Institute of Biotechnology and
Department of Biosciences, Division of Genetics
Faculty of Biological and Environmental Sciences
University of Helsinki
Finland

Computational Approaches to Biological Network Inference and Modeling in Systems Biology

Hung Xuan Ta

ACADEMIC DISSERTATION

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki in Viikki Infocentre Korona, Hall 3, on November 30th, 2012 at 14:00

Helsinki, 2012

Supervisor
Liisa Holm, University of Helsinki, Finland

Pre-examiners
Sampsa Hautaniemi, University of Helsinki, Finland
Harri Lähdesmäki, Aalto University, Finland

Opponent
Andre S. Ribeiro, Tampere University of Technology, Finland

Custos
Tapio Palva, University of Helsinki, Finland

ISSN 1799-7372
ISBN 978-952-10-8299-3 (paperback)
ISBN 978-952-10-8300-6 (PDF)
http://ethesis.helsinki.fi
Helsinki University Print (Unigrafia)
Helsinki 2012

Computational Approaches to Biological Network Inference and Modeling in Systems Biology

Hung Xuan Ta

Department of Biosciences
Division of Genetics
P.O. Box 56, FI-00014 University of Helsinki, Finland
xuanhung.ta@helsinki.fi
http://ekhidna.biocenter.helsinki.fi/users/xuan

PhD Thesis
Helsinki, June 2012, 82+25 pages
ISSN 1799-7372
ISBN 978-952-10-8299-3 (paperback)
ISBN 978-952-10-8300-6 (PDF)

Abstract

Living systems, which are composed of biological components such as molecules, cells, organisms or entire species, are *dynamic* and *complex*. Their behaviors are difficult to study with respect to the properties of individual elements. To study their behaviors, we use quantitative techniques in the "omic" fields such as genomics, bioinformatics and proteomics to measure the behavior of groups of interacting components, and we use mathematical and computational modeling to describe and predict their dynamical behavior.

The first step in the understanding of a biological system is to investigate how its individual elements interact with each other. This step consist of drawing a *static* wiring diagram that connects the individual parts. Experimental techniques that are used - are designed to observe interactions among the biological components in the laboratory while computational approaches are designed to predict interactions among the individual elements based on their properties. In the first part of this thesis, we present techniques for network inference that are particularly targeted at protein-protein interaction networks. These techniques include comparative genomics, structure-based, biological context methods and integrated frameworks. We evaluate and compare the prediction methods that have been most often used for domain-domain interactions and we discuss the limitations of the methods and data resources. We introduce the concept of the *Enhanced Phylogenetic*

Tree, which is a new graphical presentation of the evolutionary history of protein families; then, we propose a novel method for assigning functional linkages to proteins. This method was applied to predicting both human and yeast protein functional linkages.

The next step is to obtain insights into the dynamical aspects of the biological systems. One of the outreaching goals of systems biology is to understand the *emergent properties* of living systems, i.e., to understand how the individual components of a system come together to form distinct, collective and interactive properties and functions. The emergent properties of a system are neither to be found in nor are directly deducible from the lower-level properties of that system. An example of the emergent properties is synchronization, a dynamical state of complex network systems in which the individual components of the systems behave coherently, almost in unison. In the second part of the thesis, we apply computational modeling to mimic and simplify real-life complex systems. We focus on clarifying how the network topology determines the initiation and propagation of synchronization. A simple but efficient method is proposed to reconstruct network structures from functional behaviors for oscillatory systems such as brain. We study the feasibility of network reconstruction systematically for different regimes of coupling and for different network topologies. We utilize the Kuramoto model, an interacting system of oscillators, which is simple but relevant enough to address our questions.

General Terms:
thesis, computational biology, bioinformatics, systems biology

Additional Key Words and Phrases:
Enhanced Phylogenetic Tree, phase correlation, structural connectivity, functional dynamics, static network, dynamical network system, time series, synchronization, structure prediction, protein-protein interaction, protein functional linkage, phylogenetics

Acknowledgements

Over the five years of my doctoral study, I have received support and encouragement from a number of individuals. Without them, this dissertation would not have been possible.

First of all, I would like to express my deep gratitude to my supervisor, Professor Liisa Holm, for being the best supervisor one can wish for. Her guidance and support truly have developed me as a researcher.

I would like to thank Dr. Sampsa Hautaniemi at University of Helsinki for being the mentor of my study. Sampsa has given a number of valuable comments and advices which help to strengthen my dissertation manuscript.

I would like to thank Dr. Petri Törönen and Päivi Rosenström at Bioinfomatics Group for nice and helpful discussions. I am very grateful to Matti Kankainen, my friend and colleague, for giving very wise advices. I specially thank Parik and Sanna Koskinen for being nice friends. Parik, I will never forget our wonderful badminton games.

I would also like to express my great appreciation to FICS, Finnish Graduate School in Computational Sciences, for the financial fundings and a number of useful courses. I also highly appreciate the helps by Zora BioSciences Oy during I was finalizing my dissertation.

I would like to thank all my dear friends, Quang and Trang Sarah, Hai and Tho, Su and Ngoc, Trung and Trang Lennon, Dina Ngoc and Erkki, Lily Hue and Jussi, and Barbara for being beside me in five years of living in Finland.

I specially would like to thank my parents who always support, encourage and believe in me and in all my endeavors. I would also like to thank my brother Hien, his wife Lien and my niece Zin Zin for their constant encouragement. I love you all.

Finally, I wish to thank my beloved wife Hien Pham and my son Tom, without whom this effort would have been worth nothing. Your endless love, support and constant patience have made me stronger. This thesis is dedicated to you. “Anh yeu em va con nhat tren doi va mai mai”.

Original Publications of the Thesis

This thesis is based on the following peer-reviewed articles, which are referred to as Publication I-III in the text.

I. Hung Xuan Ta, Patrik Koskinen, Liisa Holm. **A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees**. Bioinformatics 2011, Volume 27, Issue 5, 700-706.

II. Hung Xuan Ta, Chang No Yoon, Liisa Holm, Seung Kee Han. **Inferring physical connectivity of complex network from functional coherent activity**. BMC System Biology 2010, 4:70

III. Hung Xuan Ta, Liisa Holm. **Evaluation of different domain based methods in protein interaction prediction**. Biochemical and Biophysical Research Communications 2009, Volume 390, Issue 3, 357-362.

List of Abbreviations

AP/MS	Affinity Purification followed by Mass Spectrometry
AS	Association
AUC	Area Under Curve
BLAST	Basic Local Alignment Search Tool
BN	Bayesian Network
Co-IP	Co-immunoprecipitation
CV	Cross-Validation
DNM	Dynamical Network Modeling
EPT	Enhanced Phylogenetic Tree
FLM	Functional Linkage Map
FPR	False Positive Rate
FSS	Finite Size Scaling
GO	Gene Ontology
LOOCV	Leave-One-Out Cross-Validation
LUCA	Last Universal Common Ancestor
MLE	Maximum Likelihood Estimation
NJ	Neighbor-joining
NRS	Negative Reference Set
NSA	Network Structure Analysis

NSI	Network Structure Inference
PE	Parsimonious Estimation
PPC	Pairwise Phase Coherence
PPI	Protein-Protein Interaction
PPIN	Protein Physical Interaction Network
PRS	Positive Reference Set
ROC	Receiver Operating Characteristic
SF	Scale-free
TPR	True Positive Rate
UPGMA	Unweighted Pair Group Method using Arithmetic averages
Y2H	Yeast-Two-Hybrid

Contents

List of Figures **xiii**

1 Introduction **1**
1.1 Systems biology 1
1.2 Promises of systems biology 3
1.3 Inference, analysis and modeling of networks 4
1.4 Main contributions 6

2 Background **9**
2.1 Biological concepts and data resources 9
2.1.1 Concept of a protein 9
2.1.2 Protein function prediction 10
2.1.3 Phylogenetics based approaches for studying protein functions 11
2.1.4 Enhanced phylogenetic tree 14
2.1.5 Protein-protein interaction 15
2.1.6 Protein-protein interaction databases 17
2.2 Classification: basic concepts 19
2.3 Dynamical network modeling: basic concepts 21
2.3.1 Networks 21
2.3.2 Dynamical network systems 23
2.3.3 Synchronization 27

3 Structural inference of protein interaction networks from high throughput biological data **31**
3.1 Comparative genomic methods 31
3.2 Structure-based methods 33
3.3 Biological context methods 34
3.4 Domain-based prediction of PPIs 34
3.4.1 Methods for inferring DDIs from known PPIs 35
3.4.2 Predicting new PPIs based on inferred DDIs 36

3.5 Phylogeny-based methods for predicting PPIs 36
3.5.1 Phylogenetic profiling 36
3.5.2 Detecting protein functional linkages with enhanced phylogenetic trees 38
3.6 Challenges of computational PPI predictions 41

4 Dynamical network modeling of complex systems using coupled phase oscillator models 43
4.1 Synchronization in complex networks 43
4.1.1 Onset of synchronization 44
4.1.2 Path to synchronization 45
4.2 Reconstruction of physical connectivity from functional dynamics 46

5 Conclusions 49

References 53

List of Figures

1.1 Life's complexity pyramid 2

2.1 Definition of ortholog, in-paralog and out-paralog 12
2.2 Example of EPT and their decompositions 15
2.3 Network topologies in generalized random networks 22
2.4 Fixed point, limit cycle and strange attractor 24
2.5 Concept of pairwise phase correlation 29

3.1 Comparative genomics approaches for PPI prediction 33
3.2 Prediction performance of the EPT method 39
3.3 Effect of selecting reference organisms to the performance of the predicting methods. 40

4.1 Global order parameter 44
4.2 Organization of synchronization in a SF network system . . 47

5.1 An integrated disease-disease, disease-gene and gene-gene interaction network 50

Chapter 1

Introduction

This chapter introduces the field of systems biology and its three important components: network structure inference, network analysis, and dynamic network modeling. The chapter ends with a summary of the author's main contributions to the field.

1.1 Systems biology

Systems biology is a field of study that is aimed at a system-level understanding of biological systems, which are composed of molecular components. The idea of applying systems theory to biology is, however, not new. Notably, in the 1960s, a number of studies attempted to view living phenomena as a map of relationships among elements and attempted to handle all living systems with an approach that is analogous to approaches in chemistry and physics (Bertalanffy, 1968; Mesarović, 1968). The term "systems biology" was born at that time, but the use of the term had essentially no impact for the first three decades (Cornish-Bowden, 2011). This delayed impact mainly occurred because of the data on which to base the theories and models was inadequate (Albert, 2007). However, the remarkable progress in molecular biology, particularly the Human Genome project, brought an abundance of data, which stimulated a revival of systems biology.

Traditional molecular biology focuses on identifying individual molecules, such as genes, mRNA, proteins and metabolites and studying their properties and specific functions (Figure 1.1). This type of part-by-part study can reveal relatively limited insights about whole biological systems such as the human body because it usually looks at only a few aspects of a system at a time. Listing all of the parts of a system; does not result in an understanding of how the system operates, and more importantly, it does not result

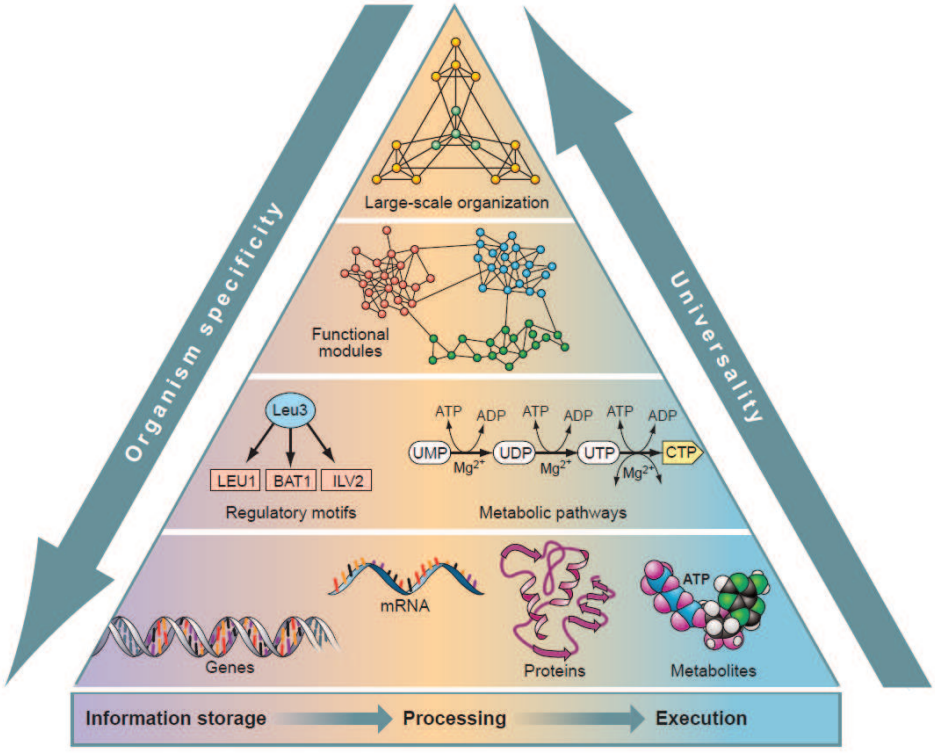

Figure 1.1: (Oltvai and Barabási, 2002). From the particular to the universal: The bottom of the pyramid shows at traditional representation of the cell's functional organization: genome, transcriptome, proteome, and metabolome (level 1). There is remarkable integration of the various layers both at the regulatory level and at the structural level. Insights into the logic of cellular organization can be achieved when we view the cell as a complex network in which the components are connected by functional links. At the lowest level, these components form genetic-regulatory motifs or metabolic pathways (level 2), which in turn are the building blocks of functional modules (level 3). These modules are nested, generating a scale-free hierarchical architecture (level 4). Although the individual components are unique to a given organism, the topological properties of cellular networks share surprising similarities with those of natural and social networks which suggests that universal organizing principles apply to all networks, from the cell to the World Wide Web (Oltvai and Barabási, 2002).

in an understanding of how the system works when some of its elements malfunction. This constrain implies that a system-wide understanding of the human body can help to better predict, prevent or remedy potential

health problems.

The post-genomic era has witnessed the development of high-throughput screening technologies such as DNA microarrays. High-throughput technologies lead to the availability of a collection of completely sequenced genomes. Currently, mass spectrometry technology can help to measure the concentration of thousands of proteins or metabolites at a time, creating a large body of biological data from diverse species. Once the full list of biological components is acquired, an understanding of how their interactions bring forth the distinctive properties of a species becomes more attainable. Such so-called "emergent properties" (Aderem *et al.*, 2011) are unpredictable and are not visible at the parts level because the parts interact with each other in nonlinear and nonadditive ways.

There has been a shift in thinking from considering a single gene, protein or metabolite at a time to thinking about multiple genes, proteins and metabolites acting in concert to form complexes, pathways or networks. Since 2000, a substantial number of studies have been published, and many institutions devoted to systems biology have been established. Systems biology combines the skills of biologists, clinical researchers, engineers, mathematicians, and computer scientists for the purpose of tackling the largest issues in understanding biological systems.

1.2 Promises of systems biology

State-of-the-art biology drives the development of new technologies and computational tools which, in turn, open new frontiers in biology. For example, the Human Genome Project motivated the development of high-throughput DNA sequencing methodologies. The need to screen the proteomes and metabolomes forced the development of mass spectrometry technologies. Consequently, these advances in biological data acquisition, together with computational strength in data management, and mathematical/statistical/modeling theories, have stimulated the reincarnation of systems biology. The appeal of problems in systems biology to engineers, mathematicians, and computer scientists will continue to spur not only the birth of new inexpensive, high-throughput, high-quality and high-speed data collection technologies but also new powerful computational tools and theories.

Systems biology is at the center of an iterative, incremental process of questioning, designing, engineering, and discovery. Systems biology makes use of the knowledge that is available from molecular biology to formulate graphical or mathematical models that are iteratively refined. New con-

cepts, findings and hypotheses acquired can assist clinical settings, and can also drive new experiments at the molecular level. Thus, systems biology is hypothesis-driven, global, quantitative, iterative, integrative, and dynamic (Aderem, 2005).

Systems biology not only aims to describe biological systems, but also targets the realm of prediction and control. Possibly the most exciting application of systems biology is to understand the relationship between health and disease and to develop more predictive, preventive and personalized medicine. A disease might be caused by internal factors inside the body such as genes and proteins, by external effects from the outside environment, or by a combination of causes. Cancer, diabetes, cardiovascular and neuro-degenerative diseases have been attracted much attention from the medical community because of their seriousness and their complexity. The 20th century brought us the Internet, cell phones, airplanes and many other advanced technologies, but we still suffer from complex diseases that remain the main causes of deaths worldwide. The reason is that we have not yet developed effective diagnostics that can help to predict disease occurrences and medicine that can help to prevent and cure the diseases.

The science underlying our traditional medical practices, from diagnosis and treatment to prevention, is based on the assumption that information about the individual parts is sufficient to explain the whole (Ahn *et al.*, 2006). Such a reductionist approach was responsible for tremendous success in medicine during the pregenomic era. However, the properties of the complex molecular networks within which diseases develop cannot be predicted by investigating the parts. For this reason, the aforementioned medical practices are, in many cases, inadequate.

Medicine of the future offers a medical model that emphasizes, in general, the customization of health-care in which all decisions and practices are tailored to individual patients. This approach constitutes personalized medicine. The technologies and tools of systems biology provide comprehensive information about a patient's proteomic, genetic and metabolic profiles which could be used to assess the patient's health status. Therefore the risk that a patient has certain diseases can be predicted at an early stage. Finally, the proper medication, with tailored dosages, is selected for each patient.

1.3 Inference, analysis and modeling of networks

The important first step of a system-level understanding of a biological system is to investigate interrelationships (organization or structure) among

molecules (Wolkenhauer, 2001). This step is called *network structure inference* (NSI). The NSI step comprises simply drawing the wiring diagram between genes, proteins, metabolites or neurons to form biological networks such as protein-protein interactions (PPI) and metabolic, signaling, transcription regulatory or neural networks. In such a graphical representation, the nodes are the system's elements (e.g., gene, proteins or other molecules) and the edges can encode their pairwise relationships (e.g., PPIs, protein-DNA interactions, co-expression relationships and functional linkages). NSI offers computational algorithms that input data and output interaction networks and that are consistent with the data that were input (Marbach *et al.*, 2010).

One type of method for biological NSIs makes use of Bayesian Network (BN) analysis (Friedman *et al.*, 2000; Emmert-Streib *et al.*, 2012). The BN methods describe biological elements as *variables* and the interaction network as *dependence* and *conditional independence* among the variables. The methods predict the interaction network as the best matches to the given data with help from prior information. The BN methods provide strong probabilistic frameworks with advantages including compact and institutive representation, the ability to capture a causal relationship, efficient model learning and the ability to address noisy data (Friedman *et al.*, 2000).

Another class of inference methods work by using data at the elements. Examples of the input data include gene/protein expression, metabolite concentration, protein domain decomposition and protein phylogenetic profiles. NSI uses computational approaches, including statistical correlation measures, such as the conditional correlation (Rice *et al.*, 2005), the co-occurrence probability (Wu *et al.*, 2003), the Pearson correlation coefficient (Ranea *et al.*, 2007), or it uses machine learning techniques such as maximum likelihood estimation (Deng *et al.*, 2002), parsimonious explanation (Guimaraes *et al.*, 2006) and to predict the structure of biological networks such as transcription-regulatory and PPI networks. In this thesis, the author investigates the computational prediction of PPIs by using comparative genomics approaches, particularly domain-based and phylogeny based methods. The reason for using comparative genomics methods is derived from the availability of the complete sequences of multiple genomes from diverse species. The limitation of the domain-based methods in predicting PPIs is addressed. A novel method for predicting protein functional linkage networks using *enhanced phylogenetic trees* (EPT) is proposed here.

The next step is called *network structure analysis* (NSA) in which graph theory is applied to extract new biological insights from a known interaction network. *Synthetic* analyses provide global information on the network,

whereas *divisive* analyses attempt to decompose or partition networks into smaller building blocks (Alon, 2006). These building blocks can be network motifs, which are topologically well-defined subgraphs that are highly enriched in the network compared to randomized networks, or clusters, which are densely connected network regions. Each type of network motif might encode a biochemical circuit that implements a specific biological function such as sign-sensitive accelerators (feed-forward loops) (Mangan and Alon, 2003). A cluster could implicate a biological machine that is composed of elements implementing shared tasks or participating in a common biological process, such as a signaling pathway.

The study of PPIs in the first part of this thesis is about inferring static interaction maps between proteins. However, living systems are dynamic. In a dynamical network system in which elements are coupled via interactions, the state of each element depends not only on itself but also on its interacting partners and how strongly it interacts with them. When we know about individual elements and the interactions among them, we consequently want to know how the system dynamically evolves under different conditions. *Dynamical network modeling* (DNM) aims to mimic and simplify complex real-life systems, using some relevant assumptions, to probe the changes in a system's behavior that arise from the perturbations to the system's elements and interactions. DNM can help to understand the dynamical aspects of systems, can generate new hypotheses and can assist experiment designs. In the second part of this thesis, we focus on understanding the relationship between the dynamical behaviors and the network structure in complex network systems. We attempt to clarify how the individual elements change coherently to form synchronization inside the system. Consequently, we study how the synchronization emerges and propagates inside systems with different network topologies, which range from regular to scale-free (SF). Finally, we test which coupling regimes and network topologies facilitate the reconstruction/inference of the network connectivity of a system from its dynamical behavior. In this thesis, we adopt the Kuramoto model, a network of oscillators in which the states of individual components are modeled by sinusoidal time series. This model is simple but relevant enough to address our questions.

1.4 Main contributions

The main contributions of this thesis are given in the original publications I-III. Below we summarize the main findings in the order of the original publications

I. Enhanced Phylogenetic Tree, a new graph presentation of evolutionary history of protein families is introduced. We present a novel method using enhanced phylogenetic trees for detecting functionally linked proteins. Our method significantly surpasses conventional phylogeny-based methods in prediction performance and potentially discovers more reliable protein functional linkages.

The author of this thesis jointly initiated and designed the work, carried out the computational research, analyzed the data and wrote the manuscript with guidance by Professor Liisa Holm at University of Helsinki.

II. This paper presents a comprehensive investigation on a family of dynamical network systems in both weak and strong coupling regimes, with the background networks that interpolates between regular and scale-free topologies. The main findings include an analysis on the path to synchronization in the complex network systems and a method for reconstructing the structure form functional dynamics.

The author of this thesis jointly initiated the work, carried out the mathematical research, implemented the numerical simulations and wrote the manuscript with guidance by Professor Seung Kee Han at Chungbuk National University and Professor Liisa Holm at University of Helsinki.

III. An systematical description of current domain-based approaches including the association method, maximum likelihood estimation and parsimonious explanation method. The performance of these methods at inferring DDIs and predicting PPIs was evaluated comparatively. The study noted artifacts that are generated by each method in certain situations and biases in the available benchmark sets.

The author of this thesis did the programming, implemented the data analysis and wrote the manuscript with guidance by Professor Liisa Holm at University of Helsinki.

Chapter 2

Background

This chapter provides background knowledge that is needed for discussions within this thesis. It introduces basic concepts, methods and data resources for biological network inference and dynamical network modeling.

2.1 Biological concepts and data resources

2.1.1 Concept of a protein

Proteins are composed of one or more polypeptides which are single linear polymer chains of amino acids. The sequence of amino acids in a protein is defined by the sequence of a gene and its transcript structures. While genes are the basic unit of heredity, proteins are the working molecules of cells; proteins perform many biological activities to keep cells functioning. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism. Proteins also display functions that include the flow of small molecules and ions (transport), sensing and reaction to the environment (signaling), control of protein activity (regulation), organization of the genome, lipid bilayer membrane and cytoplasm (structure), and the generation of force for movement (motor proteins).

Proteins were recognized as a distinct class of biological molecules in the eighteenth century by a French chemist Comte de Antoine Fourcroy and others, who distinguished by the molecules' ability to coagulate or flocculate under treatments with heat or acid. The term "protein" was proposed in 1838 by a Swedish chemist and clinician, Jöns Jacob Berzelius; protein is derived from the Greek word *proteios*, which means "primary of importance" (Wikipedia, 2004).

The primary structure of a protein is simply the linear arrangement, or sequence, of its amino acid residues. The protein secondary structure

comprises of regularly repeating local structures that are stabilized by hydrogen bonds. The most common examples are alpha helices, beta sheets and turns. The tertiary structure of a protein is the overall shape of a single protein molecule, which is stabilized by hydrophobic interactions between the non-polar side chains, hydrogen bonds between polar side chains and peptide bonds. The quaternary structure is an arrangement of multiple folded protein molecules.

A protein domain is a part of the protein sequence and structure that can evolve, function, and exist independently of the remainder of the protein chain. Each domain forms a compact three-dimensional structure and often can be independently stable and folded. The identification and characterization of protein domains and domain families is a major goal of protein science. The ADDA database is available for domain decomposition and clustering of all of the protein domain families (Heger and Holm, 2003). Protein domains can interact with each other. Domains of one protein can interact with those of another protein, mediating the interaction between those proteins (Publication III).

The proteome, termed as a portmanteau word of *proteins* and *genome*, is the entire set of proteins that are expressed by a genome, cell, tissue or organism. Proteomics is the field of study of such large-scale datasets. Key experimental technologies in proteomics include 2D electrophoresis (Görg *et al.*, 2004) to separate a large number of proteins, mass spectrometry (Conrotto and Souchelnytskyi, 2008) to identify proteins and sequence peptides in rapid high-throughput manners and protein microarrays (Joos and Bachmann, 2009) to detect the relative levels of a large number of proteins that are present in a cell. Typical techniques to screen protein-protein interactions are presented in section 2.1.5.

2.1.2 Protein function prediction

Once a new genome/proteome sequencing project has been completed, an important task that should be implemented is function prediction. Basically, the functions of a newly discovered protein can be predicted based on the relatedness between that protein and other proteins with known functions in databases. Assuming that similar protein sequences imply similar protein functions, some methods, called homology-based methods, predict protein functions by identifying the similarity in the sequences or structures at different levels, including motifs, domains, entire proteins, and secondary or tertiary structures (Eisen and Wu, 2002). Other methods group proteins in an organism or across organisms that share some common properties, for examples, a gene neighborhood (Overbeek *et al.*, 1999), into families and

then transfer functions among family members. These methods are non-homology approaches. Both homology and non-homology methods focus on detecting and quantifying the similarities or differences between species. By considering evolution, we can understand how and why those similarities and differences occurred (Eisen and Wu, 2002). Phylogenetics, therefore, became a promising approach for meeting the challenges of protein function prediction.

2.1.3 Phylogenetics based approaches for studying protein functions

Phylogenetics is the study of evolutionary relationship among species, based on molecular sequencing data or morphological data. The outcome of a phylogenetic analysis is expressed in a phylogenetic tree, which was first termed the "Tree of life" by Charles Darwin in his publication of *The Origin of Species* in 1849. A phylogenetic tree contains leaf nodes that present extant species, internal nodes that present ancestral species and a root that is the last universal common ancestor (LUCA). In an unrooted phylogenetic tree, the root is unknown. Unlike species trees in which the leaves are species, sequence trees contain leaves that are gene/protein sequences. Sequence trees describe the hierarchical relationship among the sequences and, therefore, can contain many proteins from one species.

Phylogenetics trees are also widely used to represent evolutionary relatedness among proteins. When the phylogenetics tree of a protein family is available, the function annotations can be transferred within the family. One important aspect of evolutionary relatedness among proteins is the issue of orthology and paralogy (Fitch, 1970). Two homologous proteins in two different species that evolved by speciation from a single ancestral protein are orthologs, whereas two paralogs are homologs that are derived by duplication. Paralogs predating the speciation event are denoted out-paralogs. Paralogs that were duplicated after the speciation event are denoted in-paralogs (Remm *et al.*, 2001). Figure 2.1 depicts examples of orthologs, inparalogs and outparalogs. While functional convergence tends to follow protein speciation (Peterson *et al.*, 2009), functional divergence frequently accompanies gene duplication (Lynch and Katju, 2004). Therefore, if one can determine whether the query protein is an ortholog or a paralog of a protein with known functions, one can decide whether to transfer the functions of the known protein to the query protein.

Phylogenetics approaches first build a sequence tree for the proteins that are of interest. Next, these approaches identify duplications and speciations by reconciling or mapping the sequence tree with a known species tree of the

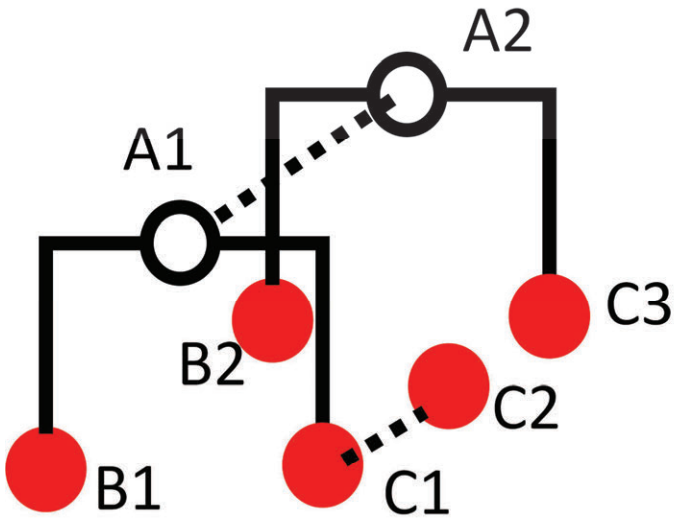

Figure 2.1: A hypothetical evolutionary relatedness among proteins B1, B2 in species B and C1, C2 and C3 in species C. The solid (dashed) edges represent speciation (duplication). C1 and C2 are inparalogs because their duplication occurred after speciation; they are co-orthologous to B1. B2 is an outparalog of the C1 and C2 genes, as are B2 and B1 (duplication and divergence prior to speciation).

organisms in which the proteins reside. In other words, the reconciliation models embed the protein tree in a species tree within which the proteins can evolve. To infer reconciliation, some parsimony (Zmasek and Eddy, 2002; Vernot *et al.*, 2008; Hahn, 2007) and probabilistic (Arvestad *et al.*, 2003; Sennblad and Lagergren, 2009; Gorecki *et al.*, 2011; Doyon *et al.*, 2012) frameworks have been developed. Parsimony methods search for an optimal reconciliation given the elementary costs of individual evolutionary events. Probabilistic methods seek reconciliation with maximum likelihood or maximum posterior probability. An excellent review of models, algorithms and programs for phylogeny reconciliation can be found in Doyon *et al.* (2011).

Construction of sequence trees

Methods that are used to construct the sequence tree for proteins of interest can be classified as either distance-based or character-based methods (Whelan *et al.*, 2001; Sleator, 2011). Distance-based methods, such as neighbor-joining (NJ) (Saitou and Nei, 1987) and the unweighted pair group method using arithmetic averages (UPGMA) (Sokal and Michener, 1958),

use a specific evolutionary model (i.e., amino acid substitutions) to calculate the evolutionary distance among proteins of interest. This distance reflects the expected average number of changes per site of sequences that have occurred since two proteins diverged from their common ancestral protein. The tree is constructed by repeatedly selecting the most closely related sequences that are distant from the others. The tree therefore minimizes the sum of the length of its branches (evolutionary distances). The simple NJ method produces unrooted trees and does not assume a constant rate of evolution through lineages. In contrast, UPGMA produces rooted trees and requires a constant-rate assumption. Distance-based methods are fast and available in some software packages such as MEGA4 (Tamura *et al.*, 2007). The main disadvantage of distance-matrix methods is their inability to efficiently use information about local high-variation regions that appear across multiple subtrees (Penny, 2004).

Character-based methods search for the most probable tree for a specific sequence set based on characters at each position of the sequence alignment and a model of evolution. The most common character-based approaches include maximum parsimony (MP) (Joseph and Felsenstein, 1996), maximum likelihood (ML) (Schadt *et al.*, 1998) and Bayesian inference (Yang and Rannala, 1997). The MP method selects the tree that requires the minimum number of character changes (mutations) to explain the given set of sequences. To detect the most parsimonious tree among a number of possible candidates, the MP method employs efficient search strategies, such as the branch and bound algorithm (Hendy and Penny, 1982), to exclude unnecessary regions of the search space from consideration.

ML methods are based on the specific probabilistic models of evolution and search for the tree with the maximum likelihood under these models. The concept of likelihood refers to the probability that a certain tree with a set of parameters (e.g., topology, branch-lengths etc.) produces a given set of data (sequences). Roughly, a tree that requires more mutations at interior nodes to explain the observed phylogeny will be assessed as having a lower probability. While the MP approach requires strict assumptions of consistency across sites and among lineages, the ML method permit varying rates of evolution across both lineages and sites. The ML method has a strong statistical foundation but it is computationally very expensive.

Bayesian inference can be used to produce phylogenetic trees by a method that is closely related to the ML methods. This method uses a prior probability distribution for the possible trees, which can be any one of the possible trees generated from the data. A Markov Chain Monte Carlo (MCMC) method is used to generate the set of trees with the highest posterior proba-

bilities (Yang and Rannala, 1997). The MCMC method can help to evaluate the posterior probabilities of trees without a need of summing over all possible topologies (Yang and Rannala, 1997).

PHYLIP (PHYLogeny Inference Package) is a free computational phylogenetics package with programs for inferring evolutionary trees (Felsenstein, 1989). Methods that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees.

2.1.4 Enhanced phylogenetic tree

In this section, we introduce the concept of the enhanced phylogenetic tree and its application to predicting functionally linked proteins. EPT is a novel graphical model of the evolution of proteins; EPT can account for the reconstructed proteomes of ancestral species and synchronous gene duplication events. EPT explicitly traces all of the descendants of proteins from the LUCA down to the extant species. EPTs are constructed based on the BLAST (Basic Local Alignment Search Tool) similarity of sequences using a hierarchical modification of the InParanoid algorithm (O'Brien *et al.*, 2005).

Figure 2.2 shows an example of EPT. In an EPT tree, nodes are proteins of extant or reconstructed species. These nodes are connected by either speciation or duplication edges. Each protein has an edge that connects to its parental protein in the parental proteomes (in the case of speciation) or in the same species (in the case of duplication). Sub-families are formed by proteins which are linked by speciation edges only. Duplication edges separate different sub-families. The corresponding occurrence profile presents the number of leaf proteins of the EPT in each species whereas the binary profile indicates the absence/presence of the leaf proteins of the EPT in each species.

In publication I, there are 91,428 protein families (EPTs) built from the proteomes of 572 complete genomes (560 prokaryotic and 12 eukaryotic organisms). The EPT method detected 2,467 subfamilies (orthologous groups) that contain both human and yeast proteins. In 2010, the Inparanoid database (Ostlund *et al.*, 2010) reported 2,154 orthologs between human and yeast species. However, the numbers of in-paralogs are not directly comparable between classification systems because there are different definitions and different clustering criteria. Namely, the above EPTs include 7,593 human and 2,514 yeast proteins that are below the common ancestor of human and yeast, while the 2,154 clusters of Inparanoid include 4,090 human and 2,534 yeast proteins.

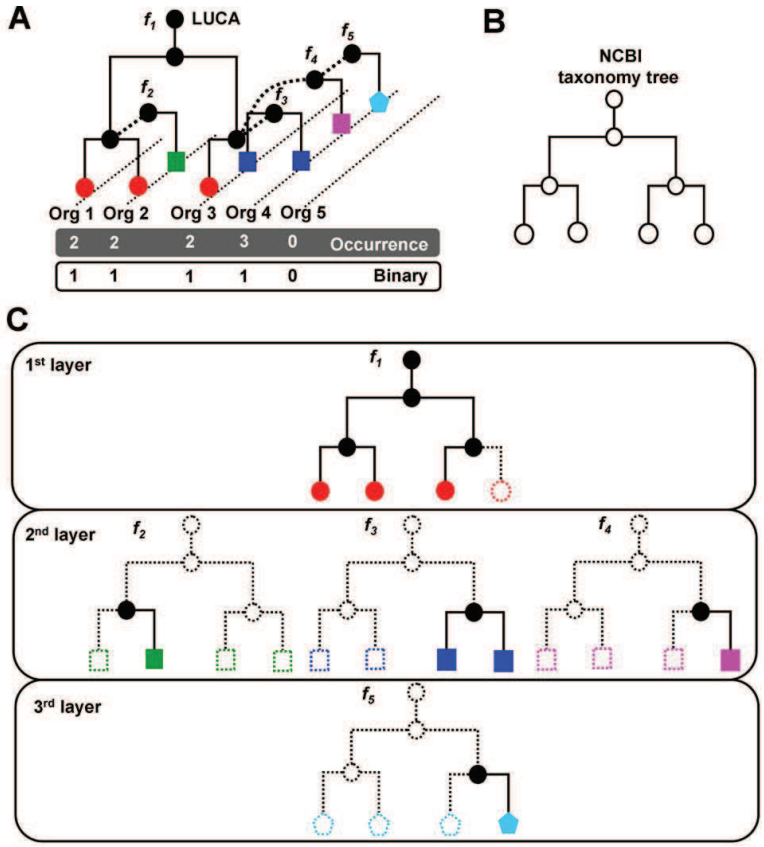

Figure 2.2: (Adapted from publication I) Example of EPT and its decompositions. (A) In the EPT, five round-dot lines represent the proteomes of five organisms. This tree has five subtrees f_1, f_2, f_3, f_4 and f_5, which form three layers. Each subtree, represents a subfamily, and is composed of the same color leaves, internal node and solid edges (speciation). The subtrees are separated by dashed edges (duplication). Below, the occurrence profile presents the number of leaf proteins in the EPT in each species whereas the binary profile indicates the absence/presence of the leaf proteins of the EPT in each species. (B) The NCBI species tree is used as a guide for constructing EPTs (see publication I for more details). (C) Three layers of the EPT. The first layer contains subtree f_1, the second layer contains f_2, f_3 and f_4, and the third layer contains f_5. In a subtree, the dashed edges (empty nodes) indicate the edges (nodes) of the NCBI taxonomy tree that do not exist in the subtree.

2.1.5 Protein-protein interaction

Proteins always interact with one another and with DNA, RNA and/or small molecules to keep cells functioning (Eisenberg *et al.*, 2000). PPIs can

be binary (physical) interactions, which refers to binding between two proteins that have residues that are in contact at some points in time, or PPIs can be functional linkages, which implies pairwise relationships between proteins that work together (i.e., participate in a common structural complex or pathway) to implement biological tasks. Understanding protein physical interaction networks (PPIN) or functional linkage maps (FLM) helps to unravel the molecular mechanisms of diseases. Protein interaction network information, therefore, has valuable applications to disease, personalized medicine, and pharmacology (Ideker and Sharan, 2008).

Traditionally, PPIs have been studied individually by biophysical and biochemical techniques which are considered to be gold standard techniques such as the co-immunoprecipitation (Co-IP) method. In a Co-IP assay, the protein (antigen) of interest and its interaction partners are co-precipitated by a specific antibody. Western blotting is then applied to identify proteins in the binding complex. The PPIs that are discovered by those small-scale assays are used to validate and assess many PPI datasets.

The low speed of traditional approaches has created a need for high-throughput screening techniques. Yeast-two-hybrid (Y2H) screening techniques have been used to generate binary PPI networks for *Saccharomyces cerevisiae* (Fromont-Racine *et al.*, 1997; Uetz *et al.*, 2000; Ito *et al.*, 2001; Yu *et al.*, 2008), *Caenorhabditis elegans* (Walhout *et al.*, 2000; Reboul *et al.*, 2003; Li *et al.*, 2004), *Drosophila melanogaster* (Giot *et al.*, 2003), and humans (Colland *et al.*, 2004; Rual *et al.*, 2005; Stelzl *et al.*, 2005). The work by von Mering *et al.* (2002) claimed that Y2H suffers from very high false positive rate. However, this evaluation appears to be excessive because the author inappropriately used "co-complex" protein sets that were derived from MIPs (Mewes *et al.*, 2004) to evaluate binary PPI sets by Y2H. Yu *et al.* (2008) proved that the Y2H technique can provide high-quality binary PPIs that cover 20% of all of the yeast binary interactions.

An alternative approach for generating co-complex interactome maps is tandem affinity purification followed by mass spectrometry (AP/MS). A number of studies using AP/MS have been performed on *Escherichia coli* (Butland *et al.*, 2005; Arifuzzaman *et al.*, 2006), *S. cerevisiae* (Gavin *et al.*, 2002, 2006; Krogan *et al.*, 2006), and humans (Ewing *et al.*, 2007).

The interactome maps discovered by high-throughput methods are still far from complete and are somewhat contradictory (von Mering *et al.*, 2002). Even when measuring the same type of interactions, different assays released datasets that suffer from a poor overlap. This outcome is skeptically interpreted as a consequence of false positive interactions caused by the technical limitations of PPI assays (von Mering *et al.*, 2002) or is optimistically ex-

plained to be a result of the low sensitivities of the methods (Yu *et al.*, 2008). Y2H and AP/MS are considered to provide orthogonal information about the interactome and both are vital to obtaining a complete picture of cellular PPI networks (Yu *et al.*, 2008).

To complement the experimental techniques, a number of computational methods have been developed (Shoemaker and Panchenko, 2007). Such in-silico approaches, including comparative genomic methods, structure-based methods and/or biological context based methods, have been used to assess or validate the available PPI datasets and to finally predict novel PPIs.

2.1.6 Protein-protein interaction databases

There are many databases of proteins and protein interactions which contain both experimental and computational data resources. Table 2.1 lists the databases of protein interactions and associations that were discovered by experimental and computational methods. These databases provide information on the interacting proteins, the method used to discover/predict the interactions/associations and the corresponding literature resources. In addition, the databases also provide tools for network visualization. Some databases are specialized for a certain organism, such as HPRD for humans, and some others, such as BioGrid, MINT, IntAct or DIP, provide data resources for most of the available species. Many databases, such as STRING, DIP or MINT, calculate the reliability of interactions that are based on experimental reproducibility or other biologically relevant information, such as gene expression profile (Deane *et al.*, 2002). Some databases, such as CORUM and MIPs, contain manually annotated/curated protein complexes. There have also been attempts to integrate data from many primary databases, such as PINA or IntAct. These databases accept manual curations by the user community.

The datasets provided by reliable databases can serve as *gold-standard* datasets of PPIs, which can be used to train and validate computational prediction models. The PPI datasets that were discovered by small-scale experiments are widely trusted by the community because it has been believed that the high-throughput technologies used to discover PPIs, such as Y2H, still suffer from a high false positive rate (von Mering *et al.*, 2002). In publication III, the benchmark dataset of PPIs contains 29,579 common PPIs of DIP and MINT. The PPIs in the intersection of the two databases are considered likely to be true positives while those that appear in only one of the databases are considered likely to be database-specific false positives. In publication I, we built a gold-standard set of co-complex proteins for humans (yeast) derived from the CORUM (MIPS) database. These manually

Table 2.1: Databases of protein-protein interactions and protein functional associations.

Database	Information	Refs
BIND	Peer-reviewed bio-molecular interaction database containing published interactions and complexes	(Bader *et al.*, 2003)
BioGRID	Protein and genetic interactions from major model species	(Stark *et al.*, 2011)
CORUM	Manually annotated protein complexes from mammalian organisms	(Ruepp *et al.*, 2010)
DIP	Experimentally determined interactions between proteins	(Salwinski *et al.*, 2004)
HPRD	Human protein domain architecture, post-translational modifications, interaction networks and disease associations	(Prasad *et al.*, 2009)
IntAct	Interaction data derived from literature curation or direct user submissions	(Aranda *et al.*, 2009)
MINT	Experimentally verified PPI mined from the scientific literature by expert curators	(Ceol *et al.*, 2010)
MIPS	Manually curated protein complexes for *Saccharomyces cerevisiae*	(Mewes *et al.*, 2004)
PINA	Protein interaction network construction, filtering, analysis, visualization and management	(Cowley *et al.*, 2012), (Wu *et al.*, 2009)
STRING	Known and predicted protein interactions including physical interactions and functional associations	(Szklarczyk *et al.*, 2011)

curated databases contain information that was obtained from individual experiments published in scientific articles, excluding the high-throughput datasets. The databases of protein complexes are appropriate for benchmarking functionally associated protein sets but not for physically interacting protein sets (Yu *et al.*, 2008). In (von Mering *et al.*, 2002), the authors used protein complexes in MIPS to evaluate physical protein interactions, which resulted in issues of bias against the physical interactions because not all of the proteins in a complex physically interact with each other, and not all physically interacting proteins occur within the same complexes.

2.2 Classification: basic concepts

In *classification* applications, it is of interest to predict a class, a category or any meaningful label using a set or sets of input features. In this thesis, typical examples include the classification of protein pairs into interacting and non-interacting categories on the basis of evolutionary or domain composition information. In general, there are two main categories of classification approaches: *supervised* and *unsupervised classification*. In the former category, all of the data must be labeled prior to the analysis; in other words, the class labels must be known prior to the construction of the classification model. After the model has been built, "unknown" or "unlabeled" samples can be classified. In unsupervised classification, information about sample classes is not analyzed or incorporated into the model construction process. Such information can be used to interpret the results and to discover potentially relevant associations between groups of samples (Azuaje, 2010).

Benchmark datasets that are used in classification are sets of data for which we know the categories. In supervised classification, benchmark datasets are used both in the training and the testing phases. In unsupervised approaches, benchmark sets serve only in the testing/evaluating phase. In binary classification approaches such as PPI predictions, benchmark datasets include *positive* and *negative reference sets* (PRS and NRS, respectively). PRSs are constructed by selecting known PPIs in reliable data sources, which were discovered by reliable experimental technologies (small-scale assays are preferred).

In publication III, the PRS of the PPIs contains 29,579 common PPIs that are in DIP (Xenarios *et al.*, 2001) and MINT (Zanzoni *et al.*, 2002), which were discovered in small-scale assays. In publication I, the PRSs of human and yeast protein functional linkages are constructed. The human PRS contains 26,813 pairs of proteins that co-occur in the same complexes derived from the CORUM database, a resource of manually annotated protein complexes from mammalian organisms (Ruepp *et al.*, 2010). The yeast PRS contains 5,888 pairs of co-complex proteins pairs, which were derived from the manually curated catalogs in MIPS (Mewes *et al.*, 2004).

An NRS is usually constructed by randomly choosing pairs of proteins that are not present in the corresponding PRS. Construction of PRSs and NRSs of the same size has been applied in prediction methodologies in bioinformatics (Ben-Hur and Noble, 2005; Yu *et al.*, 2008). Considering that the interactions are outnumbered by the number of negatives, an NRS of the same size as PRS likely contains few interacting pairs of proteins (Ben-Hur and Noble, 2005). In contrast, defining the NRS as a set of protein pairs that do not exist in the defined PRS (i.e. the complement of PRS) makes

the NRS contaminated by a number of true positives, which might have not been discovered (Ben-Hur and Noble, 2006). Taking a middle ground, some studies define reference sets that contain one positive for a specific number of negatives (Qi *et al.*, 2006), but the ratio between the sizes of the reference sets is arbitrary.

The classification performance can be accessed by the *Receiver Operating Characteristic* (ROC) curve, which depicts the relative trade-off between the costs and benefits. By *true positives* (TP), we mean samples in PRS that are predicted by the method as positives. Analogously, *true negatives* (TN) are samples in the NRS that are classified as negative predictions. *False positives* (FP) are samples in the NRS that are predicted as positives, while *false negatives* (FN) are samples that are predicted as negatives but are not present in the NRS. In the plot of the ROC curve, the x-axis represents the *false positive rate* (FPR) or 1-specificity, in other words, FP/(TN+FP), and the y-axis represents the *true positive rate* (TPR) or sensitivity, TP/(TP+FN). The classification performance is quantified by the area under the ROC curve (AUC), a measurement that is typically used for model comparison in machine learning studies (Hanley and McNeil, 1983). This measure can be interpreted as the probability that, when we randomly pick one positive and one negative sample, the classifier will assign a higher score to the positive sample than to the negative sample. An AUC of 0.5 reflects a classification by randomly choosing samples from the benchmark sets, while AUC=1 implies a perfect classification.

Cross-validation (CV) is a technique for evaluating how accurately a supervised classification would perform in practice, i.e., how the classification results will generalize to a different, independent dataset. Because we have a model with unknown parameters, we use known data to train the model, to optimize the model parameters that make the model fit the training dataset. Next, we test the trained model by using an independent dataset that was sampled from the same population as the training datasets. The model result is called *overfitting* if the model performs well on the training dataset but does not fit the testing dataset; overfitting means that the model describe noise but not the underlying relationship. Overfitting usually occurs when the number of parameters of the model is too large relative to the size of the training dataset. In one round of CV, the sample dataset is partitioned into subsets; the model fitting is performed on one subset (the training set) and then validated on the other set (the testing set). Multiple rounds of CV are implemented using independent partitions, and the validation results are averaged over rounds, which helps to reduce the variability.

Common types of CV include *k-fold* and *leave-one-out* CV (LOOCV). In k-fold CV, the sample dataset is partitioned into k subsets, where $k-1$ subsets are used as training data and one single subset is retained as testing data. The CV process is repeated k times so that every subset is used once as the testing dataset. The k validation results are averaged. LOOCV involves using a single observation in the sample dataset as the testing data and the others as the training data. This process is repeated so that each observation is used once as the testing data. When the size of the sample dataset is small, LOOCV is preferred (Azuaje, 2010).

2.3 Dynamical network modeling: basic concepts

2.3.1 Networks

Network maps are popular in many fields. Networks can be tangible, such as communication networks, electricity power grids, subway systems or systems of interest in biology and medicine, including neural networks or genetic, metabolic and protein networks. Networks can also be intangible, such as networks of acquaintances or collaborations between individuals.

Box 1. Network properties

Let $G = (V, E)$ denote a network on a set of *vertices* (or *nodes*) V and a set of *edges* E. The *degree* refers to the number of connections or interactions of a node. The *clustering coefficient* of a node is the proportion of possible connections between the neighbors of the node that are actually observed for a given node, which is a measure of the connection density around a node (Holland and Leinhardt, 1971). The *distance* between two nodes is the length of the *shortest path* connecting them in the network. The *diameter* of a network is the length or distance of the longest of all of the shortest paths between a pair of nodes in the network. The *characteristic path length* of a network is the average value of all of the shortest path lengths between all of the nodes in the network. The *betweenness centrality* of a node quantifies the number of non-redundant shortest paths passing through the node (Freeman, 1977). For more network concepts, read the book by Newman (2010)

Historically, the study of networks has mainly been the domain of a branch of discrete mathematics known as graph theory, which was initiated by the Swiss mathematician Leonhard Euler. In recent decades, there has been a movement toward performing research on complex systems. Significant progress has been made toward understanding the implications of the network's topological features on the network dynamics and function,

especially in biological networks (Zhou *et al.*, 2006). Box 1 describes basic topological properties of a network.

Network topologies

In spite of the remarkable diversity in the networks that appear in nature, their architecture is governed by a few simple principles that are common to most networks of major scientific and technological interest (Albert and Barabási, 2002; Dorogovtsev and Mendes, 2003). For decades network systems have been modeled either as chains, grids, lattices and fully-connected graphs which are completely *regular* or as *random* Erdös-Rényi network whose node degrees follow a Poisson distribution (Erdös and Rényi, 1960).

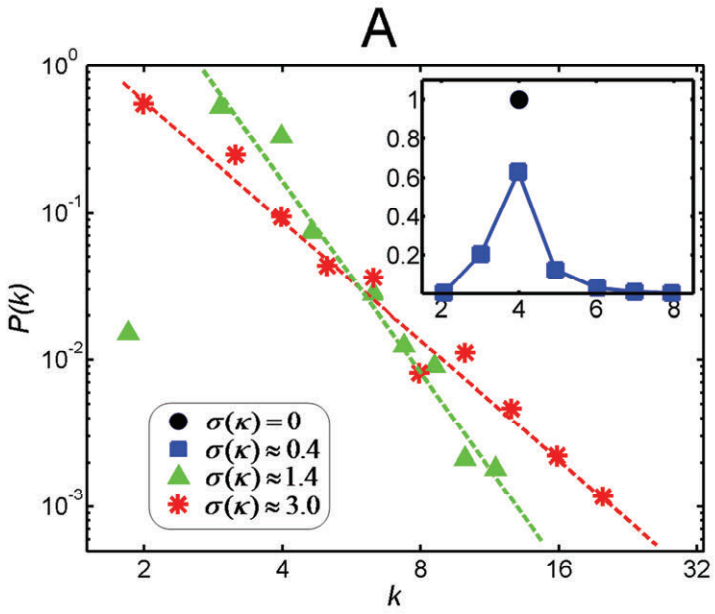

Figure 2.3: (Adapted from publication II). Degree distribution of generalized random networks with different topologies corresponding to $\sigma(\kappa) = 0$ (regular), 0.4, 1.4 and 3.0 (SF).

However, topologies of networks in real life are not trivial but complex. A number of recent findings indicate that real networks including large communication systems (Albert *et al.*, 1999; Newman *et al.*, 2002; Vázquez *et al.*, 2002), biological systems (Camacho *et al.*, 2002; Yu *et al.*, 2008), and a variety of social interaction structures (Newman, 2001; Zhou *et al.*, 2011) are characterized by a power-law degree distribution, $P(k) \propto k^{-\gamma}$, where degree k is the number of neighbors of a given node. In these *scale-free*

networks, most of the nodes have only a few links, whereas a few nodes have a very large number of links, which are often called hubs (Barabási and Albert, 1999).

Watts and Strogatz (1998) showed that many real-world networks have a small average shortest path length, but also a clustering coefficient significantly higher than expected by random chance, which is equal to the ratio of the mean degree to the network size. This feature is known as the *small-world* property. Many empirical examples of small-world networks have been documented in fields ranging from cell biology to business (Wagner and Fell, 2001; Jeong *et al.*, 2000; Basler *et al.*, 2011; Sporns, 2011; Adamic, 1999; Amaral *et al.*, 2000; Sporns *et al.*, 2000).

Generalized random graphs is a family of static random networks that satisfy the degree distribution $P(k) \propto k^{-\gamma}e^{-k/\kappa}$ with a controllable exponential cut-off scale κ (Newman *et al.*, 2001). Many real-world graphs show this exponential cut-off in the degree distribution (Amaral *et al.*, 2000; Newman, 2001). In these networks, the variance in the degree distribution, $\sigma(\kappa)$, varies as a function of κ while the mean degree remains constant. $\sigma(\kappa)$ measures the degree of heterogeneity of the networks; for example, $\sigma(\kappa) = 0$ corresponds to a regular network, $\sigma(\kappa) \approx 0.4$ corresponds to a homogeneous network and $\sigma(\kappa) \approx 3.0$ corresponds to a heterogeneous SF network (Figure 2.3). The Barabási and Albert (BA) network, which can be grown by using the preferential attachment rule (Barabási *et al.*, 1999), is a special case of this network family. The BA network has an SF degree distribution $P(k) \propto k^{-\gamma}$, with a scaling exponent of $\gamma \approx 3$.

2.3.2 Dynamical network systems

In a network system, the connectivity among the elements can be static. The network system is dynamic when the states of the elements evolve over time. In a biological context, a state might be the concentration of a molecule, the phosphorylation state of an enzyme, the expression level of a gene, the depolarization of a neuron or a circadian rhythm. In other words, a dynamical network system is a network of single interacting dynamical systems.

A single dynamical system can often be modeled by the differential equation $d\mathbf{x}/dt = \mathbf{v}(\mathbf{x})$, where $\mathbf{x}(t) = (x_1(t), ..., x_n(t))$ is a vector of state variables, t is time, and $\mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), ..., v_n(\mathbf{x}))$ is a vector of functions that encode the dynamics. One can imagine that the system $\mathbf{x}(t)$ evolves in an n-dimensional state space with axes $x_1, ..., x_n$ and is guided by the *velocity* field $d\mathbf{x}/dt = \mathbf{v}(\mathbf{x})$. Figure 2.4 shows a dynamical system that has three state variables. The system's state is presented by a vector $\mathbf{x}(t) =$

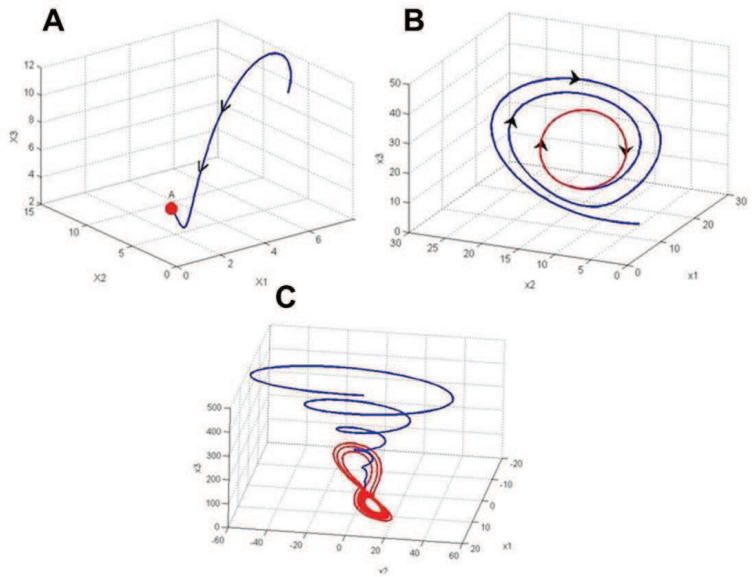

Figure 2.4: Examples of a fixed point, limit cycle and strange attractor in a 3-dimensional state space. The red curves represent attractors, and the blue curves represent the trajectory of the system. (A) The system comes to rest at point A, a fixed point. (B) The system is attracted into a limit-cycle. (C) The system settles on a space of fractional dimension, a chaotic or strange attractor.

$(x_1(t), x_2(t), x_3(t))$, which evolves in 3-dimensional space. When $\mathbf{x}(t)$ comes to rest at point $\mathbf{x}^*$, the velocity must be zero. We call $\mathbf{x}^*$ a fixed point, which is corresponding to an equilibrium state (Figure 2.4A). The fixed point is stable (a fixed point attractor) if any small disturbances damp out. An attractor that attracts systems to a cyclic path in phase space is a limit-cycle (Figure 2.4B). This attractor represents a self-sustained oscillation of the physical system (Strogatz, 2001). A chaotic or strange attractor pulls a system into a space that has a fractional dimension, where the system become caught between a two-dimensional plane and a three-dimensional solid (Figure 2.4C).

Many such dynamical systems couple together to form a dynamical network system. For example, neurons are coupled through signal transmissions from transmitters to receptors to form a neural network. In an ecological system, the dynamics of populations of species are coupled through diffusion along spatial coordinates and through trophic interactions. The stronger the coupling between the units of the network system is, the easier it is for the units to behave in a concerted fashion.

The collective behavior of dynamical systems in a network is also af-

fected by the local dynamics of the individual units. The network tends to lock into a static pattern if the dynamical system at each node has stable fixed points but no other attractors (Strogatz, 2001). At the opposite extreme, some network systems are composed of nodes that have chaotic attractors. These systems have been used mainly in the simplest mathematical settings, rather than as models of real systems (Strogatz, 2001). The intermediate case, in which each node has a stable limit cycle, has been considered in research inspired by biological examples, which range from the mutual synchronization of cardiac pacemaker cells, rhythmically flashing fireflies and chorusing crickets to wave propagation in the heart, brain, intestine and nervous system (Winfree, 2001).

Governing equations

In a network of N coupled limit-cycle oscillators, the oscillator at node i can be characterized by its phase θ_i, which evolves because of the coupling among the nodes and the difference between its intrinsic frequency ω_i and the frequencies of other nodes. Hence, the problem in terms of a large population of interacting limit-cycle oscillators, can be formulated by the following model (Winfree, 1967):

$$\frac{d\theta_i}{dt} = \omega_i + \left(\sum_{j=1}^{N} X(\theta_j) \right) Z(\theta_i), \tag{2.1}$$

where $i = 1, ..., N$. Each oscillator j exerts a phase-dependent influence $X(\theta_j)$ on all of the other oscillators; the corresponding response of oscillator i depends on its phase θ_i, through the sensitivity function $Z(\theta_i)$. Winfree (1967) showed that the system behaved incoherently, with each node oscillating at its intrinsic frequency, when the coupling is small, compared to the intrinsic frequencies.

Kuramoto (1984) modeled the systems by a firmer form as follows:

$$\frac{d\theta_i}{dt} = \omega_i + \sum_{j=1}^{N} \Gamma_{ij}(\theta_j - \theta_i), i = 1, ..., N, \tag{2.2}$$

where Γ_{ij} is the interaction function between two oscillators i and j. However, equations 2.2 is still too difficult to analyze in general because the interaction functions could have arbitrarily many Fourier harmonics. The Kuramoto model corresponds to the simplest possible case of equally weighted, all-to-all, purely sinusoidal coupling:

$$\Gamma_{ij}(\theta_j - \theta_i) = \frac{K}{N} sin(\theta_j - \theta_i), \tag{2.3}$$

where $K \geq 0$ is the coupling strength. The factor $1/N$ cancels out the dependence on the size of the system, which help to keep the coupling term to be not as intensive in the thermodynamic limit $N \to \infty$. This scenario allows the governing equations to be the following

$$\frac{d\theta_i}{dt} = \omega_i + \frac{K}{N}\sum_{j=1}^{N} sin(\theta_j - \theta_i), i = 1, ..., N \tag{2.4}$$

For simplicity, Kuramoto assumed that the intrinsic frequencies follow a unimodal and symmetric distribution around the mean frequency Ω. By redefining $\theta_i \to \theta_i + \Omega t$, one can rewrite equations 2.4 as follows:

$$\frac{d\theta_i}{dt} = \omega_i - \Omega + \sum_{j=1}^{N} \frac{K}{N} sin(\theta_j - \theta_i), i = 1, ..., N, \tag{2.5}$$

where the mean frequency Ω can be set to zero without losing generality, the ω_i's denote deviations from the mean frequency.

Kuramoto model on complex networks

The Kuramoto model can be generalized to include topological information as follow:

$$\frac{d\theta_i}{dt} = \omega_i + \sum_{j=1}^{N} \sigma_{ij} A_{ij} sin(\theta_j - \theta_i), i = 1, ..., N, \tag{2.6}$$

where σ_{ij} is the coupling strength between pairs of connected oscillators i and j, and A_{ij} is an element of the adjacency matrix that is 1 if nodes i and j are connected and 0 otherwise.

In some weighted interacting network models, σ_{ij} is randomly distributed in a range that is compatible with ω_i. It is also popular to set $\sigma_{ij} = K/k_i$, where k_i is the degree of node i. This equation means that the interaction strength of a node is averaged by the number of its neighbors, canceling out the potential effects of topological properties in heterogeneous networks. To study the dynamics of different topologies, it is more appropriate to set $\sigma_{ij} = K$, a constant coupling strength. Then, equation 2.6 reduces to

$$\frac{d\theta_i}{dt} = \omega_i + K\sum_{j=1}^{N} A_{ij} sin(\theta_j - \theta_i), i = 1, ..., N. \tag{2.7}$$

For a more realistic model, Gaussian white noise $\xi_i(t)$ with an intensity of D, satisfying $\langle \xi_i(t) \rangle = 0$ and $\langle \xi_i(t)\xi_j(s) \rangle = 2D\delta_{ij}\delta(t-s)$, can be added

to equation 2.7. Finally, the Kuramoto model on complex networks can be presented as follow

$$\frac{d\theta_i}{dt} = \omega_i + K\sum_{j=1}^{N} A_{ij} sin(\theta_j - \theta_i) + \xi_i(t), i = 1, ..., N. \tag{2.8}$$

2.3.3 Synchronization

Synchronization is a dynamical state in which two or more individuals behave coherently, almost unison. Synchronization is popularly observed in a broad range of systems, from physics and chemistry to biology and the social sciences (Pikovsky *et al.*, 2001; Kuramoto, 1984; Watts and Strogatz, 1998). In biology, examples include the synchronization of thousands of cells to keep our heart beating rhythmically or the synchronized firing of thousands of neurons to respond to external stimuli. It is observed that there is causal nexus between a sudden synchronization and some diseases such as Parkinson's disease (Hammond *et al.*, 2007) or epileptic seizures (Fisher *et al.*, 2005). Hence, the relationship between dynamical activities, such as synchronization, and the network structure is a central issue of dynamical network modeling when attempting to understand the functioning of real-world systems (Eisenberg *et al.*, 2000; Ideker and Sharan, 2008; Fromont-Racine *et al.*, 1997).

In a network of weakly coupled non-identical limit-cycle oscillators, modeled by equations 2.8, each oscillator is free to rotate at his own frequency when the coupling strength is small compared to the spread of intrinsic frequencies (Winfree, 1967). Then, as the coupling increases and crosses a certain threshold, a small group of oscillators starts to oscillate with the same rhythm. As the coupling becomes stronger, all of the oscillators become locked in phase and amplitude, and the system becomes completely synchronized. This so-called collective behavior of a system of coupled phase oscillators can be efficiently studied using the global order parameter.

Global synchronization

The global synchronization or collective behavior of the oscillator system is conventionally represented by the global order parameter defined as

$$R = \left\langle \left| \frac{1}{N} \sum_{j=1}^{N} e^{i\theta_j(t)} \right| \right\rangle_t , \tag{2.9}$$

where the brackets $\langle \ldots \rangle_t$ signify time averaging. The global order parameter measures the extent of global synchronization of the population of N oscillators (Strogatz, 2001).

The global order parameter is calculated at different regimes of coupling strength to study how the collective behavior of all of the oscillators changes between fully desynchronized and fully synchronized states (Figure 2.3B). The onset of global synchronization occurs at the *critical coupling strength*, K_c, which can be determined by a Finite Size Scaling (FSS) analysis (Hong *et al.*, 2002)

Pairwise phase correlation

The global order parameter, as a function of coupling strength, reflects the path from an incoherent to a coherent state of the system. The global order parameter, however, fails to describe where the synchronization emerges and how it propagates inside the system. Study of the local synchronization can provide more insights into the dynamical behavior of the system.

The local synchronization between two phase oscillators i and j, either connected (with physical connection) or disconnected (a non-physical connection), is quantified by pairwise phase coherence (PPC) which is defined as follows:

$$C_{ij} = \lim_{T\to\infty} \left| \frac{1}{T} \int_0^T e^{i(\theta_j(t)-\theta_i(t))} dt \right|. \tag{2.10}$$

C_{ij} , which shows how dependent the motions of two oscillators at nodes i and j are, is equal to 0 or 1, which correspond to full incoherence or coherence between nodes i and j, respectively. In a unit circle (Figure 2.5A) an arrow presents the phase difference between nodes i and j, $\delta\theta_{ij} = \theta_i - \theta_j$, at a point in time. When i and j are not synchronized, arrows for different points in time uniformly distribute on the unit circle, making the average of them is close to zero. The corresponding phase space (at the top of Figure 2.5B) shows no dependence between θ_i and θ_j. When i and j are synchronized, the arrows on the unit circle deviate slightly around a specific point, and their average, denoted by the red arrow, is significant. The corresponding phase space (at the bottom of Figure 2.5B) shows a strong dependence between θ_i and θ_j and the temporal profile of $\delta\theta_{ij}$ shows steps (at the bottom of Figure 2.5C), meaning that i and j are strongly correlated.

Local Order Parameter and Effective Coupling Strength

The governing equations in 2.8 can be rewritten approximately as follows

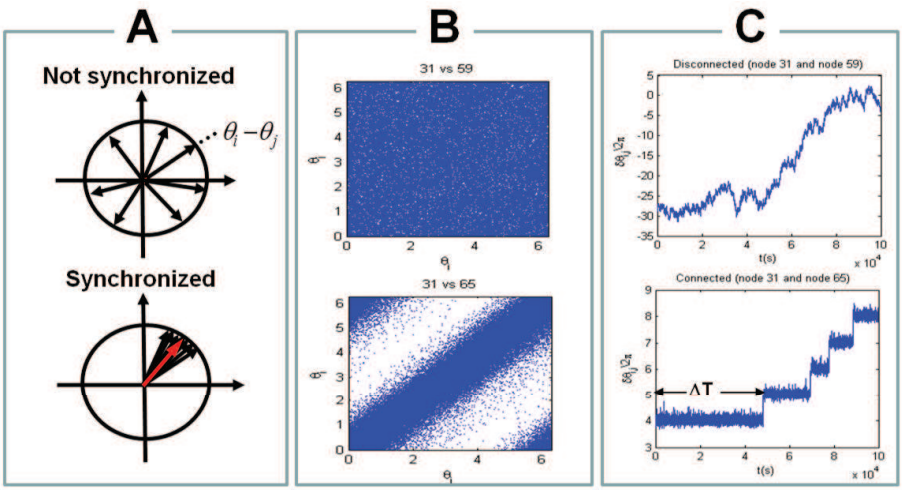

Figure 2.5: (A) Pairwise phase difference, $\delta\theta_{ij} = \theta_i - \theta_j$, in a unit circle when nodes i and j are not synchronized and they are synchronized. A black arrow presents the phase difference at a point in time and the red arrow represents the average of the phase differences over the time. The length of the red arrow presents the magnitude of C_{ij}. (B) Dependence of two phases on each other. (C) Temporal profile of the phase difference.

$$\frac{d\theta_i}{dt} = \omega_i + K k_i r_i \sin(\psi_i(t) - \theta_i(t)) + \xi_i(t), \tag{2.11}$$

where

$$r_i e^{i\psi_i(t)} = \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} e^{i\theta_j(t)}, \tag{2.12}$$

In equation 2.12, r_i is the local order parameter, which measures the coherence of k_i neighbors of node i. The oscillator at node i interacts with the network via an effective coupling strength of $\tilde{K}_i = K k_i r_i$.

Chapter 3

Structural inference of protein interaction networks from high throughput biological data

A wide range of biological molecules including genes, proteins and metabolites, interact with each other to maintain cellular functions (Eisenberg *et al.*, 2000). Examples of biological networks include protein interaction networks, whose nodes are proteins that are linked to each other by physical interactions (Ideker and Sharan, 2008; Stelzl *et al.*, 2005), metabolic networks, whose nodes are metabolites that are linked if they participate in the same biochemical reactions (Jeong *et al.*, 2000) and regulatory networks, whose directed links represent regulatory relationships between a transcription factor and a gene (Consortium *et al.*, 2005). The availability of increasing numbers of diverse 'omic' datasets together with the need to discover complex associations between genes and disease have motivated studies on inferences of biological network structure. Networks of gene-gene, gene-protein or protein-protein interactions can be reconstructed by using information that is extracted (manually or automatically) from the literature or by directly applying automated inference algorithms on large-scale experimental data, such as gene expression or phylogeny data. This chapter's focus is on discussions of computational approaches to structural inference of protein-protein interaction networks, including the author's research on this topic.

3.1 Comparative genomic methods

The complete sequencing of multiple genomes from diverse species provides an excellent opportunity to develop comparative approaches for functional

studies in proteomics. Physical interactions and functional linkages of proteins can be inferred via various patterns across many genomes. These patterns include the co-localization of genes on chromosomes (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999), the genetic fusion of two distinct proteins from one organism into a single protein in another organism (Enright *et al.*, 1999; Marcotte *et al.*, 1999), the domain composition of proteins (publication III) and phylogenetic profiles (Pellegrini *et al.*, 1999; Ranea *et al.*, 2007; Wu *et al.*, 2003; Glazko and Mushegian, 2004; Barker and Pagel, 2005; Barker *et al.*, 2007; Vert, 2002; Juan *et al.*, 2008).

Co-localization, or gene neighborhood methods detect pairs of genes that are physically close to each other on the genomes (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999) (Figure 3.1A). These pairs of genes may encode proteins that are physically interacting or functionally linked. Gene neighborhood information can be obtained from many resources, such as STRING (Szklarczyk *et al.*, 2011) or Predictome databases. These methods fail to detect interactions between proteins that are encoded by distant located genes.

Gene fusion methods (Figure 3.1B) are based on the premise that two distinct genes in a genome that are found to be fused into a common gene in another genome could encode physically interacting or functionally linked proteins (Enright *et al.*, 1999; Marcotte *et al.*, 1999). These methods are the ultimate form of gene co-localization method. Comprehensive sets of fused genes can be found at STRING (Szklarczyk *et al.*, 2011) and Predictome (Mellor *et al.*, 2002).

Domain-based and phylogeny-based methods are members of classes of methods that predict PPIs by two complementary procedures, *upcasting* and *downcasting* (Lappe *et al.*, 2001). They first perform upcasting by generalizing known individual protein-protein interactions to interactions between higher level entities such as cellular compartments, functional modules (e.g., group of orthologous proteins), or structural classes (e.g., protein domain families). Then, the methods perform downcasting the protein family interactions back to the protein level, with which they can predict new PPIs. Going up to a higher level of abstraction can gain generality but loses specificity. In the downcasting procedure, predicting interactions among all of the members of two protein families can generate false positives, especially when the two families are too large. The commonly used domain-based and phylogeny-based methods are presented in section 3.4 and 3.5, respectively.

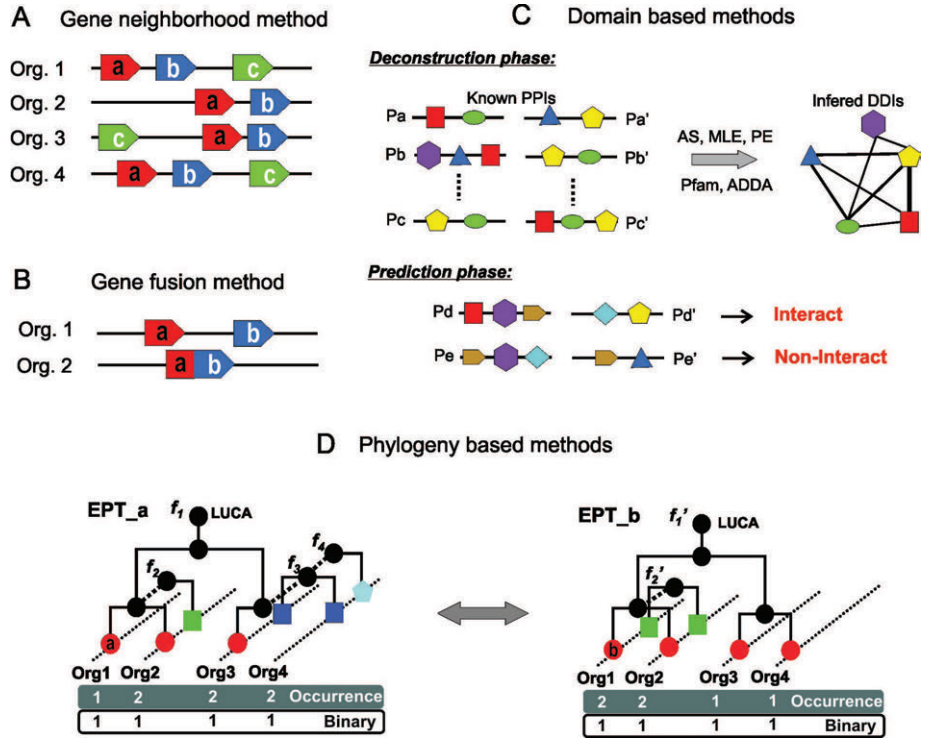

Figure 3.1: Comparative genomics approaches for PPI prediction. (A) Proteins encoded by genes **a** and **b** that are neighbors of each other in many organisms are predicted to interact. (B) Proteins encoded by genes **a** and **b** that combine (fuse) to form one gene in another organism are predicted to interact. (C) Protein **Pe** and **Pe'** (**Pd** and **Pd'**) are predicted to interact (not to interact) because they contain (do not contain any) interacting domains that are inferred from a known PPI network. (D) Phylogeny-based methods predict that two proteins that have similar binary profiles, occurrence profiles or EPTs are functionally linked proteins.

3.2 Structure-based methods

Primary protein structure methods are based on the hypothesis that PPIs might be mediated by short amino acid sequences such as motifs or k-mer compositions (Zaki *et al.*, 2009; Shen *et al.*, 2007; Ben-Hur and Noble, 2005). These methods acquire their sequence information from known experimentally discovered PPI data (the learning period) and then extrapolate to predict novel PPIs (the testing period). By making use of kernel functions, these methods might use only sequence information (Shen *et al.*, 2007) or combine with physicochemical properties, gene oncology (GO) annotations or homologous interactions (Ben-Hur and Noble, 2005).

An alternative category of structure methods predict interactions among proteins that have known three-dimensional (3D) structures (Hue *et al.*, 2010). Some approaches, such as the Protein-Protein Interaction Server (Jones and Thornton, 1996) analyze physical parameters, including the accessible surface area, the solvation potential, the residue interface potentials, hydrophobicity or planarity to identify the protein interaction sites. An alternative method, the InterPreTS, tests whether a target protein pair fits a known 3D structure of an interacting pair of proteins. 3D structure based approaches can determine not only whether two proteins interact but also the active sites and the physical characteristics of the interactions. However, these techniques are limited because only a small proportion of the proteins have accurate 3D structures.

3.3 Biological context methods

These methods make use of biological information concerning genes or proteins, such as gene expression, GO annotations or text information extracted from the literature (PubMed), which have predictive values for PPI predictions. Two interacting proteins can be co-expressed, and they could have similar GO annotations or might be mentioned in the same publication. Gene expression data can be derived from public resources, such as MIAME (Brazma *et al.*, 2001), SMD (Hubble *et al.*, 2009) or GEO (Edgar *et al.*, 2002) and GO annotations can be obtained from the Gene Ontology server (Ashburner *et al.*, 2000). These types of indirect biological information are usually integrated with direct measurements of PPIs, such as experiment-based PPI datasets, sequence information or genomic features, by using machine learning techniques that include Naive Bayesian framework (Guan *et al.*, 2010), logistic regression (Bader *et al.*, 2004), decision trees (Ferrer *et al.*, 2010), random forest (Qi *et al.*, 2009) and kernel methods (Ben-Hur and Noble, 2005; Yamanishi *et al.*, 2004). These integration studies employ various classifiers, feature sets and gold-standard datasets.

3.4 Domain-based prediction of PPIs

Domain-based methods make use of the assumption that PPIs are mediated by domains. In the first phase, these approaches infer DDIs from known PPIs by applying association (AS) analysis (Sprinzak and Margalit, 2001), maximum likelihood estimation (MLE) (Deng *et al.*, 2002; Riley *et al.*, 2005) or parsimonious explanation (PE) (Guimaraes *et al.*, 2006). In the next phase the domain-based approaches predict PPIs based on the

inferred DDIs. Namely, the methods assign an interaction to a pair of proteins if they contain domains that have been predicted as interacting. In the upcasting procedure, domain-based methods generalize known interactions between individual proteins to interactions between protein classes (i.e., a group of proteins that contain a common domain), and then the methods predict new interactions between all of the members of the interacting classes (the downcasting procedure).

3.4.1 Methods for inferring DDIs from known PPIs

In the AS method, each domain pair is assigned an $ASscore$, which is the ratio of the number of occurrences to the total number of protein pairs that contain the two domains. A high $ASscore$ indicates that two domains co-appear more frequently than expected at random, and hence, the domain pair is defined as a DDI. This method gives high scores to a very large number of potential interactions between the domain families that have very few protein members (only one or two members), which appear to be false positives. On the other hand, the AS method fails to detect hundreds of potential interactions between domain families that have many members, even though many PPIs were observed between these domain families.

The MLE method assigns interaction probabilities to all possible domain pairs to maximize the likelihood of observing the known PPIs. Similar to the AS method, the MLE method assigns high interaction probabilities to pairs of domain families that have very few protein members. Riley *et al.* (2005) introduced an $Escore$ to each domain pair, which measures the reduction in the likelihood of the observed PPIs that is caused by excluding that domain interaction. $Escore$ helps to overcome the problem of false positives and negatives by the MLE method. However, the study in publication III noted that the top predictions by the methods are determined somewhat arbitrarily.

The PE method uses an assumption that the set of correct DDIs is well approximated by the minimal set of DDIs that is necessary to justify the PPI network. By using Linear Programming optimization, this method assigns, to each domain, pair a $PEscore$, which indicates the probability that the two domains interact with each other. The method can also avoid the problem of false positives and negatives. Publication III proved that the insufficient and biased DDI benchmark sets lead to better PE method performance than the other methods.

3.4.2 Predicting new PPIs based on inferred DDIs

Our study in publication III showed that domain-based methods are limited at predicting PPIs. The reason for the weak performance of domain-based methods is that the methods predict interactions among all of the proteins containing two interacting domains. However, many proteins specifically interact with proteins that have complementary physiochemical properties. Thus, the domain-based methods overpredict the interactions, which causes the PPI networks to include many false positives.

Conservative attempts, such phylogeny-based approaches can overcome the problem of specific PPI predictions by using classes that represent orthologous proteins. Functional linkages among phylogenetics families of proteins are projected down to the protein level. In this downcasting procedure, linkages are not predicted among all of the members of two families; however, among only the members that are from the same species. This helps to improve the specificity of the predictions (see section 3.5).

3.5 Phylogeny-based methods for predicting PPIs

Phylogeny-based methods for PPI prediction are broadly applicable because there is a sufficiently large number of completely sequenced genomes. These methods are premised on the hypothesis that functionally linked or interacting proteins co-evolve, and therefore they have homologs in the same set of organisms (Pellegrini *et al.*, 1999). The term "coevolution" is often considered to be coined by Ehrlich and Raven (1964), but this term was used at least as early as 1957 (Mode, 1958). Thompson's definition of coevolution as "reciprocal evolutionary change in interacting species" (Thompson, 1994) is the most widely accepted definition. This definition implies the evolution of a biological object in response to selection imposed by a related object. The term "co-evolution" refers to the similarity of evolutionary patterns (Pazos and Valencia, 2008).

3.5.1 Phylogenetic profiling

A phylogenetic binary profile, the simplest pattern of evolution, is a binary string in which each bit indicates the presence or absence of a protein family in a different species (Pellegrini *et al.*, 1999). This approach is premised on the hypothesis that a given biological function requires the concerted action of multiple proteins. If one of the proteins is lost for any reason, there will be no selection pressure to retain the other protein that is required for that function. Going one step further, the binary string is replaced by a

weighted string in which each bit indicates the similarity score of a protein in an organism with respect to a reference organism (Date and Marcotte, 2003).

Ranea *et al.* (2007) introduced phylogenetic occurrence profiling to detect functionally related protein families in eukaryotic genomes. The phylogenetic occurrence profile of a protein family is a vector in which each element indicates the number of protein members of this family observed in one organism. Unlike prokaryotic genomes, in which a large proportion of protein families have approximately one copy per species, eukaryotic genomes show a large number of multi-protein families that have more than one member per species (Ranea *et al.*, 2007). Phylogenetic occurrence profiling, therefore, is able to detect more evolutionary signals that could not be detected by phylogenetic binary profiling.

In the phylogenetic binary profiling approach (Pellegrini *et al.*, 1999), a gene was considered to be present in another genome if there was a match above a chosen threshold using a similarity search tool such as BLAST. However, it is impossible to select a versatile cutoff value to define the presence or absence of a gene across species because evolutionary rates vary greatly among proteins. Moreover, the phylogenetic profiles cannot address the issue of orthology and paralogy very well. For a phylogenetic profile to be most useful, it should be able to distinguish orthologs from paralogs, and then genes can be grouped based on the presence and absence of orthologs (Eisen and Wu, 2002). The COG method (Tatusov *et al.*, 1997, 2001), the first massive way to determine orthology, works quite well for most bacterial genes. However, it is not the ideal way to identify orthology because it still relies on pairwise similarity scores.

Pazos and Valencia (2001) present each protein and its homologs in a hierarchical tree that indicates the sequence similarity among those proteins. This method, called *"mirrortre"*, calculates the similarity of two trees by comparing the two distance matrices that are used to build the trees. In the comparison, each matrix is transferred to a vector in which each bit is an entry of the matrix. The improved version of the "mirrortree" method, *"contextmirror"*, attempts to improve the similarity measure by applying a partial correlation (Juan *et al.*, 2008). However, the comparisons in these methods are simply the comparisons among vectors, and therefore, they are extensions of the profiling methods.

Some methods, such as (Vert, 2002; Barker and Pagel, 2005; Barker *et al.*, 2007), combine the absence/presence profiles and a species phylogenetic tree in statistical models to detect evidence of co-evolution among protein families. These studies have shown that seeking correlated gains

and losses of genes on a phylogenetic tree of species substantially improves the detection of functionally linked pairs of proteins, compared to the original across-species method (Pellegrini *et al.*, 1999). However, the patterns in these studies do not capture the speciations and duplications that occur in protein evolutionary history, which potentially contain more information.

3.5.2 Detecting protein functional linkages with enhanced phylogenetic trees

All of the phylogeny-based methods make use of the assumption that co-evolved proteins are likely to interact with each other. Therefore, the prediction performance strongly depends on how the evolutionary patterns are described. Profiling methods work fine at predicting PPIs because, for PPI prediction purpose, the family membership is already helpful. However, defining a functional hierarchy inside of families can enhance the evolutionary information and then can help to better detect the proteins that evolve in a correlated fashion, which are, by implication, functionally linked proteins. The EPT method for predicting protein functional linkages, in the first step, builds the graphical models for protein families that can capture complex evolutionary divergence and convergence events in multi-protein families.

In the second step, the method topologically compares trees to detect co-evolved families. The tree comparison algorithm rewards the common parts and penalizes the different parts between two trees. The method utilizes a layer-by-layer comparison strategy in which the layers are weighted depending on their lineage. Another approach, giving equal weights to layers, has also been tested but showed poorer results, meaning that the evolutionary events closer to LUCA are more important and, thus, have a more critical effect on the outcome. An alternative approach for comparing trees has been tested. This method transforms each EPT into a set of vectors, with each vector corresponding to a subtree and with guidance from the species tree. Such a vector contains ones for all of the common nodes and zeros for all of the different nodes between the subtree and the species tree. Then, two sets of vectors are cross-compared. The similarity of the two binary vectors is the number of mismatches subtracted from the number of matches. However, the method performs worse than the tree comparison algorithm when comparing trees.

The EPT method shows a significant improvement at predicting functional linkages compared to profiling methods. Specifically, the EPT method discovers approximately 20% (27%) of the functional linkages in the human (yeast) datasets, with a low false positive rate (approximately 5%). For the

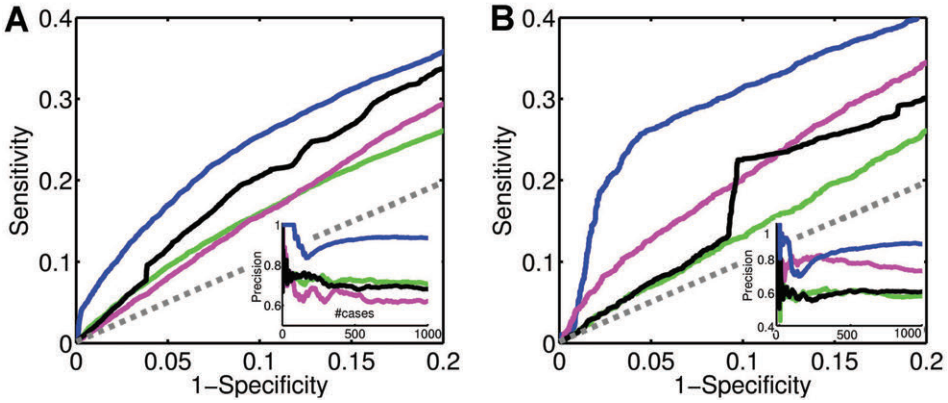

Figure 3.2: (Adapted from publication I). ROC curves of the EPT method (blue), occurrence profiling with Euclidean distance (called occ_ed) (green), occurrence profiling with Pearson correlation, (occ_ps) (magenta) and binary profiling with Pearson correlation (bin_ps) (black) for the human (A) and yeast (B) datasets. The dashed gray diagonal lines correspond to random predictions. The insets of (A) and (B) are the precisions of the methods at different numbers of top predictions (cases with the highest scores) for the human and yeast datasets, respectively. The precision is the ratio of the number of true positives over the number of top predictions, TP/(TP+FP).

top 1000 predictions, the method has a precision of more than 90% both in the human and in the yeast datasets. The method works precisely but still suffers from having a low sensitivity.

Comparatively evaluation of the EPT and profiling methods using GO data showed that the EPT method can predict better functional linkages among proteins that participate in similar biological processes or that locate in similar cellular components but do not perform the same molecular functions. This result is expected because, typically, a set of different biochemical activities (molecular functions) are required to implement biological processes.

Selection of reference organisms

The organisms that are used to construct phylogenetic profiles or trees are called reference organisms. This section discusses how the selection of reference organisms can affect the prediction performance of the EPT method, compared to the profiling methods.

It has been shown that using entirely eukaryotes, the binary profile method is limited (Snitkin *et al.*, 2006) whereas the occurrence profile

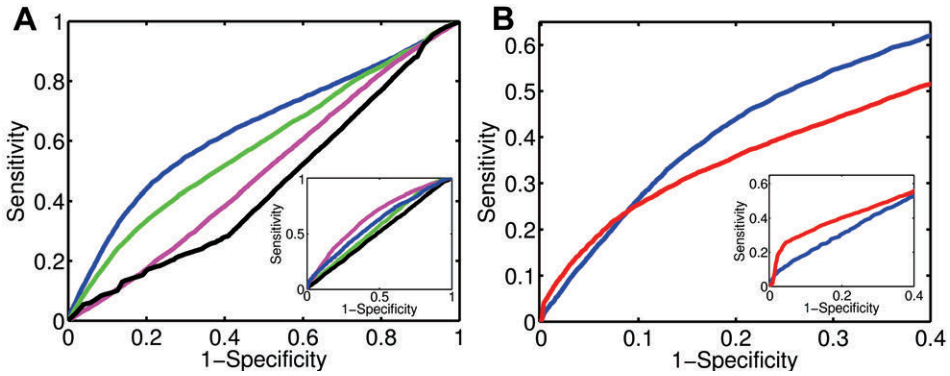

Figure 3.3: (A) ROC curves of the EPT (blue), occ_ed (green), occ_ps (magenta) and bin_ps (black) methods for human and yeast (inset) datasets when the reference organisms are all 12 eukaryotes. (B) Performance comparison between EPT-all (red) and EPT-euka (blue) for human and yeast (inset) datasets.

method performs better at predicting functional linkages (Ranea *et al.*, 2007). Because of the presence of multi-gene families in eukaryotic genomes, binary profiles cannot capture the distribution of gene copies in each genome as well as the occurrence profiles. Our study showed that the EPT method is more advanced compared to other methods when predicting human protein functional linkages, irrespective of the reference organisms (Figure 3.2A and Figure 3.3A).

Figure 3.3B presents a comparison between the method that use EPTs, which consists of 12 eukaryotic genomes (EPT-euka) and the method that use EPTs, which consist of all of the 572 genomes (EPT-all) for both the human and yeast datasets (the inset). EPT-all performs better up to an FPR of approximately 12%, whereas EPT-euka is more sensitive than EPT-all when FPR is high. This outcome indicates that, when we attempt to keep a low FPR, using EPT-all helps us to obtain a better coverage than using EPT-euka. This difference can be explained by the fact that the top predictions by the EPT method are likely the pairs of proteins that belong to large EPTs that have strong evolutionary signals on the prokaryote side (see publication I). Therefore, they lose a substantial amount of evolutionary information when they are built entirely from eukaryotes.

Limitations of the EPT method

Testing the EPT method in *Escherichia coli* datasets showed that the method's performance is poor whereas the binary profiling method performs

well. This result occurs because EPTs containing *Escherichia coli* proteins are prokaryote-biased and, therefore, contain fewer duplication events. In prokaryotic organisms, where a large proportion of the protein families have approximately one copy per species, the profile of protein copies provide a sufficient signal whereas EPTs might add more noise.

The current EPT method produces a simple clustering of proteins that is based on BLAST scores and that does not include information about lateral gene transfer (LGT). This method can be improved in a number of ways, to account for the complexity of evolutionary relations, including divergence, convergence, domain recombination and horizontal gene transfer events. For example, the choice of the descendant protein is somewhat arbitrary because BLAST scores do not reflect the complexity of these relations. Another limitation of the EPT method is that the EPT method, similar to other phylogeny-based methods, cannot infer well the physical interactions of the proteins but it is promising at predicting functionally linked proteins.

3.6 Challenges of computational PPI predictions

Many types of experimentally discovered data such as genomic, structure or gene expression data, serve as the input data for computational methods. The accuracy and sufficiency of the input data impact directly on the prediction performance. In prediction methodologies, benchmark datasets including positive and negative reference sets of PPIs are used to train the frameworks and evaluate the performances. Therefore, the setting of reliable benchmark datasets is essential to the prediction process. Recently, positive datasets that are defined by experiments have contained many false positives, especially in high-throughput experiments. PPI networks discovered by low-throughput assays are biased towards better-studied proteins. The quality of the negative references is even lower because of the lack of confirmed information about the non-interacting protein pairs. These computational approaches would inevitably benefit from the development of experimentally obtained datasets in both quality and coverage.

Chapter 4

Dynamical network modeling of complex systems using coupled phase oscillator models

Biological networks such as protein-protein interactions, metabolic, signaling, transcription-regulatory networks and neural synapses are only a few examples of large-scale dynamic systems in life. The first step toward capturing the global properties of such systems is to model them by graphs in which the nodes represent the dynamical units and the links are the interactions between them. The network structural analysis is to cope with structural issues, such as characterizing the topology of a complex wiring architecture and revealing the unifying principles that are at the basis of real networks. The study of complex network dynamics involves developing models that mimic the real-life network systems, to know how a large ensemble of dynamical systems that interact through a complex wiring topology can behave collectively. This chapter discusses how the network connectivity determines dynamical behaviors in complex network systems. We utilize the Kuramoto model of phase oscillators, which is simple but relevant enough to describe many biological examples, including the mutual synchronization of cardiac pacemaker cells, rhythmically flashing fireflies and chorusing crickets and wave propagation in the heart, brain, intestine and nervous system (Winfree, 2001).

4.1 Synchronization in complex networks

Studies on synchronization aim to clarify the conditions for units in certain systems to be synchronized. Local dynamics of the elements at nodes, coupling and network topology affect the collective behavior of the elements.

Studies that use regular network models, such as chains, grids, lattices and fully connected graphs (Erdös and Rényi, 1960), focus on the complexity that is caused by the nonlinear dynamics of the nodes, without being burdened by any additional complexity in the network structure itself. In this thesis, the author employs Kuramoto's model of phase oscillators on generalized random substrate networks: this approach helps to set dynamics aside and to turn to the effects of the coupling strength and especially the architectures.

4.1.1 Onset of synchronization

The first step is to explore the onset of global synchronization, which is characterized by the critical coupling strength separating the desynchronized and synchronized states. Kuramoto (1984) analytically resolved the critical coupling strength $K_c = 2/\pi g(\omega_0)$ for the all-to-all connected network of oscillators, where $g(\omega)$ is the distribution from which the natural frequencies are drawn, and ω_0 represents the mean frequency of the ensemble. At the stationary state ($t \to \infty$) and for large networks ($N \to \infty$), the global order parameter (Eq. 2.9) behaves as $R \sim (K - K_c)^\beta$, for all $K > K_c$, with $\beta = 1/2$.

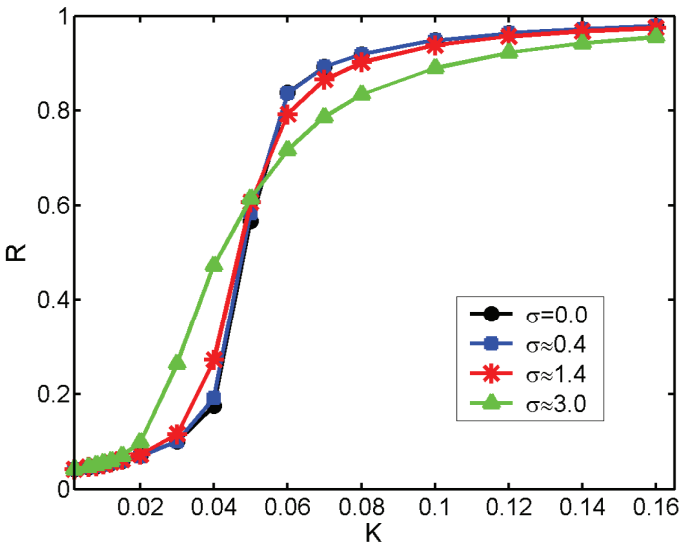

Figure 4.1: (Adapted from publication II). Global order parameter as the coupling is increased in a set of generalized random networks with different topologies, which correspond to $\sigma(\kappa) = 0$ (regular), 0.4, 1.4 and 3.0 (SF).

Investigations on synchronization in complex networks were first re-

ported for networks with SW (Hong *et al.*, 2002) and SF (Moreno and Pacheco, 2004) topologies, where the critical coupling was numerically studied by making use of finite-size scaling analysis. These studies confirmed the existence of a finite K for regular, SW and SF network systems. Moreover, the global order parameter of these complex systems turns out to be qualitatively the same as in the original Kuramoto all-to-all connected network: the square-root behavior that was found in the all-to-all architecture also holds for the SF model (Moreno and Pacheco, 2004).

The impact of the topological features of a network on the onset of synchronization has been intensively studied. Fernández *et al.* (2000) and Barahona and Pecora (2002) have shown that the small-world property of the structure of the network enhance the synchronization of the system. McGraw and Menzinger (2005) suggested that networks with a large clustering coefficient promote synchronization at lower values of the coupling strength. When the coupling is greater than the critical point, the effect of the average path length of the network dominates over the clustering coefficient. In (Oh *et al.*, 2005), the authors found that the synchronization transition crucially depends on the type of inter-modular connections.

In publication II, we continued with the above topic by studying the evolution of the global order parameter as a function of the coupling strength K for several network topologies, which range from regular to scale-free (Figure 4.1). We found that the onset of global synchronization first occurs for the SF network. The more heterogeneous the network is, the smaller the value of K that is needed for the onset of global synchronization. Conversely, the path to the complete synchronization is faster for networks with a homogeneous degree distribution. Whereas the nodes in a homogeneous network system suddenly become globally synchronized as the coupling strength increases, the path to synchronization of a heterogeneous network system is nontrivial.

4.1.2 Path to synchronization

Whereas the global order parameter can help to answer when a system starts to synchronize as the coupling increases, pairwise phase correlation can help to study where synchronization initiates and how it propagates inside the system. Our study showed that, when the coupling is very weak, there is no correlation between any two nodes in the system. The local dynamics at the nodes is so dominant that each node works in its own way. When the coupling is slightly stronger, only pairs of nodes that are physically connected to each other start to be synchronized significantly. The dynamical process in the system when the coupling is weak is driven by physical con-

nectivity, regardless of the network topology. When the coupling increases, a homogeneous network system is partitioned into many small clusters of pairwisely synchronized nodes which then merge together to form a giant cluster. Although the giant cluster contains most of the nodes, the connections among the individual clusters are still weak, which prevents global synchronization. After critical coupling, the pairwise correlation among the nodes of the giant cluster is strengthened dramatically, making the systems quickly reach a completely synchronized state.

In a heterogeneous network system, synchronization is initiated among hub nodes to form a core synchronized cluster. Figure 4.2 shows the organization of synchronization in a SF network system when it is at the critical coupling point. Clearly we observe here that the backbone of the synchronization structure is composed of the hubs excluding nodes with few neighbors. As the coupling increases, the core cluster gradually extends by recruiting more small nodes, which accounts for a slow growth of global synchronization in the SF network systems (Figure 2.3B). In the synchronized state, hub nodes are more robust under perturbations; they revert back to a synchronized cluster in a time interval that is inversely proportional to their degree (Moreno and Pacheco, 2004).

The role of the hubs on the path to synchronization and their robustness under perturbations in complex systems can be explained by the concepts of effective coupling strength and local order parameter (section 2.3.2). Directly after critical coupling, the local order parameter of every node approaches one, which makes the effective coupling strength dependent on the degree of the node. The hub nodes couple with the remainder of the system more strongly, and thus, it is easier for them to be synchronized and more stable against perturbations.

4.2 Reconstruction of physical connectivity from functional dynamics

The function-structure relationship suggests an ability to study one of the most important inverse problems: inferring physical connectivity from functional dynamics. For example, how does an electroencephalography (EEG) pattern (which is the recording of electrical activity at different positions in the brain) reflect the details of the axons among cortical neurons? Or, how can we predict physical PPIs based on gene co-expression data?

Timme (2007) studied a regular system at a synchronized state when applying an external stimulus. The responses of the system to the external inputs reveal the underlying network connectivity. Publication II studied

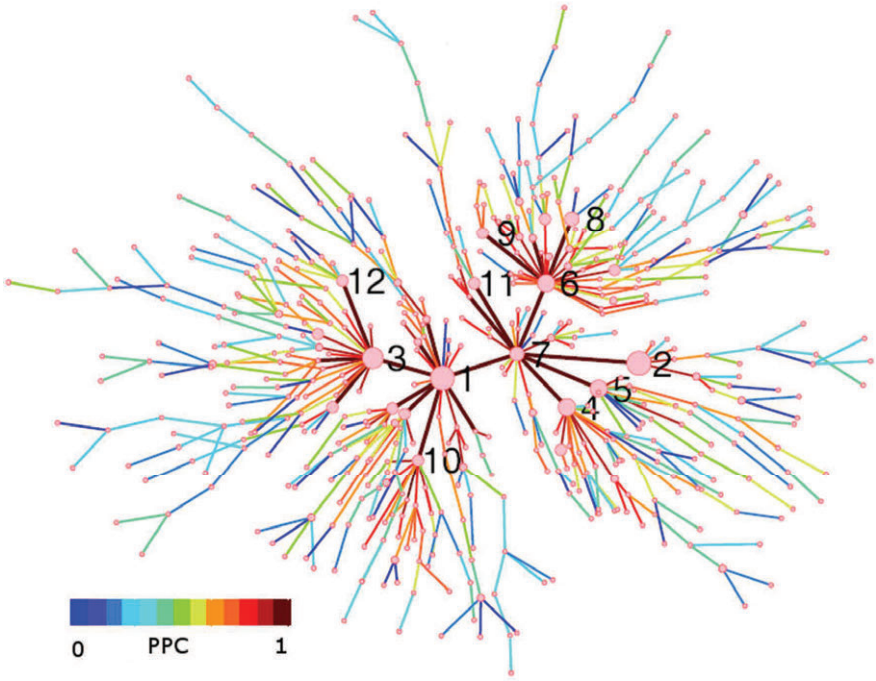

Figure 4.2: (Adapted from publication II). Synchronization in a SF network of 512 oscillators, specifically at the onset of global synchronization. A node is denoted by a circle with its size proportional to the node degree, and the color of a link between two nodes encodes the PPC of these two nodes. The 12 hub nodes with the highest node degrees are labeled

structural reconstruction in systems with network topologies that range from regular to complex at both weak and strong regime of coupling. This approach can help to discover which network topologies facilitate reconstruction and what the optimal coupling rhythms for reconstructing are.

In publication II, we proposed a method for reconstruction of the underlying connectivity of oscillatory network systems such as neural networks. Phases at nodes can be extracted from appropriate rhythms, e.g. the alpha, beta or theta rhythms of EEG data. In the next step, PPC is calculated for every pair of nodes. Averaging data that is obtained from multiple trials helps to reduce the effects of noise, and thus, enhances the reconstruction performance. The method predicts a physical connection between two nodes if the PPC among those two nodes is higher than a predefined threshold.

The reconstruction is successful in regimes of weak coupling, immediately before the onset of global synchronization, irrespective of the network topologies. This scenario implies that the method using PPC works well

for a wide range of systems that have weak coupling, which is more biologically realistic (Zhou *et al.*, 2006). As the coupling is increased, two nodes can be synchronized even if they are not connected, making no distinction between connected nodes and disconnected nodes. Therefore, in this case, the physical connectivity cannot be reconstructed by using the PPCs. The onset of global synchronization, such as an epileptic seizure in the brain, hinders the reconstruction of the physical connectivity of the systems. When the coupling is extreme, the reconstruction of connectivity in homogeneous networks is good, but the reconstruction in heterogeneous systems is poor because of the higher PPCs between hubs, even though they are not physically connected.

Chapter 5

Conclusions

The topics covered in this thesis include computational approaches to systems biology, which is particularly focused on biological network inferences and modeling. The main goals for our work on network structure inference were to study current often-used approaches and to propose efficient methods for predicting PPI networks, which are of central importance for virtually every process in a living cell. Our research on dynamical network modeling aimed to obtain an in-depth understanding of the relationship between network structure and function in complex network systems.

In this thesis, we have discussed a wide range of methods for predicting PPI networks. We especially concentrated on comparative genomic methods such as domain-based and phylogeny-based methods. Our study on domain-based methods noted some limitations of both the methods and the data resources. The domain-based methods failed to predict specific interactions among multi-protein families, which raises the requirements of conservative methods such as phylogeny-based methods.

We proposed the EPT method for assigning functional linkages to proteins. By using different evaluation approaches with high-quality datasets, we showed that our proposed method outperforms conventional phylogeny-based methods. The EPT method promisingly predicts more reliable sets of protein functional linkages in human and yeast.

PPINs and FLMs discovered by both experimental or/and computational approaches have become available for model organisms such as yeast and human, providing excellent opportunities for understanding mechanisms under biomedical phenomena, especially diseases. Human diseases tend to form an interrelated landscape, whereby different diseases are linked together based on perturbing the same biological processes. Figure 5.1 shows an integrated network of disease-disease, disease-gene and gene-gene interactions. Diseases with similar phenotypes that exhibit similar pheno-

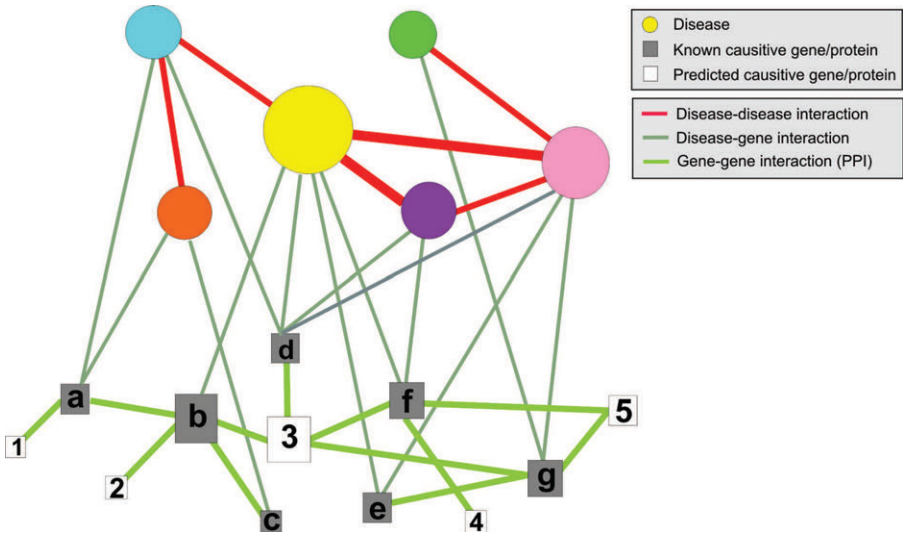

Figure 5.1: (Adapted from (Ta and Holm, 2011)) An integrated disease-disease, disease-gene and gene-gene interaction network. The size of a disease node is proportional to the number of genes that cause the corresponding disease. The size of a gene node is proportional to its degree in the gene-gene network. The thickness of a disease-disease link is proportional to the number of causative genes shared by the two diseases. A gene is predicted to be a causative gene of a disease if it interacts with a known causative gene of the disease. Known disease-gene interactions are derived from OMIM (Boyadjiev and Jabs, 2000).

type descriptions or hospital diagnosis records tend to be caused by dysfunctions of the same genes (Goh *et al.*, 2007). Diseases with dissimilar phenotypes can also be related at the molecular level. Based on these premises, PPINs and FLMs are promising for applications to the prediction of new disease-causing genes (Guan *et al.*, 2010), identification of disease-related subnetworks (Pujana *et al.*, 2007) and exploration of disease-disease associations (Goh *et al.*, 2007).

Our study in dynamical network modeling adopted a network of Kuramoto phase oscillators. This model is useful for displaying a large variety of synchronization patterns while being sufficiently flexible to be adapted to many realistic systems (Acebrón *et al.*, 2005). This approach provides a comprehensive picture of how different network topologies, which range from regular to scale-free, determine the functional activities in the systems under both weak and strong coupling regimes. The results show that, as the coupling increases from weak to strong, global synchronization first occurs in more heterogeneous network systems and gradually progress to complete synchronization, while it occurs later in homogeneous network

systems but approaches complete synchronization very quickly. In a heterogeneous network system, the synchronization occurs among hub nodes and then propagates to the remainder of the system.

The work on the reconstruction of the physical connectivity from the functional activities implies that the reconstruction method using PPC might work well with oscillatory systems that have a weak coupling, regardless of the network topology. However, when there is a strong coupling, the PPC method could not reconstruct well the heterogeneous systems but still worked well for the regular systems. The findings in this study can be applied to investigate, for example, connectivity in the human cerebral cortex, which consists of structurally segregated and functionally specialized regions that interconnect by a dense network of cortico-cortical axonal pathways (Sporns *et al.*, 2005). The PPC method might be a promising tool for reconstructing the structural connectivity among ROIs (Region of Interest) of the brain based on blood oxygenation level-dependent (BOLD) signals (Honey *et al.*, 2007).

References

Acebrón, J. A., Bonilla, L. L., Pérez Vicente, C. J., Ritort, F., and Spigler, R. (2005). The Kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of Modern Physics*, **77**(1), 137–185.

Adamic, L. (1999). The small world web. In *Research and advanced technology for digital libraries, proceeding*, volume 1696 of *Lecture notes in computer science*, pages 443–452. Springer-Verlag Berlin.

Aderem, A. (2005). Systems Biology: Its Practice and Challenges. *Cell*, **121**(4), 511–513.

Aderem, A., Adkins, J. N., Ansong, C., Galagan, J., Kaiser, S., Korth, M. J., Law, G. L., McDermott, J. G., Proll, S. C., Rosenberger, C., Schoolnik, G., and Katze, M. G. (2011). A systems biology approach to infectious disease research: Innovating the pathogen-host research paradigm. *mBio*, **2**(1).

Ahn, A. C., Tewari, M., Poon, C.-S., and Phillips, R. S. (2006). The limits of reductionism in medicine: Could systems biology offer an alternative? *PLoS Medicine*, **3**(6), e208.

Albert, R. (2007). Network inference, analysis, and modeling in systems biology. *The Plant Cell Online*, **19**(11), 3327–3338.

Albert, R. and Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**, 47–97.

Albert, R., Jeong, H., and Barabási, A. L. (1999). Internet: Diameter of the world-wide web. *Nature*, **401**, 130–131.

Alon, U. (2006). *Introduction to systems biology: Design principles of biological circuits*. Chapman and Hall, London, UK.

Amaral, L. A. N., Scala, A., Barthélémy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 11149–11152.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J.,

Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. (2009). The intact molecular interaction database in 2010. *Nucleic Acids Research*.

Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., Ara, T., Nakahigashi, K., Huang, H.-C., Hirai, A., Tsuzuki, K., Nakamura, S., Altaf-Ul-Amin, M., Oshima, T., Baba, T., Yamamoto, N., Kawamura, T., Ioka-Nakamichi, T., Kitagawa, M., Tomita, M., Kanaya, S., Wada, C., and Mori, H. (2006). Large-scale identification of protein/protein interaction of escherichia coli k-12. *Genome Research*, **16**(5), 686–691.

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2003). Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics*, **19**(suppl 1), i7–i15.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**(10), 25–29.

Azuaje, F. (2010). *Bioinformatics and biomarker discovery: "omic" data analysis for personalized medicine*. John Wiley & Sons, Ltd.

Bader, G. D., Betel, D., and Hogue, C. W. V. (2003). Bind: the biomolecular interaction network database. *Nucleic Acids Research*, **31**(1), 248–250.

Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, **22**, 78–85.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509 – 512.

Barabási, A. L., Albert, R., and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A*, **272**, 173–187.

Barahona, M. and Pecora, L. M. (2002). Synchronization in small-world systems. *Physical Review Letters*, **89**, 054101.

Barker, D. and Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Computational Biology*, **1**(1), e3.

Barker, D., Meade, A., and Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics*, **23**(1), 14–20.

Basler, G., Grimbs, S., Ebenhöh, O., Selbig, J., and Nikoloski, Z. (2011). Evolutionary significance of metabolic network properties. *Journal of The Royal Society Interface*.

Ben-Hur, A. and Noble, W. (2006). Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **7**(Suppl 1), S2.

Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21**(suppl 1), i38–i46.

Bertalanffy, L. (1968). *General system theory: foundations, development, applications*. Braziller, New York.

Boyadjiev, S. A. and Jabs, E. W. (2000). Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clinical Genetics*, **57**(4), 253–266.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, **29**(4), 365–371.

Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005). Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, **433**, 531–537.

Camacho, J., Guimerà, R., and Amaral, L. A. N. (2002). Robust patterns in food web structure. *Physical Review Letters*, **88**, 228102.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). Mint, the molecular interaction database: 2009 update. *Nucleic Acids Research*, **38**(suppl 1), D532–D539.

Colland, F., Jacq, X., Trouplin, V., Mougin, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P., and Gauthier, J.-M. (2004). Functional proteomics mapping of a human signaling pathway. *Genome Research*, **14**(7), 1324–1332.

Conrotto, P. and Souchelnytskyi, S. (2008). Proteomic approaches in biological and medical sciences: principles and applications. *Experimental oncology*, **30**(3), 171–80.

Consortium, T. F., Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R.,

Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Babu, M. M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Group, R. G. E. R., Group, G. S. G. G. N. P. C., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., and Hayashizaki, Y. (2005). The Transcriptional Landscape of the Mammalian Genome. *Science*, **309**(5740), 1559–1563.

Cornish-Bowden, A. (2011). Systems biology: How far has it come? *The Biochemist*, **33**(1), 16–18.

Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., Biankin, A. V., Hautaniemi, S., and Wu, J. (2012). Pina v2.0: mining interactome modules. *Nucleic Acids Research*, **40**(D1), D862–D865.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, **23**(9), 324 – 328.

Date, S. V. and Marcotte, E. M. (2003). Discovery of uncharacterized cellular

systems by genome-wide analysis of functional linkages. *Nature Biotechnology*, **21**(9), 1055–62.

Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions. *Molecular Cellular Proteomics*, **1**(5), 349–356.

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, **12**(10), 1540–1548.

Dorogovtsev, S. N. and Mendes, J. F. (2003). *Evolution of Networks: from Biological Nets to the Internet and WWW*. Oxford University Press, Oxford.

Doyon, J.-P., Ranwez, V., Daubin, V., and Berry, V. (2011). Models, algorithms and programs for phylogeny reconciliation. *Briefings in Bioinformatics*, **12**(5), 392–400.

Doyon, J.-P., Hamel, S., and Chauve, C. (2012). An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**, 26–39.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**(1), 207–210.

Ehrlich, P. R. and Raven, P. H. (1964). Butterflies andplants: a study in coevolution. *Evolution*, **18**(4), 586–608.

Eisen, J. A. and Wu, M. (2002). Phylogenetic analysis and gene functional predictions: Phylogenomics in action. *Theoretical Population Biology*, **61**(4), 481 – 487.

Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, **405**, 823–826.

Emmert-Streib, F., Glazko, G., Gokmen, A., and Simoes, R. D. M. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Frontiers in Genetics*, **3**(00008).

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Science*, **5**, 17–61.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee,

M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology*, **3**(7084), 89.

Felsenstein, J. (1989). PHYLIP - phylogeny inference package (Version 3.2). *Cladistics*, **5**, 164–166.

Fernández, L. F. L., Huerta, R., Corbacho, F., and Sigüenza, J. A. (2000). Fast response and temporal coherent oscillations in small-world networks. *Physical Review Letters*, **84**, 2758 – 2761.

Ferrer, L., Dale, J., and Karp, P. (2010). A systematic study of genome context methods: calibration, normalization and combination. *BMC Bioinformatics*, **11**(1), 493.

Fisher, R., Boas, V. E. W., Blume, W., Elger, C., Genton, P., Lee, P., and Engel, J. (2005). Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe). *Epilepsia*, **46(4)**, 470–472.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Biology*, **19**(2), 99–113.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, **40**(1), 35–41.

Friedman, N., Linial, M., and Nachman, I. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

Fromont-Racine, M., Rain, J.-C., and Legrain, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genetics*, **16**, 277–282.

Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**(6868), 141–147.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder,

M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**(7084), 631–636.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A protein interaction map of drosophila melanogaster. *Science*, **302**(5651), 1727–1736.

Glazko, G. and Mushegian, A. (2004). Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biology*, **5**(5), R32.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(21), 8685–8690.

Gorecki, P., Burleigh, G., and Eulenstein, O. (2011). Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinformatics*, **12**(Suppl 1), S15.

Görg, A., Weiss, W., and Dunn, M. J. (2004). Current two-dimensional electrophoresis technology for proteomics. *Proteomics*, **4**(12), 3665–3685.

Guan, Y., Ackert-Bicknell, C. L., Kell, B., Troyanskaya, O. G., and Hibbs, M. A. (2010). Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Computational Biology*, **6**(11), e1000991.

Guimaraes, K., Jothi, R., Zotenko, E., and Przytycka, T. (2006). Predicting domain-domain interactions using a parsimony approach. *Genome Biology*, **7**(11), R104.

Hahn, M. (2007). Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biology*, **8**(7), R141.

Hammond, C., Bergman, H., and Brown, P. (2007). Pathological synchronization in parkinson's disease: networks, models and treatments. *Trends Neurosciences*, **30(7)**, 357–364.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148**(3), 839–843.

Heger, A. and Holm, L. (2003). Exhaustive enumeration of protein domain families. *Journal of Molecular Biology*, **328**(3), 749 – 767.

Hendy, M. and Penny, D. (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, **59**(2), 277 – 290.

Holland, P. W. and Leinhardt, S. (1971). Transitivity in structural models of small groups. *Small Group Research*, **2**(2), 107–124.

Honey, C. J., Kötter, R., Breakspear, M., and Sporns, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(24), 10240–10245.

Hong, H., Choi, M. Y., and Kim, B. J. (2002). Synchronization on small-world networks. *Phys. Rev. E*, **65**(2), 026139.

Hubble, J., Demeter, J., Jin, H., Mao, M., Nitzberg, M., Reddy, T. B. K., Wymore, F., Zachariah, Z. K., Sherlock, G., and Ball, C. A. (2009). Implementation of genepattern within the stanford microarray database. *Nucleic Acids Research*, **37**(suppl 1), D898–D901.

Hue, M., Riffle, M., Vert, J.-P., and Noble, W. (2010). Large-scale prediction of protein-protein interactions from structures. *BMC Bioinformatics*, **11**(1), 144.

Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Research*, **18**(4), 644–652.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(8), 4569–4574.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(1), 13–20.

Joos, T. and Bachmann, J. (2009). Protein microarrays: potentials and limitations. *Frontiers in Bioscience*, **14**, 4376–4385.

Joseph and Felsenstein (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. In R. F. Doolittle, editor, *Computer Methods for Macromolecular Sequence Analysis*, volume 266 of *Methods in Enzymology*, pages 418 – 427. Academic Press.

Juan, D., Pazos, F., and Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(3), 934–939.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., PeregrÃn-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**(7084), 637–643.

Kuramoto, Y. (1984). *Chemical Oscillations, Waves and Turbulence*. Springer, Berlin.

Lappe, M., Park, J., Niggemann, O., and Holm, L. (2001). Generating protein interaction maps from incomplete data: application to fold assignment. *Bioinformatics*, **17**(suppl 1), S149–S156.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., van den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004). A map of the interactome network of the metazoan c. elegans. *Science*, **303**(5657), 540–543.

Lynch, M. and Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. *Trends in Genetics*, **20**(11), 544 – 549.

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(21), 11980–11985.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*.

Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science*, **285**(5428), 751–753.

McGraw, P. N. and Menzinger, M. (2005). Clustering and the synchronization of oscillator networks. *Physical Review E*, **72**, 015101.

Mellor, J. C., Yanai, I., Clodfelter, K. H., Mintseris, J., and DeLisi, C. (2002). Predictome: a database of putative functional links between proteins. *Nucleic Acids Research*, **30**(1), 306–309.

Mesarović, M. (1968). *Systems theory and biology: proceedings*. Springer-Verlag.

Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J., and Ruepp, A. (2004). MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, **32**, D41–44.

Mode, C. J. (1958). A mathematical model for the co-evolution of obligate parasites and their hosts. *Evolution*, **12**(2), 158–165.

Moreno, Y. and Pacheco, A. F. (2004). Synchronization of kuramoto oscillators in scale-free networks. *EPL (Europhysics Letters)*, **68**(4), 603–609.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proc Natl Acad Sci USA*, **98**, 404–409.

Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, **64**, 026118.

Newman, M. E. J., Forrest, S., and Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E*, **66**, 035101.

O'Brien, K. P., Remm, M., and Sonnhammer, E. L. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, **33**, D476–480.

Oh, E., Rho, K., Hong, H., and Kahng, B. (2005). Modular synchronization in complex networks. *Physical Review E*, **72**, 047101.

Oltvai, Z. N. and Barabási, A.-L. (2002). Life's complexity pyramid. *Science*, **298**(5594), 763–764.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., Frings, O., and Sonnhammer, E. L. L. (2010). Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, **38**(suppl 1), D196–D203.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(6), 2896–2901.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, **14**(9), 609–614.

Pazos, F. and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. *The EMBO Journal*, **27**(20), 2648–2655.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(8), 4285–4288.

Penny, D. (2004). Inferring phylogenies.-joseph felsenstein. 2003. sinauer associates, sunderland, massachusetts. *Systematic Biology*, **53**(4), 669–670.

Peterson, M. E., Chen, F., Saven, J. G., Roos, D. S., Babbitt, P. C., and Sali, A. (2009). Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Science*, **18**(6), 1306–1315.

Pikovsky, A. S., Rosenblum, M. G., and Kurth, J. (2001). *Synchronization: A Universal Concept in Nonlinear Science*. Cambrige University Press, Cambrige.

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database - 2009 update. *Nucleic Acids Research*, **37**(suppl 1), D767–D772.

Pujana, M. A., Han, J.-D. J., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, Tomas AU Gold, B., Assmann, V., ElShamy, W. M., Rual, J.-F., Levine, D., Rozek, L. S., Gelman, R. S., Gunsalus, K. C., Greenberg, R. A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Sole, X., Hernandez, P., Lazaro, C., Nathanson, K. L., Weber, B. L., Cusick, M. E., Hill, D. E., Offit, K., Livingston, D. M., Gruber, S. B., Parvin, J. D., and Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics*, pages 1338–1349.

Qi, Y., Bar-Joseph, Z., and Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, **63**(3), 490–500.

Qi, Y., Dhiman, H. K., Bhola, N., Budyak, I., Kar, S., Man, D., Dutta, A., Tirupula, K., Carr, B. I., Grandis, J., Bar-Joseph, Z., and Klein-Seetharaman, J. (2009). Systematic prediction of human membrane receptor interactions. *Proteomics*, **9**(23), 5243–5255.

Ranea, J. A. G., Yeats, C., Grant, A., and Orengo, C. A. (2007). Predicting protein function with hierarchical phylogenetic profiles: The gene3d phylo-tuner method applied to eukaryotic genomes. *PLoS Computational Biology*, **3**(11), e237.

Reboul, J., Vaglio, P., Rual, J.-F., Lamesch, P., Martinez, M., Armstrong, C. M., Li, S., Jacotot, L., Bertin, N., Janky, R., Moore, T., Hudson, J. R., Hartley, J. L., Brasch, M. A., Vandenhaute, J., Boulton, S., Endress, G. A., Jenna, S., Chevet, E., Papasotiropoulos, V., Tolias, P. P., Ptacek, J., Snyder, M., Huang, R., Chance, M. R., Lee, H., Doucette-Stamm, L., Hill, D. E., and Vidal, M. (2003). C. elegans orfeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genetics*, **34**, 35–41.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, **314**(5), 1041 – 1052.

Rice, J. J., Tu, Y., and Stolovitzky, G. (2005). Reconstructing biological networks using conditional correlation analysis. *Bioinformatics*, **21**(6), 765–773.

Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. (2005). Inferring protein domain interactions from databases of interacting proteins. *Genome Biology*, **6**(10), R89.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Research*, **38**, D497–501.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4), 406–425.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Research*, **32**(suppl 1), D449–D451.

Schadt, E. E., Sinsheimer, J. S., and Lange, K. (1998). Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Research*, **8**(3), 222–233.

Sennblad, B. and Lagergren, J. (2009). Probabilistic orthology analysis. *Systematic Biology*, **58**(4), 411–424.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(11), 4337–4341.

Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Computational Biology*, **3**(4), e43.

Sleator, R. (2011). Phylogenetics. *Archives of Microbiology*, **193**, 235–239. 10.1007/s00203-011-0677-x.

Snitkin, E., Gustafson, A., Mellor, J., Wu, J., and DeLisi, C. (2006). Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, **7**(1), 420.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, **28**, 1409–1438.

Sporns, O. (2011). The human connectome: a complex network. *Annals of the New York Academy of Sciences*, **1224**(1), 109–125.

Sporns, O., Tononi, G., and Edelman, G. (2000). Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral Cortex*, **10**(2), 127–141.

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: A structural description of the human brain. *PLoS Computational Biology*, **1**(4), e42.

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, **311**(4), 681 – 692.

Stark, C., Breitkreutz, B.-J., Chatr-aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K., and Tyers, M. (2011). The biogrid interaction database: 2011 update. *Nucleic Acids Research*, **39**(suppl 1), D698–D704.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E.

(2005). A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, **122**, 957–968.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, **410**, 268–276.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and Mering, C. v. (2011). The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, **39**(suppl 1), D561–D568.

Ta, H. X. and Holm, L. (2011). The wiring of protein networks - computational approaches for predicting protein interaction networks. *The Biochemist*, **33**(1), 8–11.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). Mega4: Molecular evolutionary genetics analysis (mega) software version 4.0. *Molecular Biology and Evolution*, **24**(8), 1596–1599.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A Genomic Perspective on Protein Families. *Science*, **278**(5338), 631–637.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. **29**, 22–8+.

Thompson, J. (1994). *The coevolutionary process*. University of Chicago Press.

Timme, M. (2007). Revealing network connectivity from response dynamics. *Phys. Rev. Lett.*, **98**(22), 224101.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Vázquez, A., Pastor-Satorras, R., and Vespignani, A. (2002). Large-scale topological and dynamical properties of the internet. *Physical Review E*, **65**, 066130.

Vernot, B., Stolzer, M., Goldman, A., and Durand, D. (2008). Reconciliation with non-binary species trees. *Journal of Computational Biology*, **15**(8), 981–1006.

Vert, J.-P. (2002). A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, **18**, S276–284.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Molecular Cellular Proteomics*, **417**, 399–403.

Wagner, A. and Fell, D. A. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **268**(1478), 1803–1810.

Walhout, A. J. M., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, **287**(5450), 116–122.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

Whelan, S., Lio, P., and Goldman, N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics*, **17**(5), 262 – 272.

Wikipedia (2004). Plagiarism — Wikipedia, the free encyclopedia. [Online; accessed 22-July-2004].

Winfree, A. (1967). Biological rhythms and the behavior of populations of coupled oscillators. *Journal of Theoretical Biology*, **16**(1), 15–42.

Winfree, A. (2001). *The geometry of biological time*, volume 12 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, second edition.

Wolkenhauer, O. (2001). Systems biology: The reincarnation of systems theory applied in biology? *Briefings in Bioinformatics*, **2**(3), 258–270.

Wu, J., Kasif, S., and DeLisi, C. (2003). Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, **19**(12), 1524–1530.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T. P., and Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nature Methods*, **6**, 75–77.

Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J., Thompson, M. J., Marcotte, E. M., and Eisenberg, D. (2001). Dip: The database of interacting proteins: 2001 update. *Nucleic Acids Research*, **29**(1), 239–241.

Yamanishi, Y., Vert, J.-P., and Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, **20**(suppl 1), i363–i370.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Molecular Biology and Evolution*, **14**(7), 717–724.

Yu, H., Braun, P., Yildim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**(5898), 104–110.

Zaki, N., Lazarova-Molnar, S., El-Hajj, W., and Campbell, P. (2009). Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics*, **10**(1), 150.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002). Mint: a molecular interaction database. *FEBS Letters*, **513**(1), 135 – 140. Protein Domains.

Zhou, C., Zemanová, L., Zamora, G., Hilgetag, C. C., and Kurths, J. (2006). Hierarchical organization unveiled by functional connectivity in complex brain networks. *Physical Review Letter*, **97**, 238103.

Zhou, T., Medo, M., Cimini, G., Zhang, Z.-K., and Zhang, Y.-C. (2011). Emergence of scale-free leadership structure in social recommender systems. *PLoS ONE*, **6**(7), e20648.

Zmasek, C. and Eddy, S. (2002). Rio: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**(1), 14.