

Computational Biology

Fabricio Alves Barbosa da Silva  
Nicolas Carels  
Marcelo Trindade dos Santos  
Francisco José Pereira Lopes *Editors*

# Networks in Systems Biology

Applications for Disease Modeling



 Springer

# Computational Biology

Volume 32

## **Advisory Editor**

Gordon Crippen, University of Michigan, Ann Arbor, MI, USA

## **Editor-in-Chief**

Andreas Dress, CAS-MPG Partner Institute for Computational Biology, Shanghai, China

## **Editorial Board**

Robert Giegerich, University of Bielefeld, Bielefeld, Germany

Janet Kelso, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

## **Editor-in-Chief**

Michal Linial, Hebrew University of Jerusalem, Jerusalem, Israel

## **Advisory Editor**

Joseph Felsenstein, University of Washington, Seattle, WA, USA

## **Editor-in-Chief**

Olga Troyanskaya, Princeton University, Princeton, NJ, USA

## **Advisory Editor**

Dan Gusfield, University of California, Davis, CA, USA

## **Editorial Board**

Gene Myers, Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

## **Advisory Editor**

Sorin Istrail, Brown University, Providence, RI, USA

## **Editorial Board**

Pavel Pevzner, University of California, San Diego, CA, USA

**Editor-in-Chief**

Martin Vingron, Max Planck Institute for Molecular Genetics, Berlin, Germany

**Advisory Editors**

Thomas Lengauer, Max Planck Institute for Computer Science, Saarbrücken, Germany

Marcella McClure, Montana State University, Bozeman, MT, USA

Martin Nowak, Harvard University, Cambridge, MA, USA

David Sankoff, University of Ottawa, Ottawa, ON, Canada

Ron Shamir, Tel Aviv University, Tel Aviv, Israel

Mike Steel, University of Canterbury, Christchurch, New Zealand

Gary Stormo, Washington University in St. Louis, St. Louis, MO, USA

Simon Tavaré, University of Cambridge, Cambridge, UK

Tandy Warnow, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Lonnie Welch, Ohio University, Athens, OH, USA

Endorsed by the *International Society for Computational Biology*, the *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

More information about this series at <http://www.springer.com/series/5769>


Fabricio Alves Barbosa da Silva ·  
Nicolas Carels · Marcelo Trindade dos Santos ·  
Francisco José Pereira Lopes  
Editors


# Networks in Systems Biology

Applications for Disease Modeling

 Springer

*Editors*

Fabricio Alves Barbosa da Silva   
Scientific Computing Program (PROCC)  
Oswaldo Cruz Foundation  
Rio de Janeiro, Brazil

Nicolas Carels   
CDTS  
Oswaldo Cruz Foundation  
Rio de Janeiro, Brazil

Marcelo Trindade dos Santos  
Department of Computational Modeling  
National Laboratory of Scientific Computing  
Petrópolis, Rio de Janeiro, Brazil

Francisco José Pereira Lopes  
Graduate Program in Nanobiosystems  
Federal University of Rio de Janeiro  
Duque de Caxias, Rio de Janeiro, Brazil

ISSN 1568-2684

ISSN 2662-2432 (electronic)

Computational Biology

ISBN 978-3-030-51861-5

ISBN 978-3-030-51862-2 (eBook)

<https://doi.org/10.1007/978-3-030-51862-2>

© Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Foreword

Since the early 60s, when Prigogine puts forward the emergent properties of complex chemical reaction networks in non-equilibrium and non-linear conditions, researchers around the world try to apply such ideas to describe the highly complex biological networks. Especially, a greater effort has been made since the beginning of the twenty-first century due to the development of omics sciences, when a huge set of global data on biological systems became available. It also appears that these technical improvements in sequencing and detecting biological molecules have propelled classical biological studies in a much more quantitative way. Furthermore, massive and universal computational resources available now have contributed crucially to what we can call a “new” model-driven quantitative biology. In this context, representation, visualization, analysis, and modeling of the topological and dynamic properties of the complex biological networks are in order.

This volume shows relevant aspects of this new area, with contributions mainly from Brazilian groups, trying to describe gene expression, metabolic, and signaling networks, as well as the brain functioning and epidemiological models. In the following chapters, complex networks are explored not only from the point of view of inference but also from dynamics and time evolution, pointing out the emerging properties of biological systems. In the first part, theoretical and computational analysis of complex biological networks is reviewed, involving visualization, inference, topological, and differential analysis, as well as modeling time evolution and sensitivity of biological processes. In the second part, the emphasis is on the application of these methods to investigate infectious and degenerative diseases, including cancer, aimed at a better understanding of the evolution of diseases and searching for relevant pharmacological targets and biomarkers.

Sophisticated mathematical and computational tools are necessary to understand the intricate processes occurring in biological networks at different levels, from the regulation of gene expression and metabolic networks into each cell, as well as signaling at the intracellular level and between cells and organisms. A broad view of modeling of these processes is presented by the authors, pointing out the importance of this approach in the rational design of new drugs, innovative gene, and immune therapies, and also in advancing the concept of personalized medicine.

Although recognized as being in the early stages, the power of this approach is well demonstrated in this volume. In fact, we can imagine much broader and new applications in this area, especially related to the new trends arising from the current revolution in information technology, such as the promising resources of the “internet of things.” Also, surprising is the number and quality of Brazilian groups involved in this area, showing a very promising evolution of research in our country.

Finally, we can point out that the current pandemic of the new coronavirus, which required rapid and accurate responses, mobilized, as expected, scientists from around the world who have largely employed tools like those discussed here, emphasizing once again the importance of these contributions.

Paulo Mascarello Bisch  
Institute of Biophysics “Carlos Chagas Filho”  
Federal University of Rio de Janeiro  
Rio de Janeiro, Brazil

# Preface

In the last decades, we have witnessed a transition from descriptive biology to a systemic understanding of biological systems that was possible due to the impressive progress in high-throughput technologies. The wave of data produced by these technologies is tremendous and offered an opportunity for big data as well as mathematical and computational modeling to take off. Systems Biology is a rapidly expanding field and comprises the study of biological systems through mathematical modeling and analysis of large volumes of biological data. Now we testify exciting times where sciences integrate one another for the benefit of solving specific problems. Of course, medical sciences do not escape this trend, and we have to follow these developments for participating and translating them into medical applications as well as for transmitting them to the next generations. Indeed, the transmission of knowledge on cutting-edge developments in System Biology was the purpose of the III International Course on Theoretical and Applied Aspects of Systems Biology, held in Rio de Janeiro in July 2019, whose contributions are now translated into the present book.

This book presents current research topics on biological network modeling, as well as its application in studies on human hosts, pathogens, and diseases. The chapters were written by renowned experts in the field. Some topics discussed in-depth here include networks in systems biology, computational modeling of multidrug-resistant bacteria, and systems biology of cancer. It is intended for researchers, advanced students, and practitioners of the field. Chapters are research-oriented and present some of the most recent results related to each topic.

This book is organized into two main sections: Biological Networks and Methods in Systems Biology and Disease and Pathogen Modeling. Although the whole book is made of contributions from researchers with a clear commitment to applied sciences, the first part brings chapters where the more fundamental aspects of biological networks in systems biology are addressed. The remaining chapters on the second part of the book deal with the application of such fundamentals in disease modeling.



We take the opportunity to acknowledge the Brazilian Coordination for Improvement of Higher-Level Personnel (CAPES), FIOCRUZ's Vice-Presidency of Education, Information and Communication, the National Laboratory for Scientific Computing, and the Computational and Systems Biology Graduate Program of IOC/FIOCRUZ that made possible this event to occur with their financial and logistics support. Finally, we cannot emphasize enough how thankful we are for all authors contributing to the book for their dedication and generosity. A special thanks to Professor Paulo Bisch for writing such a splendid foreword that much honored us and contributed to further elevate this book.

Rio de Janeiro, Brazil  
Rio de Janeiro, Brazil  
Petrópolis, Brazil  
Duque de Caxias, Brazil

Fabricio Alves Barbosa da Silva  
Nicolas Carels  
Marcelo Trindade dos Santos  
Francisco José Pereira Lopes

# Contents

## Part I Biological Networks and Methods in Systems Biology

<b>1</b>	<b>Network Medicine: Methods and Applications</b> . . . . .	<b>3</b>
	Italo F. do Valle and Helder I. Nakaya	
<b>2</b>	<b>Computational Tools for Comparing Gene Coexpression Networks</b> . . . . .	<b>19</b>
	Vinícius Carvalho Jardim, Camila Castro Moreno, and André Fujita	
<b>3</b>	<b>Functional Gene Networks and Their Applications</b> . . . . .	<b>31</b>
	Hong-Dong Li and Yuanfang Guan	
<b>4</b>	<b>A Review of Artificial Neural Networks for the Prediction of Essential Proteins</b> . . . . .	<b>45</b>
	Kele Belloze, Luciana Campos, Ribamar Matias, Ivair Luques, and Eduardo Bezerra	
<b>5</b>	<b>Transcriptograms: A Genome-Wide Gene Expression Analysis Method</b> . . . . .	<b>69</b>
	Rita M. C. de Almeida, Lars L. S. de Souza, Diego Morais, and Rodrigo J. S. Dalmolin	
<b>6</b>	<b>A Tutorial on Sobol' Global Sensitivity Analysis Applied to Biological Models</b> . . . . .	<b>93</b>
	Michel Tosin, Adriano M. A. Côrtes, and Americo Cunha	
<b>7</b>	<b>Reaction Network Models as a Tool to Study Gene Regulation and Cell Signaling in Development and Diseases</b> . . . . .	<b>119</b>
	Francisco José Pereira Lopes, Claudio Daniel Tenório de Barros, Josué Xavier de Carvalho, Fernando de Magalhães Coutinho Vieira, and Cristiano N. Costa	

## Part II Disease and Pathogen Modeling

<b>8</b>	<b>Challenges for the Optimization of Drug Therapy in the Treatment of Cancer</b> .....	163
	Nicolas Carels, Alessandra Jordano Conforte, Carlyle Ribeiro Lima, and Fabricio Alves Barbosa da Silva	
<b>9</b>	<b>Opportunities and Challenges Provided by Boolean Modelling of Cancer Signalling Pathways</b> .....	199
	Petronela Buiga and Jean-Marc Schwartz	
<b>10</b>	<b>Integrating Omics Data to Prioritize Target Genes in Pathogenic Bacteria</b> .....	217
	Marisa Fabiana Nicolás, Maiana de Oliveira Cerqueira e Costa, Pablo Ivan P. Ramos, Marcelo Trindade dos Santos, Ernesto Perez-Rueda, Marcelo A. Marti, Dario Fernandez Do Porto, and Adrian G. Turjanski	
<b>11</b>	<b>Modelling Oxidative Stress Pathways</b> .....	277
	Harry Beaven and Ioly Kotta-Loizou	
<b>12</b>	<b>Computational Modeling in Virus Infections and Virtual Screening, Docking, and Molecular Dynamics in Drug Design</b> . . . .	301
	Rachel Siqueira de Queiroz Simões, Mariana Simões Ferreira, Nathalia Dumas de Paula, Thamires Rocco Machado, and Pedro Geraldo Pascutti	
<b>13</b>	<b>Cellular Regulatory Network Modeling Applied to Breast Cancer</b> .....	339
	Luiz Henrique Oliveira Ferreira, Maria Clícia Stelling de Castro, Alessandra Jordano Conforte, Nicolas Carels, and Fabrício Alves Barbosa da Silva	
	<b>Index</b> .....	367

**Part I**  
**Biological Networks and Methods**  
**in Systems Biology**

# Chapter 1

## Network Medicine: Methods and Applications



Italo F. do Valle and Helder I. Nakaya

**Abstract** The structure and function of biological systems are determined by a complex network of interactions among cell components. Network medicine offers a toolset for us to systematically explore perturbations in biological networks and to understand how they can spread and affect other cellular processes. In this way, we can have mechanistic insights underlying diseases and phenotypes, evaluate gene function in the context of their molecular interactions, and identify molecular relationships among apparently distinct phenotypes. These tools have also enabled the interpretation of heterogeneity among biological samples, identification of drug targets and drug repurposing as well as biomarker discovery. As our ability to profile biological samples increases, these network-based approaches are fundamental for data integration across the genomic, transcriptomic, and proteomic sciences. Here, we review and discuss the recent advances in network medicine, exploring the different types of biological networks, several methods, and their applications.

**Keywords** Network medicine · Graph theory · High-throughput technologies

### 1.1 Introduction

High-throughput technologies such as next-generation sequencing, mass spectrometry, and high-dimensional flow cytometry have revolutionized medicine. By providing the molecular and cellular profile of patients, these technologies can help physicians into their medical decisions. For instance, the analysis of whole-genome

---

I. F. do Valle

Center for Complex Network Research, Department of Physics, Northeastern University, 11th Floor, 177 Huntington Avenue, Boston, MA 02115, USA

H. I. Nakaya (✉)

Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil  
e-mail: [hnakaya@usp.br](mailto:hnakaya@usp.br)

Scientific Platform Pasteur USP, São Paulo, Brazil

Av. Prof. Lúcio Martins Rodrigues, 370, block C, 4th floor, São Paulo, SP 05508-020, Brazil

© Springer Nature Switzerland AG 2020

F. A. B. da Silva et al. (eds.), *Networks in Systems Biology*, Computational Biology 32, [https://doi.org/10.1007/978-3-030-51862-2\\_1](https://doi.org/10.1007/978-3-030-51862-2_1)

sequencing allows the identification of mutations associated with a disease or a response to treatment. It is also possible to measure the activity of tens of thousands of genes, proteins, and metabolites to find the set of markers capable of predicting a medical outcome. Another example is the analysis of DNA methylation patterns in liquid biopsies that can reveal the presence of tumors in the early stages of the disease [1]. However, just having this comprehensive catalog of a patient's genes and biological components is often not sufficient to understand the mechanisms of human diseases.

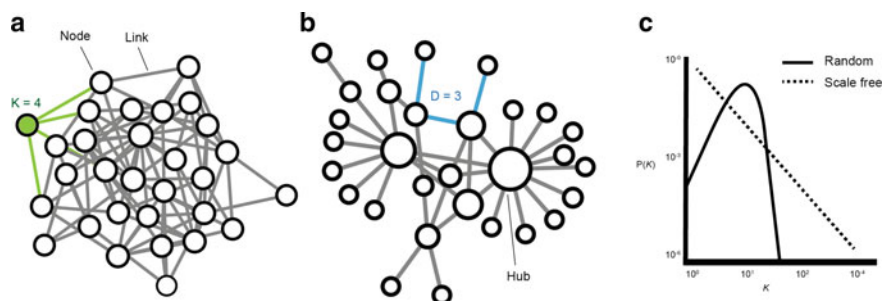
Network medicine studies the interactions among molecular components to better understand the pathogenesis of a disease. The underlying idea is that a cell can be thought as networks of interacting biomolecules, and a disease is can be seen as a “malfunctioning” in one or more regions in human biological networks [2]. A mutation that affects the correct functioning of a single protein will interfere not only with the function of that specific protein, but also with the proper functioning of many other proteins that are connected to it. Network medicine uses graph theory to analyze how networks behave in the context of a disease and one of its aims is to identify the key players related to the disease.

In this chapter, we will describe the types of biological networks and the main methods of analysis utilized in network medicine. We will also address the techniques that identify subnetworks and gene modules associated with human diseases. Finally, we will show how drug treatment can affect the network behavior.

### ***1.1.1 Basic Concepts in Graph Theory***

A network (or a graph) is a catalog of a system's components often called nodes or vertices and the interactions between them, called links or edges. In biological networks, nodes represent biomolecules, such as proteins, metabolites, and genes, while links represent different types of biological interactions between them, such as physical binding, enzymatic reaction, or transcriptional regulation. Link networks can be directed, like in the interaction where a transcription factor activates a given target gene, or undirected, where the interaction is bidirectional, like in a physical interaction between two proteins.

A key property of a node is its degree, which is equal to its total number of connections. Depending on the network, it can represent the number of proteins a given protein binds to or the number of reactions a given metabolite participates in. For directed networks, such as regulatory networks, the degree can be differentiated in outgoing degree, the number of nodes it points to, or incoming degree, and the number of nodes that point to it. The degree distribution of a network, which gives the probability  $P(k)$  that a selected node has exactly  $k$  links, is important to understand how the network works. For example, a peaked degree distribution indicates that a system has a characteristic degree from which most of the nodes do not highly deviate from (Fig. 1.1a, c). By contrast, most networks found in nature, are characterized by a power-law degree distribution, which means that most nodes have a few interactions



**Fig. 1.1** Example of a random non-scale-free network (a) and of a scale-free network (b), together with a schematic representation of their degree distributions (c). A node with degree  $K = 4$  is highlighted in green, and a shortest path of length  $D = 3$  is highlighted in blue

and that these coexist with a few highly connected nodes, the hubs, that hold the whole network together (Fig. 1.1b, c). Networks with this property are usually referred to as scale-free network. These are typical of several real-world systems, and this degree distribution implies important properties of these systems' behavior, such as the high robustness against accidental node failures [3].

Complex networks are also often characterized by the small-world property [4]. This means that any pair of nodes can be connected by relatively short paths. In biological networks, this property indicates that, for example, most proteins (or metabolites) are only a few interactions (or reactions) from any other protein (metabolite) [5–7]. Therefore, perturbing the state of a given node can affect the activity of several others in their vicinity.

## 1.2 Biological Networks

Cells are comprised of complex webs of molecular interactions between cell components [8]. These interactions form complex networks or interactomes, and many experimental approaches have been developed to completely map them. These approaches include (1) curation of existing data available in the literature (literature curation), (2) computational predictions based on different information such as sequence similarity and evolutionary conservation, and (3) systematic and unbiased high-throughput experimental strategies applied at the scale of whole genomes or proteomes. The networks derived by each of these methods have their own biases and limitations that should be carefully taken into account during computational analysis. Here, we discuss a few examples of biological networks and their respective properties.

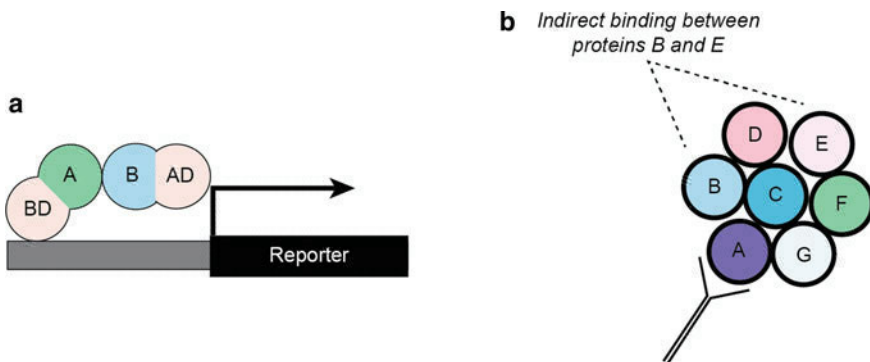
### 1.2.1 Protein–Protein Interaction (PPI) Networks

PPIs are undirected networks in which nodes represent proteins and edges represent a physical interaction between two proteins. Two main methodologies are used for large-scale interaction mapping: yeast-two hybrid (Y2H) and affinity purification followed by mass spectrometry (AP/MS) (Fig. 1.2).

In the Y2H technique, a transcription factor is split into its two components: the binding domain (BD), which binds to the DNA sequence, and the activation domain (AD), which activates the transcription. DNA recombinant tools are used to create chimeric proteins in which one protein of interest (prey) is fused to the transcription factor BD, while the other protein of interest (bait) is fused to the AD. If the prey and bait proteins physically interact, the transcription factor is reconstructed and is then able to activate the transcription of a reporter gene, which will create an indicator that the interaction occurred (Fig. 1.2a). In AP/MS, a protein of interest (bait) is purified from a cell lysate (often referred to as pull-down), and co-purified proteins (preys) are identified through mass spectrometry (Fig. 1.2b). Mappings derived from Y2H contain physical interactions, while AP/MS ones contain co-complex information—that is, the interactions can be either physical or indirect.

Several high-throughput mappings have been used to map protein interactomes in humans and model organisms. The most recent efforts for human interactomes include the Human Reference Interactome (HuRI) [9], mapped by Y2H, and BioPlex2.0 [10], and mapped using AP/MS. Several databases with literature-curated PPIs are available, and a few efforts have been made to produce high-quality interactomes derived from literature-curated data [11, 12].

Literature-curated PPIs are inherently biased toward heavily studied proteins: most interactions occur among genes characterized by many publications and their network is depleted of interactions for proteins with few or no publications [13].



**Fig. 1.2** Schematic representations of the techniques used to detect protein–protein interactions: **a** Yeast-two hybrid and **b** affinity purification of protein complexes. BD: DNA-binding domain, AD: transcription activation domain, A-B hypothetical proteins



The analysis of such networks may lead to incorrect conclusions, such as the previously reported correlation between number of interaction partners (degree) and gene essentiality [9, 13].

Small overlaps are observed among protein interactomes, even those derived by unbiased and systematic studies, which can be partially attributed to the different properties of their respective experimental strategies. For example, PPIs have different binding affinities, which may or may not be in the range of detectability for that specific method. Other factors might include fusion constructs, washing buffer, and protein expression in the cell.

### ***1.2.2 Gene Regulatory Networks***

In gene regulatory networks, nodes are transcription factors (TFs) (and/or miRNAs) and their targets, and directed edges exist between TFs (miRNA) and their targets. The most common approach for detection of regulatory interactions is chromatin immunoprecipitation (ChIP-seq)-based approaches: DNA-binding proteins are cross-linked with the DNA, an antibody is used to immunoprecipitate the protein of interest, and DNA sequencing strategies are used to identify the genomic regions where the protein binds to. The human regulatory network derived in this way by the ENCODE project revealed important features of cellular regulation: hierarchical organization of TFs in which top-level factors more strongly influence expression, while middle-level TFs co-regulate several targets. These properties avoid information-flow bottlenecks and allow the presence of feed-forward network motifs. It was also possible to observe stronger evolutionary and allele-specific activity of the most connected network components [14].

Other strategies also take advantage of DNase-Seq to identify regions that can be occupied by TFs and then identify these TFs by their binding motifs for that particular genomic region. The mapping of the regulatory networks across 41 cell lines using this technique has revealed that networks are markedly cell-specific and even TFs that are expressed across cells of a given lineage show distinctive regulatory roles in the different cells [15].

Several approaches have been developed for reverse engineering cellular networks from gene expression data. Most of these methods are based on the notion of similarity among co-expression across different experimental conditions. Methods based on measures based on correlation, mutual information and graphical models (including Bayesian networks) identify undirected edges between nodes by capturing probabilistic dependences of different kinds [16].

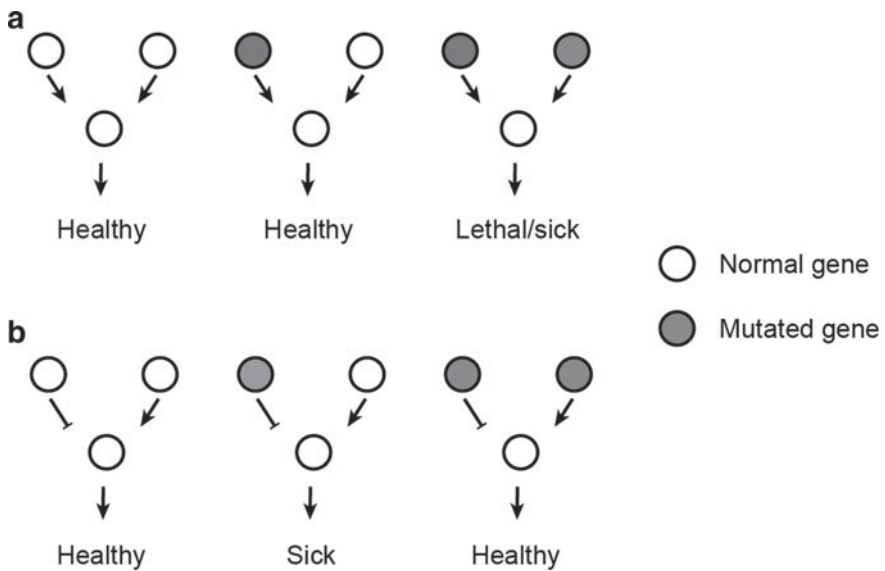
### 1.2.3 Metabolic Networks

Metabolic networks attempt to describe biochemical reactions for a particular cell or organism. In most representations, nodes are metabolites and edges are the reactions that convert one metabolite into another. In this case, edges can be directed or undirected, depending on the reversibility of the reaction. Other representations are also possible, such as nodes as metabolites and edges representing co-participation in the same biochemical reactions.

Network reconstruction involves manual curation of literature data describing experimental results on metabolic reactions as well as predicted reactions derived from orthologous enzymes experimentally characterized in other species [17].

### 1.2.4 Genetic Interaction Networks

Genetic interactions (GIs) are functional relationships between genes, and they can be classified into positive and negative interactions. In negative GIs, the observed fitness by mutating a pair of genes at the same time (double mutants) is worse than what is expected when mutating genes individually (single mutants) (Fig. 1.3a). In positive GIs, a gene mutation can mitigate the effect caused by another mutation, and the double mutant is healthier than most sick of the single mutants (Fig. 1.3b). Mapping GIs allows us to understand the mechanisms underlying robustness of



**Fig. 1.3** Schematic representation of negative (a) and positive (b) genetic interactions

biological systems and how compensatory mechanisms emerge after perturbations. Recent technological advances have made possible the systematic mapping of genetic interactions in yeast, *C. elegans* and human cells. These maps can be represented as networks in which nodes are genes and an edge exists between genes that have high similarity in their genetic interaction profiles. Genetic interaction networks have enabled the understanding of the hierarchical dependencies of cell functions, identification of functionally related processes and pleiotropic genes [18–20].

### 1.2.5 Pathogen–Host Interactomes

Pathogens have complex mechanisms to perturb host intracellular networks to their advantage and the understanding of parasite–host interactomes could provide important insights for the development of treatment strategies. For instance, it has been observed that pathogen’s proteins preferentially target hubs in human and plant interactomes [21, 22]. Systematic maps capturing viral–host protein–protein interactions have been obtained for Epstein–Barr virus [23], hepatitis C virus [23], herpesviruses [24], influenza [25], HIV [26], and others [27]. Other pathogen–host interactomes have been assembled or predicted for bacteria [28], fungi [29], worms, such as *Schistosoma mansoni* [30], and several protozoans, such as *Leishmania* [30], *Plasmodium* [30–32], and *Trypanosoma* [30].

## 1.3 Biological Networks for Functional Annotations of Proteins and Complexes

The network neighborhood of a protein reflects several of its properties: cellular localization, biological, and molecular function. Therefore, the most basic assumption is that proteins that are close to each other and/or share many neighbors in the interactomes are more likely to have a similar function. This “guilt-by-association” principle underlies many network-based methods for protein function prediction.

An example of this principle can be observed in the recent demonstration that PPI networks can be used to predict protein subcellular localization [9]. The authors showed that extracellular vesicle (EV) proteins form a significant subnetwork in the PPI network. The interaction partners of the EV subnetwork already included many proteins with established roles in EV biogenesis and cargo recruitment, and the other interaction partners with unknown subcellular localization were ranked as potential EV proteins based on the number of interactions they shared with the EV subnetwork. Experimental validation demonstrated that candidate proteins were indeed related to extracellular vesicle functions [9].

Several network-based indexes attempt to quantify the size of neighborhood that is shared between two proteins—these are often referred to in network science as node

similarity indexes. Common similarity indices for pairs of nodes take into account the number of shared interaction between the nodes normalized by their total number of interactions (Jaccard index), or by the smallest degree of either node (Simpson index) or by the product of the individual node degrees (geometric and cosine indexes) [33]. Uncharacterized proteins can be ranked based on their similarity indexes with proteins of known function, and the high-ranking proteins can be annotated to the same function. Exploiting the same principle on genetic interaction (GI) networks has been shown to be very effective for the discovery of functional complexes, control, and regulatory strategies, as well as unrecognized biosynthetic pathways [34].

Other approaches for the prediction of protein function take into account the full topology of the network [35]. Flow-based approaches consider each protein annotated to a given function as the source of a “functional flow”. After simulating the spread of this flow over time through the network, each unannotated protein is assigned a score that is proportional to the amount of flow it received during the simulation [36, 37]. A recent distance metric based on network diffusion that is able to capture similarities based on multiple paths in the network has been shown to provide finer grained distinctions when transferring functional annotation in PPI networks [38]. Other approaches integrate PPI network data with high-throughput biological data, creating functional networks for the predictions [39].

More recently, algorithms based on network embedding have also been used for the prediction of protein function. In classic versions of these approaches, matrix factorization leads to a representation of network nodes as vectors in a low-dimensional space while preserving the neighborhood similarity between nodes [40]. A network embedding algorithm has been applied in a multi-layer network, where each layer represents protein interactions in a different tissue, to provide predictions of cellular function that take into account tissue-specific protein functions [41].

## 1.4 Biological Networks and Diseases

### 1.4.1 *Disease Genes and Subnetworks*

Biological networks provide us with a unique opportunity to study disease mechanisms in a holistic manner. The interconnectivity between cellular components—genes, proteins, and metabolites—implies that the effect of specific perturbations, like mutations, will spread through the network to areas not originally affected. Biological networks provide us with the context for a given gene or protein which is essential in determining the phenotypic effects of perturbations [2].

Protein–protein interaction networks have been extensively used to study disease mechanisms. It has been observed that proteins genetically associated to a given disease tend to be colocalized in a given neighborhood of the network, forming a connected subgraph, often referred to as *disease module*. Thus, the disease module indicates a region in the network that, if perturbed, leads to the disease phenotype

[2]. In asthma, for example, out of 129 asthma-related genes, 37 formed a connected subgraph, or disease module [42]. In order to measure whether the disease module could have emerged by chance, 129 genes were randomly selected from the network and the size of the largest connected component formed by these genes is registered. This process is repeated through several iterations, usually 1,000 times, producing a null distribution. This null distribution can be then used as reference to compare the module size observed from the disease genes, providing an empirical p-value: the proportion of random iterations that produced a module size equal or greater than the real observation [42]. The biological significance of the disease modules can be verified by the fact that the asthma module contained proteins related to immune response and pathways involved in other immune-related disorders [42]. The asthma module also resulted in enriched with differentially expressed genes from normal and asthmatic fibroblast cells treated with an asthma-specific drug and close evaluation of the module revealed GAB1 signaling pathway as an important modulator in asthma [42]. In summary, this represents the simplest approach for disease module discovery: size of the largest connected component (LCC) of the subgraph formed by disease proteins, and strategy that was later demonstrated to work in hundreds of other diseases rather than asthma only [43].

However, the discovery and detection of disease modules can be often challenging, since a large proportion of the disease-associated proteins remains unknown, as well as many of the possible protein–protein interactions remain to be discovered. These limitations result in modules that are often fragmented, limited in size, and only partially describing the underlying disease mechanisms.

Several methods have been proposed for the discovery of disease-associated proteins and subnetworks. Some methods are based on the “guilt-by-association” principle and exploit the network proximity of diseased genes [44, 45]. Other methods explore the global structural and topological properties of PPI networks to identify disease-related subnetworks or disease modules. For example, some methods are based on the principle of network diffusion or random walk [37, 46]. In these methods, disease genes are starting points (seeds) of a random walker that moves from node to node along the links of a network. After a given number of iterations, the frequency in which the nodes are visited converges is used to rank highly visited subnetworks. Examples of methods based on this principle are HotNet [47] and HotNet2 [48] algorithms that aim to find modules of somatic mutations in cancer and modules of common variants in complex diseases.

The DIseAse Module Detection (DIAMOnD) algorithm introduced the concept of connectivity significance [49]. In this method, disease genes are mapped in the PPI network and, at each step of an iterative process, the node most significantly connected to the disease genes is added to the module. The connectivity significance is based on a hypergeometric test that assigns a p-value to the proteins that share more connections with the seed proteins than expected by chance [49]. However, as a cautionary note, we highlight the fact that statistical tests that assume independence of observations are not appropriate for networks [50], and degree-preserving network randomization strategies could provide better statistical support for the same

network-based principles [51, 52]. The comparison of different types of similar algorithms (i.e., node-ranking) suggests that each one has its strengths and weakness and their application might depend on the specific use case [53].

Another class of methods for module identification is based on the principle that nodes related to the same disease or function are more densely connected to each other than expected by chance (i.e., high modularity). A recent study compared different types of module identification algorithms in different biological networks [54]. Again, results showed that methods from different categories can achieve comparable performances complementary to each other [54].

Different networks can provide very different predictive performances when used with disease module identification algorithms. A comparison of different network sources indicates that the size of the network can improve performance and outweigh the detrimental effects of false positives [55]. It also showed that parsimonious composite networks, which only include edges that are also observed in other networks, can also increase performance and efficiency [55].

Integration of phenotypic data can also help in the identification of disease genes and disease modules. Caceres and Paccanaro [57] recently proposed an approach that uses disease phenotype similarity [56], to define a prior probability distribution over disease genes on the interactome. Subsequently, a semi-supervised learning method establishes a prioritization ordering for all genes in the interactome. The important advantage of this method is that it provides predictions of disease genes even for diseases with no known genes. Their method can also be used to retrieve disease modules [57].

### ***1.4.2 Disease Networks***

The interconnectivity between cellular components also implies that different diseases might be connected by the same underlying molecular mechanisms. To map these disease–disease relationships, Goh et al. [58] created the *diseasome*, a network in which nodes represent diseases; two diseases are linked if they share common genes, and these links are labeled by the number of gene-causing mutations that are shared. The network representation of disease interrelationships provides a global perspective, offering the possibility to identify patterns and principles not readily apparent from the study of individual disorders.

However, different diseases could share disease pathways and processes while not having any causal gene in common. Therefore, methods have been developed for measuring the network proximity (or overlap) between disease modules in the interactome. The  $S_{AB}$  measure compares the shortest distances between proteins within each disease (A and B, for example), to the shortest distance between A-B protein pairs [43]. It was shown that overlapping disease modules ( $S_{AB} < 0$ ) share several pathobiological properties: the respective disease proteins have similar functions and show higher co-expression across tissues, while the diseases have similar symptoms and higher risk of co-occurrence in patients [43].

It is also possible to integrate clinical information to map and understand how different diseases are related. Disease networks can map the level of disease co-occurrence, or comorbidity, from Electronic Health Records (EHR). In these networks, the nodes are usually disease codes used in clinical practice, such as ICD-9 and ICD-10, and the links are defined by a statistical measure of co-occurrence. Examples of co-occurrence measures are the phi-coefficient, a correlation measure for binary variables, and the relative risk, the ratio between observed co-occurrence and random expectation [59]. Strategies based on such maps were able to reveal comorbidities that were demographically modulated in a given population (e.g., diseases more frequently co-occurring in black males) [60], the impact of age and sex on disease comorbidities [61, 62], and to reveal temporal patterns of disease progression [63–66]. For example, the study of a disease network was able to identify patterns of disease trajectories significantly associated with sepsis mortality, which started from three major points: alcohol-abuse, diabetes, and cardiovascular diagnoses [64].

The human symptoms—disease network—that connected diseases that showed symptom similarity, indicated that strong associations in symptom similarity also reflect common disease genes and PPIs [67]. It also indicated that diseases with diverse clinical manifestations also showed diversity in their underlying mechanisms [67].

Disease networks will improve as more molecular and phenotypic data will become available for a larger number of diseases. They represent a global reference map for clinicians to better visualize and understand disease interrelationships. They might reveal principles for better treatment and prevention, as well as offer a mechanism-driven approach for the development of new disease classification guidelines [68].

## 1.5 Biological Networks and Drugs

As the study of biological networks reveals significant insights into the systemic organization of cellular mechanisms, they provide a powerful platform where we can study the interplay between drugs and diseases and identify emergent properties not apparent when single molecules are studied in isolation [69]. Biological networks have been applied for the discovery of new targets, characterization of mechanism of action, identification of drug repurposing strategies, and for prediction of drug safety and toxicity.

PPI networks have been extensively used to study drug targets and their relationship with disease proteins. The targets of most drugs tend to form connected subgraphs within the PPI that are significantly larger than expected by chance and most compound targets are characterized by significantly shorter path lengths between their associated targets [70]. Additionally, drug targets tend to be significantly proximal, in the network, to the proteins of the diseases for which they are indicated [71, 72]. These network proximity measures take into account the

shortest path lengths among drug targets and proteins, and the statistical significance is evaluated by comparing the observed distance to the distance obtained from random sets of proteins, while preserving the size and degree of the original sets. Guney et al. [72] applied these principles to study the proximity between all possible pairs among 238 drugs and 78 diseases, showing that the proximity between drug targets and disease proteins provided a good discriminating performance for distinguishing known drug–disease pairs (i.e., with clinical use) from unknown ones [72].

These observations suggest that network-based methods could aid in the identification of drugs that could be reused for conditions different from their intended indications. In particular, it has been observed that drugs often target regions and pathways that are shared across multiple conditions [73]. Potential therapeutic interventions targeting the common pathologic processes of Type 2 Diabetes and Alzheimer’s Diseases were revealed by first identifying pathways proximal to the disease modules and then ranking pathways targeted by drugs using topological information from the protein interactome [73]. In another example, transcriptomic data integrated with a protein–protein interaction network were used to identify molecular pathways shared across different tumor types, revealing therapeutic candidates that could eventually be repurposed for the treatment of a whole group of tumors [74]. Using the network proximity between drug targets and disease proteins, Cheng et al. [76] identified hydroxychloroquine, a drug indicated for rheumatoid arthritis, as a potential therapeutic intervention for coronary artery disease [75]. The authors analyzed data from over 220 million patients in healthcare databases to demonstrate that patients who happened to be prescribed for hydroxychloroquine had indeed lower risk of being diagnosed for coronary heart disease later in their lives. The study provided additional *in vitro* data suggesting that the mechanism of action for this association might involve hydroxychloroquine’s anti-inflammatory effects on endothelial cells [75].

Complex diseases tend to be associated with multiple proteins and drugs often work by targeting several proteins besides their primary target. Consequently, several approaches attempt to develop and predict drugs that target multiple proteins, as well as to identify new drug combination strategies. Recent analysis of drug targets in PPIs showed that targets are clustered in specific network neighborhoods and proximity among targets of drug pairs also correlates with chemical, biological, and clinical similarities of the corresponding drugs [70, 76]. It also showed that for a drug–pair combination to be effective, the drug targets of both drugs should overlap with the disease module of the disease for which the treatment is intended for, while not overlapping with each other [76]. Based on these principles, a network proximity method showed good accuracy on the discrimination of approved hypertensive drug combinations, and it outperformed traditional cheminformatics and bioinformatics approaches [76]. These observations also agree with experimental data evaluating morphology perturbations caused by drug combinations in cell lines [70].

In a recent application of neural networks on graphs, a multi-layer representation of protein–protein, drug–protein, and drug–drug (with links representing side effects) interaction networks was used to predict side effects with improved performance over previous methods [77]. In contrast with previous methods, this method could



not only predict a probability/strength score of a drug interaction, but could also identify which exact side effect would result from the interaction.

**Acknowledgements** We would like to thank Alberto Paccanaro for his valuable inputs.

**Funding** HIN is supported by CNPq (313662/2017-7) and the São Paulo Research Foundation (FAPESP; grants 2018/14933-2, 2018/21934-5, and 2013/08216-2).

## References

1. Shen SY, Singhania R, Fehring G, Chakravarthy A, Roehrl MHA, Chadwick D et al (2018) Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 563(7732):579–583
2. Barabási A-L, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12(1):56–68
3. Reka A, Jeong H, Barabási A-L, Albert R, Jeong H, Barabási A-L (2000) Error and Attack Tolerance of Complex Networks. *Nature* 406:378–381
4. Watts DJ, Strogatz SH (1998) Collective dynamics of “small-world” networks. *Nature* 393(6684):440–442
5. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi A-L, The large-scale organization of metabolic networks. *Nature*
6. Fell DA, Wagner A (2000) The small world of metabolism. *Nat Biotechnol* 18(11):1121–1122
7. Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411(6833):41–42
8. Vidal M, Cusick ME, Barabasi A-L (2011) Interactome networks and human disease. *Cell* 144(6):986–998
9. Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, et al (2019) A reference map of the human protein interactome. *bioRxiv* 605451.
10. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K et al (2017) Architecture of the human interactome defines protein communities and disease networks. *Nature [Internet]* 545(7655):505–509
11. Das J, Yu H (2012) HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* 6
12. Alonso-López D, Campos-Laborie FJ, Gutiérrez MA, Lambourne L, Calderwood MA, Vidal M et al (2019) APID database: redefining protein-protein interaction experimental evidences and binary interactomes. *Database (i)*:1–8
13. Luck K, Sheynkman GM, Zhang I, Vidal M (2017) Proteome-scale human interactomics. *Trends Biochem Sci* 42(5):342–354
14. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C et al (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature [Internet]* 489(7414):91–100
15. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell [Internet]* 150(6):1274–1286
16. Wang YXR, Huang H (2014) Review on statistical methods for gene network reconstruction using expression data. *J Theor Biol [Internet]* 362:53–61
17. Mo ML, Pálsson BØ (2009) Understanding human metabolic physiology: a genome-to-systems approach. *Trends Biotechnol [Internet]* 27(1):37–44

18. Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science* (80-) [Internet] 353(6306):aaf1420–aaf1420
19. Baryshnikova A, Costanzo M, Myers CL, Andrews B, Boone C (2013) Genetic interaction networks: toward an Understanding of Heritability. *Annu Rev Genomics Hum Genet* 14(1):111–133
20. Boucher B, Jenna S (2013) Genetic interaction networks: better understand to better predict. *Front Genet* 4:1–16
21. Ahmed H, Howton TC, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS (2018) Network biology discovers pathogen contact points in host protein-protein interactomes. *Nat Commun* 9(1):2312
22. Calderwood MA, Venkatesan K, Xing L, Chase MR, Vazquez A, Holthaus AM et al (2007) Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci U S A* 104(18):7606–7611
23. de Chasse B, Navratil V, Tafforeau L, Hiet MS, Aublin-Gex A, Agaugué S et al (2008) Hepatitis C virus infection protein network. *Mol Syst Biol* 4:230
24. Uetz P, Dong Y-A, Zeretzke C, Atzler C, Baiker A, Berger B et al (2006) Herpesviral protein networks and their interaction with the human proteome. *Science* [Internet] 311(5758):239–242
25. Shapira SD, Gat-Viks I, Shum BO V, Dricot A, de Grace MM, Wu L et al (2009) A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* 139(7):1255–1267
26. Jäger S, Gulbahce N, Cimermancic P, Kane J, He N, Chou S et al (2011) Purification and characterization of HIV-human protein complexes. *Methods* 53(1):13–19
27. Mendez-Rios J, Uetz P (2010) Global approaches to study protein-protein interactions among viruses and hosts. *Future Microbiol* 5(2):289–301
28. Penn BH, Netter Z, Johnson JR, Von Dollen J, Jang GM, Johnson T et al (2018) An Mtb-human protein-protein interaction map identifies a switch between host antiviral and antibacterial responses. *Mol Cell* 71(4):637–648.e5
29. Remmele CW, Luther CH, Balkenhol J, Dandekar T, Müller T, Dittrich MT (2015) Integrated inference and evaluation of host–fungi interaction networks. *Front Microbiol* 4:6
30. Cuesta-Astroz Y, Santos A, Oliveira G, Jensen LJ (2019) Analysis of predicted host–parasite interactomes reveals commonalities and specificities related to parasitic lifestyle and tissues tropism. *Front Immunol* 13:10
31. Aditya R, Mayil K, Thomas J, Gopalakrishnan B (2010) Cerebral malaria: insights from host-parasite protein-protein interactions. *Malar J* 9:1–7
32. Wuchty S (2011) Computational prediction of host-parasite protein interactions between *P. falciparum* and *H. sapiens*. *Borrmann S* (ed). *PLoS One* 6(11):e26960
33. Bass JIF, Diallo A, Nelson J, Soto JM, Myers CL, Walhout AJM (2013) Using networks to measure similarity between genes: Association index selection. *Nat Methods* 10(12):1169–1176
34. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS et al (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 446(7137):806–810
35. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. *Mol Syst Biol* 3(88):1–13
36. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21(SUPPL. 1):302–310
37. Cowen L, Ideker T, Raphael BJ, Sharan R (2017) Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet* 18(9):551–562
38. Cao M, Zhang H, Park J, Daniels NM, Crovella ME, Cowen LJ et al (2013) Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS ONE* 8(10):1–12

39. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P et al (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 38(SUPPL. 2):214–220
40. Nelson W, Zitnik M, Wang B, Leskovec J, Goldenberg A, Sharan R (2019) To embed or not: Network embedding as a paradigm in computational biology. *Front Genet* 10:1–11
41. Zitnik M, Leskovec J (2017) Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* 33(14):i190–i198
42. Sharma A, Menche J, Chris Huang C, Ort T, Zhou X, Kitsak M et al (2014) A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum Mol Genet* 24(11):3005–3020
43. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J et al (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science [Internet]* 347(6224):1257601
44. Guney E, Oliva B (2012) Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLoS One* 7(9)
45. Yin T, Chen S, Wu X, Tian W (2017) GenePANDA-a novel network-based gene prioritizing tool for complex diseases. *Sci Rep* 7:1–10
46. Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 82(4):949–958
47. Vandin F, Upfal E, Raphael BJ (2011) Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 18(3):507–522
48. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge J V, Thomas JL et al (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet [Internet]* 47(2):106–114
49. Ghiassian SD, Menche J, Barabási AL (2015) A Disease Module Detection (DIAMOND) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* 11(4):1–21
50. Croft DP, Madden JR, Franks DW, James R (2011) Hypothesis testing in animal social networks. *Trends Ecol Evol* 26(10):502–507
51. Iorio F, Bernardo-Faura M, Gobbi A, Cokelaer T, Jurman G, Saez-Rodriguez J (2016) Efficient randomization of biological networks while preserving functional characterization of individual nodes. *BMC Bioinform* 17(1):542
52. Farine DR (2017) A guide to null models for animal social network analysis. *Methods Ecol Evol* 8(10):1309–1320
53. Hill A, Gleim S, Kiefer F, Sigoillot F, Loureiro J, Jenkins J et al (2019) Benchmarking network algorithms for contextualizing genes of interest. *PLoS Comput Biol* 15(12):1–14
54. Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Fang T, Lamparter D et al (2019) Assessment of network module identification across complex diseases. *Nat Methods* 16(9):843–852
55. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P et al (2018) Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* 6(4):484–495.e5
56. Caniza H, Romero AE, Paccanaro A (2015) A network medicine approach to quantify distance between hereditary disease modules on the interactome. *Sci Rep* 5:1–10
57. Cáceres JJ, Paccanaro A (2019) Disease gene prediction for molecularly uncharacterized diseases. *PLoS Comput Biol* 15(7):1–14
58. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) The human disease network. *Proc Natl Acad Sci U S A* 104(21):8685–8690
59. Fotouhi B, Momeni N, Riolo MA, Buckeridge DL (2018) Statistical methods for constructing disease comorbidity networks from longitudinal inpatient data. *Appl Netw Sci* 3(1)
60. Hidalgo CA, Blumm N, Barabási AL, Christakis NA (2009) A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol* 5(4)
61. Chmiel A, Klimek P, Thurner S. Spreading of diseases through comorbidity networks across life and gender. *New J Phys* 16(11):115013
62. Kalgotra P, Sharda R, Croff JM (2017) Examining health disparities by gender: A multimorbidity network analysis of electronic medical record. *Int J Med Inform* 108:22–28

63. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H et al (2014) Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun* 5:1–10
64. Beck MK, Jensen AB, Nielsen AB, Perner A, Moseley PL, Brunak S (2016) Diagnosis trajectories of prior multi-morbidity predict sepsis mortality. *Sci Rep* 6:1–9
65. Giannoula A, Gutierrez-Sacristán A, Bravo Á, Sanz F, Furlong LI (2018) Identifying temporal patterns in patient disease trajectories using dynamic time warping: a population-based study. *Sci Rep* 8(1):1–14
66. Jeong E, Ko K, Oh S, Han HW (2017) Network-based analysis of diagnosis progression patterns using claims data. *Sci Rep* 7(1):1–12
67. Zhou X, Menche J, Barabási A-L, Sharma A, Zhou X (2014) Human symptoms–disease network. *Nat Commun* 5
68. Dozmorov MG (2018) Disease classification: from phenotypic similarity to integrative genomics and beyond. *Brief Bioinform* 1–12
69. Loscalzo J, Barabási A-L, Silverman EK (2017) *Network medicine*. Harvard University Press
70. Caldera M, Müller F, Kaltenbrunner I, Licciardello MP, Lardeau CH, Kubicek S et al (2019) Mapping the perturbome network of cellular perturbations. *Nat Commun [Internet]* 10(1)
71. Wang RS, Loscalzo J (2016) Illuminating drug action by network integration of disease genes: a case study of myocardial infarction. *Mol Biosyst* 12(5):1653–1666
72. Guney E, Menche J, Vidal M, Barabási A-L (2016) Network-based in silico drug efficacy screening. *Nat Commun* 7(1):10331
73. Aguirre-Plans J, Piñero J, Menche J, Sanz F, Furlong LI, Schmidt HHHW et al (2018) Targeting comorbid diseases via network endopharmacology. *Pharmaceuticals [Internet]* 11(61)
74. do Valle ÍF, Menichetti G, Simonetti G, Bruno S, Zironi I, Durso DF et al (2018) Network integration of multi-tumour omics data suggests novel targeting strategies. *Nat Commun [Internet]* 9(1):4514
75. Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási AL et al (2018) Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun* 9(1):1–12
76. Cheng F, Kovács IA, Barabási A-L (2019) Network-based prediction of drug combinations. *Nat Commun* 10(1):1197
77. Zitnik M, Agrawal M, Leskovec J (2018) Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34(13):i457–i466

# Chapter 2

## Computational Tools for Comparing Gene Coexpression Networks



Vinícius Carvalho Jardim, Camila Castro Moreno, and André Fujita

**Abstract** The comparison of biological networks is a crucial step to better understanding the underlying mechanisms involved in specific experimental conditions, such as those of health and disease or high and low concentrations of an environmental element. To this end, several tools have been developed to compare whether network structures are “equal” (in some sense) across conditions. Some examples of computational methods include DCGL, EBcoexpress, DiffCorr, CoDiNA, Diff-CoEx, coXpress, DINGO, DECODE, dCoxS, GSCA, GSNCA, CoGA, GANOVA, and BioNetStat. We will briefly describe these algorithms and their advantages and disadvantages.

**Keywords** Network science · Differential network analysis · Coexpression network · Systems biology · Network theory

### 2.1 Introduction

To understand complex systems, we need to consider the interactions between their elements. A graph is a useful tool for studying these systems due to the plasticity of network models for interpreting biological problems. In a biological context, network vertices can represent system elements such as proteins, metabolites, genes, among other examples. In coexpression networks, vertices represent genes, while edges represent coexpression between gene pairs.

---

V. C. Jardim · C. C. Moreno  
Interdepartmental Bioinformatics Program, Institute of Mathematics and Statistics - University of São Paulo, São Paulo, Brazil  
e-mail: [vinicius.jardim.carvalho@usp.br](mailto:vinicius.jardim.carvalho@usp.br)

C. C. Moreno  
e-mail: [camila.moreno@usp.br](mailto:camila.moreno@usp.br)

A. Fujita (✉)  
Department of Computer Science, Institute of Mathematics and Statistics - University of São Paulo, São Paulo, Brazil  
e-mail: [andrefujita@usp.br](mailto:andrefujita@usp.br)

We define coexpression as the statistical dependence (correlation) between the expression values of two genes. Correlation measures how coordinated the variation of expression values of two genes are in same condition samples (obtained from microarray or RNA-seq analysis). In this chapter, we use the terms *Conditions* or *experimental conditions* as synonyms of experimental treatments, such as of high, mean, and low temperatures or clinical status, such as healthy versus cancer tissues. Usually, correlation allows us to infer whether two genes belong to the same metabolic pathway or biological process. However, it does not imply that one variable influences another. Therefore, the edges that represent the correlations have no direction, constructing undirected networks.

Changes in correlations (edges) between conditions are of interest to many studies. In some cases, the aim is to verify whether the environment or genome variations affect the relationship between genes. Considering that each network represents an experimental condition, to achieve this goal, we need effective means to compare these networks. The scientific community has developed several strategies to accomplish this task, with approaches ranging from verifying the edge's existence in differing conditions to network model comparisons.

The most used correlation measure in coexpression network studies is the Pearson correlation. However, the non-parametric Spearman correlation is also frequently used since it does not demand the assumption of normality and is not limited to only detecting linear correlations. Other strategies use mutual entropy and Bayesian inference to define coexpression between genes [1].

Beyond the choice of correlation methods, it is also vital to select the threshold for a given correlation to become an edge. In this sense, we commonly use two main kinds of techniques. The most used is the hard threshold. It works as a cut-off value to remove correlations that are below a defined value (correlation threshold) or with a predetermined level of significance (p-value threshold). Another strategy is the soft threshold proposed in WCGA paper [2]. The soft threshold ponders (or rescales) the correlation values according to a power value  $\beta$ . This threshold technique works by powering the correlation to a  $\beta$  value: the higher values increase and the lower ones decrease, therefore highlighting the most relevant correlations. At the soft threshold, the network remains complete without edge removal. Once parameters for constructing networks are defined, those such as coexpression criteria and threshold technique, we can compare the resulting networks in many ways.

## 2.2 Network Comparison Methods

Many studies apply network analysis to compare different experimental conditions. One way is to quantify and compare the structural features of networks such as presence or absence of edges or the number of connections of a vertex [3, 4]. Other strategies look for edges that are exclusive of a condition [5] or identify a differential network resulting from the combination of differential expression analysis (DE) and differential coexpression (DC) [6]. Despite these methods being useful, they do not