Chapter | Ten

Network Biology Empowering Detection and Understanding of Interactions Between Genetic Factors in Development of Complex Phenotypes

Jesper Tegnér, Francesco Marabita and David Gomez-Cabrero

Department of Medicine, Unit of Computational Medicine, Karolinska Institutet, Center for Molecular Medicine, Karolinska University Hospital, Solna, Stockholm, Sweden

10.1 RISE OF BIG DATA, COMPUTING, AND PREDICTION

Our world has changed dramatically during the last decade. For example, the rise and embedding of efficient algorithms and computing power in distributed technologies, such as computers, smartphones, sensors, and the Internet, have fundamentally transformed our way of living. Large amounts of data, often referred to as Big Data, are being produced and citizens have access to an unprecedented amount of data and the challenge is to make sense and use of this data. Conceptually, such utilization requires an analysis within each data type as well as across different data types. Effectively, the task of making such data useful requires an analysis to address which parts of the data

L. Padyukov (Ed): Between the Lines of Genetic Code. DOI: http://dx.doi.org/10.1016/B978-0-12-397017-6.00010-6

correlate with each other, which combination of data parts can predict other parts of a high-dimensional data cube. In its generality, this is a very challenging task but "simple" examples are abundant. For example, the "Like" button at Facebook can reveal and predict a surprising amount of information about the user. The algorithms behind it could predict gender, political opinions, and religious beliefs in close to 90% of cases and to some lesser extent smoking and drug habits [1]. Importantly, medicine and healthcare are currently on the verge of a similar revolution. Drivers in this development include new technologies for molecular profiling and a systems analysis of disease, which has set the scene for altering medicine and healthcare from being a reactive art to becoming a preventive science. The sequencing of the human genome a decade ago and the following postgenomic acceleration of the development of technology have created a situation of immense large-scale data production [2]. Especially, the explosion in the last years of next-generation sequencing (NGS) applications and their continuing drop in price makes genome-wide, system-oriented approaches in biomedical research increasingly affordable for many molecular biology labs. Such data production at the peta/exabyte level generates enormous challenges with respect to data management, computing, security, and data analysis [3,4]. The pace of the technology development and data production has relentlessly accelerated during the last decade and will continue to do so over the next decade as genomics enters the clinic. The next step is without a doubt the application of powerful algorithms on top of these data in order to identify predictive patterns within and across different data sets.

Big Data, computing, and prediction are becoming personal. Ever since the sequencing of the human genome the notion of a personalized understanding of health and disease has been an important primus motor of the field. In particular, Leroy Hood has been an avid articulator of this ongoing transformation using the concept of P4 medicine, referring to a Personalized, Predictive, Preventive, Participatory medicine [5]. Through an integrative genomics approach, there is a promise to predict and prevent disease and to benefit from the participation of citizens and patients. Clearly, potential benefits of the concept of P4 medicine include early detection of disease, stratification of patients into subgroups that enables the selection of optimal therapy, early assessment of individual drug responses thus reducing adverse drug reactions, improvement of clinical trials by reduction of exposure time and failure rate, and development of tools enabling the clinician to shift the emphasis from treatment to prevention and from disease to wellness. Yet, to make progress towards such a vision we need powerful tools for producing, analyzing, integrating, and modeling large amounts of heterogeneous data [6,7] and to crystalize this data to personalized knowledge supporting decisions and actions. Hence, the underlying computational challenges in medicine and healthcare and the emerging amounts of data are closing in on the current situation we already can witness in other domains in our society dealing with Big Data and their analysis.

10.2 USING GENETIC VARIANTS AS INDEPENDENT FEATURES IS NOT SUFFICIENT FOR REALIZING P4 MEDICINE

During the first decade after the sequencing of the human genome, the studies of genetic variants in the DNA have been at the forefront of research. The underlying belief, early on, was that by charting single nucleotide polymorphisms (SNPs), we would be able to better understand molecular mechanisms of complex diseases and thereby improve our capability to predict disease and estimate risk of disease. Thus, large volumes of genome-wide association studies (GWAS) data have been produced during the last decade and represent a potential goldmine in conjunction with phenotype information to unravel mechanisms of complex diseases. To date, thousands of genetic variants (SNPs) have been associated to different diseases and disease-related phenotypes. It has, however, become increasingly clear that univariate SNP analyses are not sufficient for either risk prediction or for realizing a P4 medicine program. In part due to that in the aftermath of the human genome project, several layers of molecular mechanisms have been uncovered which are important for regulating the activity of genes, thus rendering the task of understanding mechanisms of disease more challenging than previously thought. Moreover, it remains unclear why the effect sizes of the genetic variants are as a rule tiny, as discussed elsewhere in this book. In particular why are the effect sizes small even for a phenotype such as height, despite the fact that it is known that there is strong genetic component? Consequently, there is still a vivid debate about how to find what has been referred to as the "*missing heritability of complex diseases*" [8]. It is a complex subject and some assumptions such as the quantitative degree of genetic component for a given phenotype might be disputed and deserve reinvestigation. Yet, this is a remaining bottleneck and we are still lacking appropriate computational tools to fully capitalize on existing GWAS data, specifically the statistical capability to analyze GWAS data benefits, and therefore it is limited by the assumption that SNPs are independent. The advantage is the assumption of independence, which increases the statistical power when testing each SNP against the phenotype (disease). Yet the limitation is that the very same assumption forces our analysis to be less comprehensible as genetic interactions will not be detectable by such design. The core problem motivating the rationale of avoiding an analysis of interactions during this first decade of genetic analysis is the effective explosion of hypotheses to be tested when searching for higher order correlations. Such an analysis investigating pairwise interactions, for example, requires correction for multiple testing which unfortunately effectively abolishes the statistical significance [9]. Yet, successful analysis of Big Data hinges upon the ability to discover predictive patterns by pooling and testing parts in the data cube, as in the Facebook example above. Furthermore, there are strong biological mechanistic justifications and medical pragmatic reasons for having a statistical framework

enabling an investigation of higher order statistical interactions. Genes do not act in isolation and there has to be a physical basis for the observed interactions. From a clinical point of view, interactions between lifestyle/environmental factors and genes are key to understand their effects on the phenotype (disease). Such an understanding of interactions has the potential to provide clinical decision support to the physician.

In the current chapter, we address how to interpret higher order statistical interactions in terms of underlying biological processes purportedly generating such dependencies. Therefore, we specifically review how biological networks, such as those generated from systems biological approaches, could possibly facilitate the interpretation of observed interactions between genes or the interactions between genes and environment. In this part, we also include and explore to what extent epigenetics could mediate and support different kind of interactions beyond linear correlations between the gene transcripts. In the following section in this chapter, we consider how to represent the problem of discovering interactions as a feature selection problem. Our rationale is to precisely understand why the problem of interactions is mathematically hard and how we could potentially empower algorithms to detect interactions. In this section, we end up essentially identifying three major challenges: (a) how to identify relevant variables (interactions) in high-dimensional data, (b) how to incorporate prior knowledge, and (c) how to develop robust methods that do not depend on fine-tuning of method specific parameters. Our discussion in this second section, emphasizing robustness and prior knowledge, motivates the final part of this chapter where we revisit biological networks and discuss how to possibly incorporate prior knowledge derived from systems biological investigations and thereby possibly increase the statistical power in the analysis of detecting interactions.

10.3 NETWORK BIOLOGY—A FRAMEWORK FOR DETECTING AND INTERPRETING GENETIC INTERACTIONS

10.3.1 Graphs—a Unifying Biological Language

Networks have proven to be the language of choice when we need to understand how to combine large and different types of data in a given biomedical problem. The usefulness of networks comes from their general capacity of capturing and representing vastly different structures and processes in the natural, social, and human sciences in the language of nodes and their connections (edges). Depending on the specific application in molecular biology within a cell, the edges can be undirected (binding between molecules), directed (molecule A has a causal effect on molecule B), and/or have a sign expressing activation or repressing in the causal action. An early quote from Laslo Barabasi (www.barabasilab.com), the physicist who has pioneered our capability of analyzing the world through the lens of networks captures here the essence of networks in biology and medicine.

> *"A key aim of postgenomic biomedical research is to systematically catalogue all molecules and their interactions within a living cell. There is a clear need to understand how these molecules and the interactions between them determine the function of this enormously complex machinery, both in isolation and when surrounded by other cells. Rapid advances in network biology indicate that cellular networks are governed by universal laws and offer a new conceptual framework that could potentially revolutionize our view of biology and disease pathologies in the twenty-first century" [10].*

Ever since this remark, almost 10 years ago, it has become evident that it is a very powerful approach to analyze living matter, such as cells, during health or disease, as interconnected molecular graphs and relate their structural properties to appropriate phenotypes where interactions play an important role [3,11]. This has become the major conceptual framework on how to organize and analyze the Big Molecular Data, which is currently being generated at increasing pace across biology and medicine. Interestingly, the core concepts of network biology as a subject have deep roots in discrete and topological branches of mathematics. In the eighteenth century, Leonhard Euler considered the problem of crossing bridges in a city in such a manner that every bridge would be crossed once and only once. This problem—referred to as the seven bridges of Konigsberg—which was proved by Euler to have no solution required the development of what turned out to be the foundation of graph theory, a corner stone of modern mathematics, and a precursor to topology. Analyzing graphs as discrete entities, rich of statistical and combinatorial enigmas, or as backbones upon which dynamical processes (equations) develop over time provide a fertile framework for understanding a surprisingly rich array of phenomena in nature. Hence, it should come as no surprise that the growing body of molecular data, which is currently being produced, could be organized and analyzed with the assistance of graphs. We will therefore first describe the different types of molecular data (nodes) and their putative interactions (edges).

10.3.2 Nodes

The sequencing of the human genome and the subsequent postgenomic acceleration of technological developments have resulted in immense large-scale data generation thus producing different types of molecular networks. These technologies have opened new windows into the cellular circuitry beyond the DNA sequence and individual SNPs as detected by GWAS [2]. One major achievement has been the explosion of the number of different types of molecular data that can be generated today. For example, NGS technologies and other high-throughput techniques produce data on DNA sequence variants, transcriptomics including different types of RNA molecules, proteins, metabolites, and epigenetic modifications, at a decreasing cost and increasing molecular resolution as exemplified by the ENCODE project [12]. Specifically, the encyclopedic analysis of genomes reveals a collection of

molecular entities, such as DNA, SNPs, copy-number variants (CNVs), DNA methylations, protein coding RNA, noncoding RNA, splice variants, RNA editing, histone modifications, nucleosome positioning, transcription factors (TFs), transcription start sites, promoters, chromatin accessible regions, localization of proteins, protein modifications (these are numerous), and metabolites. All this molecular variety creates formidable bioinformatics challenges, which essentially come in two parts—extracting the nodes and identifying their interactions. The task of extracting these nodes refers to the challenge of extracting a reliable statistically significant signal from each of these data types. This task, which requires deep expertise, is very data dependent and there is rapid progress in the field where rather mature "bioinformatics pipelines" are being produced and made publically available (bioconductor) enabling the analysis of different types of omics data, such as transcriptomics, proteomics, metabolomics, and the novel "-seq" approaches: RNA-seq, ChIP-seq, and Methyl-seq [13–15]. Yet, whereas the analysis of SNPs or RNA-seq data is comparatively mature, the bioinformatics for analyzing DNA methylations as captured using an array-based platform like the Illumina 450k array is less well understood [16,17]. To summarize, conceptually we can represent these different molecular entities as nodes in a graph. At this juncture, it is important to appreciate that the quality or reliability (false positives) of the nodes depends on the data type as well as on the specific bioinformatics pipelines which have been used to extract a set of nodes from a given data type, thus also affecting the amount of false negatives, or nodes which remain undetected. Interactions are therefore either an empirical or model-dependent phenomena—depending on viewpoint—which in any case effectively connects a node or a set of nodes with another set of nodes when considering gene–gene interactions. As for the gene–environmental interactions we will interpret or search for a corresponding physical trace between a molecular node or set of molecular nodes and an environmental factor. Therefore, it is essential to ask, how could we possibly find the edges connecting the nodes and/or the environment and the nodes? This is the second major bioinformatics challenge, which relatively speaking has been less developed as of today.

10.3.3 Edges

From a mathematical point, we have nodes with different colors corresponding to the different data types. In principle, every node having color A could be connected to another node with the same color. Moreover, an edge could theoretically be drawn between any two nodes regardless of the color. All in all, this becomes potentially a very large number of edges, thus a very complex graph including nodes of different colors. Here we will basically consider two conceptually different approaches on how to detect edges. The first idea is to observe or extract the edges directly from biological databases. We will discuss some of the major approaches and then we will discuss the evidence supporting the view that such edges can effectively become rewired in

the presence of epigenetic modifications. Following this we review some of the major network motifs and structures, which have thus far been identified in large-scale networks. In contrast, to reverse engineer the edges directly from observational data, corresponding to observing the nodes over time or during different conditions, represents the second idea with origins from engineering on how to perform system identification. We close this first section of the chapter by discussing how biological networks as defined and identified as above could provide a physical or associative basis for interpreting detected genetic or environmental interactions.

Available biological knowledge and databases now provide a rich source of putative edges. For example, there are DNA binding sites, transcription start sites, promoters, protein−protein interactions, and DNA binding proteins, thus defining special properties for a given data type as well as possible edges between the molecular nodes (Figure 10.1). Since there are over 1400 well-curated public databases [18], it is not clear how to systematically extract reliable edges from these rich resources. Moreover, there is still the challenge of how to integrate different data types, such as metabolomics, proteomics, and transcriptomics, which is relevant here to interpret genetic interactions. It is evident that we need to integrate these different molecular data types in order to understand the putative biological basis mediating such interactions as identified by the methods described in the current book. To proceed beyond genetics and interactions between only molecular nodes in order to address how to integrate different molecular data layers with external environmental factors, the concept of epigenetics is central. Epigenetics refer to the modification of DNA and/or related proteins (the nodes) without altering the nucleotide sequence.

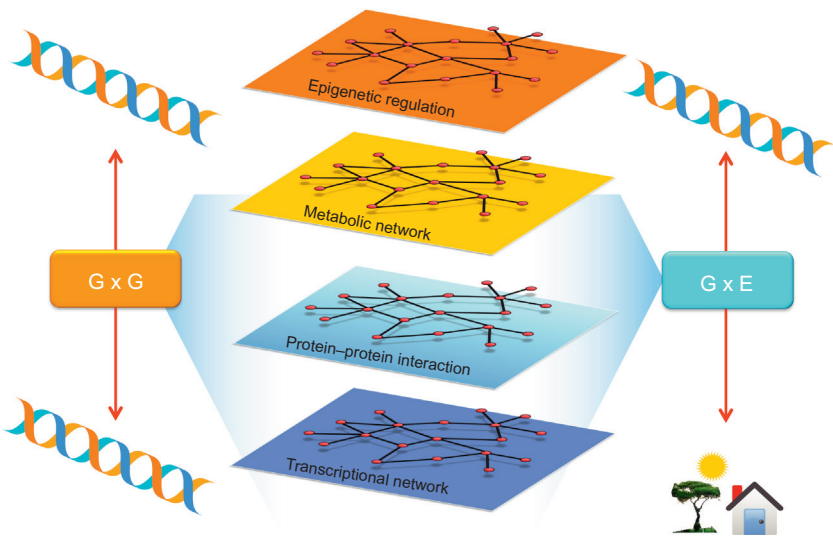

FIGURE 10.1 Schematic illustration of different kinds of networks, which can be reconstructed from different data types.

DNA methylation is currently the most studied and the best understood epigenetic modification and has been established as an additional mechanism for gene inactivation in different cell types. However, while all cell types share nearly the same genome sequence, the regulation of gene expression is not only at the level of TFs and the transcription factor binding sites (TFBS) but also at some levels where the DNA sequence is not modified, such as histone modifications or DNA methylation. Histone modification profiles have been studied at a genome-wide level and their association with gene expression has been demonstrated at promoter regions [19,20]. Furthermore, predictive computational models of gene expression on the basis of histone modifications profiles have been built [21]. Recent work has clearly shown that DNA methylation patterns differentiate among cell types therefore encoding cell and tissue-specific transcriptional programs [22–24], thus effective rewiring the graph by modifying the properties of the nodes, thereby altering the edges. Yet the precise regulatory mechanism that involves methylation is still not clear [25]. Hence, to improve the understanding the functions of DNA methylation it is useful to evaluate its distribution across the genome into CpG-rich regions known as CpG islands. Interestingly, half of the genes in vertebrates contain CpG islands, defining almost a bimodal distribution in the transcription start sites [26] and this observation suggested an association between DNA methylation and gene transcription. However, there are still numerous genomic elements, which have not been uncovered and present serious challenges. First, the definition of CpG islands is a controversial topic [27], which affects the identification of regulatory regions [28]. Second, in addition to promoter regions [29], satellite repeats [30] and CpG shores [28] have been considered to be regulatory. Third, from these considerations it follows that the characteristics of differentially methylated regions (DMRs) are not clearly defined, thus rendering the question of whether a unique CpG is regulatory or what constitutes the minimum size of a CpG regulatory region unresolved. Finally, several studies are currently investigating the mechanisms and putative functions of DNA methylation and DNA demethylation. For instance the relation of 5-methylcytosine with a recently associated modification 5-hydroxymethylcytosine [31] has been associated to both age and neurodegenerative disorders. Hence, there is still much to learn about which nodes or areas of the genome are altered by DNA methylation. The relevance of DNA methylation as regulatory mechanism is well established and may very well serve as one of possible mediators of genetic interactions. Deregulation of DNA methylation has been associated with cancers with gene body DNA methylation in tumor suppressor genes, such as TP53 [32], DNA methylation of tumor suppressors TSSs, or deregulation of DNA methylation machinery. DNA methylation can be used as a marker for several type of cancers, i.e., identification of respiratory tract cancers [33], bladder cancer [34], and solid cancer diagnostics [35]. In addition, the intensity of DNA methylation has been associated also to complex diseases, such as rheumatoid arthritis (RA) [36], chronic obstructive pulmonary disease [37], and multiple sclerosis [38]

and natural processes like aging [39,40]. These observations suggest that when gene−gene interactions are identified and correlated to diseases, it is not unlikely that DNA methylation could operate as a mechanistic mediator of the observed gene−gene interaction.

10.3.4 From Individual Edges to Networks

To integrate these different nodes (molecular data types) and their epigenetic modifications (edges), it is useful to represent this information in terms of biological networks. There has been a large number of studies, published during the last decade, defining methods for integration into networks. Here we summarize some of the major strategies to reconstruct integrated biological networks. Instead of considering a system as a set of nonrelated elements, a network summarizes the system by enumerating its relevant elements (nodes) and the interactions (edges) between the nodes [41]. Most cellular functions are organized in network-structured sets of genes and/or proteins and/or metabolites communicating through biochemical and physical interactions [10,42]. The network analysis allows the study of a system in a simplified but systematic manner capitalizing upon progress in mathematics and computational tools for analyzing graphs. Initially networks were used to study the interaction (relation) on single types of entities, such as transcript−transcript interactions and protein−protein interactions [43], but their use have been extended recently to include a wide range from different molecular elements, to even representing different diseases as nodes in a graph where the edges represent relative risk for example [11]. Interestingly, most biological networks have properties in their graphs effectively associating to being nonrandom networks, where one of the most important properties is that they have a scale-free distribution of their edges. The meaning behind a scale-free network is that the number of links or edges per node follows a power law distribution. Indeed, what has been observed in a number of biological networks is that most nodes have few connections, while a smaller number of nodes have a large number of connections, thus suggesting that such nodes are key elements in the network and whose deregulation will significantly affect the entire network [44]. The relevance of such a discovery is twofold: first it describes the necessity of identifying those highly connected nodes [45], which are candidates to operate as master regulators. Second, it provides insights into evolution as it follows that a novel connection is more likely to happen with those elements that are already highly connected. Additional properties which have been investigated include the average length of the path linking any pair of nodes among others and interestingly, biological networks turn out to have a small diameter in the sense that the distance between any two nodes is as a rule very short.

Moreover, networks do not necessarily need to be defined for a single type of element. Networks with heterogeneous node types are indeed possible, as long as the nature or semantics of the different interactions (edges) in

the networks are properly defined [46]. For example, bipartite networks with two types of nodes A and B and interactions between the two types of nodes allow the construction of two types of networks. Two nodes of the type A (A1 and A2) are connected if there is a path from A1 via a B node to A2. For example, a bipartite graph which contains diseases (node A) and genetic variants (node B) and the connections are based from GWAS studies and the corresponding ICD codes (disease). The study of such bipartite graphs allows the identification of highly connected diseases by observing the genes shared [47], by employing the logic outlined above on how to interconnect nodes of one type (diseases) by using a mapping (ICD) via the other nodes (genetic variants). Following these ideas, several investigators have recently published papers addressing the generation of networks that combine mRNA and epigenetic information. EpiRegNet [48] is a tool that allows the identification of possible epigenetic marks associated to genome-wide changes of gene expression. Here the authors constructed a bipartite network of histone−gene associations and this network was used to identify histone marks associated to a subset of genes and importantly, their methodology was validated in the analysis of an embryonic stem cell differentiation. Furthermore, the authors provide information of TF regulation by using the publically available ChIP-Seq data. Overall, explorative network based analysis of systems at the transcriptomics and epigenetic level is evidently becoming increasingly useful. As an additional example, we have the analysis performed by Ciofano et al. [49] where they identified a global Th17 transcriptional regulatory network by combining genome-wide TF occupancy data obtained from ChIP-Seq experiments, and mRNA expression of TF mutants and time series of Th17 differentiation into a biological network. Their computational analysis of the data enabled the identification of master regulators, the relevant modules (groups of genes), and the interactions (edges) between the genes and modules defining the differentiation of naïve $CD4^+$ T cells into Th17 cells. A recent paper [50] increased the temporal resolution of the Th17 network by identifying several transcriptional waves during differentiation. This clearly demonstrates the feasibility and power of a network biology approach to identify the edges between different and similar nodes. Hence, using such biological networks as a backbone to interpret the mechanistic basis underlying discovered genetic interactions promises to be a powerful methodology, potentially providing a biological basis for the observed statistical interaction.

Thus far we have considered approaches that essentially collect and integrate different data sources in order to represent them as biological networks on the basis of some rules for how to connect the data sets. Another complementary conceptual idea to identify biological networks hinges upon the insight that using a computational model it is possible to identify a biological network directly from omics measurements without depending upon current knowledge. Such an approach increases the likelihood of detecting novel edges not yet captured in current databases. The problem of identifying a system from its behavior is referred to as reverse engineering and it has been

widely used to uncover regulatory metabolic or transcriptomics networks. Many different methodologies have been developed and applied during the last decade. These include regression models (including several types of Lasso models [51]), mutual information (i.e., ARACNE [52]), correlation-based approaches, Bayesian networks, random forest algorithms, and (most recently) a combination of several different methods [53]. A recent comparison between the different methodologies made it clear that each methodology is able to capture different sets of edges of a network. The study concluded that it is more robust to use different types of information, therefore the combination of heterogeneous data types is able to uncover most efficiently the associations and minimizes the amount of false positives [53]. Despite the promising results and tools available, very few methods have been developed for integrative network analysis of heterogeneous data sets. In addition, the problem of visualizing heterogeneous networks with complex associations for explorative analysis remains an open problem despite development of state-of-the-art tools, including Cytoscape [54] and Gephi [55]. On the one hand, motif network analysis allowed the identification of small-size mechanistic relations between entities that provides properties (such as a robustness) to a system [42]. On the other hand, even though computational biology is a very active research area, much remains to be explored where for example epigenetic regulation and data have not yet been incorporated into current system identification algorithms.

In part due to these shortcomings of integrating several data types into an unbiased framework based on directly identifying biological networks directly from data, integrative bioinformatics techniques have remained useful. For example, the identification of the genetic background corresponding to epigenetic changes which correlate with a disease phenotype have recently gained momentum as a next step following the wave of pure GWAS analysis during the last 5 years. Deciphering such an additional layer of epigenetic complexity will eventually contribute to the understanding of the causal pathway from genetic variation to disease etiology, assuming that part of the heritability may be mediated by epigenetic modifications, which in turn may entail effects on the transcriptional regulation. One of the major differences between the two data types consists in the spatial and temporal variability of the marks. Whereas SNPs do not change across tissues and cannot therefore be interpreted as a consequence of a particular disease, CpG methylation is subject to spatial (tissue- or cell-specific methylation) and/or temporal variability (age-dependent, disease-associated, or environmental-mediated differential methylation). Moreover, while GWAS studies as a rule measure polymorphisms on DNA extracted from whole blood, the design of epigenetic studies is complicated by the fact that the tissue implicated in the disease pathogenesis may not be easily accessible in clinical specimens and therefore alternative tissues must be used. However, such a procedure may be still appropriate provided that the epigenetic mark is stable and has been established during developmental stages. Nevertheless, for certain disease classes,

blood-derived DNA is highly relevant due to the direct involvement of immune cells, as in the case of autoimmune diseases or liquid cancers. However, even in those latter cases, tissue heterogeneity may still represent an obstacle because (a) a specific CpG methylation may be altered only in one cell subtype, (b) the cellular composition may differ between cases and controls, and (c) the differentially methylated positions (DMPs) may be altered by the disease instead of being a direct cause. Notwithstanding the above cautions, it has been suggested that genotype–epigenotype relations exist and may contribute to the disease pathogenesis, thereby mediating the genetic risk or modulating the penetrance [56,57]. At one extreme, imprinted genes represent a straightforward example of an epigenetic mediation of the disease-predisposing variants.

One recent example where such an analysis has been performed is the identification of genotype-specific DNA methylation patterns, involved the integration of GWAS data and DNA methylation profiling from patients suffering from RA [58]. The study attempted to mitigate the gap of missed heritability in RA by identifying genotype-dependent methylated loci that represented a potential mediator of the genetic risk for this autoimmune disease. The analysis involved multiple correction and filtering stages to account for the different cellular composition in cases versus controls [59] and to filter-in only DMPs being genotype dependent thus being candidates for mediating the genetic risk. The latter was accomplished using a causal inference test [60] that was previously shown to compare favorably with Bayesian network reconstruction. Interestingly, this study utilized a special case of an inference or reverse-engineering driven approach to discover that DNA methylation, as an epigenetic process, could mediate a gene–environmental interaction. As the authors pointed out, however, the strategy and study design resulted in a list of potential mediators of the genetic risk in RA, although causal relationship cannot conclusively be obtained from case–control studies alone. However, if we do not require causality, and combine inference methods with those that use available databases it is clear that we can find a rich biological network amenable to further detailed analysis. Specifically, given a list of gene–gene interactions or gene–environmental interaction, we can investigate the network and extract putative paths in the graph which may serve as a biological mediator explaining the observed interactions. At least at one end of this complexity we can be sure about causality, since genetic markers in germline DNA are heritable and not a function of epigenetic or environmental influence.

10.4 INFERRING GENETIC INTERACTIONS OR EDGES FROM DATA IS A SPECIAL CASE OF A FEATURE SELECTION PROBLEM

Prior to the challenge of understanding and interpreting genetic interactions using biological networks we are faced with the problem of how to detect genetic interactions or epistasis from experimental data. The current book

summarizes state-of-the-art methods of how to actually statistically detect such interactions. The previous section discussed the opportunities and challenges in interpreting genetic interactions in terms of the cellular circuitry, which in part is due to our partial understanding of how a cell is working as a dynamic entity in space and time. Importantly, the task of detecting genetic interactions is a difficult statistical problem thus rendering few candidate interactions to be interpreted using a network approach. It would be useful if we could empower the detection of interactions and thereby use a large list of "pairs" which then could be probed and interpreted in terms of the underlying biological networks. Such an approach may even give us a more robust readout in terms of putative subnetworks mediating relations between genes. Therefore, in this section we discuss why the detection problem is so difficult from a mathematical perspective and given this analysis we conclude that we need useful priors in order to empower the statistics for detecting genetic interactions or gene−environmental interactions. Our analysis will then bring us back to network biology while being equipped with the idea of using such networks not primarily as vehicles for interpretation of interactions but as providing statistical priors for detecting genetic interactions.

A key challenge in statistics, data mining, and machine learning is the problem of how to select variables or features that are collectively the most informative for an outcome of interest. This is known as the variable selection problem. Here we consider the discovery of genetic interactions—that we defined as edges—as a special case of feature selection. We should define separately how strongly this definition aligns with different approaches to interaction that have been discussed in the methodological part of this book. Variable selection allows predictions based on a minimal number of measurements and simplifies construction of predictive models based on the selected variables and the features provide insight as to the quantities that are involved in predicting a genetic interaction. Feature selection is particularly difficult when searching for high-order statistical patterns such as in the case of genetic interactions since the problem is high dimensional due to the large number of possible relations. Of note is that several features could be informative for the outcome and important features (genetic variants) that are not informative individually may be informative in the context of other genetic variants, i.e., genetic interactions. This characteristic makes the problem of selecting the smallest, most-informative subset of variables computationally hard. Currently, there exist hundreds or thousands of variable selection algorithms [61]. However, most of them cannot scale up to the number of putative edges that we encounter here.

We have recently performed a detailed mathematical treatment of the variable selection problem [36,62,63]. In brief, our analysis settled an over 30-year-old consensus in the field since the classical result from 1977 by Cover and van Campenhout on the intractability of this feature selection problem. The consensus belief was that it was necessary to perform an exhaustive search of all possible combinations in order to enumerate all

relevant features for a given outcome, thus rendering the problem NP hard. Our key insight was to recast this problem into a statistical machine learning problem instead of working in a deterministic setting. In some detail we developed a statistical framework, which allows clear definitions of different types of feature sets and thereby enables us to define a rigorous separation between finding the minimal set of features for the prediction of the target T versus finding all the features that are relevant for T. This formulation made it possible to prove that for any strictly positive distribution a feature is strongly relevant if and only if it is in the Markov boundary of the target variable. This result gives a polytime algorithmic complexity for estimating the posterior. This allows us to prove that every Bayes-relevant feature is strongly relevant. However we can also prove that the opposite is false since there exist strictly positive distributions where even strongly relevant features are not relevant to the Bayes classifier. Hence, we have a mathematical basis for feature selection, which we have applied to problems in discovering features from transcriptomics data [64,65]. These techniques can also be used in the context detecting of interactions. This is a significant and central theoretical result setting the stage for progress. However, it is clear that despite the successful mathematical analysis there is an urgent need to construct suitable priors, in order to perform rigorous feature selection. Hence, to apply this reasoning to the challenge of detecting interactions we need to use rich molecular data and a network biology approach to inform mathematical algorithms for feature selection.

10.5 NETWORK BIOLOGY—A FRAMEWORK FOR DETECTING GENETIC INTERACTIONS

How to incorporate prior knowledge in a general statistical framework is still an unsolved problem. For most parts of data analysis, prior knowledge is difficult to include in a principled manner. As a rule, learning methods, both supervised and unsupervised, therefore commonly ignore prior knowledge. Typically, any prior knowledge is incorporated *ad hoc* by the human analyst in the form of selecting a suitable learning method, or a suitable version of the method. For example, an expert analyst will select an appropriate kernel when using a support vector machine or the appropriate distance function when selecting a K-Means clustering algorithm. Bayesian statisticians may claim that prior knowledge can be incorporated by an appropriate selection of priors; however, this is more easily said than done as there is no general way that can determine the prior function for many types of useful prior knowledge. Yet, it is not obvious how this type of knowledge can be incorporated in a support vector machine for example. To incorporate this knowledge in a search-n-score algorithm for Bayesian Networks, one would have to dictate higher priors to all structures where a path creating an association exists between two variables known to have association by some other study. However, how to specify such a function in a practical manner is currently unknown. Hence, in a machine learning setting it is more common to use prior

knowledge to determine which method to use or to select a subset of the output from any given method rather than directly incorporating prior knowledge in the inference procedure itself. Yet, it remains unclear how to incorporate in analysis vast public knowledge, such as the 1500 curated molecular databases [66], in a principled manner which sheds light on the problem of detecting interaction, be it purely genetic or gene−environmental interactions. Hence, given this situation we may inspect individual data types and try to assess if they could constrain or simplify the problem of identifying interactions from data. Alternatively, cellular networks could be used to reduce the number of tested interactions. Let us illustrate the first idea by asking whether chromosome−chromosome maps could suggest interactions within the genome.

A common textbook illustration captures the genome as a linear sequence of nucleotides. However, the genome is indeed a 3D structure. By the profiling of histone marks using ChIP-Seq experiments, information regarding the histone modifications and their well-known association to chromatin organization can be obtained and they show how the chromatin is open or closed but unfortunately do not provide information regarding long-range *cis* or *trans* interactions. Enhancers may be open or closed but no clear rules have yet been identified which can associate enhancers to genes, and it has been shown that the basic rule-of-thumb associating the enhancer to the closest gene is not correct most of the time. GWAS studies, where SNPs may be identified at the enhancers, would benefit from such mapping, as they would allow a deeper mechanistic understanding of a disease. One important technology is the analysis of chromatin conformation (CC) data, which aims to capture chromosome−chromosome and intra-chromosomal interactions in 3D space. Initially experimental identification of CC was performed on specific loci using the chromosome conformation capture (3C) technique, which uses spatially constrained ligation followed by locus-specific polymerase chain reaction [67]. Extensions of 3C were then developed to account for quantifying the contacts of one locus versus the entire genome [68]. Those techniques were limited to uncover contacts of one predetermined genomic regions with all potential interactions. ChIA-PET combines ChIP-based methods and 3C to find genome-wide interactions regulated by selected TFs. Recently the HighC method [69] was designed to probe interactions caught by incorporating biotin into the ends of the digested DNA before ligation, and then carrying out physical selection of these fragments. The Hi-C technology has been designed to enable the detection of all pairwise physical associations of DNA in the genome and provide quantification of contact probabilities between loci. Hence, such data could strongly suggest paths in the genome by which interactions could occur. Here we would like to remark that such a conceptual approach, illustrated by the example of Hi-C technology, assumes a physical interpretation of the notion of interaction, which is not necessarily justified since it could simply be a statistical correlation without a straightforward physical correspondence or it may represent a physical link without an interaction effect. Moreover, the analysis of the data provided by Hi-C requires that the

biases in the data must be measured, since the experimental procedures have inherent biases and experimental artifacts [70]. Several specific examples of functional interaction studies are described in Chapter 8 of this book.

As an alternative to identifying a single data type, which could serve as a direct mediator of interaction, we may ask whether cellular networks could be used to reduce the number of tests. The core idea would be to group several genetic variants on the basis of their positions in a cellular network. This would substantially reduce the required number of hypotheses (interactions) to be tested, thus reducing the multiple testing and thereby improving the statistical power. Here we may ask which networks, and how to group the genetic variants? Pathway information, from public databases, could be used to only test a representative genetic variant against other representative genetic variants. Other kinds of cellular networks, either originating from bioinformatics integration or computationally inferred, as discussed earlier in this chapter, could be used in a similar manner as illustrated schematically in Figure 10.2. A more straightforward approach for genome-wide interaction studies is described elsewhere in this book. Hence, a systems biology approach integrating networks into the problem of detecting interactions promises to empower its discovery. However, we are still in the infancy of developing tools enabling such a discovery, and we are far away from

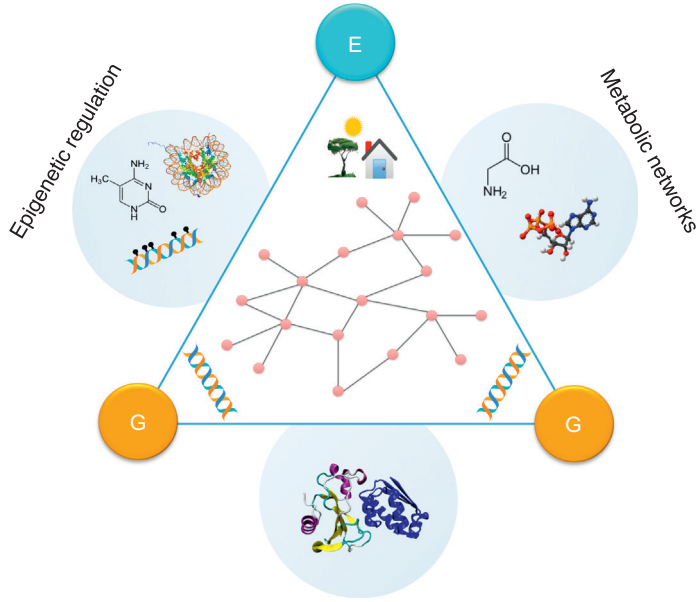

FIGURE 10.2 Schematic illustration of how putative gene—gene interactions or gene—environmental interactions could be mediated through the different types of molecular networks.

making use of this information at a personalized level in accordance with a P4 vision of medicine. Conceptually, the circle of analysis is then closed. The regulatory networks, which actually encode dependencies among genes, thus representing interactions, should definitely be exploited to efficiently detect this important feature of the genome [71].

ACKNOWLEDGMENTS

Our research is supported by Swedish Research Council (Tegnér), Swedish Research Council, CERIC (Tegnér), Swedish Research Council, SerC (Tegnér), Torsten Söderberg Foundation (Tegnér), FP7 SYNERGY-COPD (Tegnér, Gomez-Cabrero), FP7 STATegra (Tegnér, Gomez-Cabrero), and Stockholm County Council (Tegnér).

REFERENCES

[1] Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci USA 2013.
[2] Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucl Acids Res 2011;39(database issue):D19–21.
[3] Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nat Rev Genet 2010;11(9):647–57.
[4] Trelles O, Prins P, Snir M, Jansen RC. Big data, but are we ready? Nat Rev Genet 2011;12(3):224.
[5] Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. N Biotechnol 2012;29(6):613–24.
[6] Clermont G, Auffray C, Moreau Y, Rocke DM, Dalevi D, Dubhashi D, et al. Bridging the gap between systems biology and medicine. Genome Med 2009;1(9):88.
[7] Tegner JN, Compte A, Auffray C, An G, Cedersund G, Clermont G, et al. Computational disease modeling—fact or fiction? BMC Syst Biol 2009;3:56.
[8] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461(7265):747–53.
[9] Cordell HJ. Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet 2009;10(6):392–404.
[10] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5(2):101–13.
[11] Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011;12(1):56–68.
[12] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57–74.
[13] Quackenbush J. From "omes to biology". Anim Genet 2006;37(Suppl. 1):48–56.
[14] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5(7):621–8.
[15] Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet 2012;13(9):667–72.
[16] Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina Human Methylation 450 BeadChip platform. Epigenetics 2013;8:3.
[17] Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegnér J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. Bioinformatics 2013;29(2):189–96.

[18] Galperin MY, Fernandez-Suarez XM. The 2012 nucleic acids research database issue and the online molecular biology database collection. Nucl Acids Res 2012;40(database issue):D1–8.
[19] Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129(4):823–37.
[20] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 2008;40(7):897–903.
[21] Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci USA 2010;107(7):2926–31.
[22] Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, et al. Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. Proc Natl Acad Sci USA 2005;102(9):3336–41.
[23] Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 2008;454(7205):766–70.
[24] Chen PY, Feng S, Joo JW, Jacobsen SE, Pellegrini M. A comparative analysis of DNA methylation across human embryonic stem cell lines. Genome Biol 2011;12(7):R62.
[25] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012;13(7):484–92.
[26] Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci USA 2002;99(6):3740–5.
[27] Illingworth RS, Bird AP. CpG islands—"a rough guide". FEBS Lett 2009;583(11):1713–20.
[28] Doi A, Park I-H, Wen B, Murakami P, Aryee MJ, Irizarry R, et al. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat Genet 2009;41(12):1350–3.
[29] Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 2010;466(7303):253–7.
[30] Feber A, Wilson GA, Zhang L, Presneau N, Idowu B, Down TA, et al. Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors. Genome Res 2011;21(4):515–24.
[31] Song CX, Szulwach KE, Fu Y, Dai Q, Yi C, Li X, et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol 2011;29(1):68–72.
[32] Rideout 3rd WM, Coetzee GA, Olumi AF, Jones PA. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science 1990;249(4974):1288–90.
[33] Markopoulou S, Nikolaidis G, Liloglou T. DNA methylation biomarkers in biological fluids for early detection of respiratory tract cancer. Clin Chem Lab Med 2012;50(10):1723–31.
[34] Scher MB, Elbaum MB, Mogilevkin Y, Hilbert DW, Mydlo JH, Sidi AA, et al. Detecting DNA methylation of the BCL2, CDKN2A and NID2 genes in urine using a nested methylation specific polymerase chain reaction assay to predict bladder cancer. J Urol 2012.
[35] Heichman KA, Warren JD. DNA methylation biomarkers and their utility for solid cancer diagnostics. Clin Chem Lab Med 2012;50(10):1707–21.
[36] Nakano K, Whitaker JW, Boyle DL, Wang W, Firestein GS. DNA methylome signature in rheumatoid arthritis. Ann Rheum Dis 2012.
[37] Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Systemic steroid exposure is associated with differential methylation in chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2012.

[38] Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, Miller NA, et al. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature 2010;464(7293):1351–6.
[39] Johnson AA, Akman K, Calimport SR, Wuttke D, Stolzing A, de Magalhaes JP. The role of DNA methylation in aging, rejuvenation, and age-related disease. Rejuvenation Res 2012;15(5):483–94.
[40] Feinberg AP, Irizarry RA, Fradin D, Aryee MJ, Murakami P, Aspelund T, et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. Sci Transl Med 2010;2(49):49ra67.
[41] Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. Nature 2007;450(7172):973–82.
[42] Alon U. Network motifs: theory and experimental approaches. Nat Rev Genet 2007; 8(6):450–61.
[43] Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. Cell 2011;144(6):986–98.
[44] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature 2000;407(6804):651–4.
[45] Lundstrom J, Bjorkegren J, Tegner J. Evidence of highly regulated genes (in-Hubs) in gene networks of Saccharomyces cerevisiae. Bioinform Biol Insights 2008;2:307–16.
[46] Barabasi AL. Network medicine—from obesity to the "diseasome". N Engl J Med 2007;357(4):404–7.
[47] Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci USA 2007;104(21):8685–90.
[48] Wang LY, Wang P, Li MJ, Qin J, Wang X, Zhang MQ, et al. EpiRegNet: constructing epigenetic regulatory network from high throughput gene expression data for humans. Epigenetics 2011;6(12):1505–12.
[49] Ciofani M, Madar A, Galan C, Sellars M, Mace K, Pauli F, et al. A validated regulatory network for Th17 cell specification. Cell 2012;151(2):289–303.
[50] Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, Awasthi A, et al. Dynamic regulatory network controlling T17 cell differentiation. Nature 2013.
[51] Gustafsson M, Hörnquist M, Lundström J, Björkegren J, Tegnér J. Reverse engineering of gene networks with LASSO and nonlinear basis functions. Ann NY Acad Sci 2009;1158:265–75.
[52] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinf 2006;7(Suppl. 1):S7.
[53] Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods 2012;9(8):796–804.
[54] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2007;2(10):2366–82.
[55] Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. Nucl Acids Res 2010;38(Web Server issue):W96–102.
[56] Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet 2011;12(8):529–41.
[57] Bjornsson HT, Fallin MD, Feinberg AP. An integrated epigenetic and genetic approach to common human disease. Trends Genet 2004;20(8):350–8.
[58] Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol 2013;31(2):142–7.
[59] Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinf 2012;13:86.
[60] Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. BMC Genet 2009;10:23.

[61] Ben-Hur A, Guyon I. Detecting stable clusters using principal component analysis. Methods Mol Biol 2003;224:159–82.
[62] Nilsson R, Björkegren JP, Tegnér J. Consistent feature selection for pattern recognition in polynomial time. J Mach Learn Res 2007;8:589–612.
[63] Pena J, Nilsson R, Björkegren J, Tegnér J. An algorithm for reading dependencies from the minimal undirected independence map of a graphoid that satisfies weak transitivity. J Mach Learn Res 2009;10:1071–94.
[64] Nilsson R, Peña JM, Björkegren J, Tegnér J. Detecting multivariate differentially expressed genes. BMC Bioinf 2007;8:150.
[65] Nilsson R, Bjorkegren J, Tegner J. On reliable discovery of molecular signatures. BMC Bioinf 2009;10:38.
[66] Fernandez-Suarez XM, Galperin MY. The 2013 nucleic acids research database issue and the online molecular biology database collection. Nucl Acids Res 2013;41 (Database issue):D1–7.
[67] Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science 2002;295(5558):1306–11.
[68] Simonis M, Klous P, Homminga I, Galjaard R-J, Rijkers E-J, Grosveld F, et al. High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. Nat Methods 2009;6(11):837–42.
[69] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326(5950):289–93.
[70] Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 2011;43(11):1059–65.
[71] Moore JH. A global view of epistasis. Nat Genet 2005;37(1):13–4.