Chapter 2

Taking Bioinformatics to Systems Medicine

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

Abstract

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

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

1 Introduction

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

Ulf Schmitz and Olaf Wolkenhauer (eds.), *Systems Medicine*, Methods in Molecular Biology, vol. 1386,
DOI 10.1007/978-1-4939-3283-2_2, © Springer Science+Business Media New York 2016

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

2 Bioinformatics and High-Throughput Experimental Technologies

2.1 Bioinformatics in Biomedical Research

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

Table 1
Abbreviations and websites

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

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

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

2.2 New Dimensions in Biomedical Research

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

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

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

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

3 Bioinformatics and Systems Medicine

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

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

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

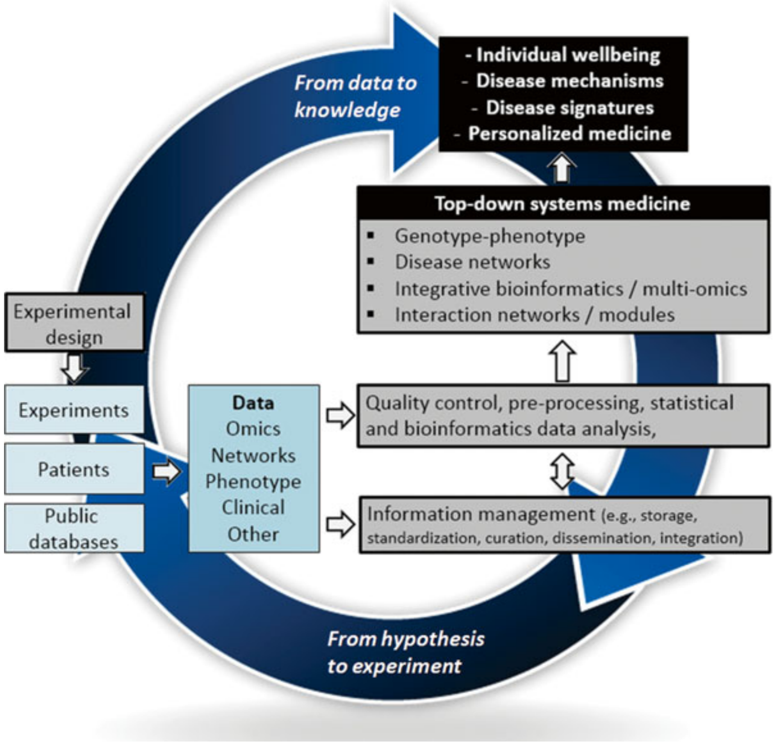
From data to knowledge
- Individual wellbeing
- Disease mechanisms
- Disease signatures
- Personalized medicine
Top-down systems medicine
▪ Genotype-phenotype
▪ Disease networks
▪ Integrative bioinformatics / multi-omics
▪ Interaction networks / modules
Experimental design
Experiments
Patients
Public databases
Data
Omics
Networks
Phenotype
Clinical
Other
Quality control, pre-processing, statistical and bioinformatics data analysis,
Information management (e.g., storage, standardization, curation, dissemination, integration)
From hypothesis to experiment

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

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

4 Information Management for Systems Medicine

4.1 Public Databases in Systems Medicine

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

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

4.2 Phenotype Databases

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

4.3 Clinical Data

To implement and advance systems medicine to the benefit of patients' health, it is crucial to integrate and analyze molecular data together with de-identified individual-level clinical data complementing general phenotype descriptions. Patient clinical data refers to a wide variety of data including basic patient information (e.g., age, sex, ethnicity), outcomes of physical examinations, patient history, medical diagnoses, treatments, laboratory tests, pathology reports, medical images, and other clinical outcomes. Inclusion of clinical data allows the stratification of patient groups into more homogeneous clinical subgroups. Availability of clinical data will increase the power of downstream data analysis and modeling to elucidate molecular mechanisms, and to identify molecular biomarkers that predict disease onset or progression, or which guide treatment selection. In biomedical studies clinical information is generally used as part of patient and sample selection, but some omics studies also use clinical data as part of the bioinformatics analysis (e.g., [9, 55]). However, in general, clinical data is unavailable from public resources or only provided on an aggregated level. Although good reasons exist for making clinical data available (Subheading 2.2), ethical and legal issues comprising patient and commercial confidentiality, and technical issues are the most immediate challenges [56, 57]. This potentially hampers the development of systems medicine approaches in a clinical setting since sharing and integration of clinical and nonclinical data is considered a basic requirement [1]. Biobanks [58] such as BBMRI [59] provide a potential source of biological material and associated (clinical) data but these are, generally, not publicly accessible, although permission to access data may be requested from the biobank provider. Clinical trials provide another source of clinical data for systems medicine studies, but these are generally owned by a research group or sponsor and not freely available [60] although ongoing discussions may change this in the future ([61] and references therein).

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

Medical patient records provide an even richer source of phenotypic information, and has already been used to stratify patient groups, discover disease relations and comorbidity, and integrate these records with molecular data to obtain a systems-level view of phenotypes (for a review see [63]). On the one hand, this integration facilitates refinement and analysis of the human phenome to, for example, identify diseases that are clinically uniform but have different underlying molecular mechanisms, or which share a pathogenetic mechanism but with different genetic cause [64]. On the other hand, using the same data, a phenome-wide association study (PheWAS) [65] would allow the identification of unrelated phenotypes associated with specific shared genetic variant(s), an effect referred to as pleiotropy. Moreover, it makes use of information from medical records generated in routine clinical practice and, consequently, has the potential to strengthen the link between biomedical research and clinical practice [66]. The power of phenome analysis was demonstrated in a study involving 1.5 million patient records, not including genotype information, comprising 161 disorders. In this study it was shown that disease phenotypes form a highly connected network suggesting a shared genetic basis [67]. Indeed, later studies that incorporated genetic data resulted in similar findings and confirmed a shared genetic basis for a number of different phenotypes. For example, a recent study identified 63 potentially pleiotropic associations through the analysis of 3144 SNPs that had previously been implicated by genome-wide association studies (GWAS) as mediators of human traits, and 1358 phenotypes derived from patient records of 13,835 individuals [68]. This demonstrates that phenotypic information extracted manually or through text mining from patient records can help to more precisely define (relations between) diseases. Another example comprises the text mining of psychiatric patient records to discover disease correlations [52]. Here, mapping of disease genes from the OMIM database to information from medical records resulted in protein networks suspected to be involved in psychiatric diseases.

5 Integrative Bioinformatics

Integrative bioinformatics comprises the integrative (statistical) analysis of multiple omics data types. Many studies demonstrated that using a single omics technology to measure a specific molecular level (e.g., DNA variation, expression of genes and proteins, metabolite concentrations, epigenetic modifications) already provides a wealth of information that can be used for unraveling molecular mechanisms underlying disease. Moreover, single-omics disease signatures which combine multiple (e.g., gene expression) markers have been constructed to differentiate between disease subtypes to support diagnosis and prognosis. However, no single technology can reveal the full complexity and details of molecular networks observed in health and disease due to the many interactions across these levels. A systems medicine strategy should ideally aim to understand the functioning of the different levels as a whole by integrating different types of omics data. This is expected to lead to biomarkers with higher predictive value, and novel disease insights that may help to prevent disease and to develop new therapeutic approaches. Integrative bioinformatics can also facilitate the prioritization and characterization of genetic variants associated with complex human diseases and traits identified by GWAS in which hundreds of thousands to over a million SNPs are assayed in a large number of individuals. Although such studies lack the statistical power to identify all disease-associated loci [69], they have been instrumental in identifying loci for many common diseases. However, it remains difficult to prioritize the identified variants and to elucidate their effect on downstream pathways ultimately leading to disease [70]. Consequently, methods have been developed to prioritize candidate SNPs based on integration with other (omics) data such as gene expression, DNase hypersensitive sites, histone modifications, and transcription factor-binding sites [71].

5.1 Data Integration

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

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

5.2 Multi-omics Disease Signatures

Single-level omics data has extensively been used to identify disease-associated biomarkers such as genes, proteins, and metabolites. In fact, these studies led to more than 150,000 papers documenting thousands of claimed biomarkers, However, it is estimated that fewer than 100 of these are currently used for routine clinical practice [81]. Integration of multiple omics data types is expected to result in more robust and predictive disease profiles since these better reflect disease biology [82]. Further improvement of these profiles may be obtained through the explicit incorporation of interrelationships between various types of measurements such as microRNA–mRNA target, or gene methylation–microRNA (based on a common target gene). This was demonstrated for the prediction of short-term and long-term survival from serous cystadenocarcinoma TCGA data [83].

6 Biological Networks

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

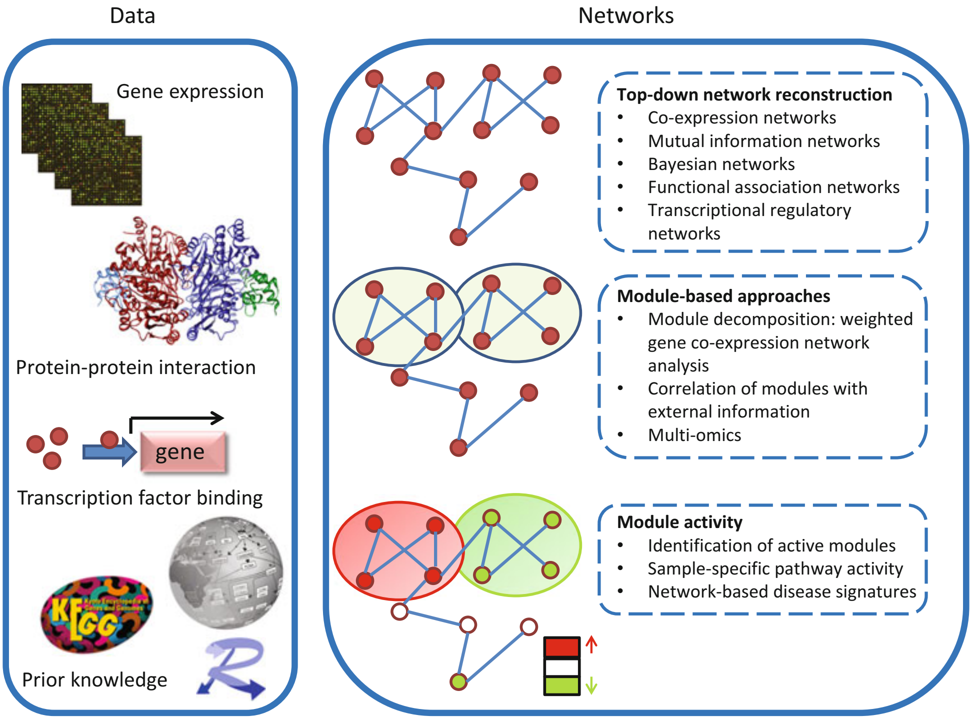
Data
Networks
Gene expression
Protein-protein interaction
gene
Transcription factor binding
Prior knowledge
KEGG
Top-down network reconstruction
• Co-expression networks
• Mutual information networks
• Bayesian networks
• Functional association networks
• Transcriptional regulatory networks
Module-based approaches
• Module decomposition: weighted gene co-expression network analysis
• Correlation of modules with external information
• Multi-omics
Module activity
• Identification of active modules
• Sample-specific pathway activity
• Network-based disease signatures

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

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

6.1 Top-Down Network Reconstruction

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

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

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

Bayesian networks allow the description of statistical dependencies between variables in a generic way [91, 92]. Bayesian networks are directed acyclic networks in which the edges of the network represent conditional dependencies; that is, nodes that are not connected represent variables that are conditionally independent of each other. A major bottleneck in the reconstruction of Bayesian networks is their computational complexity. Moreover, Bayesian networks are acyclic and cannot capture feedback loops that characterize many biological networks. When time-series rather than steady-state data is available, dynamic Bayesian networks provide a richer framework in which cyclic networks can be reconstructed [93].

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

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

Bayesian networks have also been used to integrate multi-omics data. The combination of genotypic and gene expression data is particularly powerful, since DNA variations represent naturally occurring perturbations that affect gene expression detected as expression quantitative trait loci (eQTL). *Cis*-acting eQTLs can then be used as constraints in the construction of directed Bayesian networks to infer causal relationships between nodes in the network [103].

6.2 Module-Based Approaches

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

Weighted gene co-expression network analysis (WGCNA)decomposes a co-expression network into modules using clustering techniques [109]. Modules can be summarized by their module eigengene, a weighted average expression profile of all gene member of a given module. Eigengenes can then be correlated with external sample traits to identify modules that are related with these traits. Parikshak et al. [110] used WGCNA to extract modules from a co-expression network constructed using fetal and early postnatal brain development expression data. Next, they established that several of these modules were enriched for genes and rare de novo variants implicated in autism spectrum disorder (ASD). Moreover, the ASD-associated modules are also linked at the transcriptional level and 17 transcription factors were found acting as putative co-regulators of ASD-associated gene modules during neocortical development. WGCNA can also be used when multiple omics data types are available. One example of such an approach involved the integration of transcriptomic and proteomic data from a study investigating the response to SARS-CoV infection in mice [111]. In this study WGCNA-based gene and protein co-expression modules were constructed and integrated to obtain module-based disease signatures. Interestingly, the authors found several cases of identifier-matched transcripts and proteins that correlated well with the phenotype, but which showed poor or anticorrelation across these two data types. Moreover, the highest correlating transcripts and peptides were not the most central ones in the co-expression modules. *Vice versa*, the transcripts and proteins that defined the modules were not those with the highest correlation to the phenotype. At the very least this shows that integration of omics data affects the nature of the disease signatures.

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

6.3 Network-Based Disease Signatures

Often, published disease signatures cannot be replicated [81] or provide hardly additional biological insight. Also here (modular) network-based approaches have been proposed to alleviate these problems. A common characteristic of most methods is that the molecular activity of a set of genes is summarized on a per sample basis. Summarized gene set scores are then used as features in prognostic and predictive models. Relevant gene sets can be based on prior knowledge and correspond to canonical pathways, gene ontology categories, or sets of genes sharing common motifs in their promoter regions [115]. Gene set scores can also be determined by projecting molecular data onto a biological network and summarizing scores at the level of subnetworks for each individual sample [116]. While promising in principle, it is still subject of debate whether gene set-based models outperform gene-based ones [117].

6.4 Crossing the Species Boundary

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

6.5 From Networks to Medicine

We reviewed a wide array of different approaches showing how networks can be used to elucidate integrated genetic, molecular, and cellular networks. However, in general no single approach will be sufficient and combining different approaches in more complex analysis pipelines will be required. This is fittingly illustrated by the DIGGIT (Driver-gene Inference by Genetical-Genomics and Information Theory) algorithm [127]. In brief, DIGGIT identities candidate master regulators from an ARACNe gene co-expression network integrated with copy number variations that affect gene expression. This method combines several previously developed computational approaches and was used to identify causal genetic drivers of human disease in general and glioblastoma, breast cancer, and Alzheimer's disease in particular. This enabled identification of KLHL9 deletions as upstream activators of two previously established master regulators in a specific subtype of glioblastoma.

7 Discussion

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

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

Integrative bioinformatics approaches clearly have added value for systems medicine as they provide a better understanding of biological systems, result in more robust disease markers, and prevent (biological) bias that would possibly occur from using single-omics measurements. However, such studies, and the scientific community in general, would benefit from improved strategies to disseminate and share data which typically will be produced at multiple research centers (e.g., https://www.synapse.org; [140]). Integrative studies are expected to increasingly facilitate personalized medicine approaches such as demonstrated by Chen and coworkers [141]. In their study they presented a 14-month "integrative personal omics profile" (iPOP) for a single individual comprising genomic, transcriptomic, proteomic, metabolomic, and autoantibody data. From the whole-genome sequence data an elevated risk for type 2 diabetes (T2D) was detected, and subsequent monitoring of HbA1c and glucose levels revealed the onset of T2D, despite the fact that the individual lacked many of the known non-genetic risk factors. Subsequent treatment resulted in a gradual return to the normal phenotype. This shows that the genome sequence can be used to determine disease risk in a healthy individual and allows selecting and monitoring specific markers that provide information about the actual disease status.

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

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

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

Acknowledgements

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

References

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

data analytical approaches for epigenomics. Environ Mol Mutagen 55(3):155–170

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

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

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

tion in acute myeloid leukemia. PLoS Comput Biol 10(10):e1003908

103. Zhu J, Zhang B, Smith EN et al (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 40(7):854–861

104. Zhang B, Gaiteri C, Bodea LG et al (2013) Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. Cell 153(3):707–720

105. Greenawalt DM, Dobrin R, Chudin E et al (2011) A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. Genome Res 21(7):1008–1016

106. Alon U (2007) An introduction to systems biology: design principles of biological circuits, vol 10, Chapman & Hall/CRC mathematical and computational biology. Chapman & Hall/CRC, Boca Raton, FL

107. Segal E, Friedman N, Kaminski N et al (2005) From signatures to models: understanding cancer using microarrays. Nat Genet 37(Suppl):S38–S45

108. Mitra K, Carvunis AR, Ramesh SK et al (2013) Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 14(10):719–732

109. Zhao W, Langfelder P, Fuller T et al (2010) Weighted gene coexpression network analysis: state of the art. J Biopharm Stat 20(2):281–300

110. Parikshak NN, Luo R, Zhang A et al (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell 155(5):1008–1021

111. Gibbs DL, Gralinski L, Baric RS et al (2014) Multi-omic network signatures of disease. Front Genet 4:309

112. Dittrich MT, Klau GW, Rosenwald A et al (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24(13):i223–i231

113. Vandin F, Upfal E, Raphael BJ (2011) Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol 18(3):507–522

114. Vaske CJ, Benz SC, Sanborn JZ et al (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26(12):i237–i245

115. Drier Y, Sheffer M, Domany E (2013) Pathway-based personalized analysis of cancer. Proc Natl Acad Sci U S A 110(16): 6388–6393

116. Chuang HY, Lee E, Liu YT et al (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140

117. Staiger C, Cadot S, Gyorffy B et al (2013) Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. Front Genet 4:289

118. Ala U, Piro RM, Grassi E et al (2008) Prediction of human disease genes by human-mouse conserved coexpression analysis. PLoS Comput Biol 4(3):e1000043

119. Clark C, Kalita J (2014) A comparison of algorithms for the pairwise alignment of biological networks. Bioinformatics 30(16):2351–2359

120. Berg J, Lassig M (2006) Cross-species analysis of biological networks by Bayesian alignment. Proc Natl Acad Sci U S A 103(29):10967–10972

121. Kolar M, Meier J, Mustonen V et al (2012) GraphAlignment: Bayesian pairwise alignment of biological networks. BMC Syst Biol 6:144

122. Sweet-Cordero A, Mukherjee S, Subramanian A et al (2005) An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat Genet 37(1):48–55

123. Bild AH, Yao G, Chang JT et al (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439(7074):353–357

124. Anvar SY, Tucker A, Vinciotti V et al (2011) Interspecies translation of disease networks increases robustness and predictive accuracy. PLoS Comput Biol 7(11):e1002258

125. Hu Y, Wu G, Rusch M et al (2012) Integrated cross-species transcriptional network analysis of metastatic susceptibility. Proc Natl Acad Sci U S A 109(8):3184–3189

126. Altschuler GM, Hofmann O, Kalatskaya I et al (2013) Pathprinting: an integrative approach to understand the functional basis of disease. Genome Med 5(7):68

127. Chen JC, Alvarez MJ, Talos F et al (2014) Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. Cell 159(2):402–414

128. Taylor CF, Field D, Sansone SA et al (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26(8):889–896

129. Chervitz SA, Deutsch EW, Field D et al (2011) Data standards for Omics data: the basis of data sharing and reuse. Methods Mol Biol 719:31–69

130. Rubin DL, Shah NH, Noy NF (2008) Biomedical ontologies: a functional perspective. Brief Bioinform 9(1):75–90

131. Joosten RP, Vriend G (2007) PDB improvement starts with data deposition. Science 317(5835):195–196

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