Chapter 3

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

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

Abstract

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

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

1 Introduction

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

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

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

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

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

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

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

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

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

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

disease occurrence that they could counteract through appropriate individual adaptations.

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

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

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

2   Integrative Approaches to Disease Subtype Discovery

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

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

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

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

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

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

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

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

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

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

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

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

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

3   Toward a Mechanistic Understanding of Disease Subtypes

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

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

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

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

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

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

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

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

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

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

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

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

4   Developing a Platform for Translational Informatics

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

In addition to omics datasets being collected from different platforms (such as gene expression, copy number variation methylation, etc.), phenotypic data is also becoming richer. Instead of a single endpoint status representing the phenotype (e.g., disease or non-disease), a set of measurements may be collected (e.g., mild,
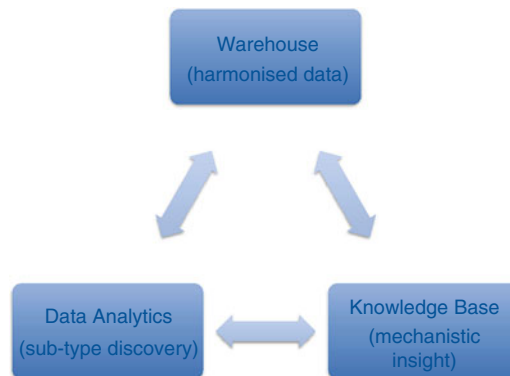
Fig. 1 Three key components for a translational medicine platform: (**a**) A data warehouse for storage and querying of clinical and multi-omics data associated with patient cohorts; data harmonization is needed to ensure that the data conforms to standards, in order to facilitate comparison across different studies. (**b**) A data analytics component for visual exploration of the data and for subtype discovery. (**c**) A knowledge base to enable experimental results to be understood in the context of known disease pathways and processes and to suggest a possible mechanistic basis for subtypes

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

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

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

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

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

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

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

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

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

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

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

5   Conclusions

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

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

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

Acknowledgments

References

1. Auffray C, Chen Z, Hood L (2009) Systems medicine: the future of medical genomics and healthcare. Genome Med 1:2. doi:10.1186/gm2

2. Hood L (2013) Systems biology and p4 medicine: past, present, and future. Rambam Maimonides Med J 4:e0012. doi:10.5041/RMMJ.10112

3. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease (2011) Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. National Academies Press, Washington, DC, USA

4. Loscalzo J, Barabasi A-L (2011) Systems biology and the future of medicine. Wiley Interdiscip Rev Syst Biol Med 3:619–627. doi:10.1002/wsbm.144

5. Menche J, Sharma A, Kitsak M et al (2015) Uncovering disease-disease relationships through the incomplete interactome. Science 347:1257601. doi:10.1126/science.1257601

6. Goh K-I, Cusick ME, Valle D et al (2007) The human disease network. Proc Natl Acad Sci U S A 104:8685–8690. doi:10.1073/pnas.0701361104

7. Hood L, Tian Q (2012) Systems approaches to biology and disease enable translational systems medicine. Genomics Proteomics Bioinformatics 10:181–185. doi:10.1016/j.gpb.2012.08.004

8. Hood L, Price ND (2014) Demystifying disease, democratizing health care. Sci Transl Med 6:225ed5. doi:10.1126/scitranslmed.3008665

9. Flores M, Glusman G, Brogaard K et al (2013) P4 medicine: how systems medicine will transform the healthcare sector and society. Pers Med 10:565–576. doi:10.2217/PME.13.57

10. Hood L, Auffray C (2013) Participatory medicine: a driving force for revolutionizing healthcare. Genome Med 5:110. doi:10.1186/gm514

11. Cesario A, Auffray C, Russo P, Hood L (2014) P4 medicine needs P4 education. Curr Pharm Des. 20(38):6071–2

12. Miller FA, Hayeems RZ, Bytautas JP et al (2014) Testing personalized medicine: patient and physician expectations of next-generation genomic sequencing in late-stage cancer care. Eur J Hum Genet 22:391–395. doi:10.1038/ejhg.2013.158

13. Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. J Hum Genet 59:5–15. doi:10.1038/jhg.2013.114

14. Sachidanandam R, Weissman D, Schmidt SC et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933. doi:10.1038/35057149

15. Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291:1304–1351. doi:10.1126/science.1058040

16. 1000 Genomes Project Consortium, Abecasis GR, Auton A et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. doi:10.1038/nature11632

17. Clarke L, Zheng-Bradley X, Smith R et al (2012) The 1000 Genomes Project: data management and community access. Nat Methods 9:459–462. doi:10.1038/nmeth.1974

18. Gerber DE, Minna JD (2010) ALK inhibition for non-small cell lung cancer: from discovery

to therapy in record time. Cancer Cell 18:548–551. doi:10.1016/j.ccr.2010.11.033

19. Shaw AT, Yasothan U, Kirkpatrick P (2011) Crizotinib. Nat Rev Drug Discov 10:897–898. doi:10.1038/nrd3600

20. Karapetis CS, Khambata-Ford S, Jonker DJ et al (2008) K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med 359:1757–1765. doi:10.1056/NEJMoa0804385

21. Auffray C, Adcock IM, Chung KF et al (2010) An integrative systems biology approach to understanding pulmonary diseases. Chest 137:1410–1416. doi:10.1378/chest.09-1850

22. Roca J, Vargas C, Cano I et al (2014) Chronic obstructive pulmonary disease heterogeneity: challenges for health risk assessment, stratification and management. J Transl Med 12(Suppl 2):S3. doi:10.1186/1479-5876-12-S2-S3

23. Gomez-Cabrero D, Lluch-Ariet M, Tegnér J et al (2014) Synergy-COPD: a systems approach for understanding and managing chronic diseases. J Transl Med 12(Suppl 2):S2. doi:10.1186/1479-5876-12-S2-S2

24. Turan N, Kalko S, Stincone A et al (2011) A systems biology approach identifies molecular networks defining skeletal muscle abnormalities in chronic obstructive pulmonary disease. PLoS Comput Biol 7:e1002129. doi:10.1371/journal.pcbi.1002129

25. Kaneko Y, Yatagai Y, Yamada H et al (2013) The search for common pathways underlying asthma and COPD. Int J Chron Obstruct Pulmon Dis 8:65–78. doi:10.2147/COPD.S39617

26. Smarr L (2012) Quantifying your body: a how-to guide from a systems biology perspective. Biotechnol J 7:980–991. doi:10.1002/biot.201100495

27. Chen R, Mias GI, Li-Pook-Than J et al (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148:1293–1307. doi:10.1016/j.cell.2012.02.009

28. McDermott JE, Wang J, Mitchell H et al (2013) Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opin Med Diagn 7:37–51. doi:10.1517/17530059.2012.718329

29. Ideker T, Dutkowski J, Hood L (2011) Boosting signal-to-noise in complex biology: prior knowledge is power. Cell 144:860–863. doi:10.1016/j.cell.2011.03.007

30. Chuang H-Y, Lee E, Liu Y-T et al (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140. doi:10.1038/msb4100180

31. Bass AJ, Thorsson V, Shmulevich I et al (2014) Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513:202–209. doi:10.1038/nature13480

32. Gomez-Cabrero D, Abugessaisa I, Maier D et al (2014) Data integration in the era of omics: current and future challenges. BMC Syst Biol 8:I1. doi:10.1186/1752-0509-8-S2-I1

33. Wheelock CE, Goss VM, Balgoma D et al (2013) Application of "omics" technologies to biomarker discovery in inflammatory lung diseases. Eur Respir J 42:802–825. doi:10.1183/09031936.00078812

34. Bel EH, Sousa A, Fleming L et al (2011) Diagnosis and definition of severe refractory asthma: an international consensus statement from the Innovative Medicine Initiative (IMI). Thorax 66:910–917. doi:10.1136/thx.2010.153643

35. Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25:2906–2912. doi:10.1093/bioinformatics/btp543

36. Lê Cao K-A, Martin PGP, Robert-Granié C, Besse P (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics 10:34. doi:10.1186/1471-2105-10-34

37. Dray S, Chessel D, Thioulouse J (2003) Co-inertia analysis and the linking of ecological data tables. Ecology 84:3078–3089

38. Culhane AC, Perrière G, Higgins DG (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics 4:59. doi:10.1186/1471-2105-4-59

39. Meng C, Kuster B, Culhane AC, Gholami AM (2014) A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics 15:162. doi:10.1186/1471-2105-15-162

40. Shen R, Mo Q, Schultz N et al (2012) Integrative subtype discovery in glioblastoma using iCluster. PLoS One 7:e35236. doi:10.1371/journal.pone.0035236

41. Curtis C, Shah SP, Chin S-F et al (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486:346–352. doi:10.1038/nature10983

42. Mo Q, Wang S, Seshan VE et al (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A 110:4245–4250. doi:10.1073/pnas.1208949110

43. Kirk P, Griffin JE, Savage RS et al (2012) Bayesian correlated clustering to integrate multiple datasets. Bioinformatics 28:3290–3297. doi:10.1093/bioinformatics/bts595

44. Lock EF, Dunson DB (2013) Bayesian consensus clustering. Bioinforma Oxf Engl 29:2610–2616. doi:10.1093/bioinformatics/btt425

45. Wang B, Mezlini AM, Demir F et al (2014) Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11:333–337. doi:10.1038/nmeth.2810

46. Le Novère N, Hucka M, Mi H et al (2009) The systems biology graphical notation. Nat Biotechnol 27:735–741. doi:10.1038/nbt.1558

47. Blinov ML, Faeder JR, Goldstein B, Hlavacek WS (2006) A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. Biosystems 83:136–151. doi:10.1016/j.biosystems.2005.06.014

48. Kitano H, Funahashi A, Matsuoka Y, Oda K (2005) Using process diagrams for the graphical representation of biological networks. Nat Biotechnol 23:961–966. doi:10.1038/nbt1111

49. Mizuno S, Iijima R, Ogishima S et al (2012) AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. BMC Syst Biol 6:52. doi:10.1186/1752-0509-6-52

50. Fujita KA, Ostaszewski M, Matsuoka Y et al (2013) Integrating pathways of parkinson's disease in a molecular interaction map. Mol Neurobiol. doi:10.1007/s12035-013-8489-4

51. Younesi E, Toldo L, Müller B et al (2012) Mining biomarker information in biomedical literature. BMC Med Inform Decis Mak 12:148. doi:10.1186/1472-6947-12-148

52. Slater T, Song D (2012) Saved by the BEL ringing in a common language for the life sciences. Drug Discovery World, Fall 2012

53. Malhotra A, Younesi E, Bagewadi S, Hofmann-Apitius M (2014) Linking hypothetical knowledge patterns to disease molecular signatures for biomarker discovery in Alzheimer's disease. Genome Med 6:97. doi:10.1186/s13073-014-0097-z

54. Deisboeck TS, Wang Z, Macklin P, Cristini V (2011) Multiscale cancer modeling. Annu Rev Biomed Eng. doi:10.1146/annurev-bioeng-071910-124729

55. Chakrabarti A, Verbridge S, Stroock AD et al (2012) Multiscale models of breast cancer progression. Ann Biomed Eng. doi:10.1007/s10439-012-0655-8

56. Ribba B, Saut O, Colin T et al (2006) A multiscale mathematical model of avascular tumor growth to investigate the therapeutic benefit of anti-invasive agents. J Theor Biol 243:532–541. doi:10.1016/j.jtbi.2006.07.013

57. Dwivedi G, Fitz L, Hegen M et al (2014) A multiscale model of interleukin-6–mediated immune regulation in Crohn's disease and its application in drug discovery and development. CPT Pharmacomet Syst Pharmacol 3:1–9. doi:10.1038/psp.2013.64

58. Przybilla J, Rohlf T, Loeffler M, Galle J (2014) Understanding epigenetic changes in aging stem cells – a computational model approach. Aging Cell 13:320–328. doi:10.1111/acel.12177

59. Burrowes KS, Doel T, Brightling C (2014) Computational modeling of the obstructive lung diseases asthma and COPD. J Transl Med 12:S5. doi:10.1186/1479-5876-12-S2-S5

60. Mardis ER (2010) The $1,000 genome, the $100,000 analysis? Genome Med 2:84. doi:10.1186/gm205

61. Barrett T, Troup DB, Wilhite SE et al (2011) NCBI GEO: archive for functional genomics data sets – 10 years on. Nucleic Acids Res 39:D1005–D1010. doi:10.1093/nar/gkq1184

62. Parkinson H, Kapushesky M, Shojatalab M et al (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35:D747–D750. doi:10.1093/nar/gkl995

63. Ballereau S, Glaab E, Kolodkin A et al (2013) Functional genomics, proteomics, metabolomics and bioinformatics for systems biology. In: Prokop A, Csukás B (eds) Systems biology. Integrative biology and simulation tools. Springer, New York, pp 3–41

64. Goecks J, Nekrutenko A, Taylor J, Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86. doi:10.1186/gb-2010-11-8-r86

65. Omberg L, Ellrott K, Yuan Y et al (2013) Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. Nat Genet 45:1121–1126. doi:10.1038/ng.2761

66. Gao J, Aksoy BA, Dogrusoz U et al (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6:pl1. doi:10.1126/scisignal.2004088

67. Henry VJ, Bandrowski AE, Pepin A-S et al (2014) OMICtools: an informative directory for multi-omic data analysis. Database 2014:bau069. doi:10.1093/database/bau069

68. Rocca-Serra P, Brandizi M, Maguire E et al (2010) ISA software suite: supporting

standards-compliant experimental annotation and enabling curation at the community level. Bioinforma Oxf Engl 26:2354–2356. doi:10.1093/bioinformatics/btq415

69. Mathew JP, Taylor BS, Bader GD et al (2007) From bytes to bedside: data integration and computational biology for translational cancer research. PLoS Comput Biol 3:e12. doi:10.1371/journal.pcbi.0030012

70. Szalma S, Koka V, Khasanova T, Perakslis ED (2010) Effective knowledge management in translational medicine. J Transl Med 8:68. doi:10.1186/1479-5876-8-68

71. Athey BD, Braxenthaler M, Haas M, Guo Y (2013) tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research. AMIA Summits Transl Sci Proc 2013:6–8

72. Schumacher A, Rujan T, Hoefkens J (2014) A collaborative approach to develop a multi-omics data analytics platform for translational research. Appl Transl Genomics 3:105–108. doi:10.1016/j.atg.2014.09.010

73. Wolstencroft K, Owen S, du Preez F et al (2011) The SEEK: a platform for sharing data and models in systems biology. Methods Enzymol 500:629–655. doi:10.1016/B978-0-12-385118-5.00029-3

74. Miralles F, Gomez-Cabrero D, Lluch-Ariet M et al (2014) Predictive medicine: outcomes, challenges and opportunities in the Synergy-COPD project. J Transl Med 12(Suppl 2):S12. doi:10.1186/1479-5876-12-S2-S12

75. Maier D, Kalus W, Wolff M et al (2011) Knowledge management for systems biology a general and visually driven framework applied to translational medicine. BMC Syst Biol 5:38. doi:10.1186/1752-0509-5-38

76. Cano I, Tényi Á, Schueller C et al (2014) The COPD knowledge base: enabling data analysis and computational simulation in translational COPD research. J Transl Med 12(Suppl 2):S6. doi:10.1186/1479-5876-12-S2-S6

77. Abugessaisa I, Saevarsdottir S, Tsipras G et al (2014) Accelerating translational research by clinically driven development of an informatics platform – a case study. PloS One 9:e104382. doi:10.1371/journal.pone.0104382

78. Hernández-de-Diego R, Boix-Chova N, Gómez-Cabrero D et al (2014) STATegra EMS: an experiment management system for complex next-generation omics experiments. BMC Syst Biol 8(Suppl 2):S9. doi:10.1186/1752-0509-8-S2-S9

79. Dudley JT, Pouliot Y, Chen R et al (2010) Translational bioinformatics in the cloud: an affordable alternative. Genome Med 2:51. doi:10.1186/gm172