17.1.2.4 Analysis Automation

Although downstream data analyses do not have a similar obvious pattern as upstream data analysis, there are still many simple and repetitive tasks in some analysis tasks. At present, the genomes of more and more species are sequenced and published. Comparative genomics has become a research hotspot. Among them, one common analysis is mining gene collinear blocks. Correspondingly, the most widely used software is MCscan (Wang et al. 2012). Though the need for analysis is common, the use of MCscan requires users to prepare rather cumbersome input files. In short, users should obtain protein sequences of two species, invoke software such as BLASTP for sequence comparison, integrate gene location information, and finally run the MCscan software. During this period, identifier mismatch may be involved, identifier naming system conflicts and file name prefixes are not unified, all of which will cause the final operation to fail or lead to incorrect results. With TBtools, users only need to place genome sequence and gene structure annotation files of two species and click the Start button.

17.1.2.5 Simplify Complex Analysis

Some data analysis tasks are not only cumbersome but also difficult to achieve. The eFP Browser development team proposed for the first time that expression value coloring on a cartoon vividly displays specific gene expression changes (Winter et al. 2007). This has been applied to a few model organisms, which could be found on the corresponding species genome website. At present, omics data is becoming more and more abundant. Cartoon-style heatmaps can be used for research and result display on all other species, and data analysis and result interpretation can be carried out more intuitively. However, eFP Browser is a browser framework, involving knowledge of computer and network configuration, which is almost impossible for scientific researchers with a pure biological background to implement in a short time. Based on similar ideas, TBtools implements a java-based non-dependency eFP graph function. Users only need to prepare a cartoon template, an expression matrix, and a color mapping relationship table to make eFP graphs. Furthermore, compared to eFP Browser, TBtools supports the output of vector graphics to ensure the clarity and interactivity of the final artwork. A similar advance can also be found on the plotting of Circos. On the whole, TBtools enables biological researchers to or even easy to complete a large number of analyses that seemed difficult before.

17.2 Overview of TBtools Functions

17.2.1 Software Acquisition, Update, and Main Interface

17.2.1.1 Software Acquisition

There are two main ways to get TBtools installation.

(a) Download it directly from the software repository at Github, https://github.com/CJ-Chen/TBtools/releases (Fig. 17.4).
(b) Obtain the latest version of the TBtools installation from the user communities at Tencent or Telegram.

Fig. 17.4 Overview of the TBtools Warehouse

17.2.1.2 Software Update

TBtools has been incorporated with background update, real-time update, and automatic update. While running, TBtools will automatically detect the current version in the background and ask the user whether to update to the latest version, ensuring that users can always use the latest version of the software with more comprehensive and more stable functions.

In an unstable network environment, users can directly obtain the TBtools main program (a jar file) elsewhere and then manually update the software in the two following ways.

(a) Go directly to the main directory of the TBtools installation and complete the program update by replacing the main program file (TBtools_JRE1.6.jar) with a newer one.
(b) Enter the "About" catalog from the main menu of TBtools, click "Update via Jar," and select the downloaded .jar file in the pop-up dialog to complete the update (Fig. 17.5).

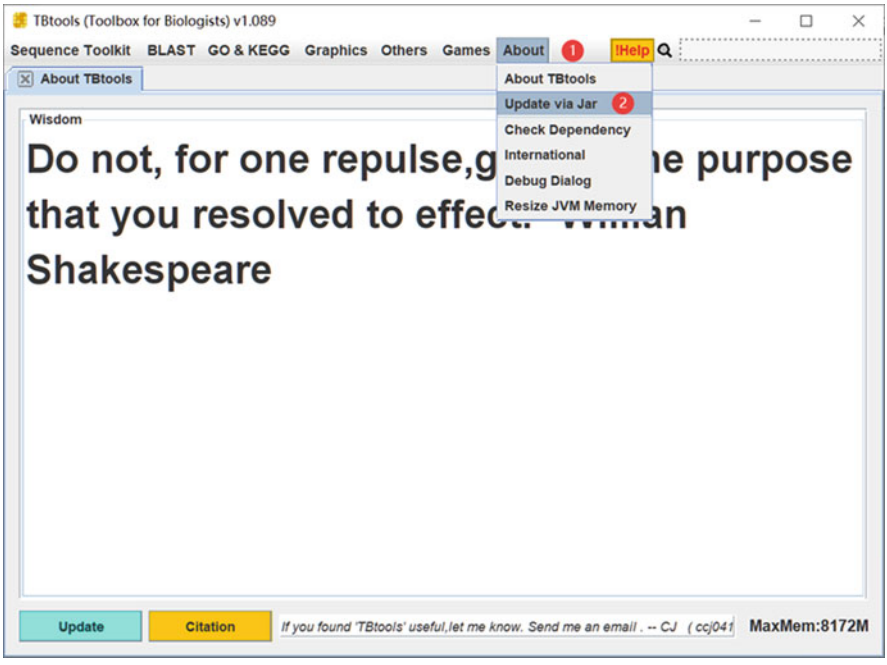

Fig. 17.5 Update TBtools from Main Menu

17.2.1.3 Main Interface of TBtools

After the TBtools software is installed, you can directly run it by a double-click on the program icon from the Start menu. The software will launch and you can see the main interface, which includes several main function catalogs (Fig. 17.6).

1. **Version of TBtools.** It can be used to check whether the current version is the latest one. Version number should be provided when communicate with us for troubleshooting.
2. **Main menu.** TBtools currently divides the main functions into six catalogs (described in detail below).
3. **About Panel.** When TBtools starts, the "About Panel" function will be automatically triggered, and a famous quote will be randomly displayed.
4. **"!Help" button.** When users are confused about a certain function, or do not know how to use or which function to used, they can click this button to get usage examples and related tutorials.
5. **Search box for functions.** There are >150 functions in TBtools. Sometimes it is hard to quickly locate a function level by level through the main menu. A more convenient way is to directly enter keywords for a specific function in the search box (e.g., "Fasta"), and then TBtools will automatically show functions related to

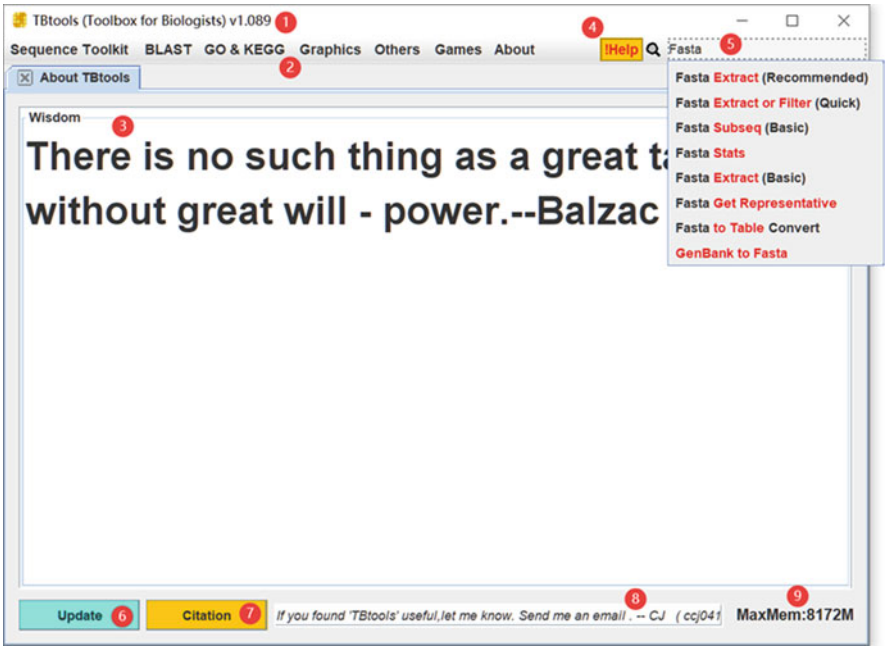

Fig. 17.6 Overview of the Main Interface of TBtools

the keywords (names containing "Fasta") and users can use the function directly by a simply click.
6. **"Update" button.** Users can manually click the "Update" button to force the update of TBtools via background update.
7. **"Citation" button.** This button directs the users to the webpage of the official TBtools publication, for a convenient citation of the software when preparing manuscripts.
8. **Message box.** The message box is used for the users to send their feedbacks (suggestions and comments, or even complimentaries) to developers via email.

17.2.2 Introduction to TBtools Function

17.2.2.1 Overview of Main Functions

The main functions of TBtools are divided into five main catalogs (Fig. 17.7).

1. **Sequence Toolkits**. This catalog mainly includes batch sequence download, bulk sequence extraction, sequence information sorting, and other sequence handling functions. Among them, the "GFF3 Sequence Extract" function is a powerful tool that could be used to extract specific feature sequences based on gene structure annotation information. Users often employ it to obtain the complete set of CDS or gene promoter sequences.
2. **BLAST.** It collects a series of functions from a stand-alone BLAST wrapper, as well as functions for format conversion, result management, and visualization.
3. **GO & KEGG.** This catalog hosts functions for gene set analyses, for example, gene ontology and KEGG pathway enrichment, and for result management and visualization as well.
4. **Graphics**. It covers functions most often used for data representation and visualization, such as venn diagram, heatmap, and seqlogo, as well as a few relatively complex graphs, for example, upset plots and circos.
5. **Others.** Functions that could not be clearly grouped in the previous four catalogs are placed under this one, including functions for text manipulation and phylogenetic analysis.

   In addition to the five main catalogs, TBtools also hosts a few games for pleasure in an extra "Games" catalog, as users may be free when use TBtools for large data analysis which take a long time to process. There is also a "About" catalog which includes options to manually update the software, adjust the maximum available memory for software operation, check whether the dependent programs are complete, view the operation information, etc. (Fig. 17.8).

TBtools currently covers more than 150 functions, and each main catalog has a series of corresponding functions. Limited by the space, we cannot introduced all

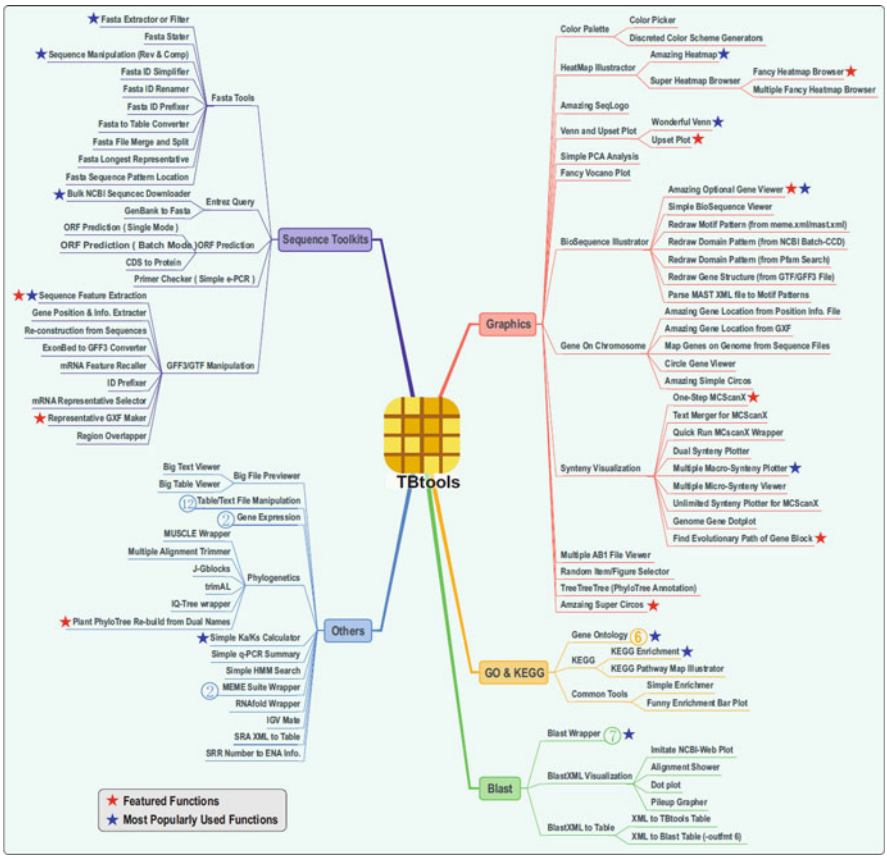

Fig. 17.7 Overview of TBtools Functions (Chen et al. 2020)

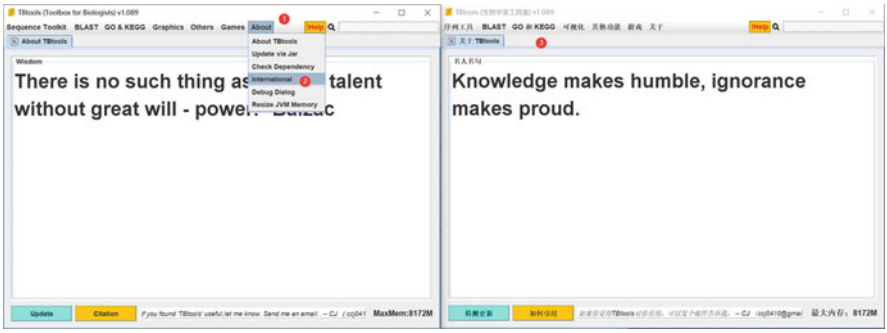

Fig. 17.8 TBtools Interface with Different Languages

of them. The following are the introduction of functions typically used for sequence manipulation and data visualization.

17.2.2.2 Functions for Sequence Manipulation

TBtools contains a large number of functions used for sequence management, mainly for files in Fasta or GFF3/GTF formats (Fig. 17.9).

Sequence Toolkit Functions are divided into five sub-menus.

1. **Fasta Tools**
   (a) **Fasta Extract / Filter.** Batch extract or filter sequence records.
   (b) **Fasta Stat.** Summarize sequence information for a sequence file, such as total sequence length, GC content, etc.
   (c) **Sequence Manipulate.** Manipulate sequences, such as reverse, complement and information sorting.
   (d) **ID Simplify/Rename/Prefix.** Manipulate sequence identifiers, such as, simplify or rename the identifier, or add prefix.
   (e) **Fasta to Table Convert.** Convert sequence file between Fasta and tab-delimited formats.

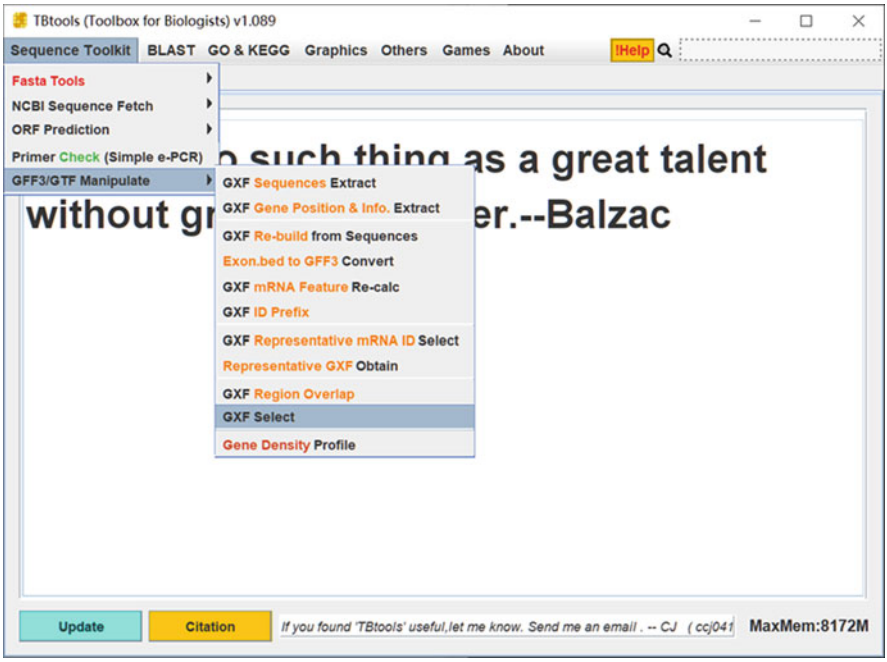

Fig. 17.9 Overview of TBtools Sequence Toolkits

(f) **Fasta Merge/Split.** Merge and split Fasta sequence files.
(g) **Fasta Get Representative.** Extract representative Fasta sequence records.
(h) **Sequence Pattern Locate.** Locate sequence regions that possess specific patterns.

2. **NCBI Sequence Fetch**

(a) **Bulk NCBI Sequence Download.** Download sequences from NCBI using accession numbers.
(b) **GenBank to Fasta.** Convert files from Genbank format to Fasta format.

3. **ORF Prediction**

(a) **Complete ORF Prediction (Single Mode).** Predict complete ORFs of a sequence in six frames.
(b) **Complete ORF Prediction (Batch Mode).** Batch predict the longest ORF in a set of sequences.
(c) **Batch Translate CDS to Protein.**

4. **Primer Check (Simple e-PCR).** Detect all possible amplified fragments for given primers in a specific sequence library.
5. **GFF3/GTF Manipulate**

GFF3/GTF files are standard annotation files storing gene structure information for any give genome sequence files.

(a) **GXF Sequence Extract.** Extract sequences of specific feature from the genome sequence file based on the GFF3/GTF file.
(b) **GXF Gene Posi. and Info. Extract.** Extract gene location and annotation information of a species.
(c) **GXF Re-build from Sequences.** Reconstruct a GFF3 file from given transcript sequences and corresponding reference genome sequences.
(d) **GXF mRNA Feature Re-calc.** Recalculate and add mRNA information to GXF files.
(e) **GXF ID Prefix.** Add specified prefixes to chromosome ID and gene ID.
(f) **GXF Representative mRNA ID Select.** Obtain all representative transcript IDs from GFF3/GTF files.
(g) **Representative GXF Obtain.** Generate GFF3 files containing only representative transcript information.
(h) **GXF Region Overlap.** Extract the annotation information that overlaps with specific intervals based on the GFF3/GTF file.
(i) **GXF Select.** Extract annotation information related to a given ID.
(j) **Gene Density Profile.** Calculate gene density for any given genome.

17.2.2.3 Graphics

To facilitate users to perform more visual data analysis, we have developed a Java plotting engine, JIGplot, from scratch. On this basis, visualization functions of a series of bio-information data analyses are implemented.

1. **Color Palette**
   (a) **Color Picker**
   (b) **Discrete Color Scheme Generator**
2. **Heatmap Illustrator**
   (a) **HeatMap**
   (b) **Cubic HeatMap**
   (c) **Layout HeatMap**
   (d) **eFP Graph**
3. **SeqLogo**
4. **Venn and Upset Plot**
   (a) **Venn**
   (b) **Upset Plot**
5. **Basic PCA Analysis**
6. **Volcano Plot**
7. **BioSequence Structure Illustrator**
   (a) **Gene Structure View (Advanced)**
   (b) **Basic BioSequence View**
   (c) **Visualize Motif Pattern**
   (d) **Visualize Domain Pattern (Batch-CDD / Pfam)**
   (e) **Visualize Gene Structure**
   (f) **Parse MAST XML File**
8. **Show Gene on Chromosome**
   (a) **Gene Location Visualize (Advanced)**
   (b) **Gene Location Visualize from GTF/GFF**
   (c) **Map Genes on Genome from Sequence Files**
   (d) **Gene Location Visualize (Basic)**
   (e) **Circle Gene View**
   (f) **Basic Circos**
9. **Advanced Circos**
10. **Synteny Visualization**
    (a) **Genome Length Filter**. Filter small sequence fragments from a genome sequence file.
    (b) **Genome Analysis Init.** Prepare files for comparative genomic analysis
    (c) **Quick Run MCScanX Wrapper**

(d) **One Step MCScanX**. Perform MCscan Analysis in One-click
(e) **Dual Synteny Plot for MCScanX**
(f) **Text Merge for MCScanX**
(g) **Multiple Synteny Plot**
(h) **Text Transformat for Micro-Synteny View**
(i) **Multiple Micro-synteny View**
(j) **Unlimited Synteny Plot for MCScanX**
(k) **Find Gene Block Evolutionary Path by Gene Pairs**
(l) **Genome Gene Dotplot**

11. **Multiple AB1 File View**
12. **Random Item/Figure Select**
13. **Tree Annotation.** Phylogenetic tree annotation and visualization.

17.2.3 Plugin Module

Functions included in the main program TBtools are those commonly used in daily bioinformatics data analysis. There are still many other functions that are useful and fit to certain needs, although not in great demand, such as the sequence conversion from Fastq to Fasta, PubMed search result management, Excel and text format conversion, and so on.

For these functions, we have developed the plugin module in TBtools, allowing users to install corresponding plugins for the functions they need (without re-installing TBtools or installing other plugins). There are currently two main acquisition modes for plugins.

17.2.3.1 User Community

Users can download plugin files with the extension .plugin in TBtools communities. After that, from the "Others" menu select "Plugin" and then "Install Plugin" (Fig. 17.10).

Select the corresponding .plugin file from the pop-up window and install it. Certainly, "drag and drop" is also supported.

17.2.3.2 Plugin Store Online

To make plugin installation more convenient, we have developed a "Plugin Store," which deposited all the available plugins. Users can find "Plugin Store" in "Plugin" under the "Others" catalog. After launching the plugin store, users can see a list of plugins (Fig. 17.11).

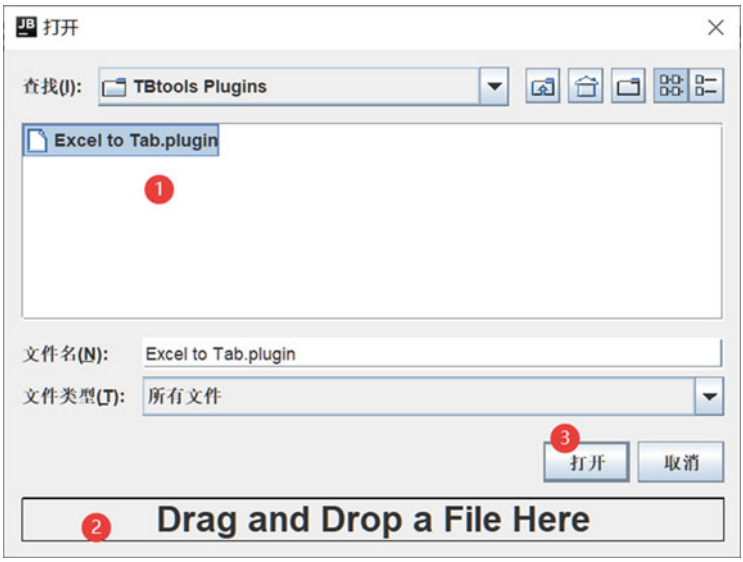

Fig. 17.10 Manually Install TBtools plugins

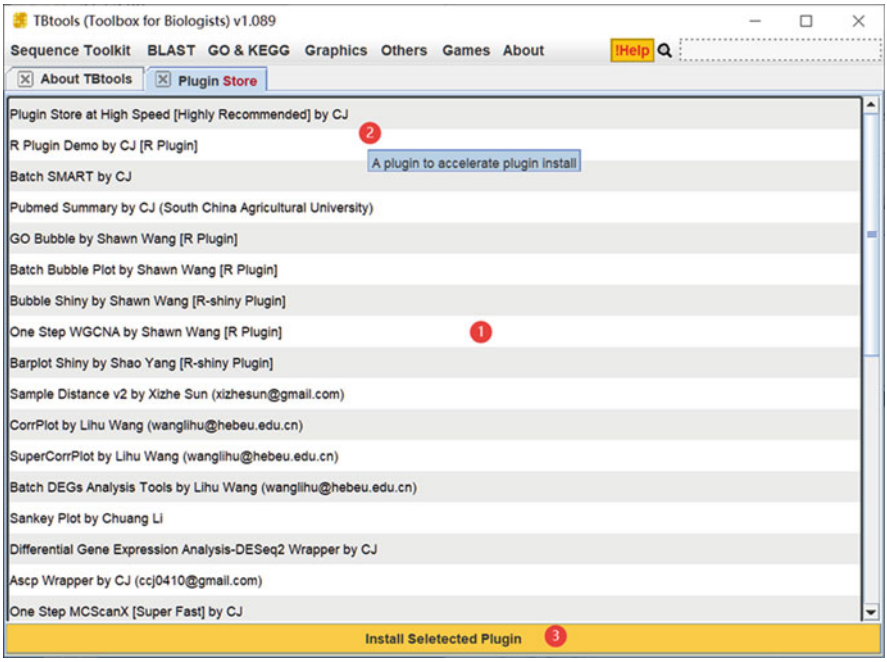

Fig. 17.11 Overview of Online Plugin Store of TBtools

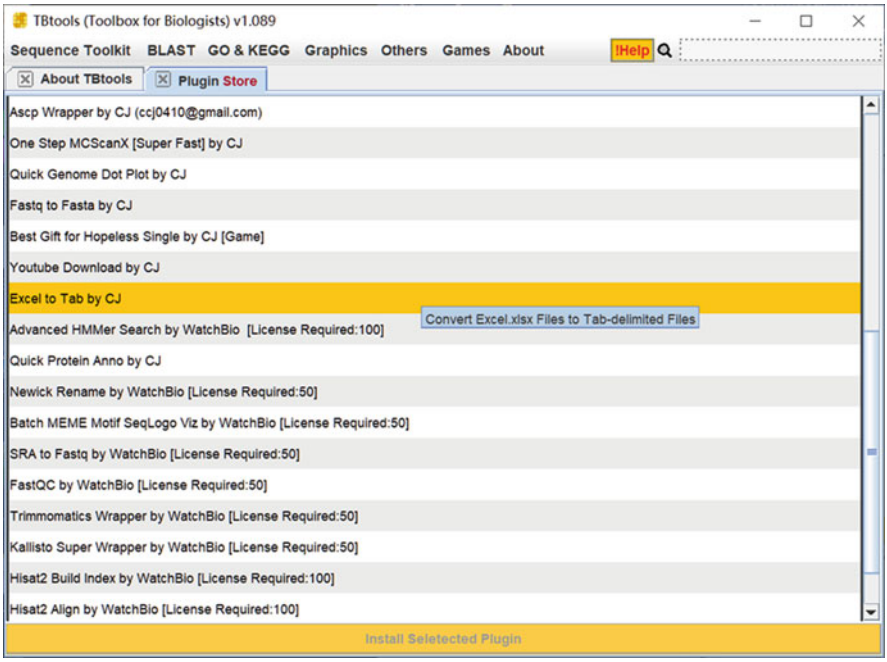

Fig. 17.12 Install a Plugin from Plugin Store

There are currently more than 30 plugins available. When the mouse hovers over a specified plugin, users can see a brief function description of the plugin. Select an ideal plugin item and click "Install Selected Plugin" to install it (Fig. 17.12).

17.2.3.3 Senior Users Participate in Plugin Development (R Plugin)

TBtools also supports users' participation in software development. Recently, we developed Rserver "plugin", an R runtime environment plugin (windows and mac version), to support direct running of R scripts. Based on this, we further developed the "R Plugin Demo" plugin. As long as users have a runnable R language script, a simple configuration can make the script into a TBtools plugin that can be used by others.

Here is an example, assuming that users currently have a script named **script.r**. Its content is

```
### Parameter acquisition
argv <- commandArgs(TRUE)
expfile <- argv[1]
title <- argv[2]
logTran <- argv[3]
colorSet <- argv[4]
titleColor <- argv[5]

### Dependent package detection and installation
if (!require('ggplot2')) install.packages('ggplot2',
repos='https://mirrors.tuna.tsinghua.edu.cn/CRAN/')
if (!require('reshape2')) install.packages('reshape2',
repos='https://mirrors.tuna.tsinghua.edu.cn/CRAN/')

### Data Processing
library(ggplot2)
library(reshape2)
expMat<-read.table(expfile,header = T,sep="\t")
head(expMat)
expMat<-melt(expMat)
if(logTran=="true") expMat$value<-log(expMat$value+1)
p<-ggplot(expMat)
p+geom_density(aes(x=value,fill=variable),alpha=I(1/4))+
  labs(title=title)+
  scale_fill_brewer(palette=colorSet)+
  theme(plot.title=element_text(size=25, hjust=0.5, face="bold", colour=titleColor,
vjust=-1))
```

The script can be invoked with the following command.

```
Rscript script.r "fpkm.xls" "R-ggplot2 BarPlot" "false" "Set1" "#e31a1c" "OutDir"
```

Users only need to prepare a few files which could be found in the "R Plugin Demo" plugin (most of them are optional) and configure the **config.txt** file to complete the plugin development (Fig. 17.13). To date, nearly ten senior users have turned their R scripts into TBtools plugins, covering a series of functions.

1. Batch Bubble Plot (desktop and shiny version)
2. Barplot (shiny version)
3. Gene co-expression Analysis (WGCNA)
4. Sample co-relation analysis
5. Differential Gene Expression Analysis (DESeq2/edgeR)
6. Sankey Plot

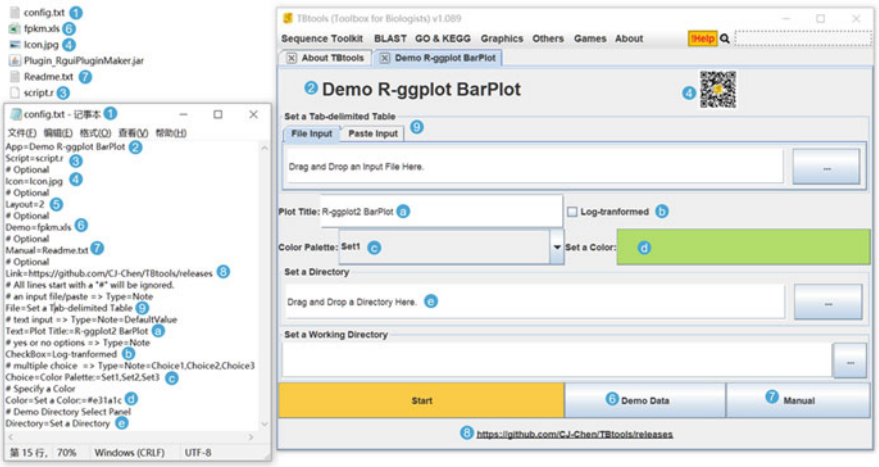

Fig. 17.13 A Demo to Make a TBtools R plugin

17.3 Demonstrations

The first step in learning swimming is getting into the water. In this section, we select four popularly used functions to demonstrate the powerful functionality of TBtools.

17.3.1 Genomic Feature Sequence Extraction Based on GFF3/GTF

With the rapid development of sequencing technologies, more and more genomes have been sequenced, which greatly promotes the scientific research on genomics. Effectively extracting and using genomic sequences becomes a routine task. TBtools provides a robust function, "GXF Sequence Extract," for quick extraction of certain sequences from genome sequences. According to the IOS logic, users only needs to set the Input and Output:

(a) **Input data.** Gene structure annotation information (GFF3/GTF format) and genome sequence (FASTA format) files of a species.
(b) **Output file.** Output file (FASTA format) path.

Most commonly, this function is used for extraction of coding sequences (CDS) or regulatory sequences (e.g., promoter).

17.3.1.1 Coding Sequence (CDS) Extraction

Open the TBtools software and select the function "Sequence Toolkit" -> "GFF3/GTF Manipulate" -> "GXF Sequence Extract" (Fig. 17.14).

1. Set the gene structure annotation information file (GFF3 or GTF format).
2. Click the "Initialize" button.
3. After that, the "Start" button will change from unavailable (gray) to available (black).
4. Select the sequence feature tag to indicate which type of sequences to be extracted. Here, it is "CDS".
5. In a GFF3 file, a sequence feature record (Feature Tag, such as CDS, EXON, mRNA) corresponds to a sequence segment of the genome and generally has grouping information with a unique ID information tag (Feature ID, such as Parent, transcript id, gene id). To obtain the complete CDS sequences of a species, we need to combine multiple CDS segment records with unified grouping information tags. Thus, we select "Parent" here.

   After the initiation is finished, then go the extraction (Fig. 17.15).

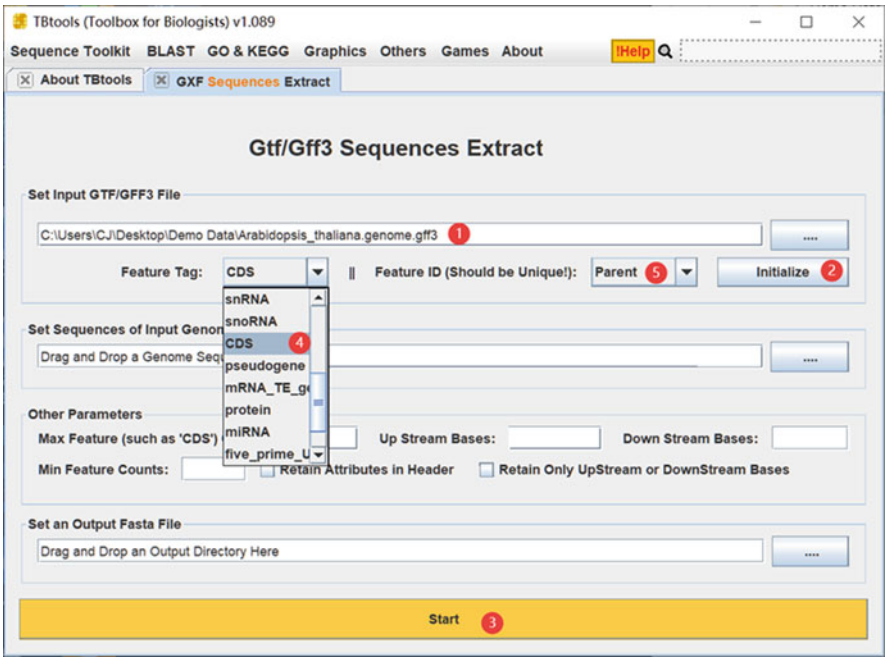

Fig. 17.14 Extraction of a Complete Set of CDS Sequences (initialization)

6. Set the genome sequence information. Please ensure that the chromosome ID is consistent between the GFF3 and the genome sequence files.
7. Set the output file path. A complete output file name is needed.
8. Click "Start" to complete the extraction. A file containing all the sequences you want will be generated in the Output directory.

17.3.1.2 Regulatory Sequence (Promoter) Extraction

In biological research, biologists usually are interested in the regulatory sequences of important genes and need to grab these sequences for further analyses. Among them, the promoter sequences of genes are the most popularly investigated. Generally, sequences of 2–3 kb upstream from the translation start site or the start codon (for these gene loci lack of UTR information) are used as a promoter sequence for analyses. TBtools can be used to get these sequences easily (Fig. 17.16).

1. Set the upstream 1000 bp sequence before CDS as the target region.
2. Check the box of "Retain only Upstream or Downstream Bases" to ensure only the specified (upstream or downstream) will be obtained; otherwise, both the upstream/downstream and CDS sequences will be extracted at the same time.
3. Set the output file path.

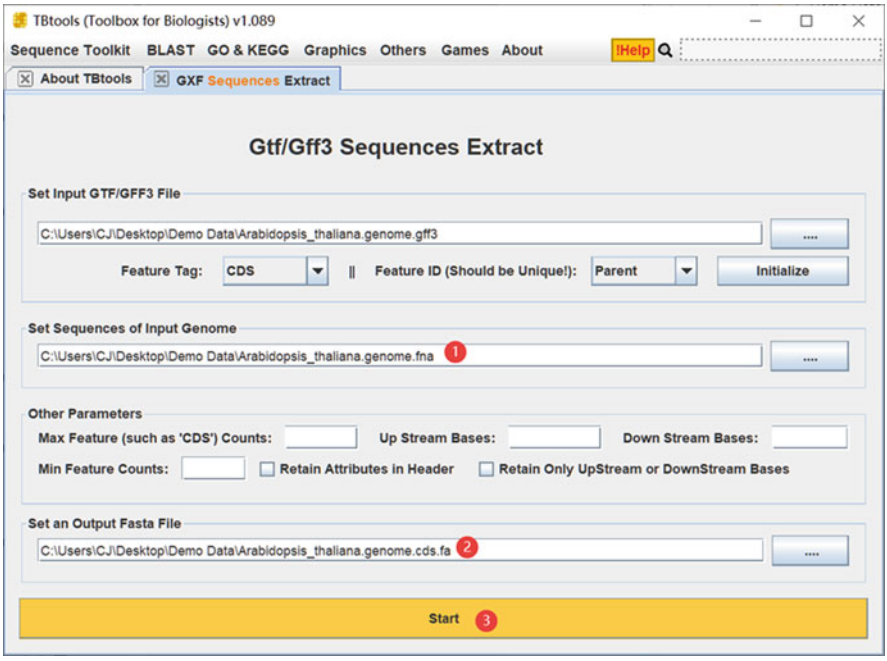

Fig. 17.15 Extraction of a Complete Set of CDS Sequences (extraction)

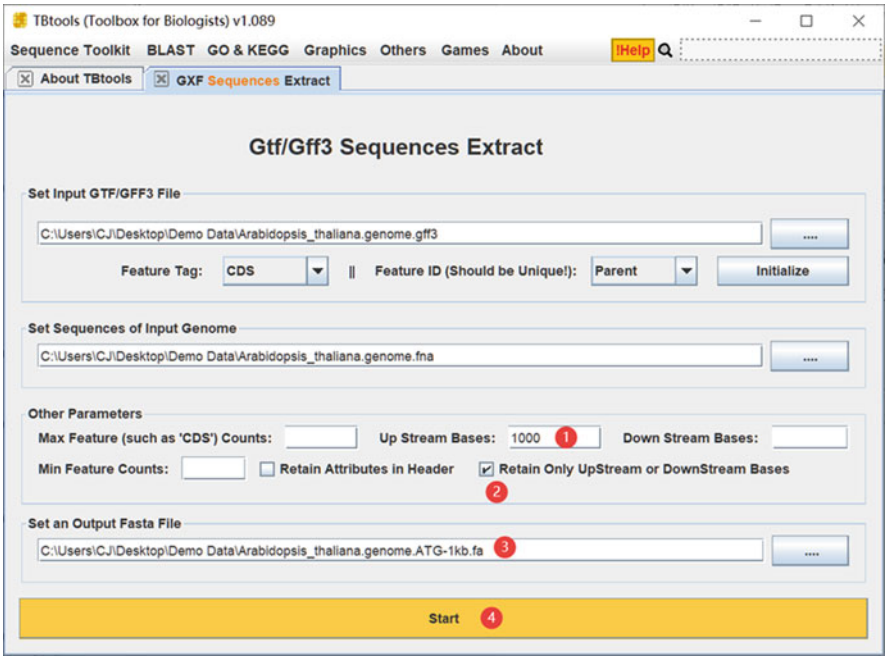

Fig. 17.16 Extraction of a Complete Set of Regulatory Sequences (promoter)

4. Click the "Start" button.

In addition, "GXF Sequence Extract" can also be used to extract a complete set of transcripts (combine multiple exon sequence together) or other sequences. The "Retain Attributes in Header" setting can be used to keep the original sequence annotation information in the output file.

17.3.2 Heatmap

Heatmap is one of the most popular graphs used for data visualization in bioinformatics data analyses. Based on its home-brew plotting engine JIGplot, TBtools provides a convenient and powerful heatmap function. Users can quickly make personalized heatmaps by using various interactive features.

17.3.2.1 Make a Simple Heatmap in a Short Time

1. Prepare a file containing a matrix of gene expression values with row and column names.
2. Paste or drag and drop the matrix file as the Input.
3. Click the "Start" button.

A heatmap plotting window will pop up instantly (Figs. 17.17 and 17.18).

17.3.2.2 Adjust the Heatmap Parameters

The heatmap graph can be personalized easily, such as data normalization, clustering of rows and columns, displaying numerical values and other information, etc. (Fig. 17.19).

1. Use the built-in color pattern to choose ideal color Scheme.
2. Use 0–1 normalization methods to format the input value matrix.
3. Cluster the data in rows and columns and display the original values in the heatmap.
4. Adjust the width of the picture (so that the number can be completely displayed in a cell).

17.3.2.3 Make a Circular Heatmap

Compared with the existing heatmap plotting tools, TBtools heatmap supports more flexible parameter adjustment. For example, users can "bend" (circularize) the heatmap to make full use of a limited space to show more information (Fig. 17.20).

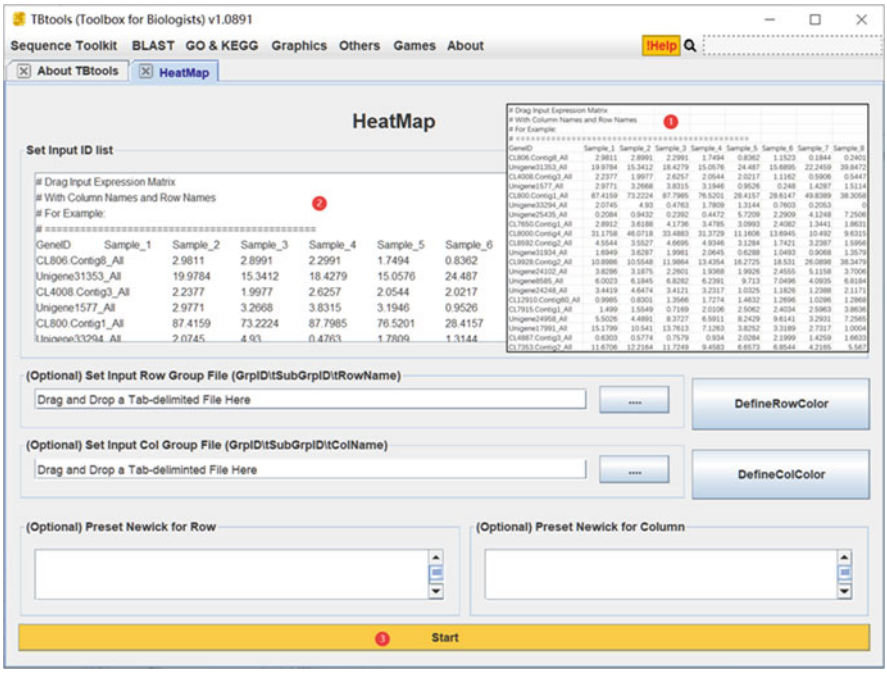

Fig. 17.17 Use TBtools to Make a Simple Heatmap

To make a circular heatmap, users only need check the "Auto Polar" box on the control panel. For further improvement, users can change the way of legend presentation by checking "Horizontal Legend".

17.3.3 Circos Plot

Circos plot is a widely used visualization approach to display large-scale genomic data. It is often used to present results at the whole genome scale to provide a comprehensive data overview. Making a Circos plot using the original package requires users to be proficient in Perl or R language programming, which inadvertently limits its application in more scientific research projects. As TBtools can easily "bend" (circularize) graphs as shown above, the Circos plotting is also supported, but in a much easier way. Users only need to prepare a few input files for graph generation without any programming or command-line operations.

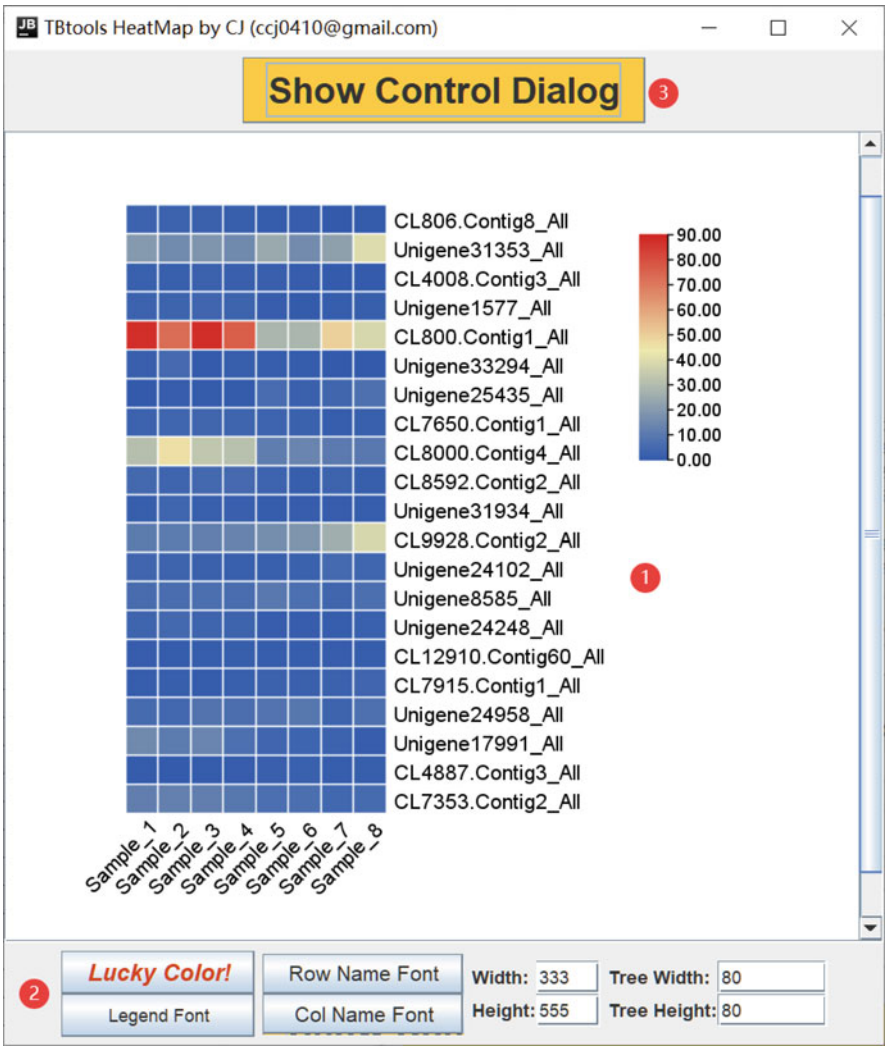

Fig. 17.18 Heatmap Plotting Window

17.3.3.1 Making the Chromosome Skeleton

With TBtools, users can make a Circos plot step by step, depending on the number of datasets to show. At first, users need to prepare the innermost chromosome skeleton track. A file containing the chromosome skeleton information ("Chromosome ID\tChromosome length" or "Chromosome ID\tChromosome starting position:Chromosome ending position") is needed, and this file can be prepared using the "Fasta Stat" function in TBtools (Fig. 17.21).

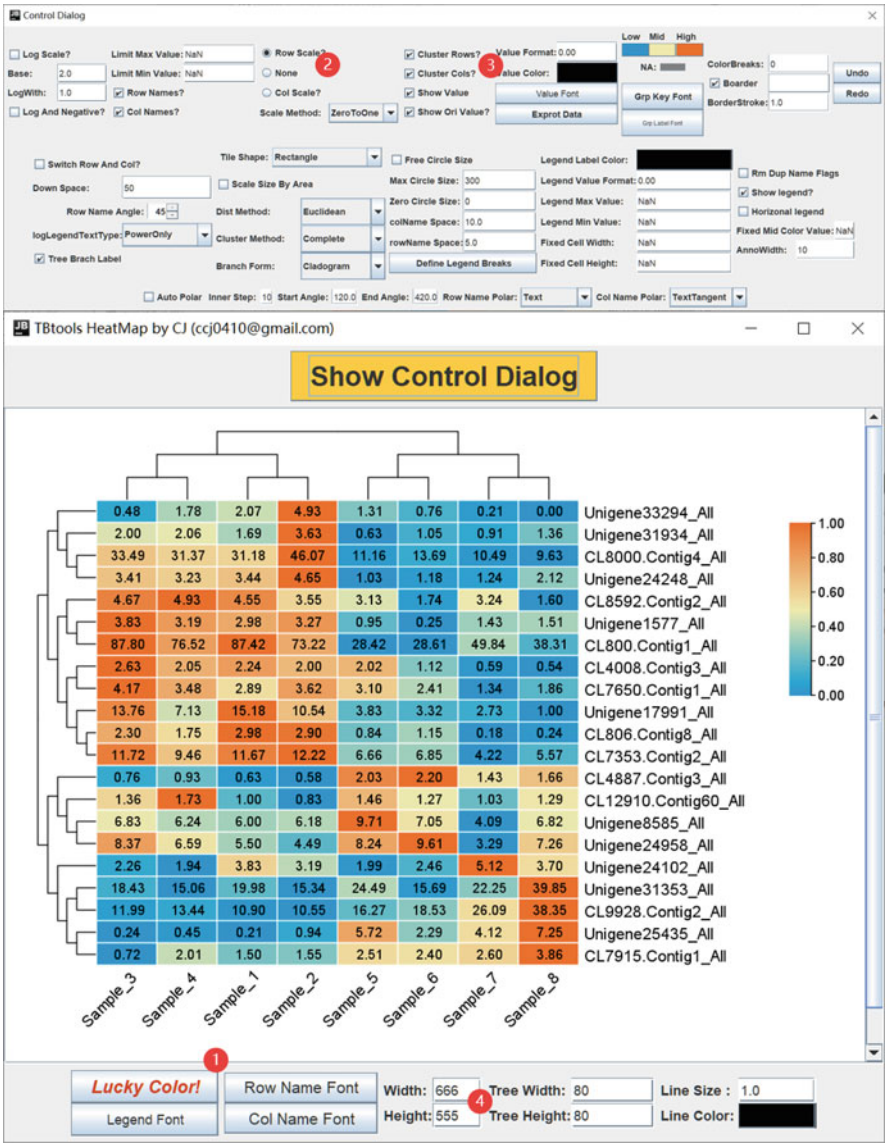

Fig. 17.19 Heatmap Parameter Adjustment (Basic)

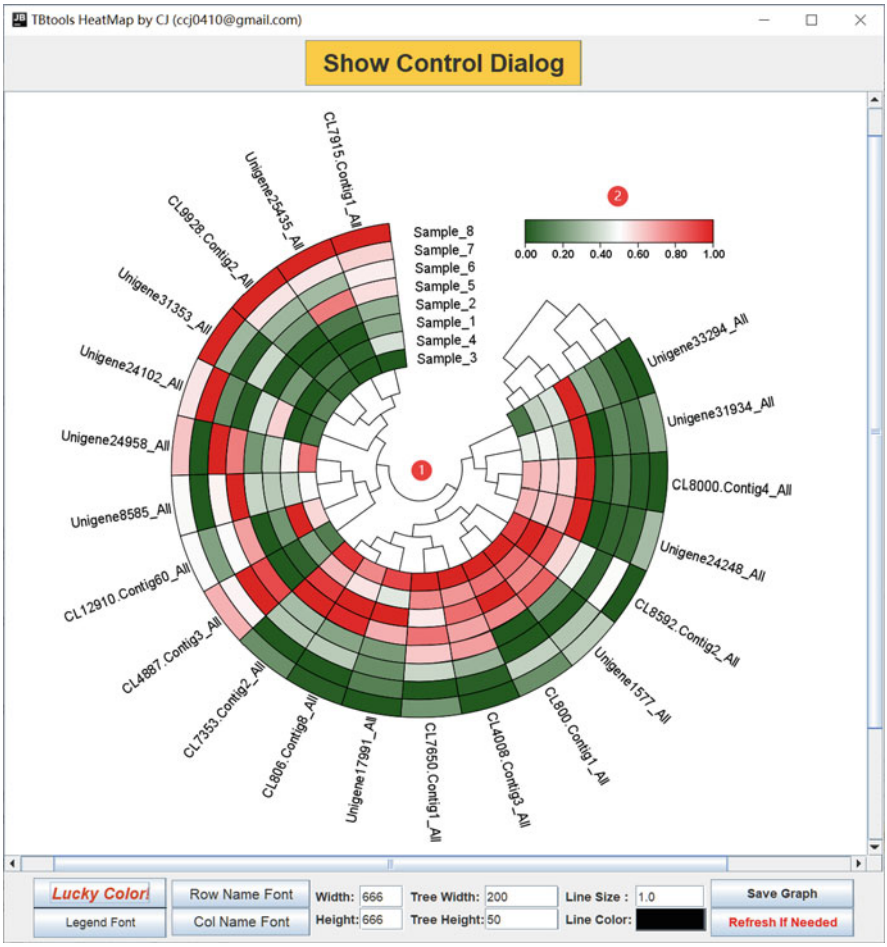

Fig. 17.20 Heatmap Parameter Adjustment (Polar Coordinates)

1. Set the genome sequence file as input.
2. Set the output file to save the length information of the chromosome sequences.
3. Check "Keep Only Sequence Length" to save only the length information to the output file.
4. Click the "Start" button.

Go to the function, "Graphics" -> "Advanced Circos", set the input file (the output file above), and "Show My Circos Plot!". Then a simple Circos graph with chromosome skeletons will be generated (Fig. 17.22).

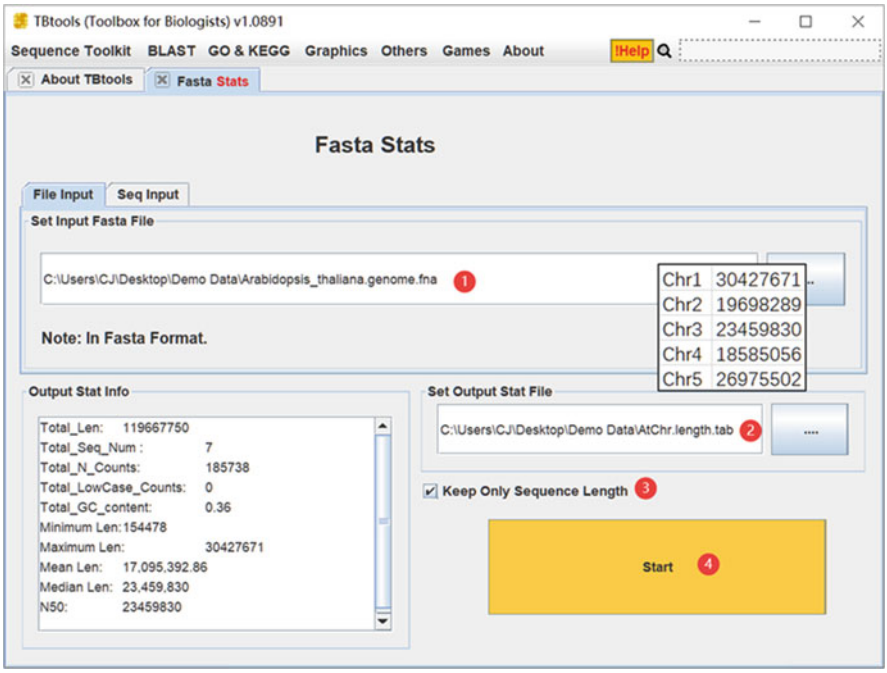

Fig. 17.21 Using "Fasta Stat" to Prepare a Chromosome Skeletons File

17.3.3.2 Display the Feature Location and Association Information of the Genome

On this simple graph, two types of information can be added: (a) Chromosome feature labels, such as the location of certain genes; (b) Chromosome segment relationships (such as large segmental duplication events). The former can be obtained through the "GXF Pos. and Info. Extract" function in TBtools; the latter can be obtained by collinearity analysis of the genome (Fig. 17.23).

1. Set the file containing information of feature locations. The format is "Chromosome ID\t Feature identifier\tChromosome starting position\tChromosome ending position\t[optional color information, R, G, B]".
2. Set the file containing interchromosomal association information. The format is "Chromosome ID\tChromosome starting position\tChromosome ending position\tChromosome ID\tChromosome starting position\tChromosome ending position\t[optional color information, R, G, B]".
3. Click "Show My Circos Plot!".
4. Users can see a Circos with more information.

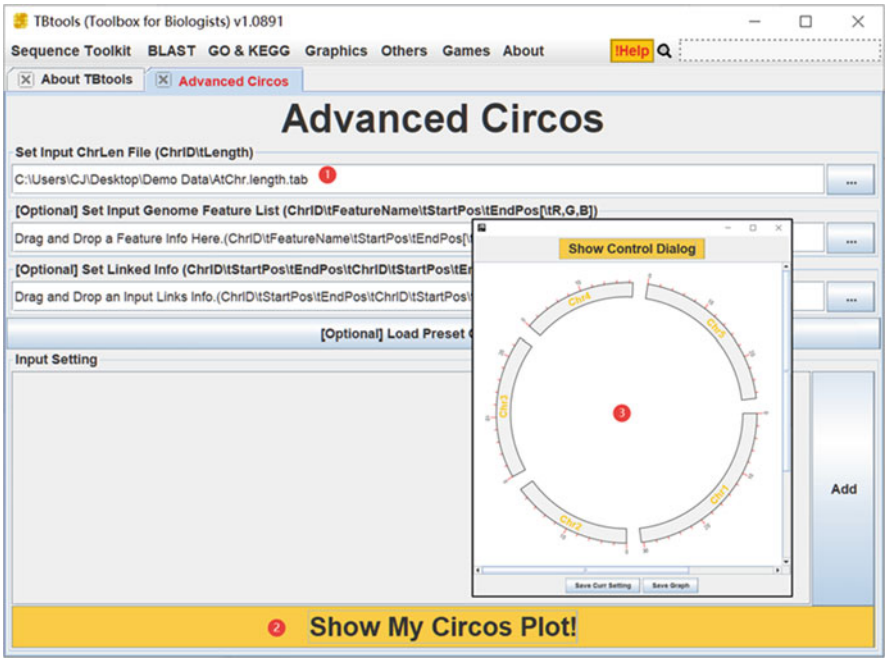

Fig. 17.22 Use "Advanced Circos" to Visualize the Chromosome Skeleton

17.3.3.3 Display Information at the Genome Scale

Usually Circos plot are used for the overview of genome-wide information, such as gene density, GC content, sequencing depth, SNP frequency, etc. These information are often recorded in a way that a chromosome region corresponds to a value. Here, we used gene density as an example, and the density information can be easily obtained using the "Gene Density Profile" function in TBtools. Open "TBtools" and select the function "Sequence Toolkit" -> "GFF3/GTF Manipulate" -> "Gene Density Profile" (Fig. 17.24).

1. Set the gene structure annotation file (GFF3/GTF) as input.
2. Set the output file path.
3. The content of the output file is formatted as "Chromosome ID\t starting position\t Chromosome ending position\tnumber of genes". As the length of each genomic interval is set to be the same, so the number of genes in each interval represents gene density.

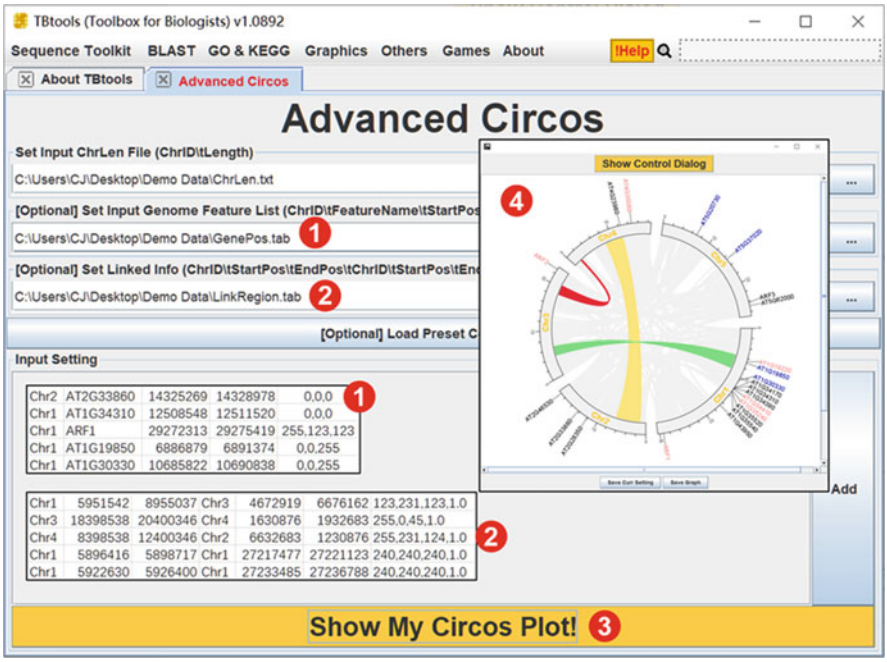

Fig. 17.23 Use "Advanced Circos" to Display Feature Locations and Collinear Regions

The resultant gene density information can be directly used for "Advanced Circos" visualization (Fig. 17.25).

1. Set the input file, which is the gene density file obtained above.
2. Adjust the Track type to "Heatmap".
3. Select the sliding window type as "None," that is, no sliding window, as the sliding window calculation has been accomplished in the "Gene Density Profile" step.
4. Gene density information is now displayed.

Data in similar formats can be displayed in different plot types. In addition to heatmap, TBtools also supports "Bar," "Line," "Point," etc. Besides, it also supports positional mark visualization, such as "Tile," "Arrow," "Triangle," etc. The input data format is "Chromosome ID\tChromosome start position\tChromosome end position\tR,G,B". Tracks of multiple data can be viewed synchronically (Fig. 17.26).

1. Overlapping of two types of tracks: "Heatmap" and "Line".
2. Use "Bar" type for the second track.
3. Support different open angles.
4. Support linear display as well.

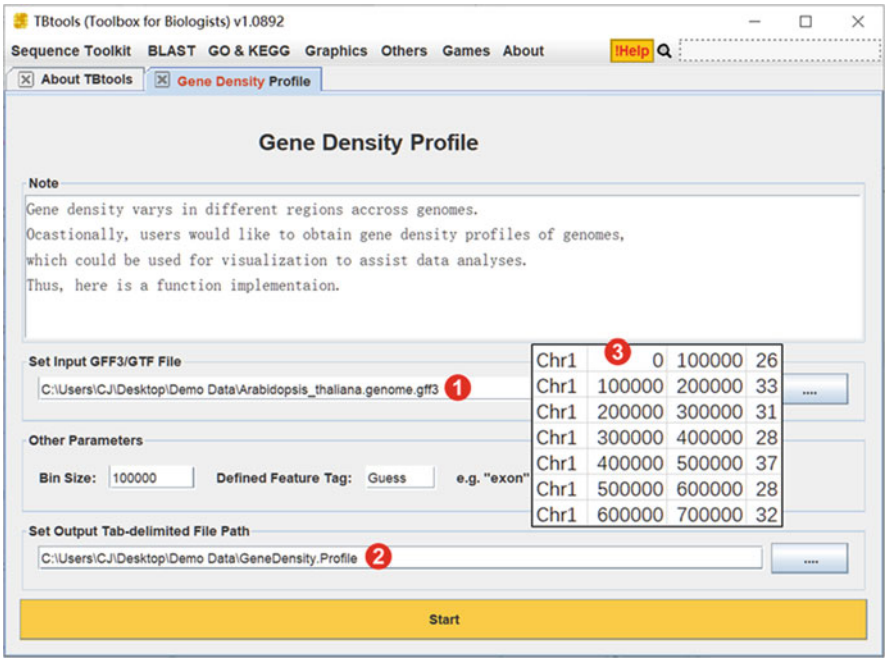

Fig. 17.24 Use "Gene Density Profile" to Get Gene Density Information

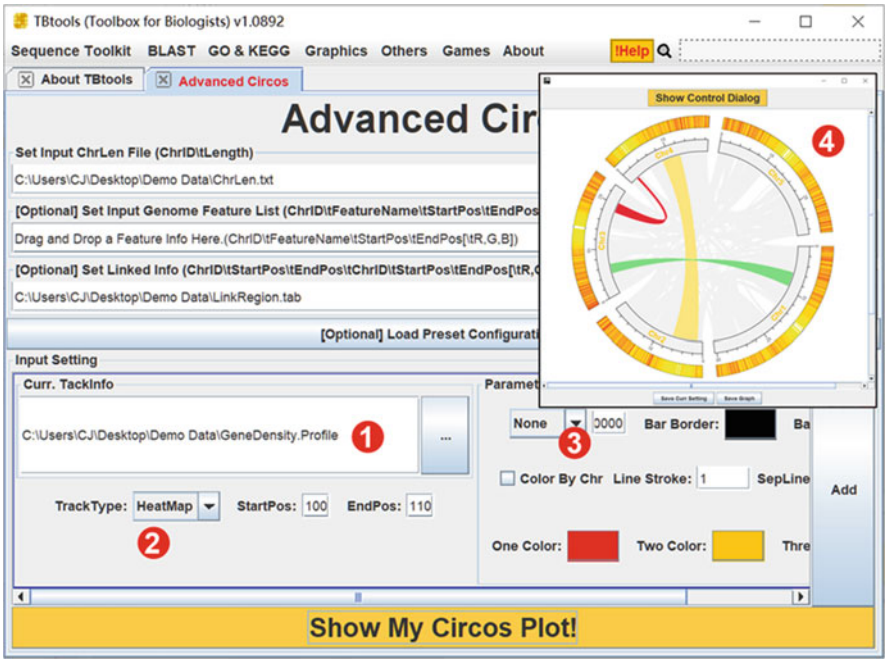

Fig. 17.25 Use "Advanced Circos" to View Gene Density over the Whole Genome

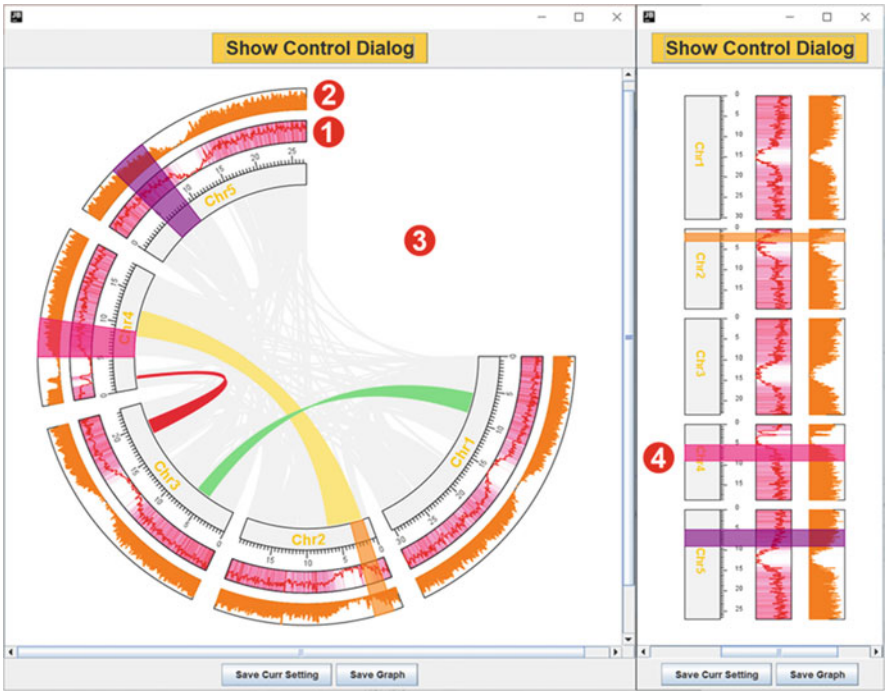

Fig. 17.26 Flexible Use of "Advanced Circos"

17.3.4 Quick Protein Functional Annotation (Plugin)

In daily data analysis, we may often obtain hundreds or thousands of new genes (unannotated), which need to be functionally annotated for biological meaning. The TBtools plugin "Quick Protein Anno" can assist users in this, which can finish the functional annotation of ~20,000 protein sequences within a few minutes and output the results into a table for further exploration. Users can install it through the plugin store. Turn to the plugin store through "Others" -> "Plugin" -> "Plugin Store".

After the installation is complete, you can run the function through the menu "Others" -> "Plugin" -> "Quick Protein Anno" (Figs. 17.27 and 17.28).

1. Set the database used for annotation. Generally, the "Swissprot" protein sequence library is used.
2. Users can click "DB Download" to download the "Swissprot" protein sequence library.
3. Set the protein sequence file to be annotated or paste the sequences directly.
4. Set an output file path.
5. The output result file formatted as "Gene ID\tHigh-frequency keyword #1\tHigh-frequency keyword #2\tOptimal comparison results".

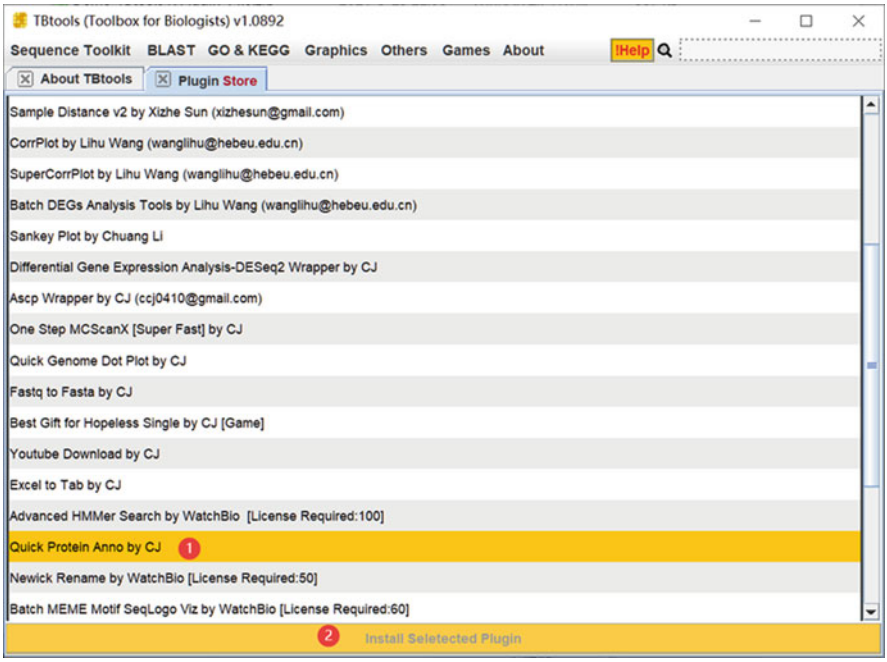

Fig. 17.27 Install the "Quick Protein Anno" Plugin through the Plugin Store

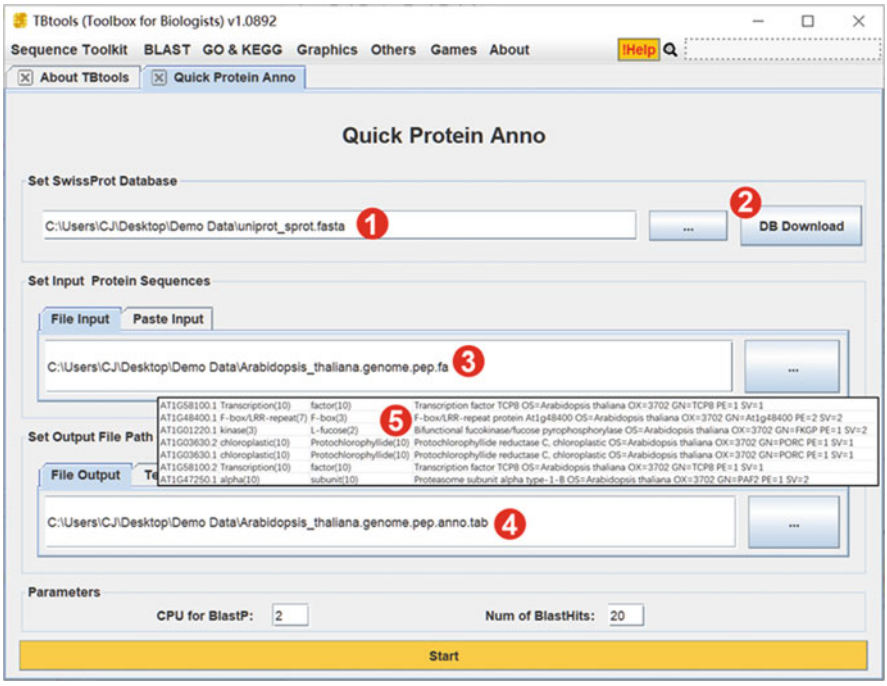

Fig. 17.28 Use "Quick Protein Anno" for Function Annotation

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R (2020) TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant 13:1194–1202

Connors J, Krzywinski M, Schein J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos : an information aesthetic for comparative genomics. Genome Res 19:1639–1645

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H et al (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40:e49

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ (2007) An 'electronic fluorescent pictograph' browser for exploring and analyzing large-scale biological data sets. PLoS One 2:e718

Chapter 18
Analyzing Multi-Omic Data with Integrative Platforms

Yan Zou

Abstract The exponential growth of molecular data put up a new challenge to the biologists. The difficulty in data storage, processing, transmission, connection, and the demand for multi-omic data analysis motivates scientists to set up integrative platforms and workflows. Here we introduce some prominent integrated bioinformatics platforms. Among them, Galaxy will be carefully discussed for its development, core values, flexible workflows, and relevant framework applications.

Keywords Multi-omic · Integrative platform · Galaxy

18.1 Integrating Diverse Tools into Bioinformatics Platforms

The availability of high-throughput sequencing technologies and high-resolution mass spectrometry in genomics, transcriptomics, proteomics, metabolomics, and phenomics promotes a large-scale multi-omic data complex. The information content is higher in integrated analysis, which requires connecting and comparing data in different omics, than in any of the molecular levels studied separately. The exponential growths of molecular data, however, put up a new challenge for biologists. The storage, processing, transmission, connection, and analysis of these data complex demand the use of disparate software programs and require computational resources beyond the capacity of many biological laboratories (Chen and Hofestädt 2014; Boekel et al. 2015). Furthermore, disparate software requires extra training time when a new analysis is conducted and standardizing diverse data formats into the identical one that program requests bring further inconvenience. For these reasons, multi-omic platforms are emerged to embrace the complexity that is associated with the exponentially increasing amounts of data.

Y. Zou (✉)
College of Life Sciences, Zhejiang University, Hangzhou, P. R. China
e-mail: 3160102154@zju.edu.cn

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_18

Some prominent software platforms have already shown their merits in coping with these problems. Some are compatible with all data regions, and some are specially designed for specific regions, such as cancer, plant cells, and viruses.

18.1.1 General Integrative Platforms and Workflows

Most bioinformatics tools used in genomics, transcriptomics, and proteomics are set in programming and command-line environments, which can be time-consuming and complex for researchers to get started. Ideal platforms and workflows such as Galaxy, Taverna, and Snakemake are created to meet the urgent need for the interactive analyses of big biological data.

18.1.1.1 Galaxy

Galaxy is a scientific workflow, data integration, and data and analysis publishing web-based platform established in 2005 (Blankenberg et al. 2011). Its graphical query interface combined with customized data storage can simplify the process Schatz (2010). Its development, core values, and flexible workflows will be carefully discussed in Sect. 18.2.

18.1.1.2 Snakemake

Snakemake (available in https://snakemake.readthedocs.io/en/stable/) is a Python-based scalable bioinformatics workflow engine published in 2012. It can scale from single-core workstations to compute clusters without modifying the workflow. It is the first system to support the use of automatically inferred multiple named wildcards (or variables) in input and output filenames (Köster and Rahmann 2018). Interaction between Snakemake and those installed in local or web-based tools is also available when both support the input and output data formats. In recent years, some Snakemake extensions, such as RASflow (Zhang and Jonassen 2020) and Sequanix (Desvillechabrol et al. 2018), build modular analysis workflow and establish graphical interfaces to help Snakemake be more flexible and user-friendly.

18.1.1.3 Taverna

Taverna (available in https://incubator.apache.org/projects/taverna.html) is a tool for the composition and enactment of bioinformatics workflows. Taverna includes a workbench application that provides a graphical user interface for the composition of workflows (Oinn et al. 2004). Scientists can organize their workflows in a new language called the simple conceptual unified flow language (Scufl). It can integrate

with the bioinformatics resource shared as Web services among the community. Taverna and Galaxy, two workflow systems widely accepted and applied by the bioinformatics community, can also be integrated into a single environment, Tavaxy (available in https://www.tavaxy.org/) (Abouelhoda et al. 2012).

18.1.2 Integrative Platforms for Specialized Data

Some integrative platforms are specially established for analyzing data from specific regions, such as cancer, virus, plants, and fungi. Combined with gene and protein expression with signaling pathways and cell characteristics, these platforms contribute professional and accessible means for biologists to process data.

18.1.2.1 Combine Integrative Platforms with Clinical Data for Cancer Research

Distinct signaling pathways and altered molecular functions in cancer cells and clusters are displayed in the integrated analyses of molecular data. The platforms built specifically for cancer cell data analysis allow cancer researchers to interactively explore altered gene sets and signaling pathways (Gao et al. 2013). What makes the platforms driven by cancer cell studies distinguished is their strong connection with clinical outcomes and potential. Genomic, metabonomics, and clinical data might be used to identify novel patient subgroups. Clinical therapy can be tailored for each patient when statistical models are produced and treatment strategies are evaluated based on stratified patient groups (Kristensen et al. 2014).

The cBioPortal for Cancer Genomics (cBioPortal, http://cbioportal.org) is one of the widely used integrative platforms specialized for analyzing cancer-related data. The cBioPortal is established for integrative analysis of cancer genomics and clinical patient profiles. With 15 provisional TCGA (The Cancer Genome Atlas) datasets and other open datasets contained, the web-based cBioPortal is uniquely designed to store every single data in the gene level and combine these data with available de-identified clinical data. The fundamental abstraction of this platform is the concept of altered genes (Cerami et al. 2012), which is used to help users simplify the mixed and complicated current datasets and develop genomic hypotheses proceeding from genetic alterations across samples, genes, and pathways.

Among other platforms targeted in cancer cell research, Web-TCGA can uniquely provide methylation analyses (Deng et al. 2016); Firebrowse (http://firebrowse.org/) can also characterize and identify genomic patterns in human cancer models through visual and programmatical tools.

18.1.2.2 Specialized Integrative Platforms in Other Fields

Integrative platforms that focus on particular fields integrate with the other database resources in their domains for further research convenience. Integrating bioinformatics resource for fungi and oomycetes, FungiDB (fungidb.org) is a free online platform for data mining and functional genomics analysis which combines Eukaryotic Pathogen Genomics Database Resource (Basenko et al. 2018).

Scientists also use bioinformatics platforms as a tool for international cooperation in urgent issues. For the pharmaceutical development and antiviral drug prediction for the COVID-19 virus, Virus-CKB (https://www.cbligand.org/g/virus-ckb) is developed as a viral-associated disease-specific chemogenomics knowledgebase (Virus-CKB), which describes the chemical molecules, genes, and proteins involved in viral-associated diseases regulation (Feng et al. 2021).

18.2 Galaxy: A Widely Accepted General Bioinformatics System

18.2.1 Introduction

As has been mentioned in Sect. 18.1.1.1, Galaxy is a bioinformatics scientific workflow and data analysis platform, which is created in 2005. It is developed by the Galaxy team at Penn State, Johns Hopkins University, Oregon Health and Science University, and the Galaxy Community using Python language.

Galaxy was initially set up for genomics and transcriptomics data analysis from the very beginning. Nevertheless, with the maturation of proteomic and metabolomic technologies, multi-omic applications started to emerge after a few years since the Galaxy was created. Now it has assembled tools in multiple domains, such as gene expression, proteomics, epigenomics, and transcriptomics. It also contains cross-domain tools, including ecology, climate science, and computational chemistry. More than 7500 tools (Jalili et al. 2020) have been contributed to the Galaxy ToolShed till January 2020.

Galaxy now has a prosperous scientific community. The community keeps organizing conferences and meetings with Galaxy-related content and sharing tool tutorials in The Galaxy Training Network. With more than 9000 total publications, including over 7500 journal articles, 500 books, and 400 conference papers by 2020 (Jalili et al. 2020), this free and open-source platform has created a community for biology researchers.

18.2.2 How to Use Galaxy

Researchers can directly use the workspace of the Galaxy platform to conduct data analysis. Galaxy can be operated both on the web and locally. Between them, the webpage is more welcomed for its convenience.

As shown in Fig. 18.1, the webpage workspace of Galaxy is separated into four parts. Right on the top of the website is the navigation bar, which provides users with easy access to data processing, including analyzing, workflows, visualization, and user function for sharing data and tutorials. On the left is the analysis tool panel in which users can apply tools to their data. The component in the middle is the detail interface, where users set up and adjust different datasets, features, and filters for analysis. The history panel on the right of the workspace shows the status of the generation of the datasets. Users can also search the analysis history and extract workflows from the histories.

Users can create workflows based on their data analysis process. A workflow is a series of tools and dataset actions that run in sequence as a batch operation (Goecks et al. 2010). In Galaxy, workflows can be created from scratch using the workflow editor or generated from the analysis already completed in history Blankenberg et al. (2010). A successfully designed workflow can be continuously reused for the future analysis, enhancing reproducibility by applying the same methods to all of the users' data.

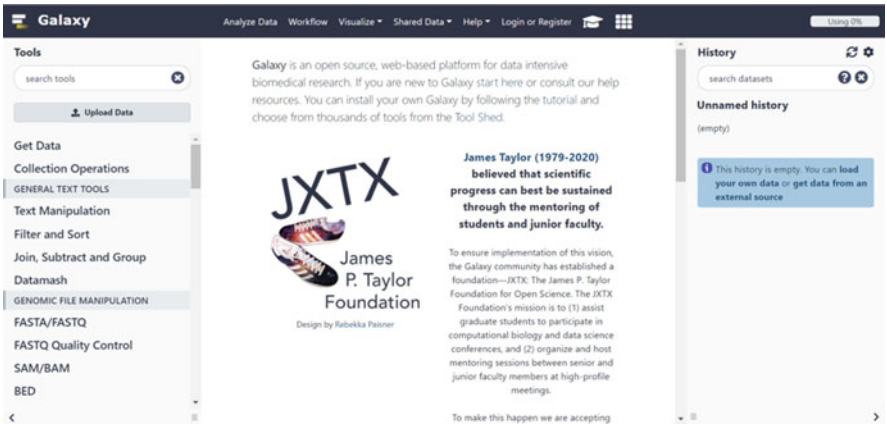

Fig. 18.1 The web-based workspace of the Galaxy platform. In the left column, users can choose the genomic tools. The detailed analysis content will be shown in the middle. The analysis history is in the right column, and users can search datasets using the search bar at the top of the panel

18.2.3 Key Requirements in Designing Galaxy

For most researchers, the use of disparate software programs and extra software training demands time and effort. The required computational resources sometimes will reach out of the capacity of most biological research laboratories. The ideal platform should meet these scientists' basic needs and keep the platform vibrant and easy to use. Five characteristics are crucial for building a thriving platform, including the flexibility to accommodate constantly evolving data types and emerging software across omics domains, reproducibility, open and free access, and long-term sustainability (Boekel et al. 2015).

Flexibility is the first general need that the developers plan to meet. Specifically, the platform needs to be open, extendable, and amenable to heterogeneous computing environments. To resolve this issue, the developer group has combined Linux-based software with Windows-based software to ensure that the platform could function well in multiple working environments.

Another distinct requirement for the platform is that it should have the ability to operate complex and multistep workflows automatically with different software. The platform can use quality control methods to evaluate the tool quality and integration efficiency.

The compatibility in high-performance computing and cloud environments makes the platform scalable to the established sequencing databases. Its large-memory allocation is integrated with multiple storage infrastructures.

One of the crucial characteristics for a bioinformatics platform to expand its lifespan is its community sharing. The publication and sharing of complete workflows not only promotes the dissemination and reproducibility of the workflows but also enhances the transparency of the data and its processing. The attention to data provenance also guarantees this.

The last essential requirement for building the desired platform is the wide adaption and sustainable user-friendly interface. Its web-based graphical user interface lowers the difficulty for the researchers with limited computational expertise to operate. The platform is sustained by the scientist's community rather than a single developer group, and each developer can publish their software and designs of workflows on their own.

18.2.4 Applications Based on Galaxy Platform

Some data analysis applications derived from Galaxy emerge to resolve typical questions. Here we introduce three applications of Galaxy: Cistrome, a new integrative platform based on Galaxy frameworks; RepeatExplorer, a computational pipeline or component aiming at repetitive DNA; and CloudMan, a cloud resource management system, as well as BioBlend, an automating pipeline analyses within Galaxy and CloudMan.

18.2.4.1 Cistrome: Galaxy-Based Integrative Platforms for Transcriptional Regulation

Chromatin immunoprecipitation (ChIP) combined with microarrays (ChIP-chip) and ChIP combined with NGS (ChIP-seq) are used for identifying cistromes, which refers to the set of cis-acting targets of a trans-acting factor on a genome-wide scale. However, the analysis of cistrome data requires both the hardware resources from the lab and the computational skills of the researchers to achieve the analyzing algorithms.

Under the conditions above, Cistrome (http://cistrome.org/ap/) has been built to provide a flexible bioinformatics workbench. Cistrome, an integrative platform for transcriptional regulation studies, is specifically designed for downstream data analysis accompanied by ChIP-chip or ChIP-seq technologies and includes fundamental analyses from peak calling to motif detection (Fig. 18.2).

To accomplish this, Galaxy framework provides a user-friendly, reproducible, and transparent workbench, on which the scientists can share, incorporate, and publish their data. Furthermore, its infrastructure makes each Cistrome tool to remember the run-time parameters in the server (Liu et al. 2011).

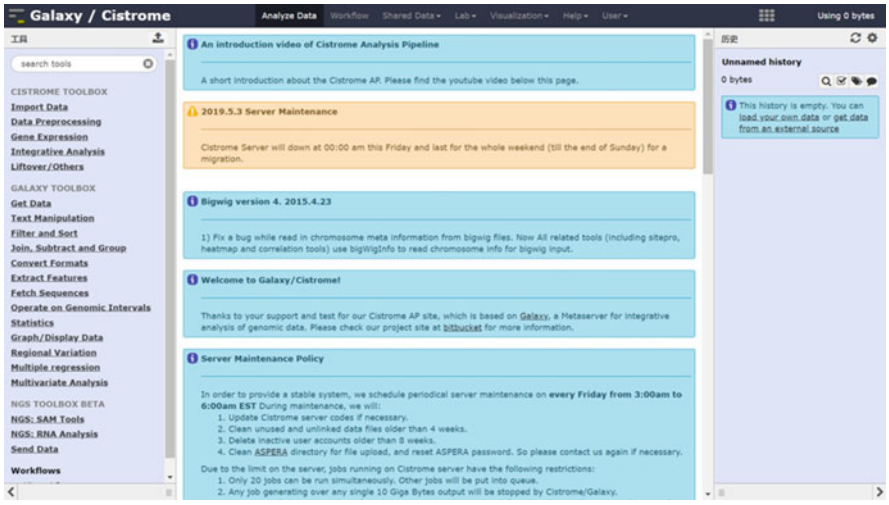

Fig. 18.2 The web-based workspace of the Galaxy/Cistrome platform. In the left column, users can choose the available tools. The messages, tool options, and detailed analysis content will be shown in the middle. The analysis history is in the right column

18.2.4.2 RepeatExplorer: A Computational Pipeline for Characterization of Repetitive Elements

Repetitive DNA makes up a large part of eukaryotic nuclear genomes. The accurate quantification and sequence characterization of repetitive DNA is complex for most researchers due to the restricted computational resources and the lack of professional analyzing tools.

With the new approach for global repeat analysis developed by Novak et al. (Novák et al. 2010) and the availability of high-throughput sequencing data, Novak et al. developed a pipeline named RepeatExplorer (http://repeatexplorer.umbr.cas.cz/) for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads (Novak et al. 2013).

The main component of RepeatExplorer is the clustering pipeline, which performs all-to-all similarity comparisons of sequence reads followed by their graph-based clustering to identify groups of reads derived from repetitive elements.

Galaxy platform provides the adaption for the tools of the RepeatExplorer pipeline. These tools can be recombined to form specialized workflows. The Galaxy platform also facilitates easy execution, documentation, and sharing of analysis protocols and results.

18.2.4.3 CloudMan and BioBlend: A Cloud Resource Management System and an Automating Pipeline Analyses within Galaxy and CloudMan

With the availability of high-throughput sequencing data and the robust research future of analyzing sequence data, the computational infrastructure and support have gradually been a problem for researchers whose laboratory cannot reach the requirement in computing. Cloud computing, a computational model, is potential in the analysis of high-throughput sequencing data. However, the established projects are only targeted at specialized problems and are unsuitable for various computing circumstances.

CloudMan is an integrated solution that the researchers could create and control fully functional compute clusters with existing tools and packages provided on cloud resources. The intricacies of cloud computing resource acquisition, configuration, and scaling could be conducted on Amazon's EC2 cloud infrastructure, and a personal computing cluster will be produced in minutes. (Afgan et al. 2010). The researchers have embedded Galaxy CloudMan on top of the Bio-Linux workstation machine image and integrated it with Galaxy.

BioBlend (http://bioblend.readthedocs.org/) is a unified API in a high-level language that wraps the functionality of Galaxy and CloudMan APIs (Sloggett et al. 2013). It is easier for researchers to automate end-to-end large data analysis using BioBlend, due to the convenient access for large datasets in the familiar Galaxy environment and the computing infrastructure provided.

References

Abouelhoda M, Issa SA, Ghanem M (2012) Tavaxy: integrating Taverna and Galaxy workflows with cloud computing support. BMC Bioinformatics 13:77

Afgan E, Baker D, Coraor N et al (2010) Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics 11:S4

Basenko E, Pulman J, Shanmugasundram A, Harb O, Crouch K, Starns D, Warrenfeltz S, Aurrecoechea C, Stoeckert C, Kissinger J, Roos D, Hertz-Fowler C (2018) FungiDB: an integrated bioinformatic resource for fungi and oomycetes. J Fungi 4(1):39

Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol 19:1–21

Blankenberg D, Coraor N, Von Kuster G, Taylor J, Nekrutenko A (2011) Integrating diverse databases into an unified analysis framework: a Galaxy approach. Database 2011:11

Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Käll L, Lehtiö J, Lukasse P, Moerland PD, Griffin TJ (2015) Multi-omic data analysis using Galaxy. Nat Biotechnol 33(2):137–139

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2(5):401–404

Chen M, Hofestädt R (2014) Approaches in integrative bioinformatics. Introduction. Springer, Berlin, pp 3–5

Deng M, Brägelmann J, Schultze JL, Perner S (2016) Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. BMC Bioinformatics 17:72

Desvillechabrol D, Legendre R, Rioualen C, Bouchier C, van Helden J, Kennedy S, Cokelaer T (2018) Sequanix: a dynamic graphical interface for Snakemake workflows. Bioinformatics 34(11):1934–1936

Feng Z, Chen M, Liang T, Shen M, Chen H, Xie X (2021) Virus-CKB: an integrated bioinformatics platform and analysis resource for COVID-19 research. Brief Bioinform 22(2):882–8955

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6(269):l1

Goecks J, Nekrutenko A, Taylor J, Galaxy T (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):86

Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Taylor J, Nekrutenko A (2020) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Res 48(1):395–402

Köster J, Rahmann S (2018) Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 34(20):3600

Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A (2014) Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer 14(5):299–313

Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y, Pape UJ, Poidinger M, Chen Y, Yeung K, Brown M, Turpaz Y, Liu XS (2011) Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol 12(8):83

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29(6):792–793

Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20(17):3045–3054

Schatz MC (2010) The missing graphical user interface for genomics. Genome Biol 11(8):128

Sloggett C, Goonasekera N, Afgan E (2013) BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics 29(13):1685–1686

Zhang X, Jonassen I (2020) RASflow: an RNA-Seq analysis workflow with Snakemake. BMC Bioinformatics 21(1):110–119