Statistical and Machine Learning Approaches for Network Analysis

Edited by
Matthias Dehmer
Subhash C. Basak

WILEY

STATISTICAL AND MACHINE LEARNING APPROACHES FOR NETWORK ANALYSIS

Edited by

MATTHIAS DEHMER
UMIT – The Health and Life Sciences University, Institute for Bioinformatics and Translational Research, Hall in Tyrol, Austria

SUBHASH C. BASAK
Natural Resources Research Institute
University of Minnesota, Duluth
Duluth, MN, USA

WILEY
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

ISBN: 978-0-470-19515-4

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Christina

PREFACE

An emerging trend in many scientific disciplines is a strong tendency toward being transformed into some form of information science. One important pathway in this transition has been via the application of network analysis. The basic methodology in this area is the representation of the structure of an object of investigation by a graph representing a relational structure. It is because of this general nature that graphs have been used in many diverse branches of science including bioinformatics, molecular and systems biology, theoretical physics, computer science, chemistry, engineering, drug discovery, and linguistics, to name just a few. An important feature of the book "Statistical and Machine Learning Approaches for Network Analysis" is to combine theoretical disciplines such as graph theory, machine learning, and statistical data analysis and, hence, to arrive at a new field to explore complex networks by using machine learning techniques in an interdisciplinary manner.

The age of network science has definitely arrived. Large-scale generation of genomic, proteomic, signaling, and metabolomic data is allowing the construction of complex networks that provide a new framework for understanding the molecular basis of physiological and pathological states. Networks and network-based methods have been used in biology to characterize genomic and genetic mechanisms as well as protein signaling. Diseases are looked upon as abnormal perturbations of critical cellular networks. Onset, progression, and intervention in complex diseases such as cancer and diabetes are analyzed today using network theory.

Once the system is represented by a network, methods of network analysis can be applied to extract useful information regarding important system properties and to investigate its structure and function. Various statistical and machine learning methods have been developed for this purpose and have already been applied to networks. The purpose of the book is to demonstrate the usefulness, feasibility, and the impact of the

methods on the scientific field. The 11 chapters in this book written by internationally reputed researchers in the field of interdisciplinary network theory cover a wide range of topics and analysis methods to explore networks statistically.

The topics we are going to tackle in this book range from network inference and clustering, graph kernels to biological network analysis for complex diseases using statistical techniques. The book is intended for researchers, graduate and advanced undergraduate students in the interdisciplinary fields such as biostatistics, bioinformatics, chemistry, mathematical chemistry, systems biology, and network physics. Each chapter is comprehensively presented, accessible not only to researchers from this field but also to advanced undergraduate or graduate students.

Many colleagues, whether consciously or unconsciously, have provided us with input, help, and support before and during the preparation of the present book. In particular, we would like to thank Maria and Gheorghe Duca, Frank Emmert-Streib, Boris Furtula, Ivan Gutman, Armin Graber, Martin Grabner, D. D. Lozovanu, Alexei Levitchi, Alexander Mehler, Abbe Mowshowitz, Andrei Perjan, Ricardo de Matos Simoes, Fred Sobik, Dongxiao Zhu, and apologize to all who have not been named mistakenly. Matthias Dehmer thanks Christina Uhde for giving love and inspiration. We also thank Frank Emmert-Streib for fruitful discussions during the formation of this book.

We would also like to thank our editor Susanne Steitz-Filler from Wiley who has been always available and helpful. Last but not the least, Matthias Dehmer thanks the Austrian Science Funds (project P22029-N13) and the Standortagentur Tirol for supporting this work.

Finally, we sincerely hope that this book will serve the scientific community of network science reasonably well and inspires people to use machine learning-driven network analysis to solve interdisciplinary problems successfully.

MATTHIAS DEHMER
SUBHASH C. BASAK

CONTRIBUTORS

Lipi Acharya, Department of Computer Science, University of New Orleans, New Orleans, LA, USA

Enrico Capobianco, Laboratory for Integrative Systems Medicine (LISM) IFC-CNR, Pisa (IT); Center for Computational Science, University of Miami, Miami, FL, USA

Christina Chan, Departments of Chemical Engineering and Material Sciences, Genetics Program, Computer Science and Engineering, and Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA

Ricardo de Matos Simoes, Computational Biology and Machine Learning Lab, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, UK

Frank Emmert-Streib, Computational Biology and Machine Learning Lab, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, UK

Damien Fay, Computer Laboratory, Systems Research Group, University of Cambridge, UK

Hirosha Geekiyanage, Genetics Program, Michigan State University, East Lansing, MI, USA

Elisabeth Georgii, Department of Information and Computer Science, Helsinki Institute for Information Technology, Aalto University School of Science and Technology, Aalto, Finland

Hamed Haddadi, Computer Laboratory, Systems Research Group, University of Cambridge, UK

Thair Judeh, Department of Computer Science, University of New Orleans, New Orleans, LA, USA

Reinhard Kutzelnigg, Math.Tec, Heumühlgasse, Wien, Vienna, Austria

Elisabetta Marras, CRS4 Bioinformatics Laboratory, Polaris Science and Technology Park, Pula, Italy

Andrew W. Moore, School of Computer Science, Carnegie Mellon University, USA

Richard Mortier, Horizon Institute, University of Nottingham, UK

Chikoo Oosawa, Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan

Matthias Rupp, Machine Learning Group, Berlin Institute of Technology, Berlin, Germany, and, Institute of Pure and Applied Mathematics, University of California, Los Angeles, CA, USA; currently at the Institute of Pharmaceutical Sciences, ETH Zurich, Zurich, Switzerland.

Kazuhiro Takemoto, Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan; PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan

Andrew G. Thomason, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, UK

Antonella Travaglione, CRS4 Bioinformatics Laboratory, Polaris Science and Technology Park, Pula, Italy

Koji Tsuda, Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology AIST, Tokyo, Japan

Steve Uhlig, School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

Tim vor der Brück, Department of Computer Science, Text Technology Lab, Johann Wolfgang Goethe University, Frankfurt, Germany

Xuewei Wang, Department of Chemical Engineering and Material Sciences, Michigan State University, East Lansing, MI, USA

Dongxiao Zhu, Department of Computer Science, University of New Orleans; Research Institute for Children, Children's Hospital; Tulane Cancer Center, New Orleans, LA, USA

CONTENTS

1

A SURVEY OF COMPUTATIONAL APPROACHES TO RECONSTRUCT AND PARTITION BIOLOGICAL NETWORKS

LIPI ACHARYA, THAIR JUDEH, AND DONGXIAO ZHU

"Everything is deeply intertwingled"
Theodor Holm Nelson

1.1 INTRODUCTION

The above quote by Theodor Holm Nelson, the pioneer of information technology, states a deep interconnectedness among the myriad topics of this world. The biological systems are no exceptions, which comprise of a complex web of biomolecular interactions and regulation processes. In particular, the field of computational systems biology aims to arrive at a theory that reveals complicated interaction patterns in the living organisms, which result in various biological phenomenon. Recognition of such patterns can provide insights into the biomolecular activities, which pose several challenges to biology and genetics. However, complexity of biological systems and often an insufficient amount of data used to capture these activities make a reliable inference of the underlying network topology as well as characterization of various patterns underlying these topologies, very difficult. As a result, two problems that have received a considerable amount of attention among researchers are (1) reverse engineering of biological networks from genome-wide measurements and (2) inference of functional units in large biological networks (Fig 1.1).

Statistical and Machine Learning Approaches for Network Analysis, Edited by Matthias Dehmer and Subhash C. Basak.

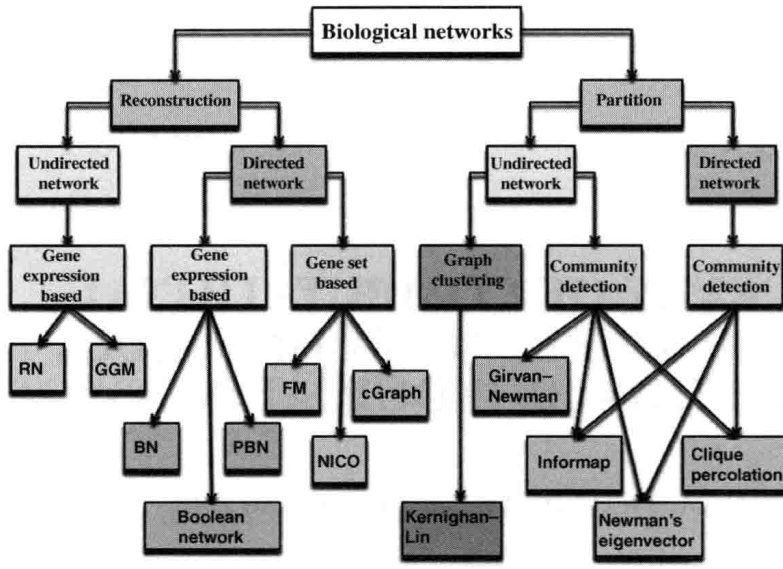

FIGURE 1.1 Approaches addressing two fundamental problems in computational systems biology (1) reconstruction of biological networks from two complementary forms of data resources, gene expression data and gene sets and (2) partitioning of large biological networks to extract functional units. Two classes of problems in network partitioning are graph clustering and community detection.

Rapid advances in high-throughput technologies have brought about a revolution in our understanding of biomolecular interaction mechanisms. A reliable inference of these mechanisms directly relates to the measurements used in the inference procedure. High throughput molecular profiling technologies, such as microarrays and second-generation sequencing, have enabled a systematic study of biomolecular activities by generating an enormous amount of genome-wide measurements, which continue to accumulate in numerous databases. Indeed, simultaneous profiling of expression levels of tens of thousands of genes allows for large-scale quantitative experiments. This has resulted in substantial interest among researchers in the development of novel algorithms to reliably infer the underlying network topology using gene expression data. However, gaining biological insights from large-scale gene expression data is very challenging due to the *curse of dimensionality*. Correspondingly, a number of computational and experimental methods have been developed to arrange genes in various groups or clusters, on the basis of certain similarity criterion. Thus, an initial characterization of large-scale gene expression data as well as conclusions derived from biological experiments result in the identification of several smaller components comprising of genes sharing similar biological properties. We refer to these components as *gene sets*. Availability of effective computational and experimental strategies have led to the emergence of gene sets as a completely new form of data for the reverse engineering of gene regulatory relationships. Gene set based approaches have gained more attention for their inherent ability to incorporate higher-order interaction mechanisms as opposed to individual genes.

There has been a sequence of computational efforts addressing the problem of network reconstruction from gene expression data and gene sets. Gaussian graphical models (GGMs) [1–3], probabilistic Boolean networks (PBNs) [4–7], Bayesian networks (BNs) [8,9], differential equation based [10,11] and mutual information networks such as relevance networks (RNs) [12,13], ARACNE [14], CLR [15], MRNET [16] are viable approaches capitalizing on the use of gene expression data, whereas collaborative graph model (cGraph) [17], frequency method (FM) [18], and network inference from cooccurrences (NICO) [19,20] are suitable for the reverse engineering of biological networks from gene sets.

After a biological network is reconstructed, it may be too broad or abstract of a representation for a particular biological process of interest. For example, given a specific signal transduction, only a part of the underlying network is activated as opposed to the entire network. A finer level of detail is needed. Furthermore, these parts may represent the functional units of a biological network. Thus, *partitioning* a biological network into different clusters or communities is of paramount importance.

Network partitioning is often associated with several challenges, which make the problem NP-hard [21]. Finding the optimal partitions of a given network is only feasible for small networks. Most algorithms heuristically attempt to find a *good* partitioning based on some chosen criteria. Algorithms are often suited to a specific problem domain. Two major classes of algorithms in network partitioning find their roots in computer science and sociology, respectively [22]. To avoid confusion, we will refer to the first class of algorithms as *graph clustering* algorithms and the second class of algorithms as *community detection* algorithms. For graph clustering algorithms, the relevant applications include very large-scale integration (VLSI) and distributing jobs on a parallel machine. The most famous algorithm in this domain is the Kernighan–Lin algorithm [23], which still finds use as a subroutine for various other algorithms. Other graph clustering algorithms include techniques based on spectral clustering [24]. Originally community detection algorithms focused on social networks in sociology. They now cover networks of interest to biologists, mathematicians, and physicists. Some popular community detection algorithms include Girvan–Newman algorithm [25], Newman's eigenvector method [21,22], clique percolation algorithm [26], and Infomap [27]. Additional community detection algorithms include methods based on spin models [28,29], mixture models [30], and label propagation [31].

Intuitively, reconstruction and partitioning of biological networks appear to be two completely opposite problems in that the former leads to an increase, whereas the latter results in a decrease of the dimension of a given structure. In fact, these problems are closely related and one leads to the foundation of the other. For instance, presence of hypothetical gene regulatory relationships in a reconstructed network provides a motivation for the detection of biologically meaningful functional modules of the network. On the other hand, prior to apply gene set based network reconstruction algorithms, a computational or experimental analysis is first needed to derive gene sets. In this chapter, we present a number of computational approaches to reconstruct biological networks from genome-wide measurements, and to partition large biological networks into subnetworks. We begin with an overview of directed and undirected networks, which naturally arise in biological systems. Next, we discuss about two

complementary forms of genome-wide data, gene expression data and gene sets, both of which can be accommodated by existing network reconstruction algorithms. We describe the principal aspects of various approaches to reconstruct biological networks using gene expression data and gene sets, and discuss the pros and cons associated with each of them. Finally, we present some popular clustering and community algorithms used in network partitioning. The material on network reconstruction and partition is largely based on Refs. [2,3,6–8,13,17–20,32] and [21–23,25–27,33–36], respectively.

1.2 BIOLOGICAL NETWORKS

A network is a graph $G(V, E)$ defined in terms of a set of vertices V and a set of edges E. In case of biological networks, a vertex $v \in V$ is either a gene or protein encoded by an organism, and an edge $e \in E$ joining two vertices v_1, $v_2 \in V$ in the network represents biological properties connecting v_1 and v_2. A biological network can be directed or undirected depending on the biological relationship that used to join the pairs of vertices in the network. Both directed and undirected networks occur naturally in biological systems. Inference of these networks is a major challenge in systems biology. We briefly review two kinds of biological networks in the following sections.

1.2.1 Directed Networks

In directed networks, each edge is identified as an ordered pair of vertices. According to the Central Dogma of Molecular Biology, genetic information is encoded in double-stranded DNA. The information stored in DNA is transferred to single-stranded messenger RNA (mRNA) to direct protein synthesis [42]. Signal transduction is the primary mean to control the passage of biological information from DNA to mRNA with mRNA directing the synthesis of proteins. A signal transduction event is usually triggered by the binding of external ligands (e.g., cytokine and chemokine) to the transmembrane receptors. This binding results in a sequential activation of signal molecules, such as cytoplasmic protein kinase and nuclear transcription factors (TFs), to lead to a biological end-point function [42]. A signaling pathway is composed of a web of gene regulatory wiring in response to different extracellular stimulus. Thus, signaling pathways can be viewed as directed networks containing all genes (or proteins) of an organism as vertices. A directed edge represents the flow of information from one gene to another gene.

1.2.2 Undirected Networks

Undirected networks differ from directed networks in that the edges in such networks are undirected. In other words, an undirected network can be viewed as a directed network by considering an undirected pair of vertices (v_1, v_2) as two directed pairs (v_1, v_2) and (v_2, v_1). Some biological networks are better suited for an undirected

representation. Protein–protein interaction (PPI) network is an undirected network, where each protein is considered as a vertex and the physical interaction between a pair of proteins is represented as an edge [43].

The past decade has witnessed a significant progress in the computational inference of biological networks. A variety of approaches in the form of network models and novel algorithms have been proposed to understand the structure of biological networks at both *global* and *local* level. While the grand challenge in a global approach is to provide an integrated view of the underlying biomolecular interaction mechanisms, a local approach focuses on identifying fundamental domains representing functional units of a biological network.

Both directed and undirected network models have been developed to reliably infer the biomolecular activities at a global level. As discussed above, directed networks represent an abstraction of gene regulatory mechanisms, while the physical interactions of genes are suitably modeled as undirected networks. Focus has also been on the computational inference of biomolecular activities by accommodating genome-wide data in diverse formats. In particular, gene set based approaches have gained attention in recent bioinformatics analysis [44,45]. Availability of a wide range of experimental and computational methods have identified coherent gene set compendiums [46]. Sophisticated tools now exist to statistically verify the biological significance of a particular gene set of interest [46–48]. An emerging trend in this field is to reconstruct signaling pathways by inferring the order of genes in gene sets [19,20]. There are several unique features associated with gene set based network inference approaches. In particular, such approaches do not rely on gene expression data for the reconstruction of underlying network.

The algorithms to understand biomolecular activities at the level of subnetworks have evolved over time. Community detection algorithms, in particular, originated with hierarchical partitioning algorithms that include the Girvan–Newman algorithm. Since these algorithms tend to produce a dendrogram as their final result, it is necessary to be able to rank the different partitions represented by the dendrogram. Modularity was introduced by Newman and Girvan to address this issue. Many methods have resulted with modularity at the core. More recently, though, it has been shown that modularity suffers from some drawbacks. While there have been some attempts to address these issues, newer methods continued to emerge such as Infomap. Research has also expanded to incorporate different types of biological networks and communities. Initially, only undirected and unweighted networks were the focus of study. Methods are now capable of dealing with both directed and weighted networks. Moreover, previous studies only concentrated on distinct communities that did not allow overlap. With the advent of the clique percolation method and other similar methods, overlapping communities are becoming increasingly popular. The aforementioned approaches have been used to identify the structural organization of a variety of biological networks including metabolic networks, PPI networks, and protein domain networks. Such networks have a power–law degree distribution and the quantitative signature of scale-free networks [49]. PPI networks, in particular, have been the subject of intense study in both bioinformatics and biology as protein interactions are fundamental for cellular processes [50].

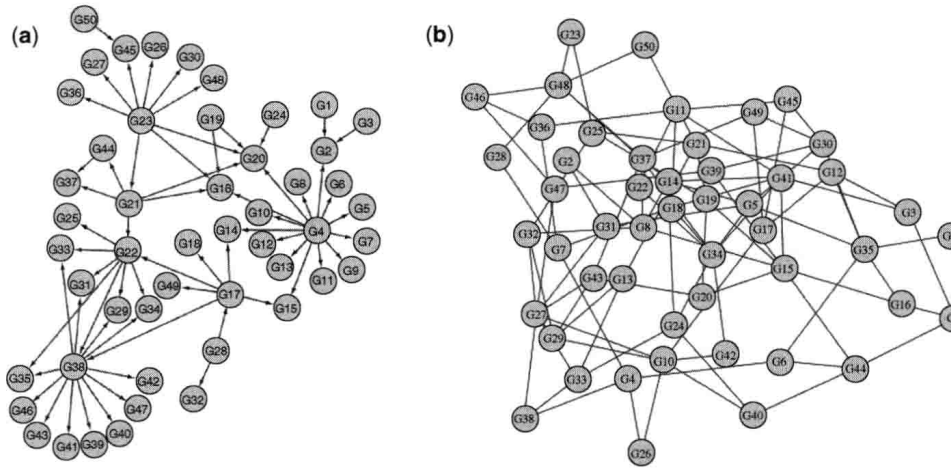

FIGURE 1.2 (a) Example of a directed network. The figure shows *Escherichia coli* gold standard network from the DREAM3 Network Challenges [37–39]. (b) Example of an undirected network. The figure shows an *in silico* gold standard network from the DREAM2 Network Challenges [40,41].

A common problem associated with the computational inference of a biological network is to assess the performance of the approach used in the inference procedure. It is quite assess as the structure of the true underlying biological network is unknown. As a result, one relies on biologically plausible simulated networks and data generated from such networks. A variety of *in silico* benchmark directed and undirected networks are provided by the dialogue for reverse engineering assessments and methods (DREAM) initiative to systematically evaluate the performance of reverse engineering methods, for example Refs. [37–41]. Figures 1.2 and 1.7 illustrate gold standard directed network, undirected network, and a network with community structure from the *in silico* network challenges in DREAM initiative.

1.3 GENOME-WIDE MEASUREMENTS

In this section, we present an overview of two complementary forms of data resources (Fig. 1.3), both of which have been utilized by the existing network reconstruction algorithms. The first resource is gene expression data, which is represented as matrix of gene expression levels. The second data resource is a gene set compendium. Each gene set in a compendium stands for a set of genes and the corresponding gene expression levels may or may not be available.

1.3.1 Gene Expression Data

Gene expression data is the most common form of data used in the computational inference of biological networks. It is represented as a matrix of numerical values,

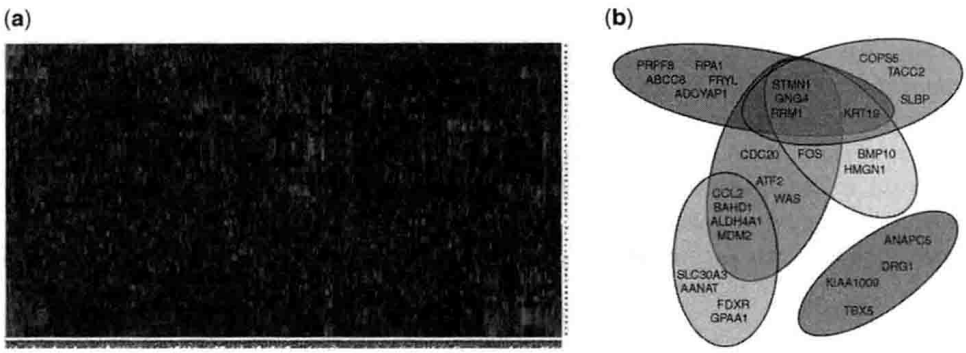

FIGURE 1.3 Two complementary forms of data accommodated by the existing network reconstruction algorithms. (a) Gene expression data generated from high-throughput platforms, for example, microarray. (b) Gene sets often resulted from explorative analysis of large-scale gene expression data, for example, cluster analysis.

where each row corresponds to a gene, each column represents an experiment and each entry in the matrix stands for gene expression level. Gene expression profiling enables the measurement of expression levels of thousands of genes simultaneously and thus allows for a systematic study of biomolecular interaction mechanisms on genome scale. In the experimental procedure for gene expression profiling using microarray, typically a glass slide is spotted with oligonucleotides that correspond to specific gene coding regions. Purified RNA is labeled and hybridized to the slide. After washing, gene expression data is obtained by laser scanning. A wide range of microarray platforms have been developed to accomplish the goal of gene expression profiling. The measurements can be obtained either from conventional hybridization-based microarrays [51–53] or contemporary deep sequencing experiments [54,55]. Affymetrix GeneChip (www.affymetrix.com), Agilent Microarray (www.genomics.agilent.com), and Illumina BeadArray (www.illumina.com) are representative microarray platforms. Gene-expression data are accessible from several databases, for example, National Center for Biological Technology (NCBI) Gene Expression Omnibus (GEO) [56] and the European Molecular Biology Lab (EMBL) ArrayExpress [57].

1.3.2 Gene Sets

Gene sets are defined as sets of genes sharing biological similarities. Gene sets provide a rich source of data to infer underlying gene regulatory mechanisms as they are indicative of genes participating in the same biological process. It is impractical to collect a large number of samples from high-throughput platforms to accurately reflect the activities of thousands of genes. This poses challenges in gaining deep biological insights from genome-wide gene expression data. Consequently, experimental and computational methods are adopted to reduce the dimension of the space of variables [58]. Such characterizations lead to the discovery of clusters