10 Network Medicine

*Pisanu Buphamalai**, *Michael Caldera**, *Felix Müller, and Jörg Menche*

10.1 Introduction

Since the publication of the first draft of the human genome less than two decades ago [1, 2], rapid technological progress has revolutionized biomedical research. Thanks to a diverse array of "omics" technologies (e.g., genome sequencing, transcriptome mapping, proteomics, metabolomics, and others), we can now quantify both healthy and disease states at molecular resolution. At the same time, it has become clear that the detailed characterization of the individual molecular components alone (genes, proteins, metabolites, etc.) does not suffice to truly understand the nature of (patho-) physiological states and how to modulate them. Indeed, biomolecules do not act in isolation, but within an intricate and tightly coordinated machinery of complex interactions, such as protein–protein, gene regulatory, or signaling interactions. Network medicine is an emerging field that aims to apply tools and concepts from network theory to elucidate this machinery. Network approaches have helped unravel the molecular mechanisms of a broad range of diseases, from rare Mendelian disorders [3], cancer [4] or metabolic diseases [5], to basic attack strategies of viruses [6], to name but a few examples. While the molecular networks that underly biological processes may be the most natural candidate for applying network concepts in biomedical research, they are certainly not the only one. Networks are used across the full spectrum of medicine, from biomarker [7] to drug discovery [8], from the spread of obesity [9] to global outbreaks of infectious diseases [10], and from characterizing the relationships among diseases [11] to those among physicians within the health care system [12].

This chapter aims to give a general introduction to the dynamic field of network medicine. We start with a broad overview of major network types that are relevant to medicine. We then discuss with more detail the cellular network of molecular interactions among proteins and other biomolecules, the perhaps most widely used network in biomedical research. In the last section, we introduce disease module analysis, an important application of network tools to elucidate the molecular mechanisms of a particular disease.

*equal contribution

10.2 Networks in Medicine

10.2.1 Overview

One can distinguish three basic network types that cover different disease-relevant relationships: (*i*) Molecular networks describing the relationships between the molecular constituents of living organisms, for example, maps of all protein–protein interactions or metabolic reactions in a cell. The observation that such molecular maps share certain universal topological features with vastly different systems, e.g., the World Wide Web, collaboration networks, power grids, and many others, was instrumental for the development of network science. Today, it seems almost trivial that networks provide the most natural way of describing and analyzing the large-scale organization of biomolecules and their interactions. (*ii*) Disease networks are a powerful tool to investigate the diverse relationships between diseases. For example, two diseases can be linked if they share genetic associations or if they have similar clinical manifestations. In contrast to molecular networks, in which links often represent direct physical interactions, disease–disease networks represent more abstract relationships. They therefore serve as beautiful examples for the power of networks as a general tool for the analysis, integration, and intuitive visualization of large and complex data. (*iii*) Population-scale networks, i.e., networks describing the complex interactions among humans have been very successful in modeling and predicting the spread of contagious diseases, for example, global swine flu or ebola pandemics. These studies show the enormous potential of networks to serve as a platform for translating exact analytical results from physics and mathematics and translating them to concrete applications in medicine. (See Box 10.1.)

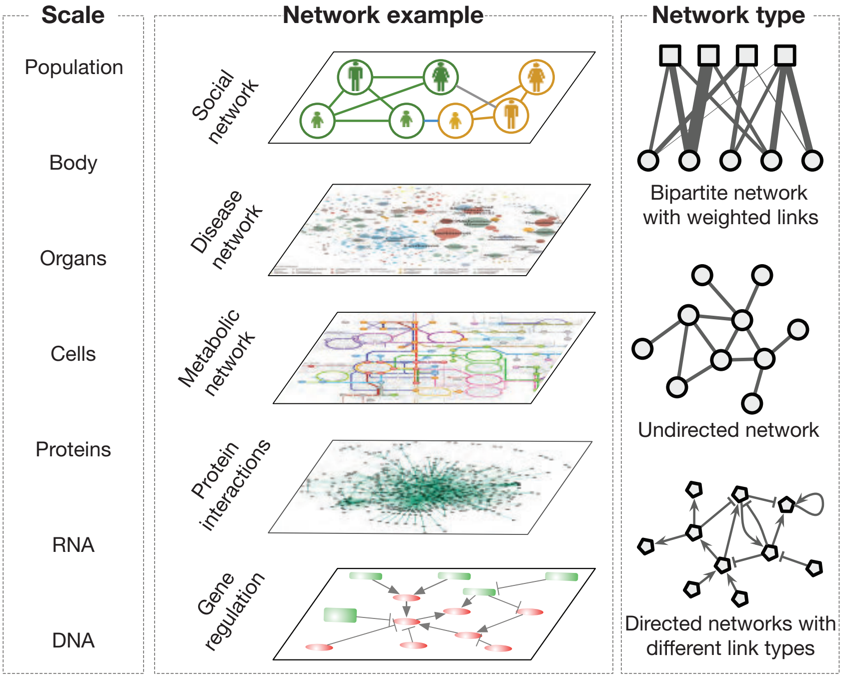
10.2.2 Molecular Networks

There are a plethora of molecular networks describing different aspects of the molecular and cellular organization of living organisms. A broad distinction can be made between physical and functional interaction networks. Physical interactions involve actual physical contact between the participating biomolecules, for example, proteins that assemble in a complex or receptor–ligand binding. Functional interaction, on the other hand, can refer to any kind of biologically relevant relationship. In co-expression networks, for example, genes are connected if their expression patterns are strongly correlated [13]. In the following we introduce the main types of molecular networks that are used to elucidate diverse disease mechanisms. Some of them were introduced in previous chapters, but we also summarize them here for completeness.

10.2.2.1 Protein–Protein Interaction Networks

Many molecular processes within a cell are performed by molecular machines consisting of a large number of protein components organized by their protein–protein interactions (PPIs). PPIs result from biochemical events steered by electrostatic forces leading to physical contacts of high specificity between two or more proteins [14]. Perturbed PPIs are involved in the pathobiology of many diseases, ranging from diabetes and obesity to Crohn's disease or cancer [15]. In analogy to the "genome" representing

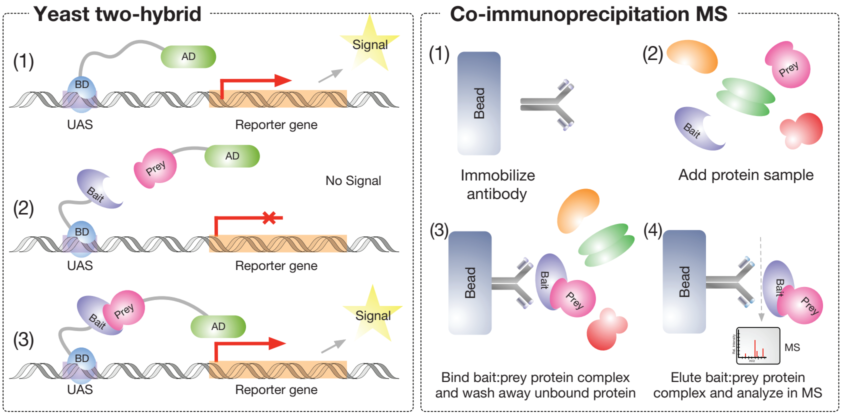
Box 10.1: Networks in medicine



The diverse networks that are studied in network medicine reflect the different levels of organization that are relevant to human disease. From the molecular level, e.g., networks of interacting biomolecules that form the basis of all cellular processes, to the level of social interactions that are involved in the transmission of infectious diseases. Depending on the particular system, different network types are used for their description. Undirected and unweighted networks represent the most basic network type. More complex types may include a link directionality, link weights or use different types of nodes, for example in bipartite networks.

the collection of all genes in an organism, the collection of all molecular interactions is often referred to as the "interactome." The interactome can be represented by a network in which the nodes are proteins and the edges correspond to physical interaction between them. Over the last decade, significant experimental efforts have been made to map out the complete human interactome. High-throughput techniques such as yeast two-hybrid (Y2H) and immunoprecipitation linked to mass spectrometry are capable of mapping thousands of interactions in parallel (see Box 10.2). There has also been substantial work in curating interactions that were identified in small-scale experiments, as well as using computational tools to predict interactions [15].

Box 10.2: Mapping the human interactome



There are two major high-throughput techniques for the identification of protein interactions:

Yeast two-hybrid: (1) the system uses a protein consisting of a DNA binding domain (BD) and an activation domain (AD) that is responsible for activating transcription of DNA. (2) In Y2H, the two domains are separated and fused to proteins whose interaction is investigated. The BD is fused to the so-called bait, the AD to the prey. (3) Upon interaction between the two proteins of interest, the AD comes in close proximity to the reporter gene and the transcription leads to a signal.

Co-immunoprecipitation coupled to mass spectrometry: (1) In a first step, a target (bait) protein-specific antibody is immobilized on beads (e.g., agarose). (2) When the cell lysate is added, the antibody will specifically bind the target protein and indirectly capture proteins (prey) that are capable of binding to it. (3) After washing away unbound proteins, (4) the proteins of interest are eluted and analyzed using mass spectrometry. In short, the sample (the proteins) is first ionized and fragmented into smaller molecules, e.g., amino acids and peptides. Their mass-to-charge ratios can then be determined by accelerating the ions and subjecting them to an electric and/or magnetic field. Finally, the proteins in the sample can be identified by comparing with databases of known masses and characteristic fragmentation patterns.

Despite these promising first steps, our knowledge of the human interactome map remains far from complete, estimates indicate that only 10–30% of the full interactome has been revealed currently [16]. Nevertheless, interactome-based studies have contributed substantially to our understanding of biological processes both in homeostasis and in disease states, see Section 10.3.

10.2.2.2 *Metabolic Networks*

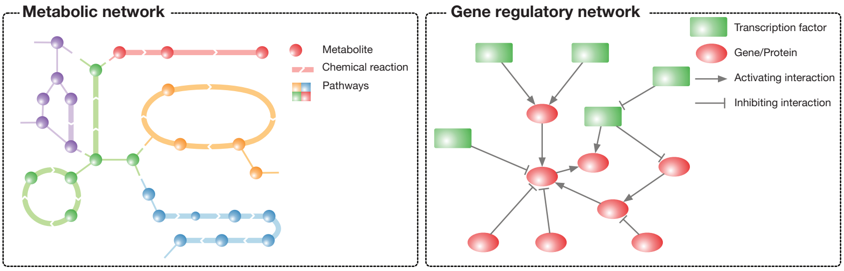
Metabolism (from Greek μεταβολή for "change") refers to the sum of all processes that are involved in assembling and disassembling the basic building blocks of cells, in particular the biochemical reactions for energy conversion. Traditionally, these reactions have been organized into specific pathways, for example the tricarboxylic acid (TCA) cycle, which corresponds to the sequence of chemical reactions in the cell that produces energy (also known as citric acid – or Krebs cycle, named after Hans Krebs, a Nobel Laureate in 1953). Metabolic networks represent collections of such pathways that connect chemical compounds (metabolites), biochemical reactions, enzymes, and genes. The relationships between the individual components of a given metabolic system can be inferred using comparative genomics combined with metabolomic data [17]. Metabolic networks are the most complete among the different biological networks, i.e., they reflect a near exhaustive knowledge of the involved biochemical processes [18]. They are available for a wide range of species and can be accessed through databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [19] or Reactome [20]. The currently most comprehensive human metabolic network, Recon 2.2 [21], includes 5,324 metabolites, 7,785 reactions, and 1,675 associated genes. Such metabolic networks do not only offer deep insights into the basic machinery of cells, but can also be used for *in silico* simulations to study how different parameters (e.g., metabolite concentrations) affect local and global properties of the biochemical network. The two most commonly used methods employ either (1) deterministic approaches (e.g., systems of ordinary differential equations) or (2) stochastic models (e.g., effect probabilities upon network perturbation) [22]. Metabolic network analyses can yield profound insights into the evolutionary emergence of complex life forms [23, 24], help understand the molecular mechanisms that drive the response to vaccination [25], or elucidate the interplay between metabolism and gene regulation [26]. (See Box 10.3.)

10.2.2.3 *Regulatory Networks*

Regulatory networks describe the complex machinery of genes and their corresponding proteins and RNAs, as well as the interactions between them that control the level of gene expression across the genome under specific conditions. Of particular importance for expression regulation are transcription factors (TFs), i.e., DNA-binding proteins that modulate the first step in gene expression [27]. In the most common representation of regulatory networks, nodes correspond to genes and links to the regulation of the expression of one gene by the product of the other. The links are typically directed and have either an activating (i.e., an increase in the concentration of one leads to an increase in the expression of the other) or inhibitory effect (increase in the concentration of one leads to decrease in the other) [28, 29]. Several experimental techniques exist to create large-scale data for building genome-wide regulatory networks, such as Chromatin-Immunoprecipitation Chip (ChIP-on-chip) [30] and ChIP-Sequencing [31]. Comprehensive databases include the Universal Protein Binding Microarray Resource for Oligonucleotide Binding Evaluation (UniPROBE) [32] or JASPAR [33].

Gene regulatory networks provide powerful tools to identify key transcription factors that control cell fate, for example in early blood development [34, 35].

Box 10.3: Metabolic and regulatory networks



Metabolic networks describe the conversion/transformation of chemicals (metabolites) within a cell, organ, or whole organism. The nodes represent specific molecules while the edges describe the chemical reactions that take place between the nodes. Often these reactions are catalyzed by enzymes. Specific routes/compartments that are known to perform a particular function are called pathways.

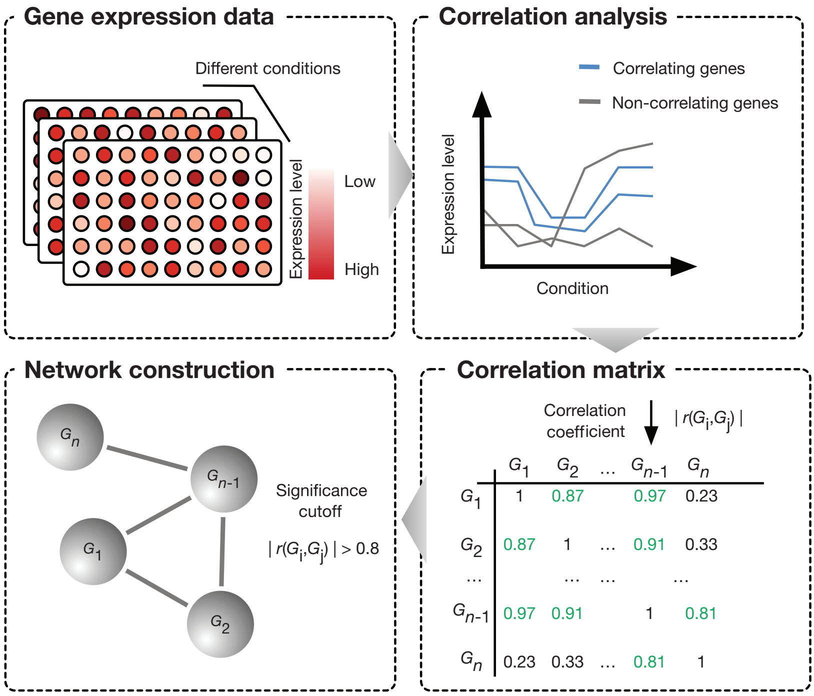
Gene regulatory networks consist of genes that regulate each other. Often these genes are transcription factors that are capable of binding to DNA. The type of interactions can be either positive leading to an increase of protein concentration of the regulated gene, or negative, which leads to a decrease in protein concentration.

They can also be used to interpret variants identified in genome-wide association studies (GWAS), as they often perturb regulatory modules that are highly specific to disease-relevant cell types or tissues [36]. Lastly, gene regulatory networks also shed light on evolutionary conditions and pathways by which new regulatory functions emerge [37]. (See Box 10.3.)

10.2.2.4 Co-Expression Networks

In co-expression networks, genes are linked if their expression levels are significantly correlated under different experimental conditions, for example over time, across different tissues or cell types, or across a patient population (see Box 10.4 for an overview of the construction process) [13, 39]. In contrast to regulatory networks, co-expression networks do not offer an immediate causal relationship between genes. They can be used, however, to identify groups of genes that are more broadly functionally related, for example, controlled by the same transcriptional regulatory program, or members of the same pathway or protein complex [40]. Network analyses have been used to identify commonly affected pathways in heterogeneous diseases like autism spectrum disorder [41] or inflammatory bowel disease [42], predict causal GWAS genes associated with bone mineral density [43], or help explain the mechanism of breast cancer development [44].

Box 10.4: Co-expression networks



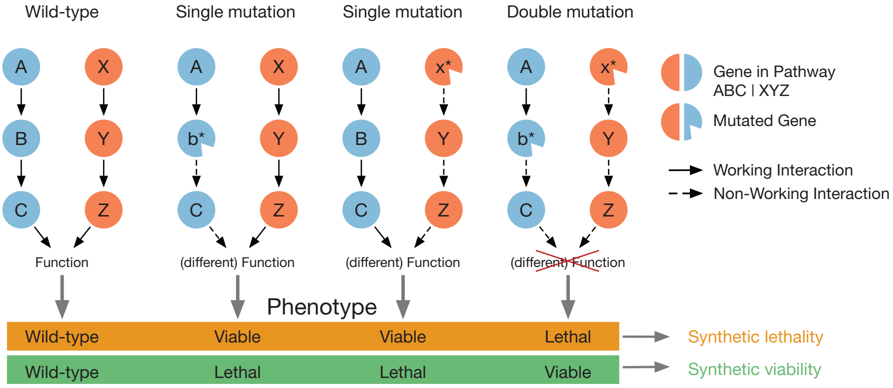
	G_1	G_2	…	G_{n-1}	G_n
G_1	1	0.87		0.97	0.23
G_2	0.87	1	…	0.91	0.33
…		…	…		…
G_{n-1}	0.97	0.91		1	0.81
G_n	0.23	0.33	…	0.81	1

Construction of a co-expression network: Creating a co-expression network requires gene expression data over several conditions, for example different treatments, across several tissues or patients. For each gene pair one can then calculate a correlation coefficient for their respective expression values across the different conditions, resulting in a correlation matrix. Extracting biologically meaningful correlations can be quite challenging, as true signals are often masked by noise that can arise, for example, from experimental confounding factors, batch effects, or sample heterogeneity. A widely used alternative to somewhat arbitrary global thresholds preserves the continuous nature of correlation scores and instead applies soft thresholding to identify network subclusters [13]. With recent large-scale resources, such as GTEx [38], noise from sample heterogeneity can be reduced and co-expression networks can be constructed in a tissue-specific manner, thus providing deeper and more robust insights onto the regulatory system in diseases.

10.2.2.5 *Genetic Interactions*

Two genes are linked by a genetic interaction if the effect of a simultaneous alteration (e.g., a mutation or the complete knock-down) of both genes differs from the

Box 10.5: Genetic interactions



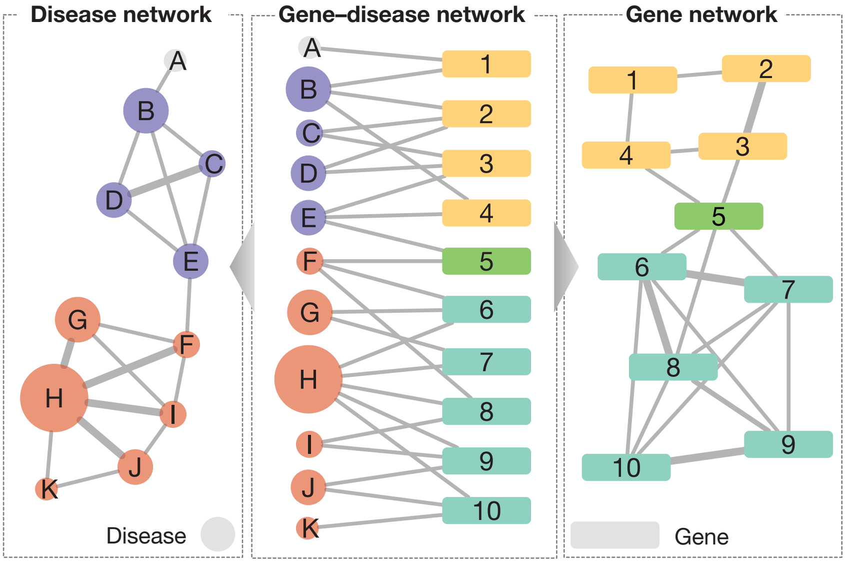
Genetic interactions occur when the phenotype of two combined mutations differs significantly from the expectation based on the individual mutations. These interactions can be either positive (combined effect stronger than expected) or negative (combined effect weaker). The two most extreme outcomes are called “synthetic lethality” and “synthetic viability.” In **synthetic lethality** the two individual mutations often occur in two independent, yet redundant pathways, so that the loss of one can be compensated for by the second. Only when targeting both pathways the systems fails. In **synthetic viability** the mutation in one pathway often leads to a toxic gene product. Only by also affecting another pathway the production of this toxic product is stopped and the resulting phenotype is again viable.

expectation based on the individual alterations [45] (see Box 10.5). The most extreme negative genetic interaction, often called “synthetic lethality,” occurs when the simultaneous mutation of two genes is lethal, while individually both mutations are viable. Conversely, the most extreme positive genetic interaction (“synthetic viability”) occurs, when a combination of two mutations is viable, while both individual mutations are lethal. Genetic interactions imply a functional relationship between the two genes, for example involvement in a common biological process or pathway, or conversely involvement in compensatory pathways with unrelated apparent function [46]. Hence, genetic interactions are an effective tool for biological discovery, e.g., for dissecting signaling pathways. They may also explain a considerable component of undiscovered genetic associations with human diseases and might help identify potential therapeutic targets. Over the last decade, genetic interactions have been investigated using mainly synthetic genetic array technology and RNA interference in yeast and *Caenorhabditis elegans*. A recent yeast based high-throughput screen [47], for example, tested all pairwise combinations of 6, 000 genes resulting in almost 1 million interactions. Such maps can be used to study the large-scale organization of functions in a cell [47], identify the hierarchical organization of specific biological processes [48], or generate hypotheses on the function of uncharacterized genes [49].

10.2.3 Disease Networks

Disease networks are a powerful framework for systematically investigating the diverse relationships among diseases. Such relationships exist on the molecular level (e.g., common genetic origin), on the phenotypic level (e.g., similar clinical manifestations) and on the population level (e.g., frequent co-occurrence in patients). A first comprehensive map of the human "diseaseome" was presented in [11], where 1,377 diseases were linked by shared genetic associations reported in the OMIM database [50] (see Box 10.6). The resulting network showed clearly that diseases can rarely be viewed as isolated quantities, each with a distinct genetic origin, but fall into highly connected clusters of disease groups with overlapping molecular roots. It was also found that diseases that are more central within the disease network tend to be more prevalent and have higher mortality rates [51]. The genetic overlap

Box 10.6: Disease networks



Disease networks in which diseases are linked if they share a genetic association are based on gene–disease association data that can be represented as a bipartite network (middle panel). This bipartite network can then be projected either onto the diseases, resulting in a disease–disease network (left) or onto a gene–gene network, in which links represent a common disease association.

among diseases also extends towards physical interactions among the respective gene products, as well as similar gene expression profiles.

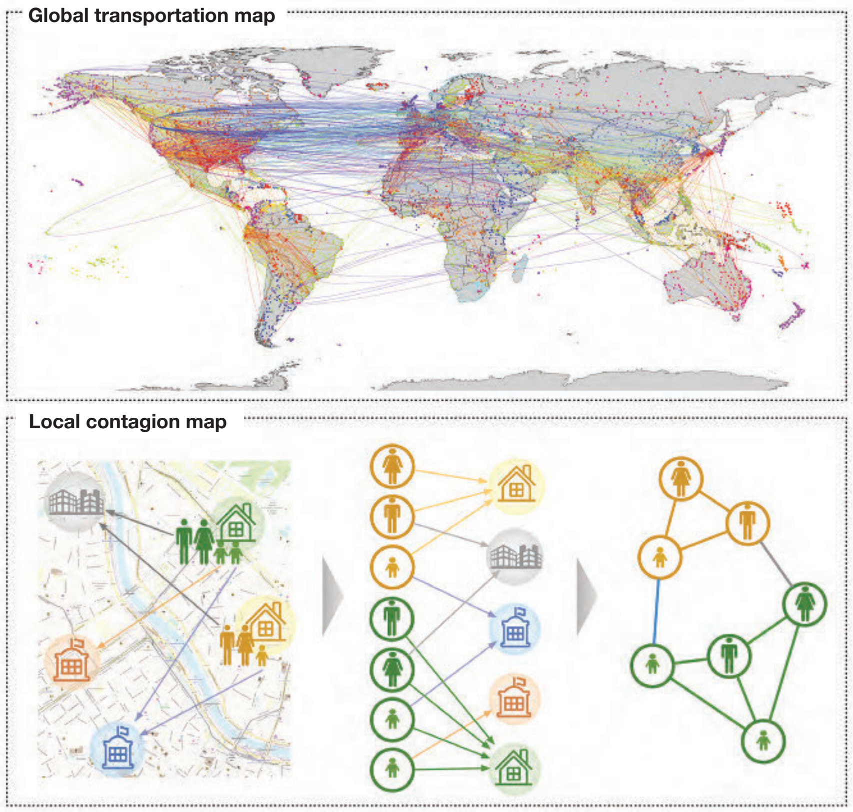
Similar results were obtained in a disease network in which diseases were linked by the similarity of their clinical manifestations [52] that were extracted from a large-scale screen of the biomedical literature and the annotated Medical Subject Headings (MeSH) metadata [53]. Confirming the strong correlation between the similarity of the symptoms of two diseases, the number of shared genetic associations and the extent to which their corresponding proteins interact, the study further revealed that the diversity of the clinical manifestations of a disease can be related to the degree of localization of the associated genes on the underlying protein interaction network. More detailed analyses that compared disease networks of different disease classes (e.g., complex diseases, Mendelian diseases, or cancer) and protein interaction networks identified interesting differences between diseases with different inheritance modes [54, 55, 56].

Networks can also be used to study comorbidity, i.e., the tendency of certain diseases to co-occur in the same patient. A disease network extracted from over 30 million patient records revealed that disease progression patterns of individual patients can be related to topological properties of the respective diseases within the co-morbidity network, for example, peripheral diseases tend to precede more central diseases [57]. These central, highly connected diseases are in turn associated with a higher mortality rate. More recently, differences in disease progression patterns that are related to age and sex have been characterized [58]. Co-morbidity networks have been used to address a wide range of further biomedical challenges, from drug repurposing [59] to the identification of potential drug side-effects [60], from biomarker identification [61] to approaches how to disentangle genetic and environmental factors of diseases [62].

10.2.4 Social Networks

A third important application of networks in medicine addresses the spread of contagious diseases, such as viral or bacterial infections (Box 10.7). Mathematical models of disease spreading go back as far as the year 1760, when Daniel Bernoulli formulated the first analytical method for quantifying the effectivity of inoculation against smallpox [64] (see Box 10.8 for an overview of important epidemiological models). Some 240 years later, the rise of complex networks made it possible to add a key ingredient to such models, namely realistic topologies of the networks on which diseases propagate, in particular global transportation maps and networks of social interactions [63] (see Box 10.7). Detailed information on interactions between humans on a local scale and on worldwide travel patterns is crucial for accurate predictions of the spatio-temporal spread of infectious diseases. Historically, the mobility of humans was largely confined by geography, such as rivers or mountains that could not be crossed easily. Such geographical borders naturally confined the propagation of epidemics. In present day, however, where both humans and goods can easily and quickly travel worldwide via air traffic, not even oceans can limit contagions [10]. As a consequence, an infection that started in a remote rural region may quickly propagate all over the world once

Box 10.7: Networks of disease spread



Global air traffic plays a major role in the spread of epidemic disease across the world. Locally, infectious diseases, but also personal traits like happiness or habits like smoking, are transmitted through social interactions. These interactions can occur, for example at home or at work, which can be represented as a bipartite network that can be mapped to a person-to-person network (illustration adapted from [63]).

it has reached an airport, leading to much faster, much wider, and seemingly more erratic patterns of global epidemics.

10.2.4.1 Transportation Networks

Network-based epidemiological models that incorporate the structure of worldwide transportation networks can shed light on the complicated propagation patterns observed in recent pandemic outbreaks, help identify the source of an outbreak, predict future highly affected areas, or design most effective immunization or prevention strategies [67, 68]. Examples for recent outbreaks of infectious diseases that were studied with the help of network models include the SARS pandemic in 2003 [69], the H1N1 outbreak in 2009 [70], the ebola crisis of 2015, or the spread of HIV in the Philippines [71].

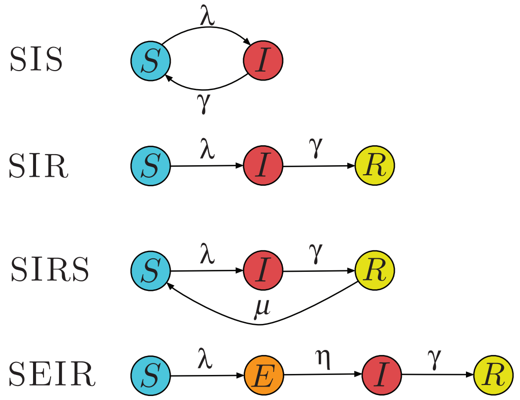
Like many other real world networks, air-traffic networks have been found to be approximately scale free [72]. Scale-free networks are therefore the prime model for analytical studies of epidemic outbreaks and for the analysis of real data from past and current epidemics [73]. Important global properties of a pandemic are directly linked to the structure of the underlying networks. For example, the characteristic (super-) hubs of scale-free networks can often be identified with large airports that play an important role in the spread of a disease, both through the large number of people gathering at such airports and through the large number of destinations that they serve. Indeed, scale-free networks are generally more prone to global infections than more regular network structures that do not exhibit the "small word effect." The critical spreading rate at which an infection is likely to propagate through the entire network is given by the ratio between the average degree and its variance. In large scale-free networks with degree distribution $P(k) \sim k^{-\gamma}$, the variance goes to infinity for power coefficients $\gamma < 3$. The critical spreading drops to zero in this case, meaning that a local infection is likely to become global, even for small infection rates [74].

10.2.4.2 Social Contagion

Approaches used to elucidate large-scale properties of infectious disease outbreaks can also be used to study the dynamics of social interactions, such as the spread of ideas, attitudes, and behaviors [75]. Reflecting the complexity of social relationships, links in social networks may represent, for example, friendship, family relationships, common work-place, shared political preferences, and many more. Collectively, these relationships not only define and shape our social relationships, but may also have concrete medical impact as shown in a seminal work on the spread of obesity [9]: The authors quantified how changes in body-mass index correlated among members of a social network of friends and family. Surprisingly, they found that obesity preferentially spreads through close social relationships. This effect is strong between men and between women, but almost negligible between man and woman. Similar studies were carried out to dissect the social component of starting to smoke [76] or of general happiness in life [77]. The results suggest that people surrounded by many happy people and those who are central in the network are more likely to become happy in the future. This effect was not observed among co-workers [77].

Recently there have also been efforts to combine global disease dynamics of transportation networks with contagion occurring on social networks. Multiplex or multilayer networks provide the analytical platform for combining several networks [78, 79, 80]. In such multilayer networks, different types of contact (at work, in the supermarket, at the airport) can be represented by distinct layers. It has been shown that the epidemic threshold is determined by the largest eigenvalue of the contact probability matrices of the different layers [78]. A powerful tool to study the full dynamics of spreading phenomena on networks, both simple or multilayered, are reaction diffusion processes [81].

Box 10.8: Basic mathematical models of disease spread



Classical epidemic models aim to determine the fraction of a population affected by a contagious disease over time. Most models represent the disease-status of an individual by one of three basic states [65]:

The **susceptible** (S) state, in which an individual can contract a disease. The **infected** (I) state, in which the individual carries the disease and can transmit it. The **recovered** (R) state, in which an individual is immune to repeated infections. More advanced models may also include further states, such as the **exposed** (E) state, in which an individual is already infected, but cannot yet transmit the disease. The microscopic dynamics of epidemiological models is given by transitions between the different states, macroscopic properties emerge from the interaction of many individuals. The most widely studied models are the following:

The **SIS model**, in which the recovery of a disease does not convey immunization, but renders an individual susceptible again, for example the common cold. The dynamics of the system are completely determined by the two rates of infection λ and recovery γ, respectively.

In the **SIR model** [66] susceptible individuals become infected with rate λ and recover with rate γ. This system exhibits an epidemic threshold $\alpha = \frac{\lambda}{\gamma}$, such that for $\alpha \leq 1$ a disease will die out in the long run, whereas for $\alpha > 1$ it will persist in the population.

The **SIRS model** contains an additional temporary immunity state, so that recovered individuals become susceptible again with rate μ. The impact of the incubation periods can be modeled by adding an exposed state (E), in which an individual has been infected, but is not yet infectious.

In network-based generalizations of these models, the individuals are identified with nodes and diseases spread along the connections of the network. In the simplest case this can be done by substituting the infection rate λ with a degree-dependent rate $\lambda = \lambda(k)$, so that the likelihood of becoming infected grows with the number of infected neighbors.

10.3 Interactome Analysis

As we have seen above, there exists a great variety of molecular interaction networks that can yield important insights into disease mechanisms. In the following, we will focus on "interactome networks" containing only physical interactions. The basic tools and concepts apply readily to other types of networks, however.

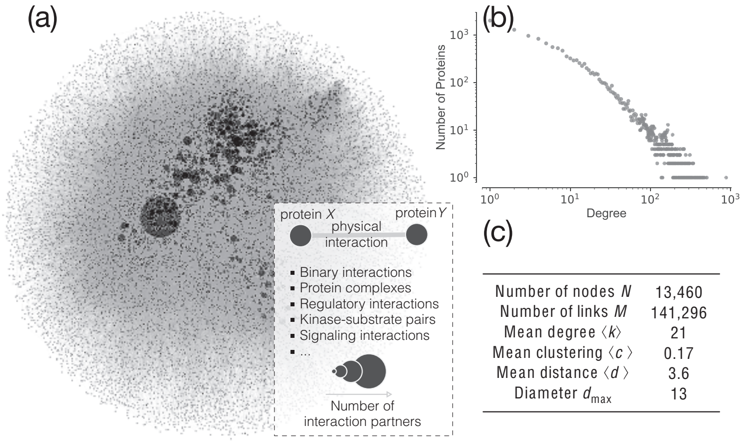
10.3.1 Interactome Construction

A large number of publicly available databases provide comprehensive collections of interactions between proteins and other relevant biomolecules (e.g. protein–DNA, protein–RNA, enzyme–metabolite interactions) in human, but also in other species, see [82] for a compendium of available resources. Among the most comprehensive, actively maintained and widely used databases are STRING [83], BioGRID [84], and MIntACT [85]. Note that they may also contain interactions that are not strictly physical, for example co-expression or other types of functional relationships among genes and their products. A well curated collection of only physical interactions has recently been published in the HIPPIE database [86]. Each interaction in HIPPIE is annotated with the original publication(s), details on the experimental protocol and an aggregated confidentiality score, thus allowing the user to adapt the final interactome network to specific requirements and preferences.

Generally, one can distinguish between three main sources of PPIs: (1) **interactions curated from the scientific literature** and typically derived from small-scale experiments, for example using co-immunoprecipitation, X-ray crystallography, or nuclear magnetic resonance. (2) **Interactions from systematic, proteome-scale mapping efforts**. The two main techniques are yeast two-hybrid (Y2H) assays [87] and binding affinity purifications coupled to mass spectrometry (MS) [88, 89], which produce rather different, yet complementary results (see Box 10.2). Y2H can map out precise, binary protein interactions, yet without biological context. It is not guaranteed, for example, that an experimentally observed interaction is biologically relevant, or whether the two respective proteins are in fact never expressed at the same time in the same cell. Co-complexes observed in MS experiments, on the other hand, are derived from a specific biological sample, yet are more difficult to translate into precise pairwise interactions [14]. (3) **Interactions from computational predictions**, for example based on protein structure [90] or other genomic data [91]. All three sources of PPIs have strengths and limitations in terms of comprehensiveness, noise and biases [92], such as biases in the selection of protein pairs [93] or experimental biases, for example towards highly expressed genes [87].

10.3.2 Basic Interactome Properties

Figure 10.1 gives a visual impression of a manually curated interactome from [16] and summarizes its global topological properties. In total, it contains 13, 460 proteins connected via 141, 296 physical interactions, so on average each protein has about 21 interaction partners. Characteristic not only to this, but also to many other complex



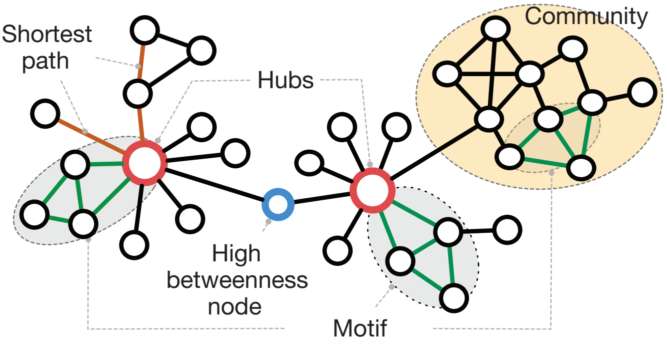
Number of nodes N	13,460
Number of links M	141,296
Mean degree $\langle k \rangle$	21
Mean clustering $\langle c \rangle$	0.17
Mean distance $\langle d \rangle$	3.6
Diameter $d_{\max}$	13

Figure 10.1: (a) A global picture of the interactome (original data curated by [16], figure adapted from [94]). The network consists of 13,460 proteins and 141,296 interactions that have been collected from different sources with various kinds of physical interactions, including binary interactions from systematic yeast two-hybrid screens, protein complexes, kinase-substrate pairs and others. (b) The overall topology is characterized by a highly heterogeneous degree distribution that follows approximately a power-law. (c) Other important structural properties of the interactome.

networks, is the high heterogenity among the degrees of the nodes, i.e., in the number of connections they have to other nodes differs widely (see Box 10.9 for an overview of important terms in network science). While the vast majority of proteins have only few neighbors (more than $2,000$ have only a single link), there is also a considerable number of nodes with hundreds of connections, such as *GRB2* (degree $k = 872$), *YWHAZ* ($k = 502$) and *TP53* ($k = 450$), so-called "hubs." The histogram of all nodes' degrees shows "scale-free" properties,[1] i.e., $P(k)$ follows approximately a power-law $P(k) \sim k^{-\gamma}$. As laid out in more detail in Chapter 3, the broad degree distribution and, as a consequence, the presence of hubs have a profound impact on many network properties. Hubs serve as shortcuts that connect distinct parts of the network, resulting in a network property often referred to as the "small word effect" [96] (in some cases of scale-free networks even "ultra-small" [97]). In the interactome, for example, it takes on average less than four steps ($\langle d \rangle = 3.6$) to reach any other protein from any given starting point. This high degree of connectedness is also associated with a remarkable resilience of the overall network structure against random failure of individual nodes and/or edges. Scale-free networks can maintain global connectedness even upon removal of a considerable fraction of nodes and edges [98, 99, 100, 101]. The flipside of this robustness towards random failure, however, is a particular vulnerability towards targeted attack against the hubs [102]. For the interactome, for example,

[1] How accurately this and other networks can be described by a power-law is subject to some debate, see [95] for a thorough discussion. For our purposes, however, the precise mathematical nature of the degree distribution plays only a secondary role.

Box 10.9: Basic topological characteristics of networks



- The degree of a node is the number its direct neighbors. The degree distribution across all nodes is an important global network characteristic.
- Scale free networks are characterized by a degree distribution that follows a power law: While most nodes have few neighbors, there are also a few highly connected hubs with a large number of neighbors.
- A path between two nodes is a sequence of links connecting the two. The minimum number of links needed to connect the two is called shortest path length and represents their network distance.
- Centrality measures quantify the topological importance of a node within the network. There are different types of centrality measures, the betweenness centrality, for example, quantifies how many shortest paths of the full network cross through a certain node.
- Clustering describes a tendency observed in many biological (and other) networks that two neighbors of a node are often also connected to each other, thus forming a triangle.
- Motifs are small recurrent subgraphs in a network that occur particularly frequently.
- Network communities are groups of tightly interconnected nodes that have more connections among themselves than to the rest of the network.

the removal of $\sim 30\%$ of the most highly connected nodes is sufficient to completely destroy the network, leaving only disconnected fragments.

10.3.3 Interactome Topology and Biological Function

The degree of connectedness of a protein is directly related to its biological importance: As first shown for the yeast *Saccharomyces cerevisiae* [103], and later confirmed also in human cell lines [49], the products of essential genes, i.e., genes that are critical for the survival of an organism, tend to have a high number of interaction partners and take on central positions in the interactome. In contrast, genes whose loss of function can be more easily compensated for tend to have fewer interactions and are situated at the periphery of the interactome.

Interactome networks have also important structural features that go beyond the degree (or other measures of centrality) of individual nodes: "Network modules," i.e. groups of nodes that are densely interconnected among themselves, but sparsely connected to the rest of the network, can often be identified with proteins that jointly perform a certain function [104, 105, 106]. This relation between functional similarity of genes (see ahead to Box 10.14) and their closeness in interactome networks has also been found for shared pathway membership, co-localization in the same cellular component or co-expression [87, 89]. The local aggregation of cellular function within interactome networks represents a fundamental biological organization principle that forms the basis for many important applications, ranging from the prediction of protein function to disease gene identification and drug target prioritization.

10.3.4 Diseases in the Interactome

The observation that functionally similar proteins are often densely interconnected can be generalized also to other relationships among genes, in particular to shared disease associations. Genes that are implicated in the same disease tend to have more interactions among each other than expected for completely randomly distributed genes [107]. Note, however, that this does not necessarily imply particularly densely interconnected network patterns as those observed for genes involved in the same function. Indeed, *dys*function is typically distributed among several, often only loosely connected functional modules within the interactome [108]. A systematic study on $\sim$ 300 complex diseases showed that currently available interactome networks offer sufficient coverage to identify these "disease modules," thereby confirming a fundamental hypothesis of interactome-based approaches to human disease [16]. The specific topological properties of disease modules differ between classes of diseases (e.g., complex diseases, Mendelian diseases, or cancer) and inheritance modes (autosomal dominant or recessive). Cancer driver genes are often highly central, while recessive disease genes tend to be more isolated at the periphery of the interactome [56].

10.3.5 Localization in Networks

As shown above, network-based localization of (dys)function is a central part of many interactome-based studies. In network science, the identification of densely connected groups of nodes is known as "community detection" [109]. While numerous algorithms exist for this task, they are usually not well suited for the identification of only weakly connected local network neighborhoods such as disease modules [108]. In order to quantify the tendency of a given set of disease genes to be localized in a certain neighborhood, we first need to inspect different possibilities for **measuring distances among a set of nodes in a network**. The simplest way to summarize the localization of a set **S** consisting of s nodes into a single quantity is to compute the network distance d_{ij} for all $\binom{S}{2} = \frac{s(s-1)}{2}$ pairs of nodes i and j and take the average:

$$d_{\text{av}}(\mathbf{S}) = \tfrac{2}{s(s-1)} \sum_{ij} d_{ij}\,, \tag{10.1}$$

which can be interpreted as a diameter of the set **S**. As a consequence of the "small-world" nature of many relevant networks, differences in the absolute values of d_{av} for different gene sets are often relatively small. Several variations and extensions of Equation 10.1 have therefore been proposed [110]. For example, instead of taking the average over all possible node pairs, one can consider only the distance to the next closest node, respectively:

$$d_{\text{close}}(\mathbf{S}) = \tfrac{1}{s} \sum_{i} \min_{j \in \{\mathbf{S} \backslash i\}} (d_{ij})\,. \tag{10.2}$$

This gives different results as d_{av} in situations where a module is split into several "islands," for example due to network incompleteness. Whereas d_{close} correctly reflects the high degree of localization within the individual islands, it is diluted when the distances of all pairs are averaged. Other variations include adding weights to different path lengths d_{ij}, see Box 10.10 for more examples. Complementary to such distance-based measures, one can also use **connectivity-based measures** to determine the degree of connectedness among a set of nodes. The simplest way is to consider the number of links between them. A perhaps more intuitive measure is given by the size of the largest connected component, i.e., the highest number of nodes that are directly connected to one another. We can apply tools from statistical physics to understand many of its properties analytically [111]. It is, however, relatively sensitive to data incompleteness. In extreme cases, a single missing link in the network or a missing node from the set **S**, e.g., a protein, whose disease association is yet unknown, can fragment the connected component into isolated nodes.

The concepts introduced above can be readily extended to measure distances between two node sets **S** and **T**, for example, for quantifying the interactome-based similarity between two diseases [16]. The equivalent of Equation 10.1, i.e., the average over all possible pairs of nodes between two node sets is given by

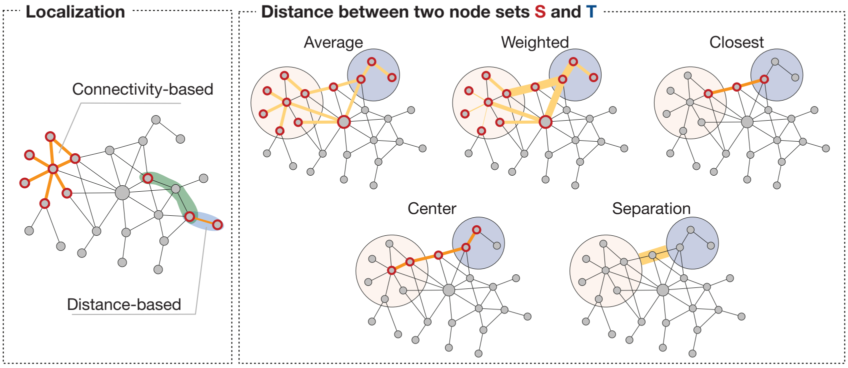
$$d_{\text{av}}(\mathbf{S}, \mathbf{T}) = \frac{1}{s} \sum_{i \in \mathbf{S}} \frac{1}{t} \sum_{j \in \mathbf{T}} d_{ij}\,. \tag{10.3}$$

Similarly to different linkage methods in hierarchical clustering algorithms, there are different ways to compute the distance between two sets of nodes, see Box 10.10 for a number of frequently used options.

10.3.6 Randomization of Network Properties

By themselves, the absolute values of localization or distance as introduced above bring few insights. To judge whether an observed clustering of a particular node set is significant, we need to compare it to suitable random models. Many quantities that

Box 10.10: Distance measures in networks



There are different ways to quantify the degree of "localization" of a given set of nodes **S**, i.e., whether or not they aggregate in a certain network neighborhood. Distance-based localization measures are based on different averages over pairwise distances d_{ij} between all nodes in the set, e.g.:

$$d_{\text{av}}(\mathbf{S}) = \tfrac{2}{s(s-1)} \sum_{ij} d_{ij} \tag{10.4}$$

$$d_{\text{close}}(\mathbf{S}) = \tfrac{1}{s} \sum_{i} \min_{j \in \{\mathbf{S} \setminus i\}} (d_{ij}) \tag{10.5}$$

$$d_{\text{exp}}(\mathbf{S}) = -\tfrac{2}{s(s-1)} \ln \sum_{ij} \exp\left(-d_{ij}\right) \tag{10.6}$$

These measures can be generalized to two node sets **S** and **T**:

$$d_{\text{av}}(\mathbf{S},\mathbf{T}) = \tfrac{1}{s} \sum_{i \in \mathbf{S}} \tfrac{1}{t} \sum_{j \in \mathbf{T}} d_{ij} \tag{10.7}$$

$$d_{\text{close}}(\mathbf{S},\mathbf{T}) = \tfrac{1}{s+t} \Big[\sum_{i \in \mathbf{S}} \min_{j \in \mathbf{T}}(d_{ij}) + \sum_{i \in \mathbf{T}} \min_{j \in \mathbf{S}}(d_{ij}) \Big] \tag{10.8}$$

$$d_{\text{exp}}(\mathbf{S},\mathbf{T}) = -\tfrac{1}{s} \sum_{i \in \mathbf{S}} \tfrac{1}{t} \ln \sum_{j \in \mathbf{T}} \exp\left(-d_{ij}\right) \tag{10.9}$$

Nodes that are common to both sets **S** and **T** are usually taken to contribute with $d_{ij} = 0$ in the above formula. Instead of averaging over all pairs of nodes between **S** and **T** one can also define a center for each and use the distance between them:

$$d_{\text{center}}(\mathbf{S},\mathbf{T}) = d\left(\text{center}(\mathbf{S}), \text{center}(\mathbf{T})\right) \tag{10.10}$$

Another option is the separation parameter introduced in [16]:

$$\text{sep}(\mathbf{S},\mathbf{T}) = d_{\text{close}}(\mathbf{S},\mathbf{T}) - \tfrac{1}{2}\left(d_{\text{close}}(\mathbf{S}) + d_{\text{close}}(\mathbf{T})\right) \tag{10.11}$$

Negative values $\text{sep}(\mathbf{S},\mathbf{T}) < 0$ suggest overlapping network modules, while $\text{sep}(\mathbf{S},\mathbf{T}) > 0$ indicates separated modules. Note, however, that the separation parameter is not an intensive quantity, i.e., its magnitude depends on the number of nodes in the respective sets.

occur in the context of network analyses do not follow normal (Gaussian) distributions, such as the scale-free degree distribution, and therefore require particular care when choosing statistical tests. Comparisons with ensembles of randomized networks obtained from simulations are often the best choice. In general, we can distinguish two types of randomizations: (1) **Randomizing the network topology**, for example the interaction partners of a particular protein, and (2) **randomizing node attributes**, such as the disease associations of a group of genes.

10.3.6.1 Randomizing the Network Topology

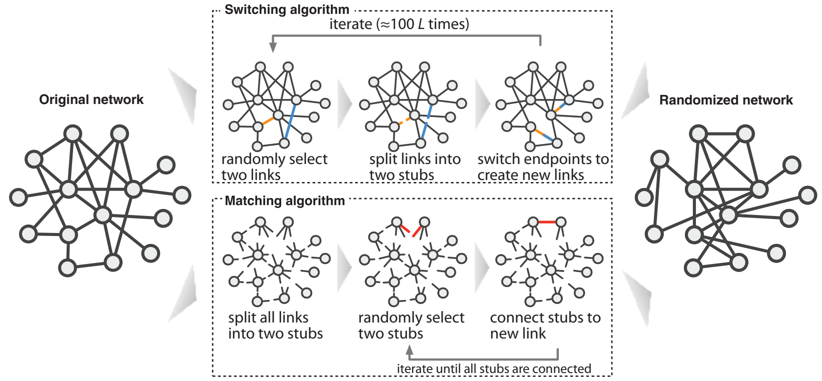
To exclude that a seemingly interesting observation, for example, the local aggregation of disease genes in the interactome, could be a generic consequence of the overall topology of the underlying network, we need to compare our results from the original network with those obtained from networks with randomized topology. There are numerous randomization procedures. Which one is most suited, depends on the particular reference that is needed for a specific observation. The simplest method is to fix only the number of nodes N and the number of links L of the original network and to redistribute the links completely at random among the nodes. As shown in Chapter 3, this procedure results in an Erdős-Rényi network. Many properties of Erdős-Rényi networks can be calculated analytically and without extensive computer simulations, for example the expected clustering or the size of the largest connected component. However, the topology of most real world networks differs substantially from the one of a corresponding complete random graph, for instance hubs are completely absent in the latter. Hence, comparisons between the two are rarely meaningful and can in fact be rather misleading.

A more adequate reference that is suitable for most applications is given by networks in which the number of neighbors of every node are kept constant, but the specific interaction partners are completely randomized. This ensures that important structural features, in particular the degree distribution and presence of hubs, are preserved in the ensemble of randomized networks. Box 10.11 introduces the two main algorithms that are used to generate such randomized networks: The "switching algorithm" [112], is an iterative method, where at each step two links are selected at random and their endpoints are swapped. For example, the links connecting the nodes $n_1 \leftrightarrow n_2$ and $n_3 \leftrightarrow n_4$, respectively, can be reconnected to $n_1 \leftrightarrow n_3$ and $n_2 \leftrightarrow n_4$. Note that this may result in multiple links between two nodes or self-loops. In an application where such links are not meaningful, the original link pairs should be restored. As we repeatedly apply this procedure, the interactions of the network become more and more randomized, without altering the degree of each node. A drawback of this simple method is that no precise criteria exist as to how many switches should be performed to ensure a good mixing. Empirical results suggest $100\,L$ switching attempts, which can be computationally rather expensive for large networks [113].

A more efficient method for generating random networks with a prescribed degree sequence is to apply a variation of the "configuration model" [114, 115]. The second algorithm introduced in Box 10.11 is the "matching algorithm," in which all links of a given network are broken at once and then randomly reassembled one by one. As in the switching algorithm, the potential creation of self-loops and multiple links may

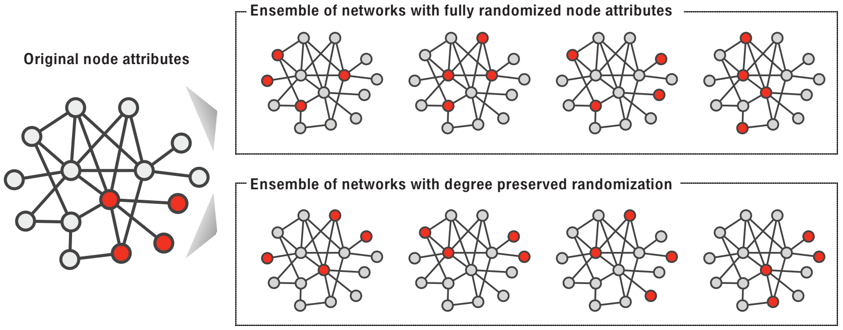
Box 10.11: Network randomization

Randomizing the network topology


Original network
Switching algorithm
iterate (≈100 L times)
randomly select two links
split links into two stubs
switch endpoints to create new links
Matching algorithm
split all links into two stubs
randomly select two stubs
connect stubs to new link
iterate until all stubs are connected
Randomized network

There are two frequently used algorithms to generate an ensemble of randomized networks with fixed degree distribution. In the *switching algorithm*, two links are chosen at random and their endpoints switched. Repeating this procedure will eventually lead to a fully randomized version of the original network. In the **matching algorithm**, all links of the given network are broken at once and then one by one reconnected at random.

Randomizing node attributes


Original node attributes
Ensemble of networks with fully randomized node attributes
Ensemble of networks with degree preserved randomization

The most basic procedure to randomize node attribues (e.g., disease associations of genes) is to redistribute them completely at random on the network. For more restricted random controls, one can also keep specific topological properties of a node attribute constant, in particular the degree of the annotated node. In this case, only nodes with the same (or at least similar) properties are allowed choices.

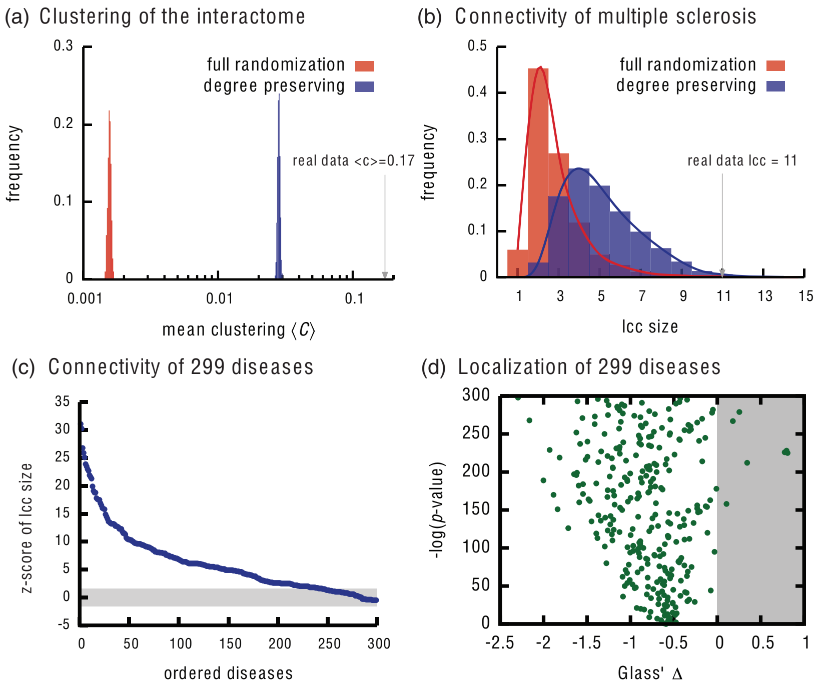


Figure 10.2: Network randomization. (a) Comparison of the clustering coefficient of the interactome (see Figure 10.1) with distributions obtained from complete randomization and degree-preserving randomization. (b) Comparison of the size of the largest connected component (lcc) of proteins associated with multiple sclerosis in the interactome with two distributions obtained from full and degree preserving randomization, respectively. (c) Sorted z-scores of the lcc size of 299 diseases in the interactome. (d) Significance and effect size of the observed localization $d_{\text{av}}(\mathbf{S})$ of 299 diseases compared to randomized gene sets. (Data from [16].)

need to be prevented in certain applications. Note that in this case the ensemble of the generated networks is no longer completely unbiased, but the effects are usually small and can often be neglected for large networks [113].

Figure 10.2a shows an application of the two randomization strategies to evaluate the observed mean clustering coefficient $\langle C \rangle = 0.17$ of the interactome. As expected, we find excellent agreement between the values observed in 10,000 simulations of a full random model corresponding to an Erdős-Rényi network and the respective analytical value $\langle C \rangle = p = \frac{2L}{N(N-1)} = 0.0016$. Simulations of the degree preserving matching algorithm yield the considerably higher mean value $\langle C \rangle = 0.03$, which is still significantly smaller than the originally observed clustering, indicating that the clustering of the interactome could not have emerged by chance.

10.3.6.2 Randomizing Node Properties

Instead of rewiring the structure of the network itself, it is often useful to consider randomizing certain node attributes, for example disease associations of individual genes in the interactome. In the simplest case of **random label permutation**, we detach the attribute of interest from their original nodes and redistribute them completely at random among all nodes of the network. For example, to investigate the connectivity of N_d disease proteins in terms of their largest connected component (lcc), we select the same number of proteins randomly from the network and measure their lcc. Repeating this procedure yields a random control distribution that can then be used to determine the statistical significance of the original lcc. According to data from [16], multiple sclerosis has $N_d = 69$ known associated proteins in the interactome that form an lcc of size $S = 11$. Figure 10.2 (b) shows the lcc distribution for 69 randomly picked proteins from 10,000 simulations. The distribution has a mean of $\langle S_{rand}^{full} \rangle = 2.9$ and a standard deviation of $\sigma = 1.4$. The statistical significance of the observed lcc size can be quantified using the z-score

$$z\text{-score} = \frac{S - \langle S_{\text{rand}}^{\text{full}} \rangle}{\sigma}, \tag{10.12}$$

yielding z-score $= 5.8$. For normal distributions, z-scores > 1.65 correspond to a p-value < 0.05 (corresponding to a right-sided test, left- or two-sided tests are also possible) and are considered to be statistically significant. The empirical p-value, i.e., the fraction of all random simulations with $S_{\text{rand}}^{\text{full}} \geq S$ was found to be p-value $= 0.003$. Taken together, we conclude that the connected component for multiple sclerosis is unlikely to have emerged by chance or as a trivial consequence of the network topology, indicating the potential presence of a disease module.

10.3.6.3 *Degree Preserving Label Permutation*

There are also stricter attribute randomization procedures that impose certain constraints on the allowed set of nodes among which an attribute can be distributed. Prominent cancer genes, for example, tend to have a large number of interactions in literature-curated interactome networks, simply because they have been investigated more intensively than other genes. To test whether the high connectivity among such genes can be explained by their high degree alone, we need to generate random distributions of node attributes that maintain the degree of the individual nodes carrying the original annotation. Note that swapping only between nodes of exactly the same degree will be problematic for high-degree nodes, as there may be only few, or even a single node in the entire network that have a certain degree. It is therefore useful to relax the requirement of having exactly the same degree and work with bins of nodes with comparable degree instead. Figure 10.2 (b) shows the distribution $S_{\text{rand}}^{\text{degree}}$ obtained using such an approach. The mean value $\langle S_{\text{rand}}^{\text{degree}} \rangle = 5.1$ is larger than the one obtained from the full randomization, but still significantly smaller than the value $S = 11$ from the original data (z-score $= 3.1$, empirical p-value $= 0.009$), indicating that the high degree of the disease proteins alone does not explain their observed high connectivity.

These randomization procedures can also be applied to evaluate the distance-based localization measures introduced above, for example $d_{\text{av}}(\mathbf{S})$. From each random simulation we can extract $d_{\text{av}}^{\text{rand}}$ and then compute the mean $\langle d_{\text{av}}^{\text{rand}} \rangle$ and corresponding standard deviation $\sigma\left(d_{\text{av}}^{\text{rand}}\right)$. In analogy to the z-score introduced above, we can use Glass' Δ to quantify the effect size of any difference observed between the true value $d_{\mathrm{av}}(\mathbf{S})$ and the values obtained in the respective randomization simulations:

$$\Delta = \frac{d_{\mathrm{av}}(\mathbf{S}) - \langle d^{\mathrm{rand}} \rangle}{\sigma\left(d^{\mathrm{rand}}\right)}\,. \tag{10.13}$$

The statistical significance of an observed difference in the respective means $d_{\mathrm{av}}(\mathbf{S})$ and $\langle d^{\mathrm{rand}} \rangle$ can be obtained from a Mann–Whitney U test, for example. Figure 10.2 (c–d) shows the results for the randomization valuation of the localization observed among 299 diseases on the interactome.

Numerous more advanced randomization procedures exist that can preserve topological features beyond the degree distribution. For example, there are algorithms to generate randomized networks that maintain the mean clustering coefficient of the original network [116] or the correlation structure between the degrees of adjacent nodes [117, 118]. Another level of sophistication needs to be applied when randomizing metabolic networks, where simple link rewiring would likely generate reactions that are biochemically impossible [119, 120].

10.4 Disease Module Analysis

10.4.1 Overview

Sequencing technology has accelerated the discovery of disease associated genetic variations significantly. For most diseases, however, we are still far from a complete understanding of the underlying molecular mechanisms. Most complex diseases, such as cardiovascular diseases, cancer, or diabetes mellitus (the three most frequent causes of death worldwide), involve hundreds of genes and their complex interactions. It has been estimated, for example, that more than 2,000 genes are involved in intellectual disabilities, yet our current knowledge includes only around 800 genes [121]. The situation is similar for rare Mendelian disorders. Estimates for the total number of rare genetic disorders range from 6,000 to 8,000, a majority of which likely to be caused by a single genetic aberration. Despite this simple genetic architecture, less than half of all suspected diseases and corresponding disease genes are currently known.

Network-based **disease modules** offer a general framework for investigating how the pathobiology of a particular disease may arise from a combination of many genetic (but also epigenetic, environmental, behavioral etc.) variations. Succesful applications range from rare Mendelian disorders [3], to cancer [4] and other complex disorders, like metabolic [5], inflammatory [42], or developmental diseases [122]. A disease module is loosely defined as the comprehensive set of cellular components associated with a certain disease and their interactions. More specifically, the term refers to a connected subgraph of the interactome, whose perturbation causes the disease [18]. Figure 10.3 gives an overview of the disease module analysis process. The first step is to construct an interaction network and collect genes known to be associated with the particular disease of interest. These "seed genes" will serve as starting point for network-based gene prioritization algorithms. The resulting network module can then be validated and enriched with various additional datasets that will also be used in the biological interpretation of the final disease module.

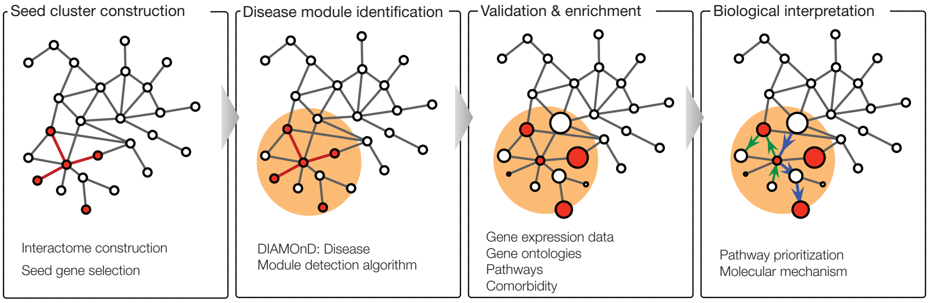


Figure 10.3: The basic steps of a disease module analysis process: First, interactome and seed gene data are collected. Next, a network-based disease gene prioritization method is employed. The performance of the predictions is then validated through comparison and enrichment with independent external data. In the last step, the module is explored for important biological pathways, overlap with other disease modules etc. (Figure adapted from [123].)

10.4.2 Seed Cluster Construction

The first step of the disease module analysis is the construction of a seed cluster, i.e., the curation of a suitable molecular interaction network and a set of genes known to be associated with the particular disease of interest. Box 10.12 lists a number of resources that may serve as a starting point.

10.4.2.1 Interactome Construction

As introduced above, one can make a broad distinction between physical interactions, e.g., protein co-complexes or binary protein–protein interactions, and functional interactions, e.g., genetic interactions or co-expression. By definition, physical interactions represent a direct molecular relationship, thus facilitating the identification of causal molecular mechanisms. Functional interactions, on the other hand, offer a much broader spectrum of potentially relevant associations between genes and gene products and can often be more easily adapted to a particular diseases, for example by incorporating tissue-specific expression data. Incorporating such information can considerably improve disease gene prioritization [124, 125, 126], see also Chapter 11. The choice of interaction type and used data sources will affect coverage (number of contained genes/proteins and their interactions), biases (for example, towards well-studied genes) and signal to noise ratio (number of false positive interactions) of the final interactome. Physical interactions offer more control over biases and signal to noise ratio, but often at the cost of lower coverage. Biases can be reduced by relying only on data obtained from systematic high-throughput studies, e.g., from [87, 89]. False positive interactions can be reduced by filtering for interactions that have been reported by several studies and by different experimental techniques. Several databases, such as HIPPIE [86] or STRING [83] offer integrated interaction scores for this purpose.

Box 10.12: Resources for disease module analyses

Interactome databases:

BIOGRID	thebiogrid.org
BioPlex	bioplex.hms.harvard.edu
HIPPIE	cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/
IntAct	www.ebi.ac.uk/intact
MatrixDB	matrixdb.univ-lyon1.fr
MINT	mint.bio.uniroma2.it
STRING	string-db.org

A more comprehensive list can be found on EBI's PSICQUIC view that also offers programmatic acces, see www.ebi.ac.uk/Tools/webservices/psicquic/view/

Disease genes:

DGA	dga.nubic.northwestern.edu
GWAS Catalog	www.ebi.ac.uk/gwas
Gene2Mesh	gene2mesh.ncibi.org
HGMD	hgmd.cf.ac.uk
OMIM	omim.org
OrphaNet	www.orpha.net

Integrated and functional web-based services:

DisGeNet	disgenet.org
GeneMANIA	genemania.org
HumanBase	hb.flatironinstitute.org

Ontologies:

Disease ontology (DO)	disease-ontology.org
Gene ontology (GO)	www.geneontology.org
Human phenotype ontology (HPO)	human-phenotype-ontology.github.io
Mammalian phenotype ontology (MPO)	www.informatics.jax.org/vocab/mp_ontology

A comprehensive list of biological ontologies can be accessed from EBI's Ontology Lookup Service under https://www.ebi.ac.uk/ols/ontologies

10.4.2.2 Seed Gene Selection

There are numerous resources that collect genes associated with diseases (see Box 10.12). Note that the term "disease associated gene" itself is only loosely defined and covers a wide spectrum from high penetrance dominant mutations to GWAS variants of rather small effect size or genes observed to be differentially regulated in patient subgroups. Similarly, the level of evidence for reported disease associations may differ greatly, from rare gene variants with a known and experimentally validated functional mechanism, to genes with unknown mechanism, yet repeatedly confirmed in multiple patient cohorts, to rather speculative associations inferred solely from text mining.

10.4.2.3 Evaluation of the Seed Cluster

Both the interactome construction and the seed gene selection involve a certain tradeoff between using only highest-confidence data and achieving the highest possible coverage. There is no simple and universally applicable solution to this challenging problem that requires a certain amount of experimentation, ideally guided by a domain expert for the specific disease under study. From a network perspective, however, localization measures introduced above can be used as a rough indicator whether a particular combination of interactome and seed gene data meets the minimal criteria for a meaningful disease module analysis. Figure 10.4 shows the seed cluster for an asthma disease module from [123]. From a total of 129 seed genes that could be mapped to the interactome, 37 form the largest connected component, indicating a highly significant (z-score $= 10.7$) network localization. This suggests that the seed cluster has sufficient "signal" pinpointing the network neighborhood of the complete asthma module that can then be identified through a network-based expansion algorithm.

10.4.3 Network-Based Disease Gene Prioritization

Network-based disease gene prioritization methods build on the observation that genes associated with the same disease tend to be localized in the same interactome neighborhood. We can therefore use the network topology to extrapolate from a given set of seed genes to identify other genes that are likely to be also involved in the disease or at least strongly affected by the local interactome perturbation. Over the last years, numerous algorithms have been developed for this purpose. They can be broadly classified into three major categories: (1) connectivity based methods (2) path-based methods and (3) diffusion-based method (see Box 10.13).

10.4.3.1 Connectivity-Based Methods

Connectivity-based methods exploit the observed propensity among disease genes to interact with each other. Early pioneering approaches considered all direct neighbors of seed genes as potential candidate genes [127]. As more and more interactome and seed gene data become available, such approaches tend to generate an increasing number of false positives. More recent algorithms therefore utilize more advanced connectivity patterns, such as graphlets [128], or take the degree heterogeneity of the interactome explicitly into account [129]. Indeed, hubs in the network are expected to

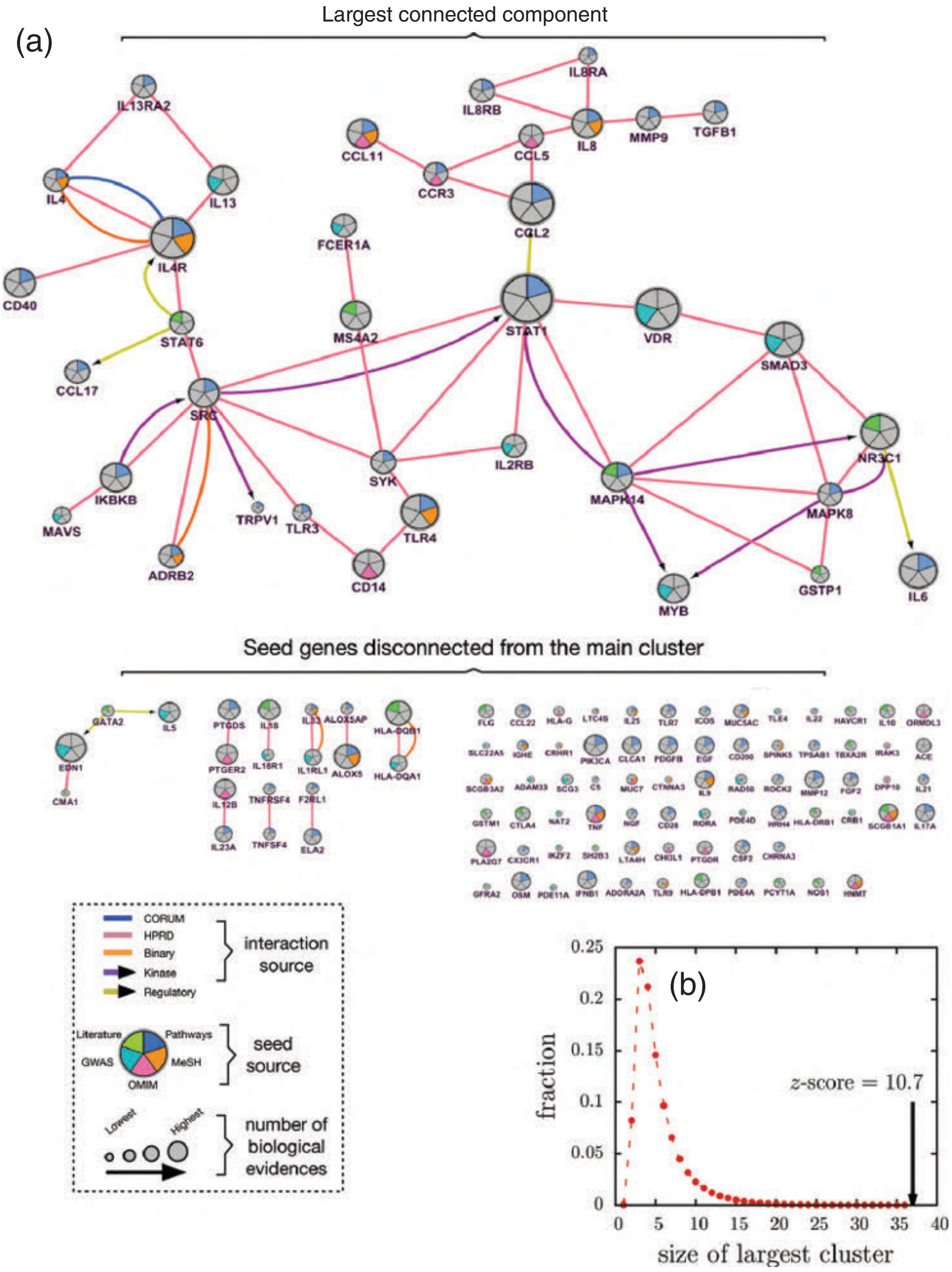


Figure 10.4: Seed cluster of an asthma disease module analysis from from [123]. (a) Of the 129 expert curated seed gene, 37 form the largest connected component, the rest are scattered throughout the interactome. (b) The size of the largest connected component is highly significant (z-score = 10.7) compared to random expectation.

Box 10.13: Network-based disease gene prioritization

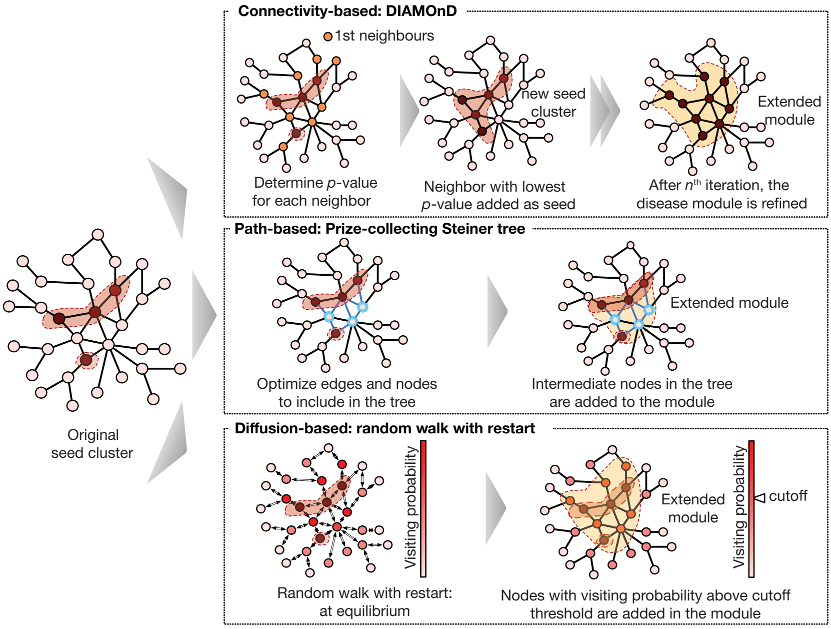

Connectivity-based: DIAMOnD
1st neighbours
Determine *p*-value for each neighbor
new seed cluster
Neighbor with lowest *p*-value added as seed
Extended module
After *n*th iteration, the disease module is refined
Path-based: Prize-collecting Steiner tree
Optimize edges and nodes to include in the tree
Extended module
Intermediate nodes in the tree are added to the module
Original seed cluster
Diffusion-based: random walk with restart
Visiting probability
Random walk with restart: at equilibrium
Extended module
Visiting probability
cutoff
Nodes with visiting probability above cutoff threshold are added in the module

Illustration of three different methodologies for network-based disease gene prioritization: (1) **Connectivity-based** methods evaluate the direct neighbors of seed genes. (2) **Path-based** methods evaluate candidate genes based on their network distance to seed genes. (3) **Diffusion-based** methods use a dynamical process to rank candidate gene according to how strongly they are influenced by the seed genes.

also interact with a large number of seed genes without necessarily implying a disease-association. To correct for these effects, the DIAMOnD algorithm [108, 123] evaluates the *significance* of a given number of connections k_s to s seed genes with respect to the total degree k of a given candidate gene. In a network of size N, with s randomly distributed seed genes, the probability that a gene with degree k connects to exactly k_s seed genes is given by the hypergeometric distribution

$$P(X = k_s) = \frac{\binom{s}{k_s}\binom{N-s}{k-k_s}}{\binom{N}{k}} \,. \tag{10.14}$$

The significance of a given number of connections is therefore given by the p-value

$$p\text{-value} = \sum_{n=k_s}^{k} P(X = n) \,, \tag{10.15}$$

which can then be used to iteratively rank all genes in the network. Note that the resulting disease module may consist of genes without direct connectivity to the initial seed genes.

10.4.3.2 Path-Based Methods

Instead of using the direct connectivity to seed genes, candidate genes can also be ranked according to their network distance to the set of seed genes (compare also with Box 10.10). A versatile set of algorithms that combines different distance measures for prioritizing candidate genes has been proposed in [130]. Instead of ranking the genes iteratively, it is also possible to search for an optimal set of candidate genes that collectively minimize the path lengths between the seed genes. Such approaches often implement variations of minimum spanning tree (or "Steiner tree") search algorithms [131, 132, 133]. Basically, the algorithm will construct a tree consisting of a minimum amount of edges while connecting all the seeds into a single cluster.

10.4.3.3 Diffusion-Based Methods

The methods described above rely only on the static topology of the network. It is also possible to use dynamical models to explore the network neighborhood around the seed genes for gene prioritization [3, 4, 134, 135, 136, 137]. Among the most widely used dynamical models are diffusion processes, such as the random walk with restart (RWR) [138]: Here, the seed genes serve as starting points for a random walk process along the links of the network. At every time step, the walker either proceeds to a randomly picked neighboring gene, or returns with restart probability r to one of the seed genes. The restart ensures that the local neighborhood around the seed genes is emphasized by the walker, otherwise all seed gene information would be lost in the long run of the process. The frequencies with which the individual nodes in the network are visited will eventually converge to a steady state and can then be used to rank all genes in the network according to their "dynamical closeness" to the seed genes. The process can be formalized as follows: Consider the vector $\mathbf{p}_t$ whose elements $p_i \ldots p_N$ represent the probability of the walker visiting node i at time t. The visiting probability at time t can be derived from the visiting probability at time $t-1$ via

$$\mathbf{p}_t = \mathbf{W}\mathbf{p}_{t-1}\,, \tag{10.16}$$

where $\mathbf{W}$ is the so-called transition matrix and defined as the column normalized adjacency matrix $\mathbf{A}$ with $W_{i,j} = \frac{A_{i,j}}{\sum_i k_i}$. At time t_0, only seed genes have (uniform) non-zero probability p, as well after each restart, which happens at a rate r. Equation 10.16 then becomes

$$\mathbf{p}_t = (1-r)\mathbf{W}\mathbf{p}_{t-1} + r\mathbf{p}_0\,. \tag{10.17}$$

The steady-state solution for Equation 10.17 is given by

$$\mathbf{p}_\infty = r(\mathbf{I} - (1-r)\mathbf{W})^{-1}\mathbf{p}_0\,. \tag{10.18}$$

The genes in the network can then be ranked according to the visiting probability p_∞. The restarting probability r can be used to adjust the influence of the seed genes on the diffusive process, from free diffusion (walker is not restricted by seed genes, $r = 0$) to no diffusion at all (walker remains at seeds, $r = 1$).

10.4.4 Validation and Enrichment

After completion of the preferred candidate gene ranking procedure, we first need to evaluate its performance. A second, closely related task is to determine a sensible cutoff, i.e. how many ranked genes should be considered for the final disease module, as most prioritization methods rank all genes in the network without offering an intrinsic stopping criterion. There are two complementary approaches: (1) Estimating the predictive power of the disease gene predictions using cross-validation methods. (2) Comparison with independent biological data.

10.4.4.1 Cross-validation of Prediction Performance

In principle, cross-validation of disease gene prioritization algorithms works in the same way as with other classification tasks (compare also with Chapters 6–8): For a basic k-fold cross-validation, the set of seed genes is first randomly divided into k groups (the special case where k equals the number of seed genes is often referred to as "leave-one-out" cross-validation). One of the groups can then serve as the "test-set" of true positives, while the remaining $k-1$ groups are used as modified seed gene pool. The gene prioritization algorithm is then run on this modified pool to test how well the method is able to retrieve the left out genes in the test set. Repeating this procedure k times with each of the k groups serving as test set yields a statistic on the expected average performance of the method. The choice of k determines the trade-off between high bias (large k) and high variance (small k). An important difference to many other classification tasks is the lack of clear true negatives, i.e., genes that we know not to be involved in the disease. Several proxies have been proposed, for example essential genes, genes of high genetic variability or manually curated genes that are unlikely to be involved in a particular disease according to their expression patterns. These gene sets can only offer approximations and remain necessarily incomplete, making the interpretation of standard performance measures difficult, such as receiver operating characteristic curves.

10.4.4.2 Enrichment with Independent Biological Data

A complementary approach for estimating the performance is to test for enrichment of the ranked genes with independent biological data (see Box 10.12). Figure 10.5 shows the biological enrichment of the top 400 ranked genes from an asthma disease module analysis [123]. To compare the biological signal of the ranked genes with the one of the manually curated seed genes, the authors chose a sliding window of ranked genes with the same size of the seed genes and within each window computed the enrichment with five different datasets: (1) Genes differentially expressed in a relevant case/control study, (2) genes participating in expert curated relevant biological pathways, (3) genes contained in general pathways that were found enriched in the seed genes, (4) genes annotated to similar biological processes as the seed genes according the gene ontology (GO, see Box 10.14) and (5) genes that are known to be implicated in

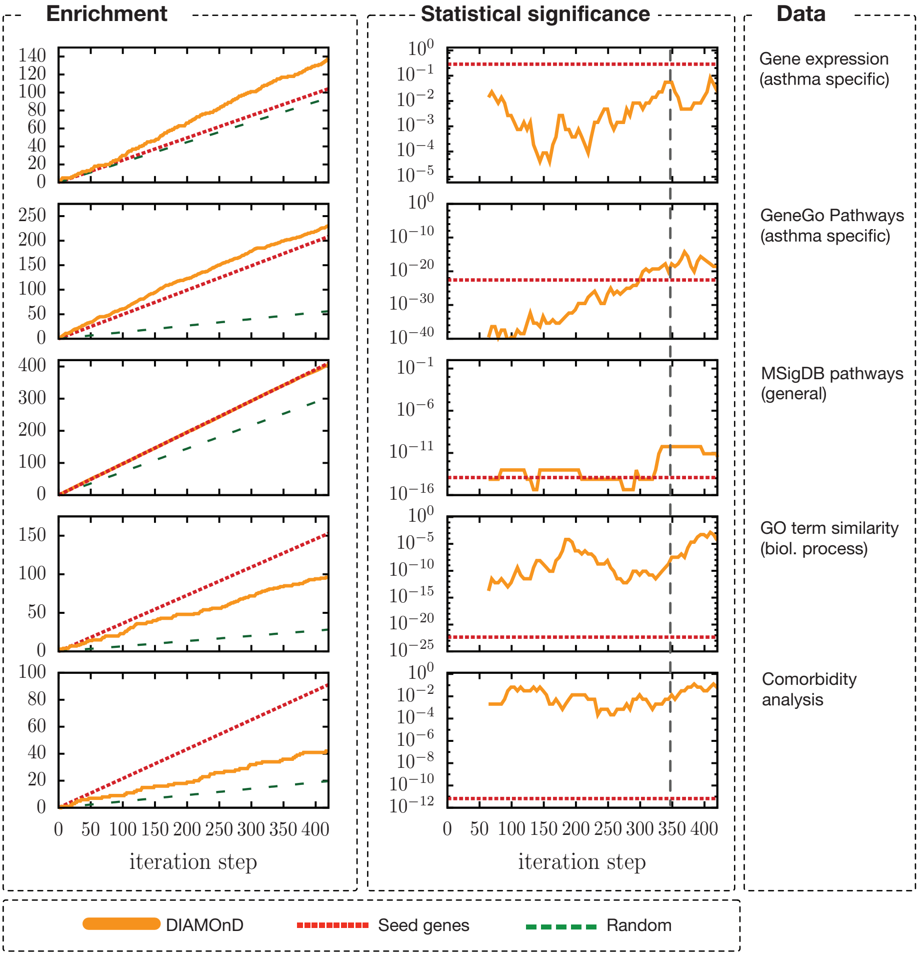


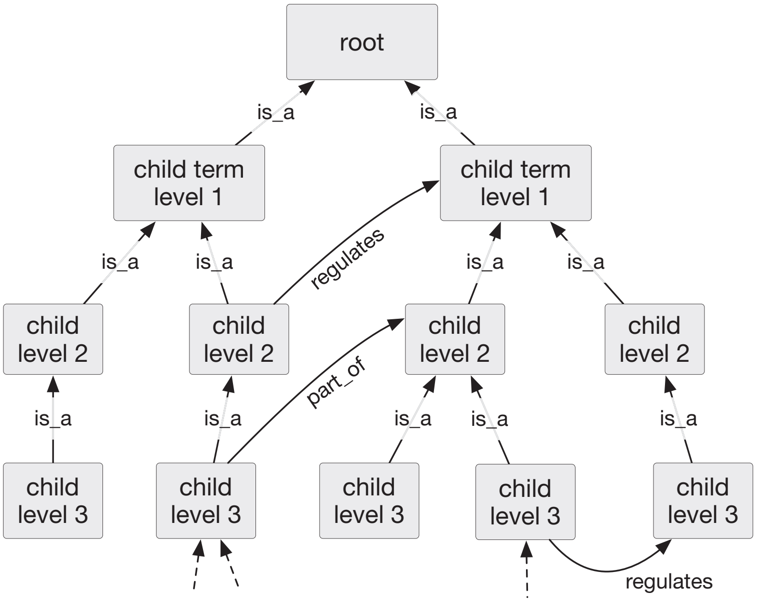
Figure 10.5: Biological enrichment of the asthma disease module in [123]. The first two columns show the number (and the corresponding statistical significance, respectively) of the identified candidate genes that were found in the different validation datasets indicated in the third column. The values for the candidate genes are show in orange, the values for seed genes and random expectation in red and green, respectively.

diseases that show high co-morbidity with asthma. A comparison of the enrichments across different datasets allows for an evaluation of the general plausibility of the ranked genes, but also for an estimation of the border of the disease module.

10.4.5 Biological Interpretation

The data collected for the performance evaluation can further be used for an integrated analysis of the biological mechanisms represented in the disease module. Figure 10.6

Box 10.14: Ontologies



Ontologies are controlled vocabularies to organize the knowledge of a specific field, for example biological pathways, diseases or phenotypes (see Box 10.12 for a list of biomedical ontologies). These vocabularies are usually manually curated by an authoritative consortium of domain experts. An important vocabulary is the gene ontology (GO). It consists of three separate branches: (1) "cellular component" (4,195 terms), (2) "molecular function" (11,120 terms), and (3) "biological process" (29,682 terms), each forming a hierarchical, acyclic tree. The root term at the top is the most general, increasingly specific terms are connected by either `is_a`, `part_of` or `regulates` links that describe the particular relationship between the respectively linked terms.

Ontologies are not only useful for systematic annotation and collection of knowledge, but can also be used to assess the "semantic similarity" among different terms according to their relative position in the tree [139]. A common approach relates the specificity (tree depth) of a term to its information content (IC). The similarity between two terms can then be calculated from the IC of their most informative (i.e., highest IC) common ancestor. Note that most biological entities, such as gene products, are usually annotated with several terms and different strategies can be used to aggregate the similarity among several terms, see [139] for a detailed discussion.

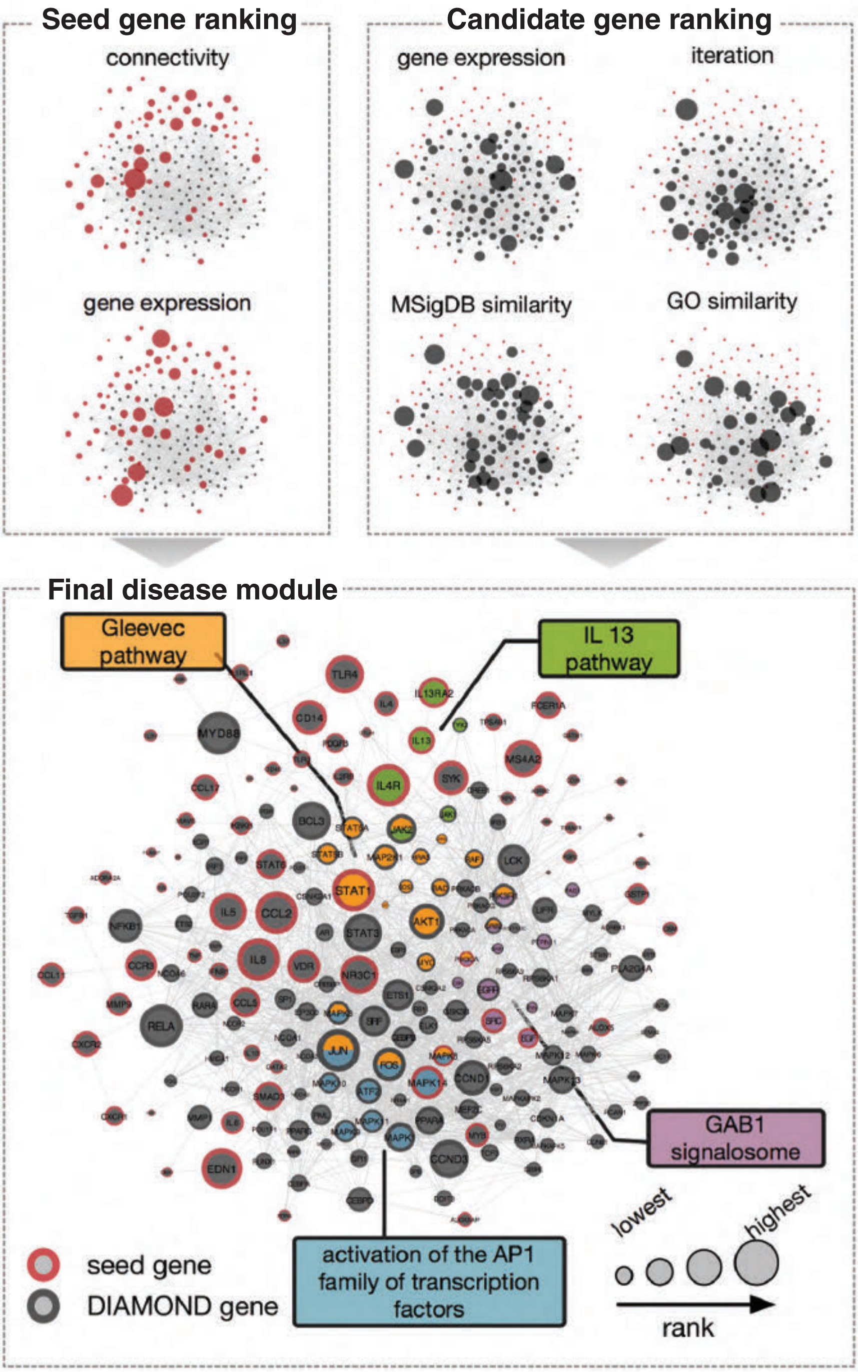


Figure 10.6: Illustration of the ranking procedure (top) and the final asthma disease module (bottom) from [123]. Seed genes and candidate genes are first ranked separately according to their enrichment with different biological datasets. The individual rankings are then combined into a final score for each gene in the disease module, which can then be used to prioritize pathways within the module.

illustrates how the different data are combined into a final score for each gene in the asthma disease module, which in turn can be used to prioritize pathways within the module. The first step is to create a ranking of all genes for each individual data source. Seed genes and candidate genes are often examined separately, which has the advantage that they can be given different weights when they are combined later on. Depending on the particular data type, the ranking can be based on fold-change for differential expression data, GWAS p-value or functional similarity with known processes (compare with Box 10.14), for example. The individual rankings can then be combined into a single score, e.g., using the so-called Borda-count [140]: The score of a gene is taken to correspond to its inverted rank and the scores of different rankings are simply added. Finally, the integrated gene score can be used to prioritize pathways within the module, thus complementing commonly used measures, such as coverage of genes in the pathway. The integrated biological relevance of a pathway within the module can be quantified by the average score of its genes. Additional potentially interesting network-based analyses that can be performed with the disease module include identifying overlaps with other diseases or with network modules known to be modulated by drugs, for example using the distance measures above, or applying community detection to identify potential submodules, for example for patient stratification.

10.5 Summary and Outlook

Network medicine is a highly dynamic and rapidly expanding field covering virtually all areas of biomedical research. This brief introduction can therefore only provide a necessarily incomplete and highly subjective selection. We hope that the references we provide may serve as a starting point for further reading and also recommend a recently published textbook focusing exclusively on this subject [141].

An important challenge in current biomedical research is to integrate the ever growing amount of "omics" data (e.g., genomics, epigenomics, proteomics, metabolomics, lipidomics). Network approaches are inherently holistic and integrative, and particularly multilayer networks are very promising candidates for addressing this challenge [79]. First analytical analyses of multilayer networks highlight the importance of a detailed, context-aware mapping of different types of interactions to fully understand the interplay between structure and dynamics of such complex networks [142]. So far, most studies on biomolecular networks focus on structural network properties and a thorough understanding of their dynamical properties remains an important issue. The concept of dynamic controllability, for example, is well established in network theory [143, 144] and could in principle be applied to driving a cell from a disease state to a healthy state [143]. We expect that such network approaches will be key to designing advanced therapeutics for complex diseases that cannot be understood, nor treated, by a simple mono-causal molecular mechanism. The ultimate goal of network medicine is of course to contribute not only to basic research, but to the translation to benefit patients. Based on the pace at which network medicine is progressing, we are confident that this exciting and challenging goal will be reached rather sooner than later.

10.6 Exercises

To familiarize yourself with some basic network-based approaches to human diseases we will perform a rudimentary disease module analysis. The exemplary solution we provide is based on the programming language python and utilizes heavily the excellent networkx module, but of course other programming languages offer similar functionalities.

10.1 Constructing the interactome

- **(a)** Use one of the databases listed in Box 10.12 to construct an interactome network. We sugget using HIPPIE, as it allows for both programmatic access via an API or download of the entire dataset in an easy to parse text format.
- **(b)** Construct different networks with different parameters, such as different confidence scores or different experimental sources.
- **(c)** Perform a basic characterization of the overall topology of each network, e.g., overall coverage, degree distribution, number of isolated components, distribution of shortest pathlengths, clustering coefficient, etc.

10.2 Constructing a seed cluster for a particular disease

- **(a)** Use one of the databases listed in Box 10.12 to assemble a set of seed genes for a specific disease.
- **(b)** Place the seed genes on the interactome and determine the degree of localization using different measures from Box. 10.10.
- **(c)** Assess the statistical significance of the measured localization using different randomization schemes, both for the network topology and the seed genes (see Box. 10.11).

10.3 Constructing a disease module

- **(a)** Implement two different network-based gene prioritization algorithms introduced in Box 10.13.
- **(b)** Rank all genes in the interactome using both methods and with varying parameters of the respective algorithms.
- **(c)** Evaluate how the results change when removing various fractions of the seed genes.

10.4 Perform an enrichment analysis of the disease module

- **(a)** Use the databases listed in Box 10.12 to assemble an independent set of genes with potential relevance to the the disease, e.g., genes found to be differentially expressed in a patient cohort.
- **(b)** Test whether the ranked candidate genes are enriched for the genes of the independent validation set.
- **(c)** Perform a gene set enrichment analysis of the disease module using gene ontology to identify prominent biological processes within the module.

Note: Solutions are available to instructors at www.cambridge.org/bionetworks.

References

[1] Craig Venter J, Adams MD, Myers EW, et al. The sequence of the human genome. *Science*, 2001;291(5507):1304–1351.

[2] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*, 2001;409(6822):860–921.

[3] Smedley D, Köhler S, Czeschik JC, et al. Walking the interactome for candidate prioritization in exome sequencing studies of mendelian diseases. *Bioinformatics*, 2014;30(22):3215–3222.

[4] Leiserson MDM, Vandin F, Wu HT, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 2015;47(2):106–114.

[5] Chen Y, Zhu J, Lum PK, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 2008;452(7186):429–435.

[6] Pichlmair A, Kandasamy K, Alvisi G, et al. *Nature*, 2012;487:486–490.

[7] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 2007;3:140.

[8] Csermely P, Korcsmfiaros T, Kiss HJM, London G, Nussinov R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology & Therapeutics*, 2013;138:333–408.

[9] Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 2007;357:370–379.

[10] Colizza V, Barrat A, Barthfielemy M, Vespignani A. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences USA*, 2006;103:2015–2020.

[11] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proceedings of the National Academy of Sciences USA*, 2007;104(21):8685–8690.

[12] Landon BE, Keating NL, Barnett ML, et al. Variation in patient-sharing networks of physicians across the United States. *Journal of the American Medical Association*, 2012;308(3):265–273.

[13] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 2005;4:Article 17.

[14] De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLOS Computational Biology*, 2010;6:e1000807.

[15] Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. *Cell*, 2011;144:986–998.

[16] Menche J, Sharma A, Kitsak M, et al. Disease networks. Uncovering disease–disease relationships through the incomplete interactome. *Science*, 2015;347(6224):1257601.

[17] Thiele I, Palsson, BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 2010;5(1):93–121.

[18] Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 2011;12:56–68.

[19] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research,* 2017;45:D353–D361.

[20] Fabregat A, Jupe S, Matthews L, et al. The reactome pathway knowledgebase. *Nucleic Acids Research*, 2016;44:D481–D487.

[21] Swainston N, Smallbone K, Hefzi H, et al. Recon 2.2: From reconstruction to model of human metabolism. *Metabolomics*, 2016;12:109.

[22] Chan SY, Loscalzo J. The emerging paradigm of network medicine in the study of human disease. *Circulation Research*, 2012;111:359–374.

[23] Goldford JE, Hartman H, Smith TF, Segre D. Remnants of an ancient metabolism without phosphate. *Cell*, 2017;168:1126–1134.e9.

[24] Josephides C, Swain PS. Predicting metabolic adaptation from networks of mutational paths. *Nature Communications*, 2017;8:685.

[25] Li S, Sullivan NL, Rouphael N, et al. Metabolic phenotypes of response to vaccination in humans. *Cell*, 2017;169:862–877.e17.

[26] Klosik DF, Grimbs A, Bornholdt S, Hutt MT. The interdependent network of gene regulation and metabolism is robust where it needs to be. *Nature Communications*, 2017;8:534.

[27] Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. *Science* 2005;309(5740):1559–1563.

[28] Zhang Y. Gene regulatory networks: Real data sources and their analysis. In Iba H, Noman N, eds., *Evolutionary Computation in Gene Regulatory Network Research*. John Wiley & Sons, Inc.;2016, pp. 49–65.

[29] Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 2008;9:770–780.

[30] Blat Y, Kleckner N. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell*, 1999;98:249–259.

[31] Furey TS. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, 2012;13:840–852.

[32] Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 2015;43: D117–D122.

[33] Mathelier A, Fornes O, Arenillas DJ, et al. JASPAR 2016: A major expansion and update of the open access database of transcription factor binding profiles. *Nucleic Acids Research*, 2016;44:D110–D115.

[34] Moignard V, Woodhouse S, Haghverdi L, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, 2015;33:269–276.

[35] Goode DK, Obier N, Vijayabaskar MS, et al. Dynamic gene regulatory networks drive hematopoietic specification and differentiation. *Developmental Cell*, 2016;36:572–587.

[36] Marbach D, Lamparter D, Quon G, et al. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, 201613:366–370.

[37] Friedlander T, Prizak R, Barton NH, Tkačik G. Evolution of new regulatory functions on biophysically realistic fitness landscapes. *Nature Communications* 2017;8:216.

[38] GTEx Consortium. Human genomics: The Genotype-Tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 2015;348:648–660.

[39] De Smet R, Marchal K. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology* 2010;8:717–729.

[40] Weirauch MT. Gene coexpression networks for the analysis of DNA microarray data. In Dehmer M, Emmert-Streib F, Graber A, Salvador A, eds., *Applied Statistics for Network Biology*. Wiley-VCH Verlag GmbH & Co. KGaA;2011, pp. 215–250.

[41] Parikshak NN, Swarup V, Belgard TG, et al. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* 2016;540:423–427.

[42] Peters LA, Perriogue J, Mortha A, *et al.* A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nature Genetics*, 2017;49:1437–1449.

[43] Calabrese GM, Mesner LD, Stains JP, et al. Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. *Cell Systems*, 2017;4:46–59.e4.

[44] Guo X, Xiao H, Guo S, Dong L, Chen J. Identification of breast cancer mechanism based on weighted gene coexpression network analysis. *Cancer Gene Therapy*, 2017;24:333–341.

[45] Boucher B, Jenna S. Genetic interaction networks: Better understand to better predict. *Frontiers in Genetics*, 2013;4:290.

[46] Srivas R, Shen JP, Yang CC, et al. A network of conserved synthetic lethal interactions for exploration of precision cancer therapy. *Molecular Cell*, 2016;63:514–525.

[47] Costanzo M, VanderSluis B, Koch EN, et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 2016;353:aaf1420.

[48] Kramer MH, Farré JC, Mitra K, et al. Active interaction mapping reveals the hierarchical organization of autophagy. *Molecular Cell*, 2017;65:761–774.e5.

[49] Blomen VA, Májek P, Jae LT, et al. Gene essentiality and synthetic lethality in haploid human cells. *Science*, 2015;350:1092–1096.

[50] Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online mendelian inheritance in man (OMIM R), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 2015;43:D789–798.

[51] Lee DS, Park J, Kay KA, et al. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences USA*, 2008;105:9880–9885.

[52] Zhou X, Menche J, Barabasi, AL, Sharma A. Human symptoms–disease network. *Nature Communications*, 2014;5:4212.

[53] NIH: US National Library of Medicine. Medical subject headings. Available online at www.nlm.nih.gov/mesh/.

[54] Barrenas F, Chavali S, Holme P, Mobini R, Benson M. Network properties of complex human disease genes identified through genome-wide association studies. *PLOS ONE*, 2009;4:e8090.

[55] Zhang M, Zhu C, Jacomy A, Lu LJ, Jegga AG. The orphan disease networks. *American Journal of Human Genetics*, 2011;88:755–766.

[56] Pinero J, Berenstein A, Gonzalez-Perez A, Chernomoretz A, Furlong LI. Uncovering disease mechanisms through network biology in the era of next generation sequencing. *Science Reports*, 2016;6:24570.

[57] Hidalgo, CA, Blumm, N, Barabasi, AL, Christakis, NA. A dynamic network approach for the study of human phenotypes. *PLOS Computational Biology*, 2009;5:e1000353.

[58] Chmiel A, Klimek P, Thurner S. Spreading of diseases through comorbidity networks across life and gender. *New Journal of Physics*, 2014;16:115013.

[59] Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics,* 2016;17: 615–629.

[60] Duran-Frigola, M, Rossell, D, Aloy, P. A chemo-centric view of human health and disease. *Nature Communications*, 2014;5:5676.

[61] Gomez-Cabrero, D. Menche J, Vargas C, et al. From comorbidities of chronic obstructive pulmonary disease to identification of shared molecular mechanisms by data integration. *BMC Bioinformatics*, 2016;17:441.

[62] Klimek, P, Aichberger, S, Thurner, S. Disentangling genetic and environmental risk factors for individual diseases from multiplex comorbidity networks. *Science Reports*, 2016;6:39658.

[63] Pastor-Satorras, R, Castellano, C, Van Mieghem, P, Vespignani, A. Epidemic processes in complex networks. *Reviews of Modern Physics*, 2015;87: 925–979.

[64] Bernoulli, D. Essai d'une nouvelle analyse de la mortalité causée par la petite verole et des avantages de l'inoculation pour la prevenir. *Histoire de l'Academie Royale des Sciences (Paris) Avec les Mémoires de Mathematique & de Physique*, 1760;1:1–45.

[65] Hethcote HW. Three basic epidemiological models. *Applied Mathematical Ecology*, 1989;18:119–144.

[66] Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1927;115:700–721.

[67] Longini Jr IM, Nizam A, Xu S, et al. Containing pandemic influenza at the source. *Science*, 2005;309:1083–1087.

[68] Granell C, Gomez S, Arenas A. Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Physical Review Letters*, 2013;111:128701.

[69] Hufnagel L, Brockmann D, Geisel T. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences USA*, 2004;101:15124–15129.

[70] Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 2013;342:1337–1342.

[71] Verdery AM, Siripong N, Pence BW. Social network clustering and the spread of HIV/AIDS among persons who inject drugs in two cities in the Philippines. *Journal of Acquired Immune Deficiency Syndromes*, 2017;76:26–32.

[72] Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*, 1999;286:509–512.

[73] Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. *Physical Review Letters*, 2001;86,:3200–3203.

[74] Pastor-Satorras R, Vespignani A. Immunization of complex networks. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2002;65: 036104.

[75] Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. *Reviews of Modern Physics*, 2009;81:591–646.

[76] Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 2008;358:2249–2258.

[77] Fowler JH, Christakis NA. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham heart study. BMJ, 2008;337: a2338.

[78] Cozzo E, Banos RA, Meloni S, Moreno Y. Contact-based social contagion in multiplex networks. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2013;88:050801.

[79] Boccaletti S, Bianconi G, Criado R, et al. The structure and dynamics of multilayer networks. *Physics Reports*, 2014;544:1–122.

[80] Kivelä M, Arenas A, Barthelemy M, et al. Multilayer networks. *Journal of Complex Networks*, 2014;2:203–271.

[81] Noh JD, Rieger H. Random walks on complex networks. *Physical Review Letters*, 2004;92:118701.

[82] Bader GD, Cary MP, Sander C. Pathguide: A pathway resource list. *Nucleic Acids Research*, 2006;34:D504–D506.

[83] Szklarczyk, D. et al. The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 2017;45:D362–D368.

[84] Chatr-Aryamontri A, Oughtred R, Boucher L. et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 2017;45:D369–D379.

[85] Orchard S, Ammari M, Aranda B, et al. The MIntAct project: IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 2014;42:D358–63.

[86] Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: Enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 2017;45:D408–D414.

[87] Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. *Cell*, 2014;159:1212–1226.

[88] Huttlin EL, Ting L, Bruckner RJ, et al. The BioPlex network: A systematic exploration of the human interactome. *Cell*, 2015;162:425–440.

[89] Huttlin EL, Bruckner RJ, Paulo JA, et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 2017;545:505–509.

[90] Zhang QC, Petrey D, Deng L, et al. Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 2012;490:556–560.

[91] Jansen, R. Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, 2003;302:449–453.

[92] Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein–protein interaction networks and biology: What's the connection? *Nature Biotechnology*, 2008;26:69–72.

[93] Gillis J, Ballouz S, Pavlidis P. Bias tradeoffs in the creation and analysis of protein–protein interaction networks. *Journal of Proteomics*, 2014;100:44–54.

[94] Caldera M, Buphamalai P, Muller F, Menche J. Interactome-based approaches to human disease. *Current Opinion in Systems Biology*, 2017;3:88–94.

[95] Clauset A, Shalizi C, Newman M. Power-law distributions in empirical data. *SIAM Review*, 2009;51:661–703.

[96] Watts DJ, Strogatz SH. Collective dynamics of "small-world" networks. *Nature*, 1998;393:440–442.

[97] Cohen R, Havlin S. Scale-free networks are ultrasmall. *Physical Review Letters*, 2003;90:058701.

[98] Callaway DS, Newman ME, Strogatz SH, Watts DJ. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 2000;85:5468.

[99] Newman ME, Strogatz SH, Watts, DJ. Random graphs with arbitrary degree distributions and their applications. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2001;64:026118.

[100] Cohen R, Erez K, Ben-Avraham D, Havlin S. Resilience of the internet to random breakdowns. *Physical Review Letters*, 2000;85:4626.

[101] Dorogovtsev SN, Mendes JF. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press;2003.

[102] Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. *Nature*, 2000;406:378–382.

[103] Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature,* 2001;411:41–42.

[104] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature*, 1999;402:C47–52.

[105] Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences USA*, 2003;100:12123–12128.

[106] Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 2004;5:101–113.

[107] Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences USA*, 2008;105:4323–4328.

[108] Ghiassian SD, Menche J, Barabasi AL. A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLOS Computational Biology,* 2015;11:e1004120.

[109] Fortunato S. Community detection in graphs. *Physics Reports*, 2010;486:75–174.

[110] Guney E, Menche J, Vidal M, Barabasi AL. Network-based in silico drug efficacy screening. *Nature Communications*, 2016;7:10331.

[111] Newman ME. The structure and function of complex networks. *SIAM Review*, 2003;45:167–256.

[112] Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science*, 2002;296, 910–913.

[113] Milo R, Kashtan N, Itzkovitz S, Newman MEJ, Alon U. On the uniform generation of random graphs with prescribed degree sequences. arXiv preprint available at https://arxiv.org/pdf/cond-mat/0312028.pdf. 2004.

[114] Bender EA, Canfield ER. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 1978;24:296–307.

[115] Bollobás B. Random graphs. In *Graph Theory*, 123–145 (Springer, 1979).

[116] Serrano MA, Boguna M. Tuning clustering in random networks with arbitrary degree distributions. *Phys. Rev. E* 72, 036133 (2005).

[117] Boguna, M, Pastor-Satorras, R. Class of correlated random networks with hidden variables. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2003;68:036112.

[118] Weber S, Porto M. Generation of arbitrarily two-point-correlated random networks. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 2007;76:046111.

[119] Samal A, Martin OC. Randomizing genome-scale metabolic networks. *PLOS ONE*, 2011;6:e22295.

[120] Basler G, Ebenhoh O, Selbig J, Nikoloski Z. Mass-balanced randomization of metabolic networks. *Bioinformatics*, 2011;27:1397–1403.

[121] Vissers LELM, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. *Nature Reviews Genetics*, 2016;17:9–18.

[122] Krishnan A, Zhang R, Yao V, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, 2016;19: 1454–1462.

[123] Sharma A, Menche J, Chris Huang C, et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Human Molecular Genetics*, 2015;24: 3005–3020.

[124] Barshir R, Shwartz O, Smoly IY, Yeger-Lotem E. Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLOS Computational Biology,* 2014;10:e1003632.

[125] Magger O, Waldman YY, Ruppin E, Sharan R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLOS Computational Biology,* 2012;8:e1002690.

[126] Li M, Zhang J, Liu Q, Wang J, Wu FX. Prediction of disease-related genes based on weighted tissue-specific networks by using DNA methylation. *BMC Medical Genomics*, 2014;7(Suppl 2):S4.

[127] Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein–protein interactions. *Journal of Medical Genetics*, 2006;43:691–698.

[128] Wang XD, Huang JL, Yang L, et al. Identification of human disease genes from interactome network using graphlet interaction. *PLOS ONE* 2014;9:e86142.

[129] Erten S, Bebek G, Ewing RM, Koyuturk M, et al. DADA: Degree-aware algorithms for network-based disease gene prioritization. *BioData Mining*, 2011;4:19.

[130] Guney E, Oliva B. Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLOS ONE*, 2012;7:e43557.

[131] Bailly-Bechet M, Borgs C, Braunstein A, et al. Finding undetected protein associations in cell signaling by belief propagation. *Proceedings of the National Academy of Sciences USA*, 2011;108:882–887.

[132] Tuncbag N, McCallum S, Huang SSC, Fraenkel E. SteinerNet: A web server for integrating ’omic’ data to discover hidden components of response pathways. *Nucleic Acids Research*, 2012;40:W505–W509.

[133] Tuncbag N, Gosline SJC, Kedaigle A, et al. Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLOS Computational Biology,* 2016;12:e1004879.

[134] Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular triangulation: Bridging linkage and molecular-network information for

identifying candidate genes in Alzheimer's disease. *Proceedings of the National Academy of Sciences USA*, 2004;101:15148–15153.

[135] Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLOS Computational Biology*, 2010;6:e1000641.

[136] Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Compututational Biol*ogy, 2011;18:507–522.

[137] Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: A universal amplifier of genetic associations. *Nature Reviews Genetics*, 2017;18:551–562.

[138] Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *The American Journal of HumanGenetics* 82, 949–958 (2008).

[139] Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. *PLOS Computational Biology,* 2009;5:e1000443.

[140] Van Erp M, Schomaker, L. Variants of the borda count method for combining ranked classifier hypotheses. In Schomaker L, Vuurpijl L, eds., *Proceedings 7th International Workshop on Frontiers in Handwriting Recognition*. International Unipen Foundation;2000, pp. 443–452.

[141] Loscalzo J, Barabasi AL, Silverman EK, eds. *Network Medicine: Complex Systems in Human Disease and Therapeutics*. Harvard University Press;2017.

[142] De Domenico M, Granell C, Porter MA, Arenas A. The physics of spreading processes in multilayer networks. *Nature Physics*, 2016;12:901–906.

[143] Liu YY, Slotine JJ, Barabasi AL. Controllability of complex networks. *Nature*, 2011;473:167–173.

[144] Liu YY, Slotine JJ, Barabasi AL. Observability of complex systems. *Proceedings of the National Academy of Sciences USA*, 2013;110:2460–2465.