

Imperial College London
Department of Department of Computing

Uncovering Disease Associations via Integration of Biological Networks

Kai Sun

February 2014

Supervised by Dr. Nataša Pržulj

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Department of Computing of Imperial College
London
and the Diploma of Imperial College London

Declaration of Originality

I declare that the work presented in this thesis is my own, except where acknowledged. For all work based on collaborative research, the author's individual contributions are indicated in the end of corresponding chapter.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

Current understanding of how diseases are associated with each other is mainly based on the similarity of clinical phenotypes. However, without considering the underlying biological mechanisms of diseases, such knowledge is limited and can even be misleading. With a growing body of transcriptomic, proteomic, metabolomic and genomic data describing diseases, we proposed to gain insights into diseases and their relationships in the light of large-scale biological data. We modelled these data as networks of inter-connected elements, and developed computational methods for their analysis.

We exploited systematic measures based on graphlets to uncover biological knowledge from network topology. Since recently some doubt had arisen concerning the applicability of graphlet-based measures to low edge density networks, we first evaluated the use of graphlet-based measures and demonstrated their suitability for biological network comparison. We also validated the use of graphlet-based measures for finding well-fitting random models for protein-protein interaction (PPI) networks, and demonstrated that five viral PPI networks are well fit by several theoretical models not previously tested.

To gain novel insights into diseases and their relationships, we integrated different types of biological data and developed computational approaches to compare diseases based on their underlying mechanisms. We applied several similarity measures including standard methods and two novel network-based measures to estimate disease association scores. We showed that disease associations predicted by our measures are correlated with associations derived from standard disease classification systems, comorbidity data, genome-wide association studies and literature co-occurrence data significantly higher than expected at random, demonstrating the ability of our measures to recover known disease associations. Furthermore, we presented case studies to validate the use of our measures in identifying previously

undiscovered disease associations. We believe novel associations uncovered in our studies can enhance our knowledge of disease relationships, and may further lead to improvements in disease diagnosis and treatment.

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor Dr. Nataša Pržulj, without whom this dissertation would not have been possible. Dr. Pržulj has led me into the exciting world of bioinformatics, and has provided me with valuable advices and inspiration on my research projects. She has taught me to think critically, write academically, and be an independent researcher. Her guidance and support have been invaluable over the three years of my Ph.D study.

I also thank current and previous members of Dr. Pržulj's research group: Dr. Noël Malod-Dognin, Dr. Joana Gonçalves, Dr. Ömer Yaveroğlu, Vuk Janjić, Anida Sarajlić, Vladimir Gligorijević, Miles Mulholland and Yulian Ng, not only for being friends, but also for their useful comments and discussions on my research projects. I also thank Prof. Wayne Hayes for his help and collaboration in a research project.

A special thanks goes to my industrial supervisor Dr. Chris Larminie for all his guidance and support during the three years. I also thank members of GlaxoSmithKline (GSK) Computational Biology group, especially Dr. Natalie Buchan, Dr. Peter Woollard and Dr. Hannah Tipney, for their valuable discussions and feedbacks on my research projects.

Furthermore, I thank Prof. Marek Sergot, Prof. Duncan Gillies and Dr. Amani EL-Kholy for their encouragement and concerns during my Ph.D. I also thank Brett Harmony for his helpful comments and feedbacks on academic writing.

Finally, I would like to thank GSK Research & Development Ltd and Department of Computing at Imperial College London for their financial support. My research has also partially been supported by Dr. Pržulj's European Research Council Starting Independent Researcher Grant 278212 and National Science Foundation Cyber-Enabled Discovery and Innovation OIA-1028394.

This dissertation is dedicated to my parents, Zhiliang Sun
and Chen Chen, for their constant love and support
throughout my life.

Contents

1. Introduction	15
2. Background	18
2.1. Biological data	19
2.1.1. Protein-protein interactions	19
2.1.2. Biological pathways	21
2.1.3. Disease-gene associations	23
2.1.4. Drug-target associations	24
2.2. Graph theory for network analysis	25
2.2.1. Graph theoretic terminology	25
2.2.2. Network comparison	26
2.2.3. Network models	30
2.2.4. Graphlet-based measures for network analysis	33
2.3. Disease-disease associations	38
2.3.1. Disease classification	38
2.3.2. Computational approaches for uncovering disease-disease associations	39
3. Evaluation of graphlet-based measures	43
3.1. Motivation	44
3.2. Methods	45
3.2.1. Data: PPI networks	45
3.2.2. Network models and their utility	47
3.3. Results and discussion	48
3.3.1. Sensitive vs. unstable	48
3.3.2. The size of the ‘unstable’ region	52
3.3.3. Effect of more up-to-date PPI data	54
3.3.4. Local density vs. average density	55
3.3.5. Interpreting the results of non-parametric tests	56

3.3.6. Modelling viral PPI networks	58
3.4. Conclusions	60
3.5. Author's contributions	61
4. Novel insight into disease-disease associations	62
4.1. Motivation	63
4.2. Methods	66
4.2.1. Biological data	66
4.2.2. Disease similarity measures	68
4.2.3. Evaluation	71
4.3. Results and discussion	73
4.3.1. Comparison of disease-gene association datasets	73
4.3.2. Evaluation of similarity measures	75
4.3.3. Comparison of similarity measures	81
4.3.4. Case studies	82
4.4. Conclusions	83
4.5. Author's contributions	85
5. Biological network integration	86
5.1. Motivation	87
5.2. Methods	89
5.2.1. Data collection	89
5.2.2. Enrichment analysis	91
5.2.3. Disease similarity measure	91
5.3. Results and discussion	93
5.3.1. Integration of disease-related data	93
5.3.2. Evaluation of similarity measure	95
5.3.3. Crohn's disease: a case study	99
5.4. Conclusions	104
5.5. Author's contributions	104
6. Conclusions	106
6.1. Summary of the dissertation	106
6.2. Future work	108
6.2.1. Modelling biological networks	108
6.2.2. Integration of molecular interaction networks	109

6.2.3. Constructing and analysing weighted biological networks	110
6.2.4. Combining similarity measures	110
6.2.5. Disease reclassification	110
Bibliography	111
A. Appendix	136
A.1. Supplementary information for Chapter 2	136
A.2. Supplementary information for Chapter 3	138
A.3. Supplementary information for Chapter 4	141
A.4. Supplementary information for Chapter 5	144

List of Tables

2.1. Comparison between different network models based on global network properties.	33
3.1. The eighteen PPI networks we analysed	46
3.2. Graphlet frequencies in model and PPI networks	57
3.3. Overlap amounts for histograms	60
4.1. Comparison of studies on inferring disease association	65
4.2. Evaluation of similarity measures against ICD-9 classification	76
4.3. Evaluation of similarity measures against comorbidity	79
4.4. Evaluation of similarity measures against GWAS	80
4.5. List of the top 30 diseases associated with DM	84
5.1. The statistics of the IDN	94
5.2. Details of drugs developed for the treatment of Crohn's disease.	102
5.3. Details of potential drug repositioning opportunities for Crohn's disease.	104
A.1. Three-digit level ICD-9 classification.	136
A.2. Top level MeSH tree categories.	137
A.3. Details of PPI data obtained from different versions of the BioGRID database	144
A.4. Comparison of AUC values obtained by the topology-based similarity measure using different PPI data	144
A.5. List of the top 30 diseases associated with PD	145
A.6. AUC values obtained by evaluating the integration-based similarity measure against disease co-occurrence in the literature.	146
A.7. Similarity scores between Crohn's disease and its top 20 associated diseases.	147

List of Figures

2.1. A schematic representation of a PPI network and an example of an <i>Arabidopsis</i> PPI network	20
2.2. A schematic representation of a metabolic network	22
2.3. Three types of graph representation.	27
2.4. All 13 types of 3-node connected subgraphs	29
2.5. Examples of model networks.	31
2.6. The 30 graphlets with 2-5 nodes and their 73 automorphism orbits.	34
2.7. Construction of the human disease network and the disease gene network.	41
3.1. GDD agreement scores vs. edge density when comparing model networks to each other	48
3.2. GDD agreement scores vs. edge density when comparing 30 random model networks to each other with 50 nodes and 100 nodes	49
3.3. Depiction of the sensitivity of GDD agreement as a function of edge density when comparing 30 GEO model networks to each other	50
3.4. Robustness of GDD agreement with respect to noise	51
3.5. Empirical distributions of GDD agreement for ER-vs-ER comparisons	53
3.6. Empirical distributions of GDD agreement for GEO-vs-GEO comparisons	53
3.7. A timeline of PPI data available since 2000	55
3.8. Comparing six random network models to viral PPI networks using the non-parametric test	59
4.1. The overlap of four disease-gene association datasets	74

4.2. ROC curves obtained by evaluating the topology-based measure against comorbidity	78
5.1. Flow chart of data integration.	94
5.2. The disease-disease network constructed based on the integration-based similarity scores.	96
5.3. Evaluation against disease co-occurrence in the literature. . .	98
5.4. Top 20 diseases associated with Crohn's disease.	99
5.5. The most common genes between Crohn's disease and its top 7 associated diseases.	101
A.1. GDD agreement scores vs. edge density when comparing 10, 20, 30, 40, 50 and 100 random model networks to each other with 2000 nodes	138
A.2. Visualisation of the human PPI network HS and one of its highly connected subgraphs.	138
A.3. Depiction of the sensitivity of GDD agreement and RGF distance as a function of edge density	139
A.4. Comparing six random network models with PPI networks of different species	140
A.5. ROC curves obtained by evaluating the annotation-based measure against comorbidity	141
A.6. ROC curves obtained by evaluating the function-based measure against comorbidity	142
A.7. Degree distributions of diseases of the four disease-gene association datasets	143
A.8. The distribution of ϕ -correlation scores	143
A.9. The overlap of predictions	144
A.10. The overlap of three disease-gene association datasets	146
A.11. The most common biological pathways shared between Crohn's disease and its top 7 associated diseases.	148
A.12. Associations between the most common genes and the most common pathways	149

List of Abbreviations

Network modelling

ER	Erdős-Rény random graph
ER-DD	ER random graph with the same Degree Distribution as a data network
GEO	GEOmetric random graph
GEO-GD	GEOmetric Gene Duplication and mutation model
SF	Scale-Free network
STICKY	STICKiness index-based network model

Network analysis

GDD	Graphlet Degree Distribution
RGF	Relative Graphlet Frequency
GRAAL	GRAph ALigner
HCS	Highly Connected Subgraph
ROC	Receiver Operating Characteristic
AUC	Area Under Curve

Biological data and repositories

PPI	Protein-Protein Interaction
GWAS	Genome-Wide Association Study
GO	Gene Ontology
OMIM	Online Mendelian Inheritance in Man
CTD	Comparative Toxicogenomics Database
FunDO	Functional Disease Ontology annotations
HuGENet	Human Genome Epidemiology Network

HuDiNe	Human Disease Network
BioGRID	Biological General Repository for Interaction Datasets
HPRD	Human Protein Reference Database
KEGG	Kyoto Encyclopedia of Genes and Genomes

Disease classification

ICD	International Classification of Diseases
MeSH	Medical Subject Heading
DO	Disease Ontology
UMLS	Unified Medical Language System
SNOMED	Systematized Nomenclature of Medicine

1. Introduction

Correct classification of diseases is critical for effective diagnosis, treatment and prevention of diseases. Therefore, disease classification has become key to modern biology and medicine. The current standard disease classification system is the International Classification of Diseases (ICD), which has been used worldwide for general epidemiological and health management purposes [1]. Other commonly used classifications include Medical Subject Headings (MeSH) [2], Systematized Nomenclature of Medicine (SNOMED) [3] and Unified Medical Language System (UMLS) [4]. However, current knowledge of disease classification is still mainly based on the similarity of clinical phenotypes, focusing on symptoms of affected organs. Without considering the molecular mechanisms driving diseases, such knowledge is limited and can even be misleading. For example, a common phenotype can be caused by different underlying mechanisms, while a common mechanism may lead to various phenotypes. Further understanding of how diseases are associated with each other based on molecular-level biological data is expected to lead to improvements in disease diagnosis, prognosis and treatment.

During the past decade, advances of ‘omics’ technologies has lead to a wealth of molecular-level biological data. These data, including genetic data (e.g., genome-wide association study [GWAS] data), genomic data (e.g., microarray data), proteomic data (e.g., protein-protein interaction [PPI] data), metabolomic data (e.g., metabolic pathways) and transcriptomic data (e.g., RNA-Sequencing data), have provided great opportunities for us to improve current understanding of cellular systems as well as the underlying mechanisms driving diseases. Many of these biological data have been modelled as *networks*, in which *nodes* represent elements of biological systems such as proteins and genes, and *edges* between nodes represent the associations between these biological elements. For example, a PPI network models the physical interactions between proteins in a cell, where a node represents a

protein and an edge exists between two nodes if the corresponding proteins can physically bind to each other. As most biological networks are large and complex, it is necessary to develop efficient and biologically meaningful algorithms for their analysis.

Systematic approaches based on *graphlets* (small, connected, induced subgraphs of large networks) have been introduced to measure the structure (also called *topology*) of biological networks [5, 6]. Graphlet-based measures are proving useful in many network research tasks, including network modelling [7, 8], network comparison [5, 6], and network alignment [9, 10, 11, 12]. Recently some doubt has arisen concerning the applicability of graphlet-based measures for biological network comparison: it has been claimed that the measures are ‘unstable’ in regions of low edge density [13]. In addition, it has also been claimed that there is no existing network model that matches the structure found in PPI networks [13].

In the rest of the thesis, we first give a review of related works in Chapter 2. Topics such as biological data, network modelling and comparison, related studies on disease associations, are explained in that chapter. We then examine the use of graphlet-based measures for biological network comparison in Chapter 3. We identify the edge density regions in which the topology of model networks is ‘unstable’, and show how graphlet-based measures correctly detect this topological instability. Moreover, by investigating recent PPI networks of different species, we show that these ‘unstable’ regions do not affect the analysis of current PPI networks, as current PPI networks are dense enough to avoid these regions. Furthermore, we validate the use of graphlet-based measures for finding well-fitting random models for PPI networks, and show for the first time that five viral species, possessing the latest and most complete PPI networks, are well-fit by several network models.

To gain novel insights into diseases and their relationships, we take advantage of diverse biological data and develop computational methods to compare diseases by considering their underlying mechanisms. In Chapter 4, we first analyse four publicly available disease-gene association datasets, and apply three disease similarity measures, namely annotation-based measure, function-based measure and topology-based measure, to estimate the association scores between diseases. By systematically evaluating our similarity measures against the ICD classification, a statistical measure of comorbidity and disease associations derived from GWAS data, we show that

disease associations predicted by our measures are correlated with known disease associations significantly higher than expected at random.

In Chapter 5, we further our results by integrating different types of large-scale biological data, including GWAS data, disease-chemical associations, biological pathways and Gene Ontology (GO) annotations, with the aim to gain more complete understanding of the relationships between diseases. In particular, we integrate these data into a heterogeneous network, where nodes are bio-entities (e.g., genes, diseases, chemicals, pathways and GO terms) and edges between nodes represent their associations. We extend our disease similarity measures to an integration-based measure to predict novel disease associations by mining the heterogeneous network. Our results are systemically evaluated against the MeSH classification and a statistical measure of disease co-occurrence in PubMed. In addition, we show that our similarity measure achieves better performance by using the integrated data than solely using a specific type of data. To illustrate the use of our approach to uncover novel diseases associations, we present a case study of Crohn’s disease, in which we identify its top associated diseases and validate our predicted associations via mining the literature. We also give examples to demonstrate how our approach can be used for drug repositioning.

Finally, in Chapter 6 we summarise the conclusions of this thesis. Discussions of future research plans are also covered in that chapter. Other related work of the thesis, such as supplemental materials of Chapter 3, Chapter 4 and Chapter 5, can be found in the appendix.

2. Background

Networks have been used to represent many large-scale biological data. In graph theory, a *network* (or a *graph*) consists of a set of objects, called *nodes*, along with a set of links connecting them, called *edges*. There exist many different types of biological networks, such as PPI networks which model the physical interactions among proteins and gene regulation networks which model the regulation of gene expression in a cell. As most of these biological networks are large and complex, it is necessary to develop efficient and biologically meaningful algorithms for their analysis.

In this chapter, we first give an overview of different types of biological data, including PPIs, biological pathways, disease-gene associations and drug-target associations, along with explanations of how these data can be modelled as biological networks. We also introduce graph theoretic terminologies and concepts, and review the major challenges of biological network research. In particular, we discuss the global network properties and local network properties, and show the use of these network properties for network modelling and network comparison. Furthermore, we discuss current disease classification systems and review the literature of related studies on predicting disease-disease associations.

2.1. Biological data

2.1.1. Protein-protein interactions

Proteins are large biological molecules that play a central role in most molecular processes in a cell. They catalyse metabolic reactions, replicate DNA, transport molecules, defend antigens and transmit messages from cell to cell. An interaction between two proteins occurs when they physically bind to each other to perform a biological function. Studying PPIs is expected to provide valuable insights into human diseases, which may lead to the development of new therapeutic approaches [14].

Two methods that have commonly been used to detect PPIs are yeast two-hybrid (Y2H) screening [15, 16, 17, 18, 19, 20, 21] and Mass Spectrometry (MS) of purified complexes [22, 23, 24, 25]. Compared to traditional small-scale biochemical techniques, they are more standardised, and provide a more comprehensive view of the entire interactome. Using these methods, recent studies have published partial PPI networks for many organisms, including viruses (e.g., herpes simplex virus [26]), prokaryotes (e.g., *Escherichia coli* [27]), and eukaryotes (e.g., yeast [16, 25, 24], nematode worm [28, 20], fruit fly [29], and human [18, 19]). These PPI data are now publicly available in several databases including the Biological General Repository for Interaction Datasets (BioGRID) [30], the Human Protein Reference Database (HPRD) [31], the MIPS Mammalian Protein-Protein Interaction Database (MIPS) [32], the Molecular Interaction database (MINT) [33], the Interologous Interaction Database (I2D) [34, 35], the Database of Interacting Proteins (DIP)[36] and IntAct [37].

PPIs can be modelled as networks. In a PPI network, a node represents a protein and an undirected edge exists between a pair of nodes if the corresponding proteins can physically bind to each other (Figure 2.1). During the past decade, the topology of PPI networks has become a popular subject of research. It has been shown that the topology of PPI networks could be used to uncover novel insights into evolution [38, 39, 11], diseases [14, 40, 41, 42], and gene functions [43, 44, 45, 46, 39].

Due to limitations in experimental techniques, current PPI networks are still noisy and largely incomplete. Studies have suggested that current PPI networks have high false-positive rates and even higher false-negative rates [49, 50, 51]. For example, it is claimed that the estimated false positive rates

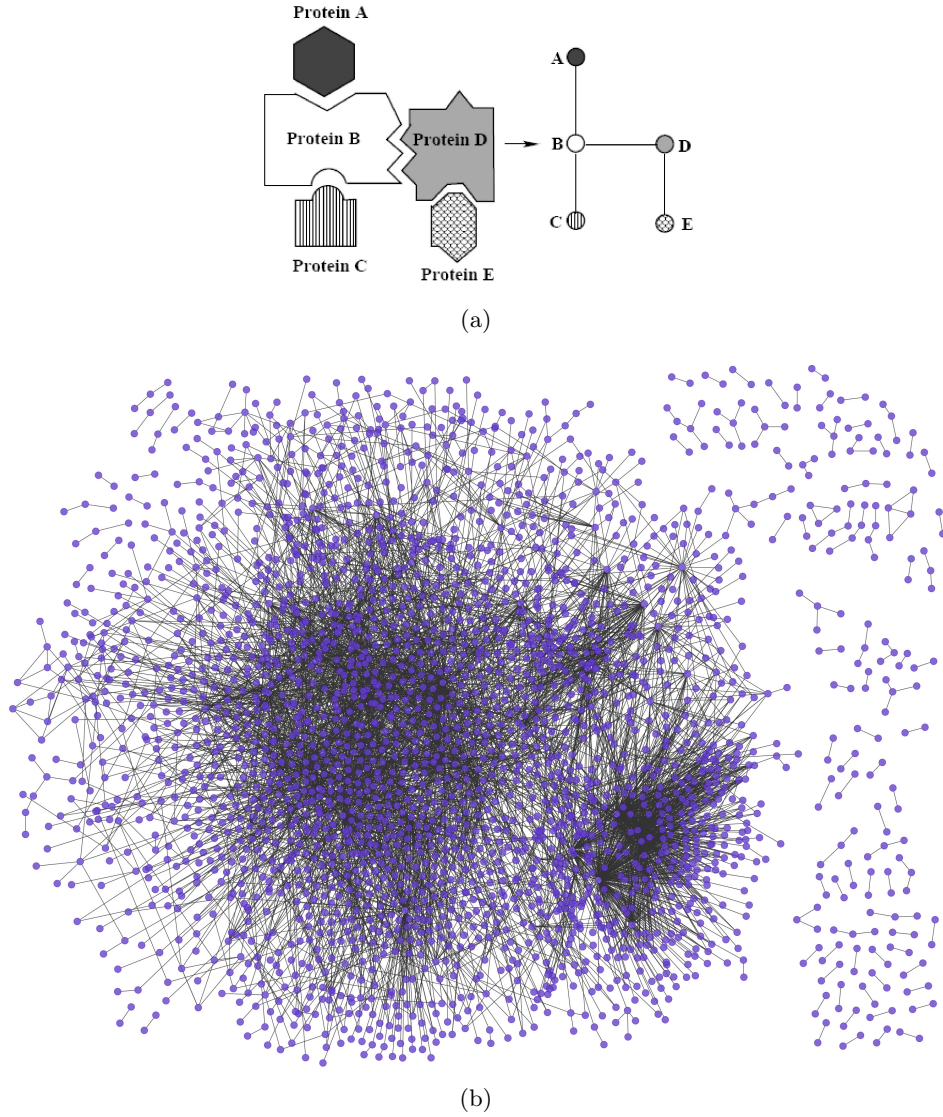


Figure 2.1.: (a) A schematic representation of a PPI network. This figure is taken from [47]. (b) A PPI network of *Arabidopsis thaliana* with 2,634 nodes and 5,529 edges. PPIs are obtained from Arabidopsis Interactome Mapping Consortium (2011) [21]. Cytoscape 2.8.1 [48] was used for the visualisation.

of PPI networks constructed from Y2H screening data are in the range of 25% to 45%, and the estimated false negative rates are in the range of 75% to 90% [51]. Current PPI networks also contain the sampling and data collection biases [52, 53, 47]. For example, some proteins that are relevant to

human diseases are better studied than the others [47]. For these reasons, there are still many difficulties and challenges in uncovering biologically meaningful knowledge from PPI networks.

2.1.2. Biological pathways

A biological pathway represents an ordered series of actions among molecules (mostly proteins and small chemical compounds) in a cell. It models how molecules interact with each other to carry out biological functions. Current knowledge of biological pathways is mainly extracted from the scientific literature by experts, and stored in repositories such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [54], Reactome [55], and WikiPathways [56]. Knowledge of biological pathways allows researchers to identify pathways involved in a disease and identify which step of a pathway is affected in a patient who has the disease, thus pathway data have been widely used to provide insights into the underlying mechanisms of human diseases (e.g., [57, 58]).

There exists different types of biological pathways. Three of the most common pathways are those involved in metabolism (represented by metabolic networks), those involved in gene expression and regulation (represented by gene regulation networks) and those involved in signal transduction (represented by cell signalling networks).

Metabolic network

Metabolic pathways model the metabolism, which is the set of biochemical reactions that allow living organisms to grow and reproduce, maintain their structures, and respond to their environments. In a metabolic pathway, a primary metabolite (a small molecule such as Amino acid) is transformed into another metabolite(s) through a series of chemical reactions catalysed by *enzymes*.

Metabolic pathways can be represented by metabolic networks. There exists several different representations for metabolic networks, including substrate network, reaction network, enzyme-centric network and substrate-enzyme network (see [59] for a review of metabolic networks). For example, in a metabolic network represented as a substrate network, a node represents a metabolite and an edge between a pair of nodes represents the biochemical

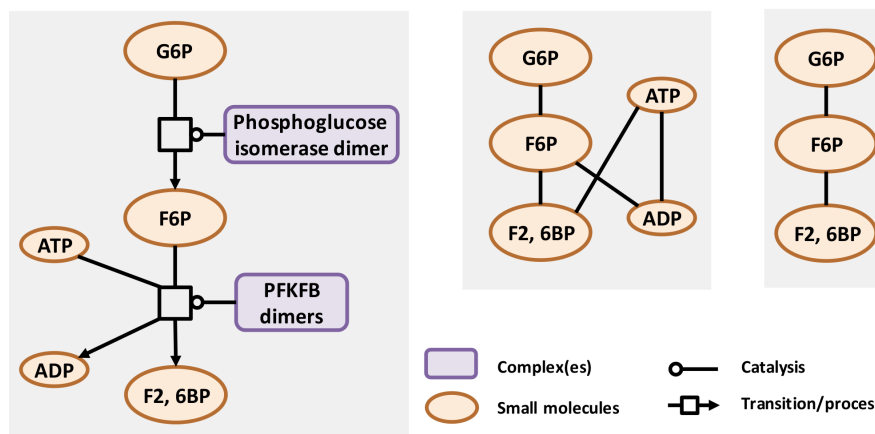


Figure 2.2.: An example of modelling a metabolic pathway (left) as an undirected network (center and right). The pathway illustrates the first two steps of glycolysis. The main metabolites are alpha-D-glucose 6-phosphate (G6P), D-fructose 6-phosphate (F6P) and D-fructose 2,6-bisphosphate (F2, 6BP). ATP and ADP are currency metabolites thus can be ignored when constructing the substrate representation of the network (right). The two reactions are catalysed by phosphoglucose isomerase and PFKFB dimers. The pathway data is obtained from Reactome [55].

reaction that converts one metabolite to the other. Figure 2.2 illustrates an example of how to model a simple metabolic pathway as a substrate network. The pathway can be first modelled as an undirected network if all interacting metabolites are considered equally. Since currency metabolites such as water, oxygen, adenosine triphosphate (ATP) and adenosine diphosphate (ADP) normally perform as carriers, for many applications it is useful to ignore them in order to produce biological meaningful results from network analysis [60]. A simplified substrate representation of the metabolic network can be further constructed by removing currency metabolites. Only connections between the main metabolites are presented in the simplified network.

Gene regulation networks

Gene regulation pathways model the regulation of gene expression in a cell. Regulation of gene expression includes a number of cellular processes that control the increase or decrease of the production of gene products (proteins

or RNAs). Several steps in the gene expression process can be regulated, including transcription, RNA sliding, translation and post-translational modification. For example, in a transcription step, genes can interact with each other indirectly to regulate the expression levels of mRNAs. This process can be modelled as a transcriptional regulation network, in which nodes are genes, and directed edges are interactions through which the products of one gene affect those of another. For example, if a transcription factor, which is the protein product of gene X , binds regulatory DNA regions of a gene Y to regulate the production rate of gene Y 's protein product, this process can be modelled as a simple network which contains a directed edge from node X to node Y (this example was first illustrated by Milo *et al.* (2002) [61]). Due to limitations in experimental techniques, current gene regulation networks are still largely incomplete. For example, it has been estimated that only $< 3\%$ of the expected gene regulatory interactions of human have been discovered [62].

Cell signalling network

Cell signalling pathways are ordered sequences of signal transduction reactions in a cell. They model the complex communication system that governs basic cellular activities. Signal transduction occurs when an extracellular signal (e.g., a chemical) activates a receptor on the surface of a specific cell. The receptor interacts with the external signal and alters intracellular molecules (e.g., proteins) to create a response. Errors in cell signalling pathways may lead to diseases such as cancer and diabetes. Cell signalling pathways can be modelled as a network in which nodes are genes and edges represent the order of signal transduction reactions in the cell.

2.1.3. Disease-gene associations

One of the most commonly used forms of biological data for studying human diseases is disease-gene association. A gene is said to be associated with a disease, if mutations in that gene may lead to that disease. For example, genes BRCA1 and BRCA2 are associated with inheritable breast cancer, since it has been shown that up to 90% of inheritable breast cancers are due to mutations in these two genes [63]. Disease-gene associations can be obtained from large-scale knowledge-bases such as the Online Mendelian

Inheritance in Man (OMIM) [64]. OMIM contains information about genetic disorders and human genes, along with phenotype-genotype relationships between them. It has been used in various disease-related studies (e.g., [65, 66, 67]). OMIM mainly focuses on mendelian disorders (i.e., disorders caused by a mutation in a single gene), and until recently it started to collect information about complex diseases. There also exist other disease-gene association repositories that focus on different aspects of phenotype-genotype relationships, such as the Comparative Toxicogenomics Database (CTD) [68], which focuses on environmental chemicals' effects on human diseases, and Pharmacogenomics Knowledge Base (PharmGKB) [69], which focuses on the impact of genetic variation on drug response.

Recent advances in GWAS studies have enabled the exploration of disease-gene associations in a systematic way on a genome scale [70, 71, 72]. GWAS examines the genome for single-nucleotide polymorphisms (SNPs) that occur more frequently in people with a particular disease than in people without it. GWAS is considered to be one of the most robust routes for identifying genetic variations associated with diseases. Recently, several publicly available repositories such as the National Human Genome Research Institute (NHGRI) GWAS catalog [73] and the GWASdb [74] have been developed to collect and integrate GWAS data obtained from various studies conducted by different research groups.

Disease-gene associations can be modelled as a bipartite network (see Section 2.2.1 for the definition of 'bipartite network'). In a disease-gene association network, a node represents a known human disease or a disease-related gene, and an edge exists between a disease node and a gene node if that gene is known to be associated with that disease. Disease-gene associations play an important role in improving our current understanding of molecular causes and underlying mechanisms of human diseases. Details of how disease-gene associations have been used to uncover novel biological knowledge will be discussed in Section 2.3.2.

2.1.4. Drug-target associations

Drugs are chemical substances that bind to and modify the actions of specific proteins [75]. These specific proteins are called *targets*. In the area of drug discovery, a protein is *druggable* if it is known to or is predicted to

bind to a drug. The subset of the 21,000 genes in the human genome that code for druggable proteins are called *the druggable genome* [75]. Sometimes a drug designed for a target may also bind to other unintended proteins, causing *side effects*.

The knowledge of drug-target associations is of utmost importance in drug discovery and development. Drug-target association data are available in several repositories, including Drugbank [76], PharmGKB [69], and ChEMBL [77]. Similar to disease-gene associations, drug-target associations can be modelled as a bipartite network (see Section 2.2.1), in which a node is a drug or a protein, and an edge exists between a drug and a protein if that protein is a target of that drug. Yildirim *et al.* [78] constructed the first drug-target network by using associations obtained from Drugbank. By analysing the drug-target network along with a human PPI network, they claimed that most drugs were palliative as they did not bind to proteins corresponding to the underlying causes of human diseases [78]. Drug-target associations have also been used for target identification [79, 80], drug repositioning [81, 82, 83, 84], and providing insights into molecular mechanisms of drug actions and side effects [85, 86].

2.2. Graph theory for network analysis

2.2.1. Graph theoretic terminology

Theoretical insights from graph theory have been successfully applied to various biological network modelling and analysis tasks (e.g., [87, 5, 67, 6, 78, 88]). In graph theory, a *network* or a *graph* is usually denoted by $G(V, E)$, where V is a set of n (or $|V|$) nodes and $E \subseteq V \times V$ is a set of m (or $|E|$) edges connecting them. Nodes adjacent to a node v are called v 's *neighbours*, and the number of edges connecting v is called v 's *degree*.

A *path* between two nodes u and v in a network is a sequence of edges connecting them. A graph is said to be *connected* if there exists a path between every pair of its nodes. The distance of nodes u and v , denoted by $d(u, v)$, is defined as the length (i.e., the number of edges) of the *shortest path* between nodes u and v .

A graph is *undirected* if its edges have no direction; otherwise, it is *directed*. A graph can be either *weighted* or *unweighted* depending on

whether there is a weight assigned to each of its edges. A graph is *bipartite* if its nodes V can be partitioned into two disjoint sets such that in each set, the nodes belonging to it are not connected. For an undirected graph, if it has no self-loops (i.e., edges connected at both ends to the same node) and multiple edges (i.e., more than one edge between any two different nodes), it is said to be a *simple* graph.

A *subgraph* S of a graph G is a graph whose nodes and edges belong to G . A subgraph S of a graph G is *induced* if S contains all edges that appear in G over the same subset of nodes. Given two graphs $G(V, E)$ and $G'(V', E')$, they are *isomorphic* if there is a bijective function $f : V \mapsto V'$, such that for all $u, v \in V$, $\{u, v\} \in E \Leftrightarrow \{f(u), f(v)\} \in E'$.

In practice, different data structures can be used for the representation of graphs. Three most commonly used representations are *edge list*, *adjacency list* and *adjacency matrix* (Figure 2.3). An edge list of a graph G is simply the list of G 's edges. An adjacency matrix of G is an $n \times n$ matrix, in which an entry a_{ij} indicates the presence or absence of an edge between nodes i and j . An adjacency list is an array of n lists, and a list i in the array contains all neighbours of the node i . Generally speaking, an edge list or an adjacency list is preferred to represent sparse graphs, and an adjacency matrix is preferred to represent dense graphs.

2.2.2. Network comparison

One of the major challenges in biological network research is *network comparison*. It aims to identify similarities and differences between networks. The comparison can be between two data networks (i.e., networks that represent data) or between a data network and a theoretical model (see Section 2.2.3 for details of network models). Network comparison is an essential part of biological network analysis. However, since it relies on the *subgraph isomorphism* problem, comparing large networks is computationally intensive. The subgraph isomorphism problem is a computational task to identify whether there is a subgraph S of a graph G that is isomorphic to another graph H , and it has been proved that this problem is *NP-complete* [89]. For this reason, some computable approximate measures of network topology, commonly called *network properties*, have been used to compare large networks. Network properties can be roughly divided into two categories:

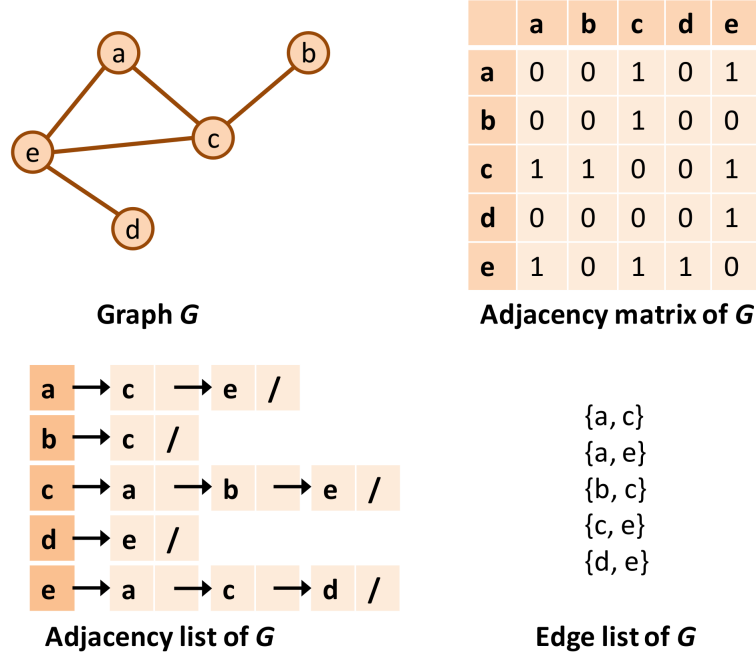


Figure 2.3.: An example of a graph and its adjacency matrix, adjacency list and edge list representations.

global network properties and *local network properties*.

Global network properties

Global network properties are statistical measures that describe an overview of a given network. The most commonly used global network properties include *edge density*, *degree distribution*, *average diameter*, and *clustering coefficient*.

The *edge density* of a network is defined as the proportion of the number of edges to the maximum possible number of edges. For a given network G with n nodes and m edges, the maximum number of edges is $\binom{n}{2} = n(n-1)/2$, thus its edge density is $m/\binom{n}{2}$. The *degree distribution* is the probability distribution of nodes' degrees over the whole network. Thus if n_k denotes the number of nodes in the network G with degree k , the degree distribution of G is $P(k) = n_k/n$, for each value of k . The *average diameter* (or *average path length*) of a network is defined as the average distance over all pairs of nodes in a network.

The *clustering coefficient* of a node is defined as the proportion of the

number of edges between that node’s neighbours to the maximum possible number of edges between the neighbours. Let $\deg(u)$ be the degree of a node u and E_u be the number of edges between u ’s neighbours, thus the clustering coefficient of u is $C_u = E_u / \binom{\deg(u)}{2}$. The *average clustering coefficient* of a network is defined as the average of clustering coefficients over all nodes in the network [90]. The distribution of average clustering coefficients of all nodes with degree k in a network is called the *clustering spectrum* of the network.

Global network properties of real-world networks have been extensively studied as they are relatively easy to compute. For example, it has been shown that many real-world networks have degree distributions that approximately follow power law distributions (so called *scale-free* [91, 92], also see Section 2.2.3 for details). Meanwhile, many large real-world networks tend to have small diameters and large clustering coefficients (so called *small-world* [90]). However, these global network properties can not precisely describe a network’s topology. In particular, networks with exactly the same value of a network property can have very different structures. For example, a network with 3 triangles and a network with a 9-node circle have the same number of nodes, the same number of edges and the same degree distribution, but obviously these two networks have very different structures (this example was illustrated in [47]). Furthermore, due to the incompleteness and the sampling and data collection biases in current biological data, global network properties of biological networks may even be misleading (discussed in Section 2.1). Therefore, it is essential to perform local statistics for large biological networks.

Local network properties

Local network properties measure local structures of a network. There are mainly two types of local network properties, *network motifs* and *graphlets*. *Network motifs* are defined as subgraphs that recur in a data network significantly more often than in randomised networks [61, 87]. *Graphlets* are defined as small, connected and induced subgraphs of a larger network [5]. Both network motifs and graphlets can be considered as small building blocks of large, complex networks, and have been successfully applied to various network analysis tasks [61, 87, 5, 6, 46]. For example, Milo *et al.*

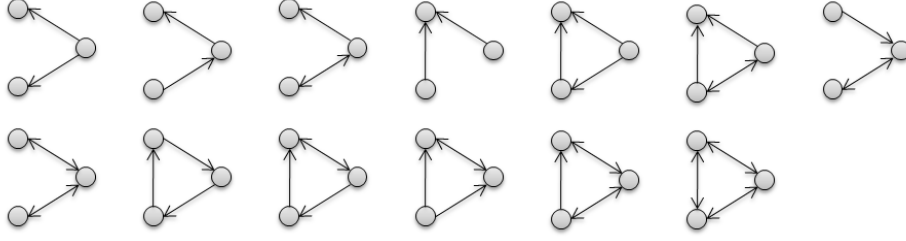


Figure 2.4.: All 13 types of 3-node connected subgraphs. This figure is reproduced from [61].

(2002) [61] identified network motifs in several real-world networks, including transcriptional regulation network, food webs, neuron connectivity network, electronic circuits network and World Wide Web, by scanning for all possible 3-node and 4-node subgraphs (see Figure 2.4 for examples of 3-node subgraphs). They showed that different types of networks had different motif sets, and thus suggested that network motifs could be used to classify networks. Moreover, Shen-Orr *et al.* (2002) [87] showed that network motifs of a transcriptional regulation network had specific biological functions, for example, the feed-forward loop motif represented the response of external signal of a system.

Network motifs have been used to compare the local structure of complex networks [93, 94, 95]. An approach based on *significance profile* was proposed by Milo *et al.* (2004) [93] for network comparison. To compute the significance profile of a given network, Milo *et al.* first generated a set of randomised networks with the same degree distribution as the given network, then compared the given network with randomised networks according to the occurrence of 3-node and 4-node subgraphs in these networks. For each subgraph i , the statistical significance was computed by a Z score:

$$Z_i = (N_{real_i} - \langle N_{rand_i} \rangle) / std(N_{rand_i}), \quad (2.1)$$

where N_{real_i} is the frequency of the subgraph i appears in the given network, and $\langle N_{rand_i} \rangle$ and $std(N_{rand_i})$ are the mean and the standard deviation of the frequency of i appears in randomised networks. The significance profile for the given network was defined as a vector of normalised Z score, $SP_i = Z_i / (\sum Z_i^2)^{1/2}$. By using this method, Milo *et al.* computed significance profiles for different types of networks, and showed that several previously

unrelated networks (such as a mammalian cell signalling network and a neuronal synaptic network of *Caenorhabditis elegans*) were topologically similar [93].

However, it has been noticed that motif-based approaches may lead to wrong results, since the detection of network motifs highly depends on the choice of the underlying null model [94]. For example, if the Erdős-Rényi (ER) random graph (see Section 2.2.3 for details) is chosen as the null model, it is very likely that dense subgraphs would be identified as network motifs since they rarely occur in an ER random graph. Systematic approaches based on another local network properties, graphlets, are free from the biases introduced by null model selection. Details of graphlets and graphlet-based approaches are discussed in Section 2.2.4.

2.2.3. Network models

Network models provide us a way to describe the structural features of a system. With a precise mathematical model representing the structure of biological networks, one can not only reproduce the observed biological data, but also predict the system’s behaviour to gain novel insights.

One of the first steps in modelling biological systems is to develop mathematical models that have similar statistical properties to the real-world networks. Several network models have been proposed for this purpose, including Erdős-Rényi (ER) random graph [96, 97], ER random graph with the same degree distribution as a data network (ER-DD), scale-free (SF) network [91, 98], geometric random graph (GEO) [99], geometric gene duplication and mutation (GEO-GD) model [8] and stickiness index-based (STICKY) network model [7]. These network models have been proved useful in identifying network motifs [61, 87, 93], finding cost-effective strategies for interactome detection [100], and predicting PPIs [101].

The ER model (Figure 2.5a) is one of the earliest network models. It can be generated by fixing the number of nodes n in the network, and adding edges uniformly at random until a given edge density is reached. If the probability of adding an edge between any pair of nodes is ρ (i.e., ρ is the edge density), an ER model created by this process has approximately $n(n - 1)\rho/2$ edges. Despite the simplicity of the model, the ER model is considered as one of the standard models to compare the data with, as many



Figure 2.5.: Examples of model networks. Left: An ER model. Center: An SF model. Right: A GEO model. Figures are taken from [47].

of its properties have been proven theoretically [102].

The degree distribution $P(k)$ (i.e., $P(k)$ is the fraction of nodes in the network having degree k , see Section 2.2.1) of an ER model follows a binomial distribution $P(k) = \binom{n-1}{k} \rho^k (1-\rho)^{n-1-k}$. This distribution becomes a Poisson distribution $P(k) = (np)^k e^{-np} / k!$ in the limit of large n . However, most real-world networks have degree distributions very different from this [91, 92]. An ER-DD model is a variation of the ER model, in which the degree distribution is forced to be the same as that of a data network. ER-DD models can be generated by assigning ‘stubs’ to each node in the network according to the degree distribution of the data network, and adding edges to pairs of stubs at random (see [103] for details). The ER-DD model preserves the degree distribution of the data network, while other properties such as clustering coefficient are still the same as the ER model.

An SF model (Figure 2.5b) is a network model in which the degree distribution follows a power law distribution $P(k) \sim k^{-\gamma}$, where γ is a parameter typically in the range $\gamma \in (2, 3)$ [91]. Many real-world networks, including PPI networks, metabolic networks, World Wide Web networks and social networks, have been conjectured to follow power law distributions [92]. An SF model can be generated by a preferential attachment process proposed by Barabási and Albert (1999) (so called SF-BA model) [91]. Starting with an initial connected network, new nodes are added to the network and connected to existing nodes with a probability that is proportional to the degree of existing nodes until a given density is reached. As a result of this process, the degree of a few nodes (hubs) is significantly higher than that of

the other nodes, as illustrated in Figure 2.5a. An SF model can also be generated based on a gene duplication and mutation model (SF-GD) [98]. In the duplication step, a parent node is chosen uniformly at random from the existing network and a new child node that connects to all neighbours of the parent node is added. An edge between the parent node and the child node is added with a given probability p . In the mutation step, for each node that connects to both the parent node and the child node, one of the two edges is removed with a given probability q . If the degree distribution of the data network follows a power law distribution, an ER-DD model generated from the data network is also an SF model. This is a special case of an ER-DD model called random SF (SF-RND).

SF models have small average diameters and power law degree distributions similar to that of the real-world networks, and they were used to model early PPI networks of *Saccharomyces cerevisiae* (yeast) [104, 105]. However, it was argued that the power law degree distribution in these PPI networks was an artifact of sampling: these early PPI networks were largely incomplete, thus the observed scale free topology could not be confidently extrapolated to complete interactomes [52]. Evidences have shown that PPI networks can be better modelled by other models, in particular, GEO models, than SF models [5, 6].

In a GEO model (Figure 2.5c), nodes represent uniformly randomly distributed points in a metric space, and nodes are connected by edges if the distance of their corresponding points is within some radius threshold r [99]. A GEO model has a degree distribution that follows a Poisson distribution, which is different from the real-world networks. However, GEO models reproduce high clustering coefficients and small average diameters of real-world networks, and have been shown to provide better fit to the currently available PPI networks than ER and SF models [6]. GEO-GD models are GEO models that incorporate the principles of gene duplications and mutations [8]. To generate a GEO-GD model, one chooses a parent node from the existing network uniformly at random, and adds a child node in a random direction at a randomly chosen distance either up to r with probability p or up to $10r$ with probability $1 - p$ (see [8] for more details). GEO-GD models have degree distributions that follow power law distributions, small average diameters and high clustering coefficients, thus they preserve the global properties of the real-world networks.

	Degree distribution	Clustering coefficient	AVG diameter
Real-world	Power law	High	Small
ER	Poisson	Low	Small
ER-DD	Power law	Low	Small
SF	Power law	Low	Small
GEO	Poisson	High	Small
GEO-GD	Power law	High	Small
STICKY	Power law	High	Small

Table 2.1.: Comparison between different network models based on global network properties.

A STICKY model is based on a stickiness index that is proportional to the degree of a node in a real network [7]. This model was proposed particularly for modelling PPI network, under the assumption that the higher the degrees of two proteins in the PPI network, the more likely they interact with each other. The STICKY model also preserves the global properties of the real-world networks, and it has been shown that currently available PPI networks are well-fit by the STICKY model [7, 106].

Table 2.1 summaries a comparison between different network models based on their global network properties. Since real-world networks are shown to be small-world and scale-free, they have small average diameters, high clustering coefficients, and power law degree distributions. In this regard, real-world networks seem to be better fit by GEO-GD and STICKY models than other models.

2.2.4. Graphlet-based measures for network analysis

Systematic approaches based on *graphlets* have been introduced to measure the structure (or *topology*) of large networks. As described in Section 2.2.2, a *graphlet* is defined as a small, connected and induced subgraph of a larger network [5]. Figure 2.6 depicts all 30 of the possible graphlets that have 2, 3, 4, and 5 nodes. Topologically identical nodes within a graphlet are said to belong to the same *automorphism orbit* [6]. In total there are 73 different automorphism orbits in the 30 graphlets, as shown in Figure 2.6.

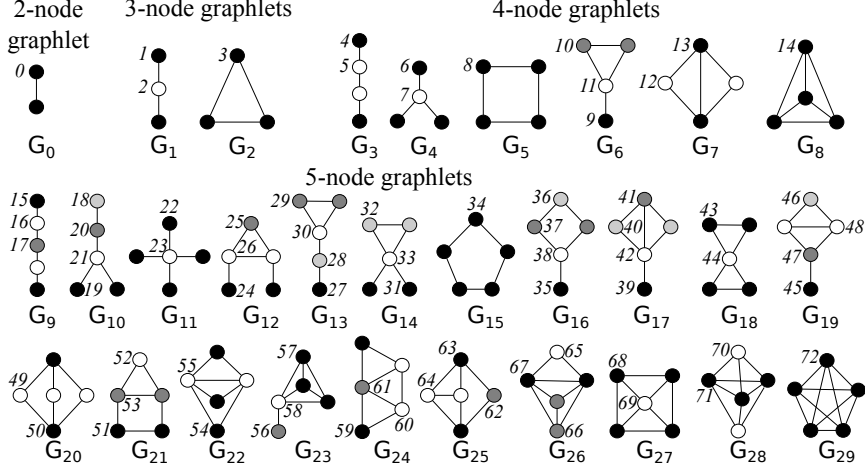


Figure 2.6.: Graphlets with 2-5 nodes $G_0, G_1, \dots, G_{29}$. The automorphism orbits are numbered from 0 to 72, and the nodes belonging to the same orbit are of the same shade within a graphlet [6].

RGF distance and GDD agreement

Graphlets have been used to measure the similarity of the topology of PPI networks and network models [5, 6]. In particular, two measures have been proposed for this purpose. One is based on comparing graphlet frequencies between two networks, and is termed *relative graphlet frequency* (RGF) distance [5]. The other is based on comparing *graphlet degree distributions* (GDDs) between two networks, and is termed *GDD agreement* [6]. Both measures allow quantification of the similarity between the topology of two data networks, or comparison of a data network with a network model.

The RGF distance between two networks G and H is computed as follows. Let $N_i(G)$ be the number of graphlets of type i ($i \in 1, \dots, 29$) in the network G . The relative frequency of graphlets of type i is defined as $N_i(G)/T(G)$, where $T(G) = \sum_{i=1}^{29} N_i(G)$ is the total number of graphlets of G . The RGF distance between two networks G and H is computed as:

$$RGF(G, H) = \sum_{i=1}^{29} |F_i(G) - F_i(H)|, \quad (2.2)$$

where $F_i(G) = -\log(N_i(G)/T(G))$. This measure has been applied to examine the fit of ER, ER-DD, SF-BA, and GEO models to PPI networks of *Saccharomyces cerevisiae* (yeast) and *Drosophila melanogaster* (fruit fly)

[5]. With respect to the measure, it has been shown that the topology of PPI networks is better modelled by a GEO model than a SF-BA model [5].

GDD agreement was later introduced as another graphlet-based measure for network comparison [6]. The notion of the *degree* has been generalised to *graphlet degree*: the *degree distribution* over degrees k of a graph is a distribution of how many nodes have degree k ; in turn, a *GDD* is an analogous distribution for an automorphism orbit that counts the number of nodes touching this orbit k times over all k . For example, the GDD corresponding to orbit 3 in Figure 2.6 measures how many nodes touch k triangles for each value of k . There are 73 GDDs for graphlets with up to 5 nodes, and the degree distribution is the first one of them.

More specifically, let d_G^j be the j^{th} GDD of a network G and $d_G^j(k)$ be the number of nodes in a network G that are touched k times by a particular automorphism orbit j shown in Figure 2.6. To compute the normalised graphlet degree for orbit j , $d_G^j(k)$ is first scaled by $1/k$ to decrease the contribution of larger degrees in a GDD, and then normalised to give a total sum of 1,

$$N_G^j(k) = \frac{d_G^j(k)/k}{\sum_{l=1}^{\infty} d_G^j(l)/l}. \quad (2.3)$$

For two networks G and H and a particular orbit j , the distance $D^j(G, H)$ between their normalised j^{th} GDD is defined as:

$$D^j(G, H) = \frac{1}{\sqrt{2}} \left(\sum_{k=1}^{\infty} [N_G^j(k) - N_H^j(k)]^2 \right)^{\frac{1}{2}}, \quad (2.4)$$

where the $1/\sqrt{2}$ is a normalisation constant that ensures the distance is always less than 1. Finally, the GDD agreement between networks G and H is defined as:

$$\text{GDDA}(G, H) = \frac{1}{73} \sum_{j=0}^{72} (1 - D^j(G, H)). \quad (2.5)$$

If two networks G and H have a GDD agreement close to 1, that means G and H are considered to be topologically similar as their GDDs, scaled appropriately to their network size, are statistically similar. GDD agreement has been used to compare 14 PPI networks of eukaryotic organisms with 4

random network models, and the fit between the PPI networks and the model networks has been assessed [6].

Graphlet signature similarity

Graphlet-based measures have also been developed to measure the topological similarity between nodes in a network. The *graphlet signature similarity* compares nodes in a network based on the local structure of node neighbourhoods [46]. The *graphlet signature* is a generalisation of the notion of the degree: the degree of a node u counts how many edges it touches; in turn, the *graphlet signature* of u is a 73-dimensional vector that counts how many times u is touched by each of the automorphism orbits shown in Figure 2.6 [46].

The graphlet signature similarity of a pair of nodes u and v is computed as follows. Let u_i be the i^{th} element of the graphlet signature of u , i.e., u_i is the number of times u is touched by the particular automorphism orbit i . Then the distance between the i^{th} orbit of nodes u and v is defined as:

$$SigDist_i(u, v) = w_i \times \frac{|\log(u_i + 1) - \log(v_i + 1)|}{\log(\max(u_i, v_i)) + 2}. \quad (2.6)$$

where w_i is a weight assigned to orbit i which is defined as $w_i = 1 - \log(o_i)/\log(73)$, and o_i is the dependency count of orbit i . For example, for orbit 4, its dependency count o_4 is 3, as orbit 15 is affected by orbits 0, 1, and itself (see [46] for details). Finally, the graphlet signature similarity between u and v is computed as:

$$SigSim(u, v) = 1 - \frac{1}{\sum_{i=0}^{72} w_i} \left(\sum_{i=0}^{72} SigDist_i(u, v) \right) \quad (2.7)$$

Signature similarities have been applied to measure the topological similarities between proteins in a PPI network [46]. It has been shown that topologically similar proteins are very likely to belong to the same protein complexes, perform the same biological functions, be localised in the same subcellular compartments, and have the same tissue expressions [46]. Furthermore, topological similarities of proteins in a PPI network have been considered as a complementary information to sequence similarities, and have been applied to detect homologous proteins [46], align PPI networks of different species [9, 10, 11, 12], uncover melanogenesis regulatory network

components, and identify disease genes [40, 107].

Graphlets for network alignment

Graphlet-based measures discussed above have also been applied for many biological network alignment tasks [9, 10, 11, 12, 108]. Given two or more networks of the same type, network alignment allows us to compare networks and identify conserved subnetworks within them. For example, network alignment can be used to measure the global similarity between PPI networks of different species, further to infer phylogenetic relationships [11]. Such alignment could also provide transfer of knowledge between species, as well as insights into evolution, disease and gene function [47].

Analogous to sequence alignment, network alignment can be *local* or *global*. Local network alignment algorithms, such as the BLAST family (PathBLAST [109], NetworkBLAST [110] and NetworkBLAST-M [111]), MAWISH [112], NetAlign [113] and Graemlin 1.0 [114], aim to match subnetworks from one network to subnetworks in another network. However, the alignments produced by these algorithms may be ambiguous, as local alignment algorithms map each subnetwork of similarity independently, thus allow one node to have different pairings. Global network alignment algorithms uniquely map each node in one network (the smaller network) to exactly one node in the other network (the larger network), even though this may lead to suboptimal matchings in some local regions. These global network alignment algorithms include the IsoRank family (pairwise IsoRank [115], multiple IsoRank [116] and IsoRankN [117]), Graemlin 2.0 [118], GA and PATH [119], HopeMap [120], NATALIE [121, 122], PISwap [123], GEDEVO [124], SPINAL [125], NETAL [126] and graphlet-based network alignment algorithms [9, 10, 11, 12]. So far four graphlet-based algorithms have been proposed for global network alignment, including GRAPh ALigner (GRAAL) [9], Hungarian-algorithm-based GRAAL (H-GRAAL) [10], Matching-based Integrative GRAAL (MI-GRAAL) [11] and Common-neighbors-based global GRAAL (C-GRAAL) [12]. These graphlet-based algorithms align two networks by considering the graphlet signature similarities of network nodes. Graphlet signature similarity has also been extended for fast and flexible alignment of protein 3D structures [108].

Software for graphlet-based measures: GraphCrunch

Several graphlet-based measures, including RGF distance, GDD agreement, signature similarity, and the GRAAL network alignment algorithm, have been implemented in the open source software GraphCrunch [127, 128]. In addition, GraphCrunch also contains random network generators of ER, ER-DD, GEO, SF-BA, GEO-GD, SF-GD and STICKY models. GraphCrunch can be used to solve many network analysis and modelling tasks, such as finding the best fitting model for a given network, comparing different networks based on their global and local network properties, aligning two networks using GRAAL (see [9] for details of network alignment and the GRAAL algorithm), as well as identifying clusters in a given network based solely on network topology.

2.3. Disease-disease associations

2.3.1. Disease classification

Diseases can be classified based on different criteria. Two of the most widely used criteria are topography and anatomy. In the topographic classification, diseases are classified based on the affected region or system (e.g., gastrointestinal disease and vascular disease), while in the anatomic classification, diseases are classified based on the affected organ or tissue (e.g., heart disease and liver disease). Other criteria such as physiology, pathology, etiology and epidemiology, can also be used to classify diseases.

Published and maintained by the World Health Organization (WHO), the ICD classification system is considered to be the international standard of disease classification [1]. It provides a system of diagnosis codes for identifying and grouping diseases, and has been used worldwide for general epidemiological and health management purposes. In the ICD classification system, diseases are classified into several major categories (see Table A.1), which can be further divided in subgroups represented by 3-5 digit level diagnosis codes. For example, according to the 9th revision of ICD (ICD-9)¹, diabetes mellitus (ICD-9 code: 250) is classified under the major category ‘endocrine, nutritional and metabolic diseases, and immunity disorders’, and it can be further divided into subgroups represented by 4-digit level codes

¹<http://www.who.int/classifications/icd/en/>

(e.g., ICD-9 code 250.1 for diabetes with ketoacidosis) or 5-digit level codes (e.g., ICD-9 code 250.10 for type 2 diabetes with ketoacidosis).

Another commonly used disease classification is the MeSH thesaurus, a controlled vocabulary produced by the National Library of Medicine (NLM) [2]. The MeSH thesaurus is used by the NLM for indexing, cataloging, and searching for biomedical and health-related information and documents collected in databases such as the PubMed². It consists of sets of terms that are arranged in a 12-level hierarchical tree structure³. As shown in Table A.2, the most general level of the MeSH tree contains broad headings such as ‘Neoplasms’ (MeSH tree ID: C04) and ‘Mental Disorders’ (MeSH tree ID: F03). More specific headings can be found at more narrow levels of the MeSH tree, such as ‘Inflammatory Breast Neoplasms’ (MeSH tree ID: C04.588.180.576).

Since there exist different disease classification systems, the Disease Ontology⁴ (DO) has been developed recently with the aim to provide a unifying representation of human diseases for the biomedical community [129]. It also contains cross mappings between several common biomedical vocabularies including ICD, MeSH, OMIM, SNOMED and UMLS. Such mappings are essential for integrating disease-related data obtained from different sources such as GWAS studies, patient records and the literature.

2.3.2. Computational approaches for uncovering disease-disease associations

During the past decade, various large-scale genomic, proteomic, transcriptomic and metabolomic studies have provided an abundance of biological data describing diseases, prompting the scientific community to gain insights into diseases and their relationships at a molecular systems-level. As a result, several computational approaches have been developed and applied to uncover disease-disease associations from various types of biological data.

Text mining is one of the most commonly used methods to uncover phenotype associations. van Driel *et al.* (2006) [65] inferred similarities between human phenotypes by an automated text mining of the OMIM database, and found those similarities were positively correlated with a number of mea-

²<http://www.ncbi.nlm.nih.gov/pubmed>

³<http://www.nlm.nih.gov/mesh/trees.html>

⁴<http://disease-ontology.org>

asures of gene function. Phenotype associations identified by van Driel *et al.* (2006) were further used to investigate the modularity of phenotypes and its correlation with disease classification, functional genomics and drug-target associations by Jiang *et al.* (2008) [130]. Also by text mining the OMIM database, Lage *et al.* (2007) [66] derived phenotype similarity scores and integrated them with human PPI data to predict disease gene candidates. A recent study by Xu *et al.* (2013) [131] extracted disease-manifestation relationships by text mining the literature and inferred disease-disease associations based on manifestation similarities.

Apart from large-scale disease-gene association databases like OMIM, other resources such as electronic patient records have also been used for studying disease-disease associations. A type of disease-disease association that can be derived from medical records is disease *comorbidity*, which indicates the potential for the co-occurrence of diseases in the same individual. Hidalgo *et al.* (2009) [132] built a phenotypic database summarising comorbidity data obtained from the disease history of 32 million American patients. Two statistical measures were used to quantify the strength of comorbidity association between a pair of diseases: the Relative Risk (RR) and ϕ -correlation. They showed that both measures captured statistically significant comorbidity associations, but RR tended to prioritise associations between diseases within the same ICD-9 category while ϕ -correlation tended to prioritise associations between diseases across different ICD-9 categories. This phenotypic database was further analysed to examine the correlation between comorbidity and disease progression [132], the underlying structure of cellular networks [133] and the evolution of human disease genes [134]. Comorbidity associations between diseases have also been used to infer genetic overlaps between phenotypes [135] and investigate possible molecular basis for diseases [136].

Disease-disease associations have also been inferred by other computational approaches, including analysing disease-associated SNPs from GWAS data [137, 138], fusing systems-level molecular data [139], measuring the correlation between disease-related gene expression data [81, 140] and comparing annotations in biomedical ontologies such as GO [141], DO [142] and Human Phenotype Ontology (HPO) [143]. Recently, network-based approaches have been successfully applied to represent and analyse the interconnectedness of human diseases. Goh *et al.* (2007) [67] constructed

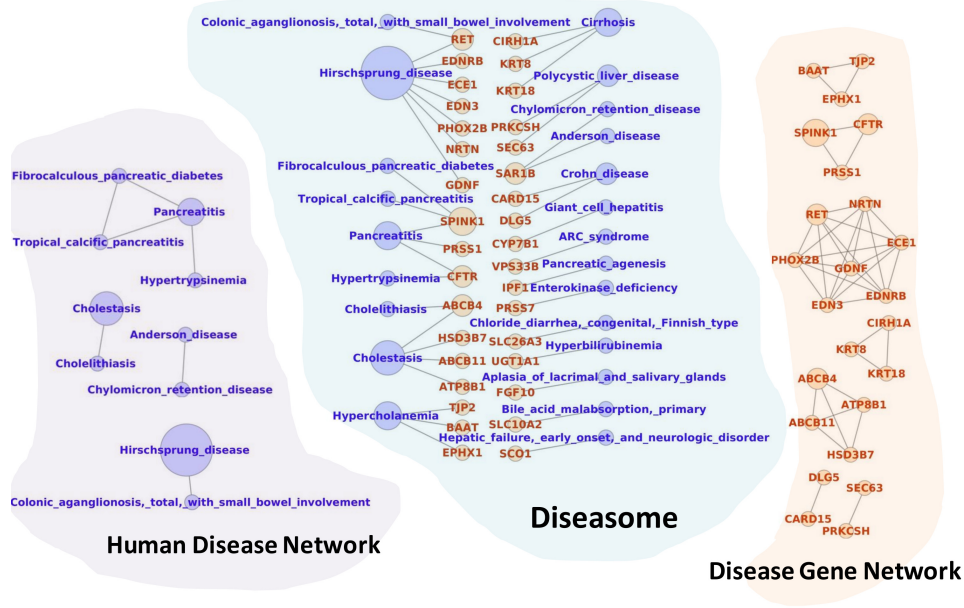


Figure 2.7.: An example of how to construct the human disease network (left) and the disease gene network (right) from the diseasome bipartite network (center). The diseasome network was constructed from the disease-gene association data collected by Goh *et al.* (2007) [67]. Only diseases classified as ‘Gastrointestinal diseases’ and their associated genes were shown in the figure. Blue and orange nodes represent diseases and human genes respectively. The size of a node is proportional to its degree in the diseasome network.

the first human *diseasome*, a bipartite network representing disease-gene associations obtained from OMIM (discussed in Section 2.1.3). They also generated two network projections: one was the ‘human disease network’, in which nodes were diseases and two diseases were connected by an edge if they shared at least one gene; the other was the ‘disease gene network’, in which nodes were genes and two genes were connected by an edge if they were involved with the same disease. Figure 2.7 demonstrates the construction of the human diseasome and its two projections. By analysing these networks with a human PPI network, Goh *et al.* (2007) showed that genes associated with the same disease were functionally related, and the majority of disease genes were nonessential and showed no tendency to encode ‘hubs’ in the PPI network.

A human disease association network can also be constructed based on other systems-level data [144, 57, 140, 145]. For example, Lee *et al.* (2008) [144] used metabolic data along with OMIM to construct a metabolic disease network in which two diseases were connected if their associated enzymes catalysed adjacent metabolic reactions; Li and Agarwal (2009) [57] associated diseases to biological pathways where disease genes were enriched, and constructed a pathway-based disease network by connecting diseases that shared at least one pathway. Analysis of human disease networks have been shown useful for building novel hypotheses about disease mechanisms and identifying new drug targets and biomarkers for complex diseases [41].

3. Evaluation of graphlet-based measures

Gaining biological insights from PPI networks requires the development of efficient and biologically meaningful algorithms for their analysis. Systematic measures based on graphlets are proving useful in this regard. Recently, the use of graphlet-based measures for biological network comparison has been questioned: it has been claimed that the measures are ‘unstable’ in regions of low edge density.

In this chapter, we demonstrate that graphlet-based measures are suitable for biological network comparison. Starting by generating empirical distributions of GDD agreement scores, we identify the edge density regions in which the topology of model networks is ‘unstable’, and show how graphlet-based measures correctly detect this topological instability. We also demonstrate that this ‘instability’ does not affect the analysis of recent PPI networks. Furthermore, we show that data networks have local densities much higher than a model network would have, since models are uniformly dense. Hence, graphlet-based measures are stable in regions that are of interest in real networks. Finally, we validate the use of graphlet-based measures for finding well-fitting network models for PPI networks by using a recently devised non-parametric statistical test. We show for the first time that five viral species, possessing the latest and most complete PPI networks, are well-fit by several network models.

The work presented in this chapter is in collaboration with Prof. Wayne Hayes and it is published in the following paper:

- W. Hayes, K. Sun, and N. Pržulj, “Graphlet-based measures are suitable for biological network comparison,” *Bioinformatics*, vol. 29, no. 4, pp. 483-491, 2013.

3.1. Motivation

Networks have been used to represent interactions in molecular biology, such as those between genes in gene regulatory networks and those between proteins in PPI networks (discussed in Section 2.1). A PPI network models physical interactions between proteins in a cell, where a node represents a protein and an undirected edge exists between two nodes if the corresponding proteins can physically bind to each other. The whole set of such interactions in an organism forms the *interactome*. Understanding the structure of PPI networks is an integral step towards understanding cellular systems.

Recent advances in high-throughput PPI detection methods, such as Y2H screening [15, 16, 18, 19, 20, 21] and MS of purified complexes [22, 23, 25, 24] have provided an abundance of network data. As these PPI networks are large and complex, it is necessary to develop efficient and biologically meaningful algorithms for their analysis. Since exactly solving many network analysis problems (e.g., network comparison) is computationally intractable, computable approximate measures have been developed to analyse these networks (discussed in Section 2.2.2).

Systematic measures based on graphlets have been introduced to analyse the topology of large networks [5, 6]. In particular, two measures have been proposed for network comparison, namely RGF distance and GDD agreement. RGF distance is based on comparing graphlet frequencies between two networks [5], while GDD agreement is based on comparing *graphlet degree distributions* (GDDs) between two networks [6]. Both measures allow quantification of the similarity between the topology of two data networks, or comparison of a data network to a model network. For example, GDD agreement has been used to compare fourteen PPI networks of eukaryotic organisms with four different random network models, and assess the fit between PPI networks and model networks [6]. Details of RGF distance and GDD agreement can be found in Section 2.2.4.

Recently, the use of graphlet-based measures for network comparison has been questioned: it has been claimed that the measures are ‘unstable’ in regions of low edge density [13]. In the study of Rito *et al.* (2010) [13], empirical distributions of GDD agreement scores were calculated and a novel non-parametric test for assessing the statistical significance of the fit be-

tween PPI networks and theoretical models was introduced. Six PPI networks of yeast and human were tested, and it was found that none of these PPI networks were well-fit by the three random network models tested, i.e., ER, ER-DD, and GEO models [13].

In the rest of this chapter, we first give details of PPI networks we analysed, along with methods we used to model and analyse these PPI networks. We then demonstrate graphlet-based measures are suitable for biological network comparison, based on the following observations. First, RGF distance and GDD agreement are not ‘unstable’ for low edge density networks; instead, they correctly detect instability that is present in the structure of low edge density model networks. Second, the ‘unstable’ low-density region described in Rito *et al.* (2010) [13] is small in large networks, and current PPI networks have densities outside this region. Third, PPI networks used in Rito *et al.* (2010) [13] are not only out-of-date now, but were also out-of-date at the time of their study. Fourth, local densities of natural networks are much higher than those of model networks, since models are uniformly dense. Finally, we use the non-parametric test proposed in Rito *et al.* (2010) [13] to demonstrate that PPI networks of many species, particularly five viral species, are well-fit by several existing network models.

3.2. Methods

3.2.1. Data: PPI networks

We collected and analysed eighteen PPI networks of various species, including five PPI networks of herpesvirus, four PPI networks of prokaryotes and nine PPI networks of five eukaryotes (Table 3.1).

Viral PPI networks were collected from Fossum *et al.* (2009) [26], including PPI networks of herpes simplex virus type 1 (HSV-1), Kaposi’s sarcoma-associated herpesvirus (KSHV), varicella-zoster virus (VZV), Epstein-Barr virus (EBV) and murine cytomegalovirus (mCMV). These viral PPI networks are considered to be the most complete PPI networks, as they were produced by complete genome-wide Y2H experiments, which means all possible pairs of proteins were tested for interaction.

The four PPI networks of prokaryotes we used are bacterial PPI networks collected from different studies, including PPI networks of *Campy-*

PPI	Nodes	Edges	Density	Reference
HSV-1	47	100	0.09251	Fossum <i>et al.</i> (2009) [26]
KSHV	50	115	0.09388	Fossum <i>et al.</i> (2009) [26]
VZV	57	160	0.10025	Fossum <i>et al.</i> (2009) [26]
EBV	60	208	0.11751	Fossum <i>et al.</i> (2009) [26]
mCMV	111	393	0.06437	Fossum <i>et al.</i> (2009) [26]
CJJ	1,111	2,988	0.00485	Parrish <i>et al.</i> (2007) [146]
HS	1,529	2,667	0.00228	Stelzl <i>et al.</i> (2005) [18]
MZL	1,804	3,094	0.00190	Shimoda <i>et al.</i> (2008) [147]
HG	1,873	3,463	0.00198	Rual <i>et al.</i> (2005) [19]
SPP	1,920	3,102	0.00168	Sato <i>et al.</i> (2007) [148]
ECL	1,941	3,989	0.00212	Peregrin-Alvarez <i>et al.</i> (2009) [27]
AT	2,634	5,529	0.00159	Arabidopsis Interactome Mapping Consortium (2011) [21]
WB	2,817	4,527	0.00114	BioGRID (ver. 3.1.74) [30]
YB	5,607	57,143	0.00364	BioGRID (ver. 3.1.74) [30]
FB	7,372	24,063	0.00089	BioGRID (ver. 3.1.74) [30]
HB	8,920	35,386	0.00089	BioGRID (ver. 3.1.74) [30]
HR	9,141	41,456	0.00099	Radivojac <i>et al.</i> (2008) [149]
HH	9,465	37,039	0.00082	HPRD (ver. 9) [150]

Table 3.1.: The eighteen PPI networks we analysed, ordered by size.

lobacter jejuni (CJJ) [146], *Mesorhizobium loti* (MZL) [147], *Synechocystis sp. PCC6803* (SPP) [148] and *Escherichia coli* (ECL) [27]. SPP and MZL were both constructed from PPIs produced by Y2H experiments [148, 147]. CJJ is the high confident part of the PPI network of *C. jejuni* produced by Y2H experiments [146]. ECL is the high confident functional interaction network of *E.coli* inferred from both experimental interaction datasets and computational predication datasets [27].

We also analysed nine PPI network of five eukaryotes, including *Arabidopsis thaliana*, *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (nematode worm), *Drosophila melanogaster* (fruit fly), and *Homo sapiens* (human). As can be seen from Table 3.1, HS was constructed by using the human PPIs reported by Stelzl *et al.* (2005) [18], and HG was constructed by using the human PPIs reported by Rual *et al.* (2005) [19]. Since Stelzl *et al.* (2005) [18] and Rual *et al.* (2005) are the first Y2H studies of the human interactome, both HS and HG are largely incomplete. Note that

these two PPI networks were analysed both by Przulj (2007) [6] and Rito *et al.* (2010) [13]. HH is the human PPI network obtained from HPRD version 9 [150] (released in April 2010). HR is the human PPI network constructed by Radivojac *et al.* (2008) [149]. Finally, HB, WB, FB and YB are PPI networks of human, nematode worm, fruit fly and yeast obtained from BioGRID version 3.1.74, released in March 2011 [30]. The PPI network of *Arabidopsis* (AT) was constructed by using PPIs published by the Arabidopsis Interactome Mapping Consortium (2011) [21].

Note that for all PPI networks, we removed self-loops, duplicate interactions and inter species interactions from the PPI data, as we model these PPI data as simple, undirected networks (discussed in Section 2.1.1 and Section 2.2.1).

3.2.2. Network models and their utility

Random network models discussed in this chapter include ER, ER-DD (SF-RND), GEO, GEO-GD, SF-GD and STICKY models. Details of these network models and their properties can be found in Section 2.2.3. Models play a key role in understanding the behaviour of complex systems. A precise model that represents the structural features of a system can provide us with a way to describe the observed data, and more importantly, it allows us to predict the system's behaviour to gain novel insights. Thus, modelling biological networks is fundamental to understanding complex biological systems.

One of the first steps in modelling biological networks is to develop mathematical models that have similar statistical properties to the real-world networks. However, current biological networks are noisy and largely incomplete (discussed in Section 2.1). In addition, they contain sampling biases as well as other biases in data collection, handling and interpretation [151, 152, 52, 153, 53]. Thus, due to the noise, incompleteness and biases in the data, it is very difficult to find the model that perfectly fit the data. Instead, our goal is to find the model that better fit the data to guide the development of network models in the right direction.

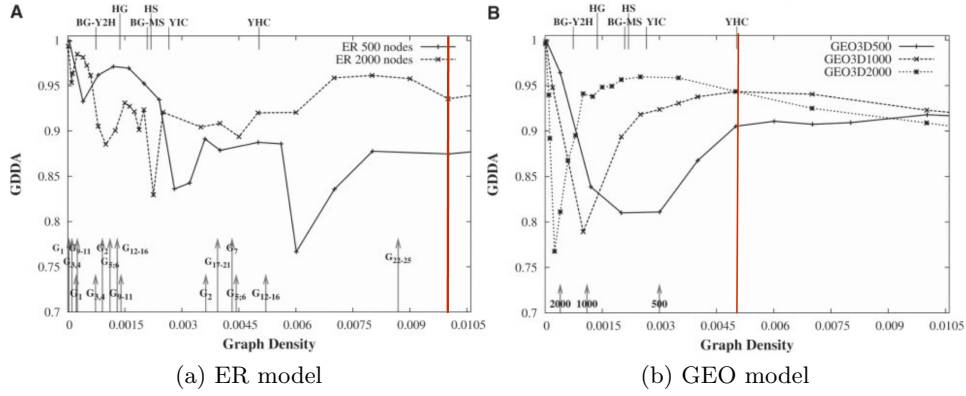
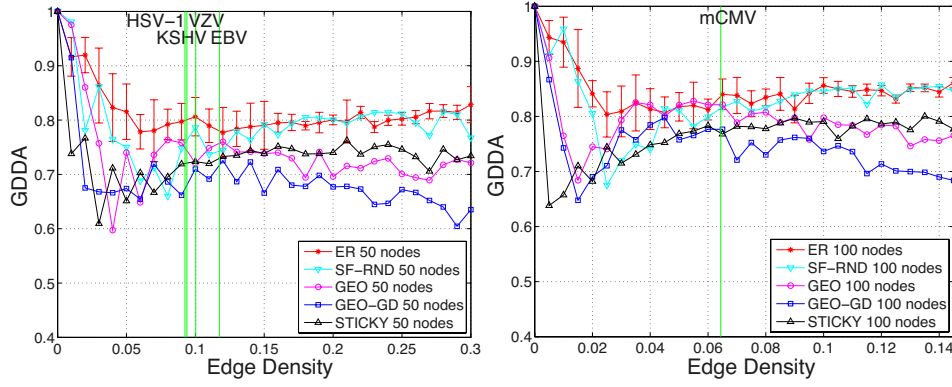


Figure 3.1.: GDD agreement scores vs. edge density when comparing (a) 50 ER model networks to each other, (b) 50 GEO model networks to each other. Edge densities of the PPI networks analysed by Rito *et al.* (2010) are indicated by bars on the top x-axis. These PPI networks are discussed in Section 3.3.3. In (a), arrows along the x-axis indicate the approximate edge density at which Rito *et al.* (2010) expected the first graphlet to appear in an ER model network with 500 nodes (indicated by short arrows) and 2000 nodes (indicated by long arrows). In (b), arrows along the x-axis indicate the approximate edge density at which Rito *et al.* (2010) expected 3-node graphlets to appear in a GEO model network with 500, 1000 and 2000 nodes. These two figures are taken from Rito *et al.* (2010) [13]. Note that two red lines are added into the figures to indicate the ‘unstable’ regions of ER and GEO model (will be discussed in Section 3.3.2).

3.3. Results and discussion

3.3.1. Sensitive vs. unstable

It has been concerned that graphlet-based measures are ‘unstable’ in regions of low edge density [13]. In particular, Rito *et al.* (2010) [13] questioned the use of GDD agreement (GDDA) for network comparison. They claimed that ‘GDDA has a pronounced dependency on the number of edges and vertices of the networks being considered’, and ‘GDDA score is not stable in the region of graph density relevant to current PPI networks’ [13]. Their arguments can be illustrated by Figure 3.1, which depicts the dependency of GDD agreement scores on edge density (i.e., the proportion of the number of edges to the maximum possible number of edges, discussed in Section



(a) 50-node model-vs-model comparisons (b) 100-node model-vs-model comparisons

Figure 3.2.: GDD agreement scores vs. edge density when comparing 30 random model networks to each other with (a) 50 nodes, (b) 100 nodes. Error bars are standard deviations of GDD agreement scores. Since all models have about the same variance, we show the error bars only for ER models for better visualisation. Green lines mark the edge densities of the five viral PPI networks discussed in Section 3.2.1. In all cases, the viral PPI networks have densities outside the region of sensitivity.

2.2.2) when comparing two model networks to each other. The abrupt drop of GDD agreement curves at low edge density regions shown in these figures was referred to as ‘unstable GDDA’ in Rito *et al.*’s study [13].

It is crucial to distinguish between a measure being unstable, versus a measure being *sensitive* when the structure of a network is unstable (i.e., network properties have a statistically high variance). We claim that graphlet-based measures (both GDD agreement and RGF distance) are sensitive that they correctly detect differences between the structure of various networks. For example, consider two networks: one contains four isolated edges and the other contains a square and four isolated nodes. Both networks have eight nodes and four edges, but obviously, their structures are very different from each other. Both RGF distance and GDD agreement would correctly detect this difference. Thus, it is not the measure that is unstable, but the network structure itself is unstable at low edge density regions. Graphlet-based measures correctly measure the topological instability in networks with a low edge density.

For this reason, we believe that the abrupt drop of GDD agreement curves

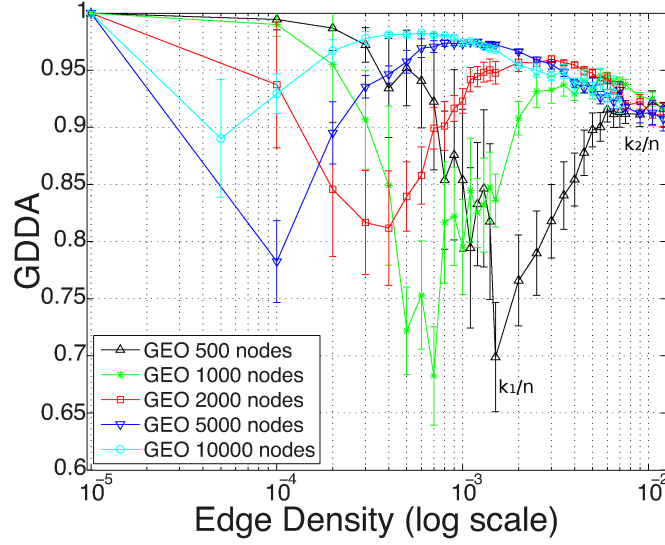


Figure 3.3.: GDD agreement scores vs. edge density when comparing 30 GEO model networks to each other, with 500, 1000, 2000, 5000 and 10000 nodes. Similar plots that illustrate the sensitivity of GDD agreement and RGF distance when comparing other random model networks (ER, SR-RND, GEO-GD and STICKY) can be found in Figure A.3. k_1/n indicates the edge density where the minimum model-vs-model GDD agreement score for GEO models with 500 nodes occurs, and k_2/n indicates the edge density where the model-vs-model GDD agreement score for GEO models with 500 nodes recovers.

at low edge densities is due to network properties having a statistically high variance at low edge densities. This high variance is shown in Figure 3.2 and Figure 3.3, as the error bars which represent the standard deviations of GDD agreement scores are obviously larger at low edge densities than they are at high edge densities. The network's topological instability causes a low GDD agreement score when comparing two model networks that have low edge densities. The network properties of model networks become more similar at higher edge densities, resulting in a higher GDD agreement score when comparing two networks generated from the same model.

Note that in Figure 3.2 and Figure 3.3, a point in a curve represents an average GDD agreement score when comparing 30 random model networks to each other. We chose 30 as the number of networks we used since a sample size of 30 networks is more statistically significant than any smaller

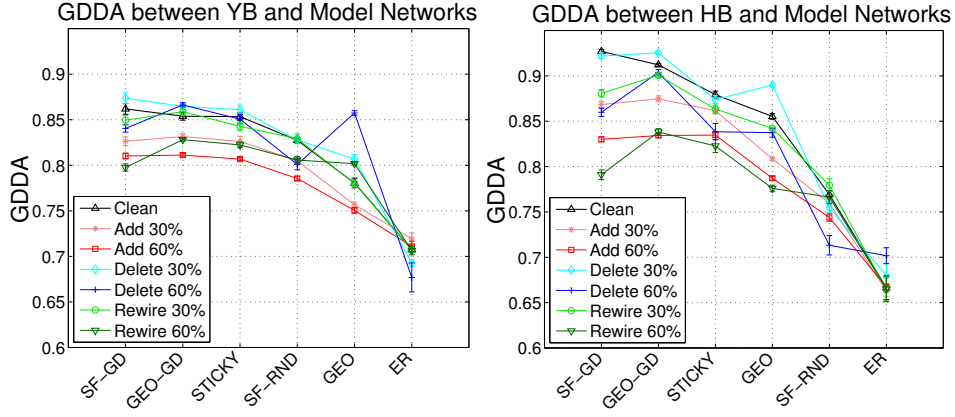


Figure 3.4.: GDD agreement is robust with respect to noise. In particular, we see that adding, deleting, or re-wiring as many as 60% of the edges in the model networks has little effect on which model best fits the data. Thus, conclusions about which models better fit the data are relatively insensitive to the fact that real-world network data contain substantial amounts of noise.

values (e.g., 10 or 20 networks). Meanwhile, the shape of the curves does not change with the increased number of model networks (see Figure A.1 in the Appendix). Using a large sample size (e.g., 100 networks) is computationally expensive and also unnecessary as demonstrated in Figure A.1.

GDD agreement is sensitive to noise, whereas conclusions about network fit using GDD agreement are robust to noise (see Section 2.1 as well as [5, 52, 152, 6, 101] for discussions of noise). Figure 3.4 shows how GDD agreement scores change when we add, delete, and rewire up to 60% of edges in the model networks. As can be seen, GDD agreement is sensitive as it generally decreases with increasing noise level. For example, GDD agreement between the yeast PPI network YB and an SF-GD model network is about 0.86, while the score drops to 0.8 when we rewire 60% of edges in the model network. Meanwhile, we can also observe from Figure 3.4 that the *ordering* of which model best fits the data undergoes only small changes when noise is added. In particular, all curves in Figure 3.4 generally have a decreasing tendency, indicating that the orderings of models on the horizontal axis remains the same, independent of noise level.

3.3.2. The size of the ‘unstable’ region

To examine whether the ‘unstable’ regions affect the analysis of current PPI networks, we analysed eighteen PPI networks of various species (discussed in Section 3.2.1). We found that even small real-world networks, such as the viral PPI networks, have edge densities outside the topologically ‘unstable’ region. Figure 3.2 depicts the edge densities of viral PPI networks (green vertical lines) and shows that they are outside the ‘unstable’ region, regardless of the assumed theoretical model. A similar effect for larger PPI networks is demonstrated in Figure 3.5 and Figure 3.6. In these two figures, we plot a surface depicting the dependency of GDD agreement on the number of nodes n and the edge density ρ for ER and GEO models. The non-viral PPI networks listed in Table 3.1 are plotted in the (n, ρ) plane according to their sizes and edge densities. As can be seen from Figure 3.6, none of these PPI networks are in the ‘unstable’ region of GEO models. Only a few PPI networks (such as HS and HG, which are the earliest human PPI networks) are in the ‘unstable’ regions of ER models (Figure 3.5). Note that the size of the ‘unstable’ regions varies is not the same in different random models. For example, the ‘unstable’ region of a GEO model is much smaller than that of an ER model with the same number of nodes and edge density. Furthermore, the ER model is the simplest random model and it is well known that the structure of real PPI networks is very different from that of an ER model (e.g., [5, 6]; also discussed in Section 2.2.3).

In order to decide whether a given network has an edge density in the ‘unstable’ region, we need to quantify the size of the ‘unstable’ region for model networks. As can be seen from Figure 3.1, it is clear that the range of the ‘unstable’ region for ER models with 500 nodes is in $[0, 0.01]$, and the range of the ‘unstable’ region for GEO models with 500 nodes is in $[0, 0.005]$. From Figure 3.3 and Figure 3.1b, we can infer that the minimum GDD agreement score for GEO-vs-GEO comparisons occurs at an edge density that is proportional to k_1/n for some value of k_1 , where n is the number of nodes in the network. Furthermore, the ‘recovery’ of GDD agreement scores referred by Rito *et al.* (2010) occurs at an edge density that is proportional to k_2/n for some value of k_2 ($k_2 > k_1$) (see Figure 3.3 for an example of the scaling for GEO models with 500 nodes). This scaling is also valid for all other random models: as can be observed from Figure 3.2, the size of the

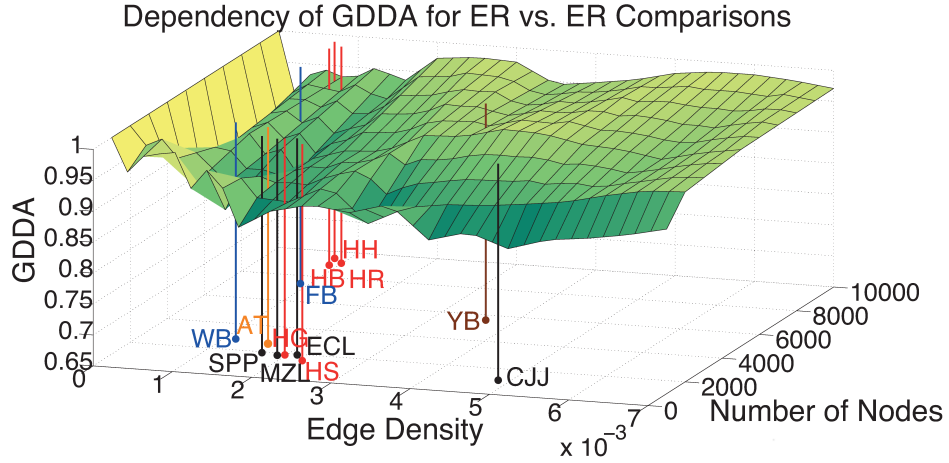


Figure 3.5.: Empirical distributions of GDD agreement for ER-vs-ER comparisons. The surface defined by 30x30 ER-vs-ER GDD agreement scores as a function of number of nodes n and edge density ρ . Each line across the surface represents the (n, ρ) of a real PPI network listed in Table 3.1. These PPI networks are grouped according to the kingdom (e.g., Bacteria, Plantae and Animalia) that the species belongs to, shown by lines with different colours.

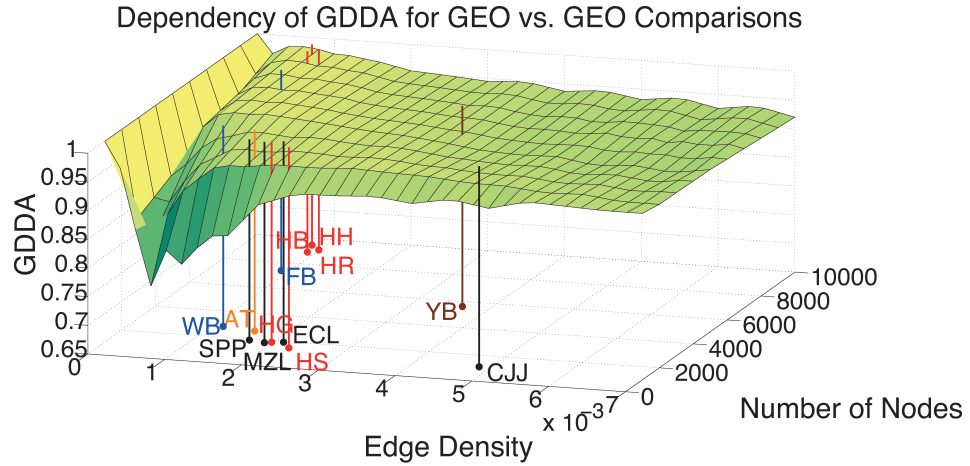


Figure 3.6.: Empirical distributions of GDD agreement for GEO-vs-GEO comparisons (see the legend of Figure 3.5 for details). As can be seen, none of the PPI networks are in the ‘unstable’ region of GEO models.

‘unstable’ region for all models is doubled when we increase the number of nodes from 50 to 100. Since the ‘unstable’ region for GEO models with 500 nodes spans the edge density region $\rho \in [0, 0.005]$ (see Figure 3.3 and Figure 3.1b), we infer that $k_2 \approx 2.5$ for GEO models. Similarly, we infer that $k_2 \approx 5$ for ER models (see Figure 3.2 and Figure 3.1a). Generally speaking, it is reasonable to hypothesise that the ‘unstable’ region spans the edge density region $\rho \in [0, k_2/n]$ where the value of k_2 depends on the network model but is likely in the range $k_2 \in [2, 5]$.

It is important to note that the width of the ‘unstable’ region asymptotically shrinks with increasing n (Figure 3.2 and Figure 3.3). As can be seen from Table 3.1, the current largest PPI network contains more than 9000 nodes. Assuming that k_2 is in the range $k_2 \in [2, 5]$, we find that most of the real-world networks listed in Table 3.1 have edge densities higher than $4/n$, and all of them have edge densities higher than $3/n$. Considering the *local* densities of real-world networks are much higher than those of uniformly dense random models (as will be discussed in Section 3.3.4), it is certain that these networks are dense enough that this ‘instability’ does not affect their analysis.

3.3.3. Effect of more up-to-date PPI data

Rito *et al.* (2010) analysed six PPI networks of yeast and human, but all of these networks are not only out-of-date now, but was also out-of-date at the time of their study (Figure 3.7). For yeast, Rito *et al.* (2010) used two PPI networks: one was the earliest Y2H data collected by Ito *et al.* (2001) [17] with only about 800 nodes and 800 edges; the other was the earliest MS-based high confidence PPI data collected by von Mering *et al.* (2002) [151] with about 1,000 nodes and 2,500 edges. For human, they used the first two Y2H studies of the human interactome: one was collected by Stelzl *et al.* (2005)[18] with only about 1,700 nodes and 3,000 edges, and the other was collected by Rual *et al.* (2005) [19] with about 3,000 nodes and 6,700 edges. Rito *et al.* (2010) also analysed PPI data obtained from BioGRID, but they split the data up into two PPI networks: one with only PPIs from Y2H experiments and the other with only PPIs from MS-based experiments. By doing this, they decreased the coverage and density of the currently available human PPI network, while increasing false negative

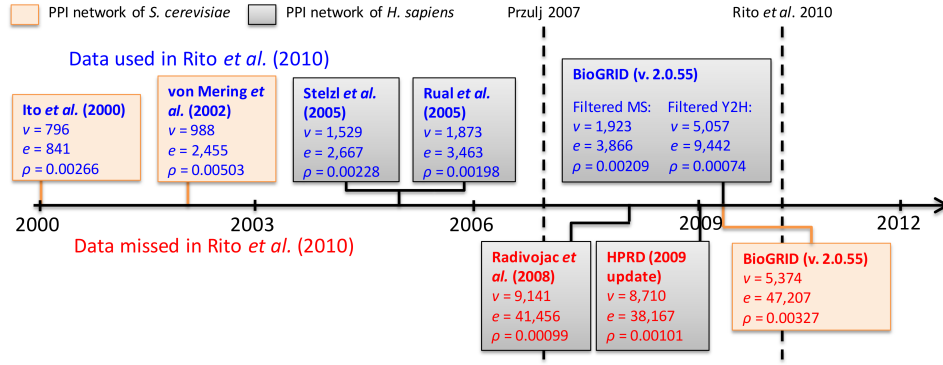


Figure 3.7.: A timeline of PPI data available since 2000. Above the timeline are PPI data used by Rito *et al.* (2010) [13]; below the timeline are data that could have been used by them but were not, resulting in lower edge densities in the PPI networks they studied.

rates of the two PPI networks they constructed. For these reasons, PPI networks analysed by Rito *et al.* (2010) have much lower edge density than recently available PPI networks. Thus their conclusion does not hold for more up-to-date PPI data.

3.3.4. Local density vs. average density

It is crucial to distinguish between *local* and *average* density. Rito *et al.* (2010) analysed random model networks with the same average density (i.e., the edge density of the whole network) of data networks. However, data networks have *local* densities much higher than a uniformly dense ER or GEO model would have [154, 5, 155]. Table 3.2 shows the edge density at which Rito *et al.* (2010) [13] expected the first graphlet to appear, along with the actual graphlet frequencies of two human PPI networks HS and HG. As can be seen, even the earliest human PPI networks have much more large-, dense-graphlets than one would expect in an ER or a GEO model with the same average densities of these PPI networks. For example, the PPI network HG contains 19 copies of graphlet G_{28} (see Figure 2.6 for all 30 graphlets with 2-5 nodes). However, this graphlet is expected to appear in an ER model with 2000 nodes until the edge density reaches 0.02495. Since HG has about 2,000 nodes and an edge density of 0.00198, this expected edge density is about 12 times higher than the actual density of HG. Even for a

GEO model, the expected edge density of one graphlet G_{28} to appear in the network is about 0.0025, which is still much higher than HG’s edge density. In addition, HG contains millions of distinct 5-node graphlets, while model networks are expected to contain only a few of them. Similar observations hold for the PPI network HS. These observations demonstrate that PPI networks are not uniformly dense, but instead contain highly dense sub-regions where the topology (and hence graphlet-based measures) is stable. This is also confirmed by a well established observation that PPI networks have higher clustering coefficients than model networks [5].

We further illustrate this effect on the human PPI network HS by finding dense sub-regions of the network. To do this, we used the Highly Connected Subgraph (HCS) algorithm proposed by Hartuv and Shamir (2000) [156] for graph partition. The largest subgraph of HS found by the HCS algorithm contains 158 nodes (that is, over 10% of the nodes in HS) and 744 edges (about 28% of the edges in HS); it has an edge density of 0.05999, which is about 26 times higher than the density of HS (Figure A.2). Assuming that k_2 is in the range $k_2 \in [2, 5]$ (discussed in Section 3.3.2), for a network of $n = 158$ nodes, the ‘unstable’ region would span the edge density region $\rho \in [0, 0.03165]$ (the worst-case, when $k_2 = 5$). A density of 0.05999 is obviously much higher than the upper bound of the ‘unstable’ region, thus GDD agreement is stable when measuring the properties of this dense subgraph.

3.3.5. Interpreting the results of non-parametric tests

Graphlet-based measures provide a way to compare two networks according to the statistical distribution of graphlets. Based on GDD agreement, Rito *et al.* (2010) [13] provided a non-parametric test for assessing the fit between random model networks and real PPI networks. Two distributions of GDD agreement scores are generated in this test: one is for model-vs-model comparisons and the other is for data-vs-model comparisons. To obtain the distribution of GDD agreement scores for model-vs-model comparisons, we generated a number of model networks with the same size and density as the PPI network, and computed GDD agreement scores across all pairs of these model networks. We also calculated GDD agreement scores between the PPI network and these random model networks to get the distribution of data-vs-model comparisons. The difference between these two distributions

	ER 1000	ER 2000	GEO 1000	GEO 2000	HG	HS
G_1	0.00008	0.00003	~ 0.0011	~ 0.0004	76807	36776
G_2	0.00182	0.00091	"	"	394	86
G_3	0.00029	0.00011	~ 0.0033	~ 0.0013	675350	296346
G_4	0.00029	0.00011	"	"	2018299	435650
G_5	0.00221	0.00111	"	"	10143	9559
G_6	0.00221	0.00111	"	"	38731	6185
G_7	0.00752	0.00432	"	"	2053	154
G_8	0.01698	0.01070	"	"	20	1
G_9	0.00059	0.00025	~ 0.0060	~ 0.0025	9056316	3438523
G_{10}	0.00059	0.00025	"	"	32102612	7462899
G_{11}	0.00059	0.00025	"	"	54733928	5067616
G_{12}	0.00261	0.00130	"	"	915264	123138
G_{13}	0.00261	0.00130	"	"	265484	43393
G_{14}	0.00261	0.00130	"	"	1433133	118173
G_{15}	0.00261	0.00130	"	"	19242	5491
G_{16}	0.00261	0.00130	"	"	1466355	740181
G_{17}	0.00702	0.00394	"	"	237849	9495
G_{18}	0.00702	0.00394	"	"	3743	274
G_{19}	0.00702	0.00394	"	"	36440	4754
G_{20}	0.00702	0.00394	"	"	54512	44519
G_{21}	0.00702	0.00394	"	"	18534	4585
G_{22}	0.01426	0.00869	"	"	17671	177
G_{23}	0.01426	0.00869	"	"	2196	148
G_{24}	0.01426	0.00869	"	"	2420	179
G_{25}	0.01426	0.00869	"	"	2162	463
G_{26}	0.02426	0.01573	"	"	409	13
G_{27}	0.02426	0.01573	"	"	103	9
G_{28}	0.03667	0.02495	"	~ 0.0025	19	0
G_{29}	0.05104	0.03609	"	"	0	0

Table 3.2.: Graphlet frequencies in model and PPI networks. To the left of the double vertical line, we reproduce parts of Tables 2 and 3 from Rito *et al.* (2010) [13]. These numbers represent the approximate edge density at which Rito *et al.* (2010) expected the first graphlet of the type specified in column 1 to appear in a model network of either ER (columns 2 and 3) or GEO (columns 4 and 5) type with 1000 (columns 2 and 4) or 2000 (columns 3 and 5) nodes. To the right of the double vertical line are the actual frequencies of each graphlet type in human PPI networks HG (column 6, $n = 1873, \rho = 0.00198$) and HS (column 7, $n = 1529, \rho = 0.00228$). The highlighted numbers in row G_{28} are discussed in the text.

can be used to evaluate the model fit. Thus if a particular PPI network is truly well-fit by a particular model, we would expect the GDD agree-

ment distribution for data-vs-model comparisons to be the same as the one for model-vs-model comparisons. This non-parametric test allows us to see whether the data network is statistically distinguishable from the model networks based on a given measure of network similarity.

However, Rito *et al.* (2010) used the amount of overlap between the two distributions (i.e., the model-vs-model GDD agreement distribution and the data-vs-model GDD agreement distribution) as a binary criterion for whether the data network is fit by the model, without considering the actual values of GDD agreement scores. We believe that this interpretation is too restrictive. For example, consider the following two cases for a given data network: (A) Model A has model-vs-model GDD agreement scores distributed normally around $99\% \pm 0.5\%$ and has data-vs-model GDD agreement scores distributed normally around $97\% \pm 0.5\%$; (B) Model B has model-vs-model GDD agreement scores distributed normally around $80\% \pm 15\%$ and has data-vs-model GDD agreement scores distributed normally around $70\% \pm 15\%$. By considering only the amount of overlap between the two distributions, the non-parametric test would suggest that Model B better fits the data than Model A, since case (B) has substantial overlap between distributions while case (A) has none. However, a GDD agreement score of 97% (case (A)) is certainly (assuming the model is derived independently of the data) better than a GDD agreement score of just 70% (case (B)). The absolute value of GDD agreement between the data network and the model network is an important criterion of the topological similarity between them. Since the models we analysed are still preliminary and purely empirical, it is possible that the model may be too narrowly defined thus the two distributions have no overlap. For example, adding a small amount of noise to Model A (to mimic noise in the data) might lead to higher data-vs-model GDD agreement scores and lower model-vs-model scores thus provide a large amount of overlap as well as a high GDD agreement score. It is not clear to us that overlap between two distributions should be the sole criterion for judging how well a model fits the data.

3.3.6. Modelling viral PPI networks

Current PPI networks are still noisy and largely incomplete (discussed in Section 2.1.1). For example, bacterial PPI networks SPP and MZL (Table

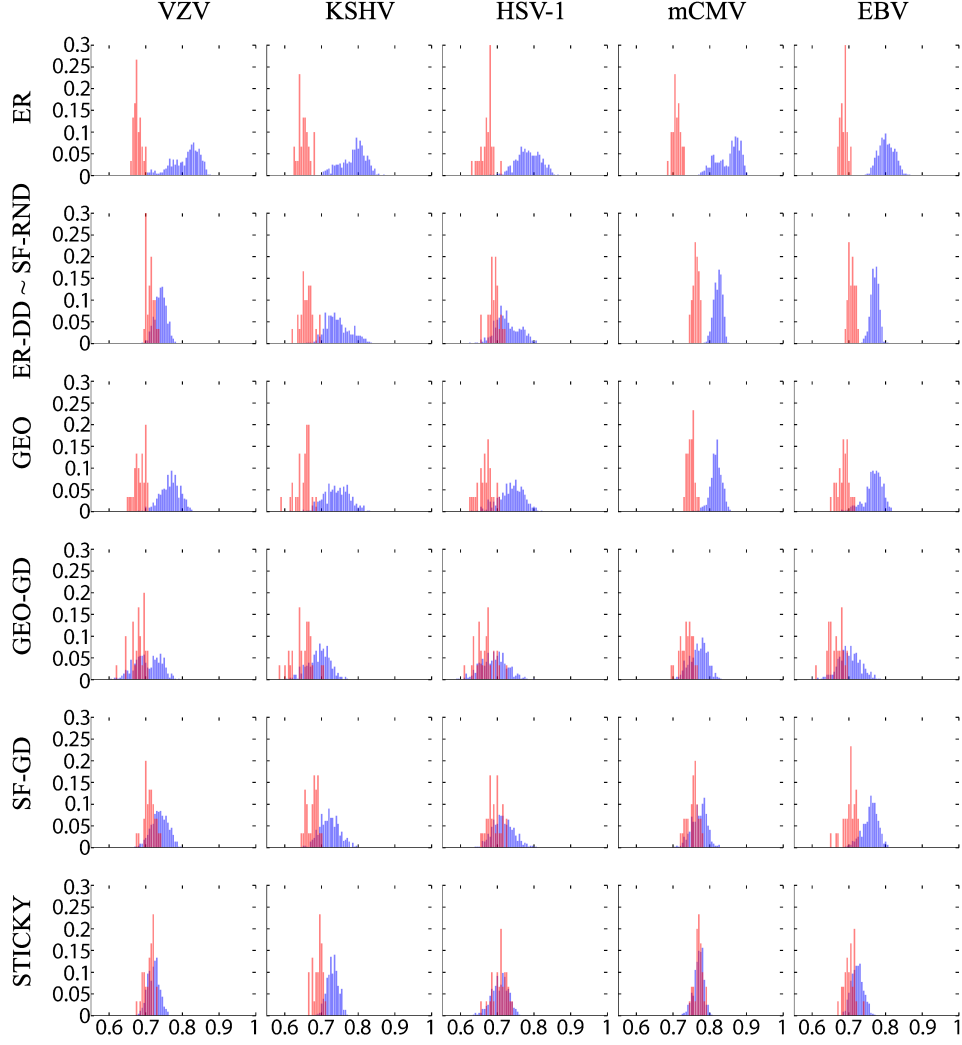


Figure 3.8.: Comparing six random network models to the PPI networks of five species of herpesvirus [26] using the non-parametric test described in Rito *et al.* (2010) [13]. Each row represents one theoretical model, and each column represents one viral PPI network. The horizontal axis is the GDD agreement score, and the vertical axis is measured probability density. In each figure, the blue bars represent a histogram of GDD agreement scores across all pairs of 30 randomly generated model networks with the same size and density as the corresponding PPI network. The red bars represent a histogram of the GDD agreement scores of the viral PPI network compared to the same 30 model networks.

	VZV	KSHV	HSV-1	mCMV	EBV	Mean	StdDev
ER	0.0000	0.0000	0.0092	0.0000	0.0000	0.0018	0.0041
SF-RND	0.2667	0.0299	0.369	0.0000	0.0000	0.1331	0.1729
GEO	0.0161	0.0368	0.1782	0.0069	0.0529	0.0581	0.0694
GEO-GD	0.3931	0.2839	0.5264	0.3586	0.3897	0.3903	0.0878
SF-GD	0.4299	0.2046	0.5000	0.5483	0.0966	0.3558	0.1958
STICKY	0.5517	0.1138	0.7207	0.7414	0.4299	0.5115	0.2564

Table 3.3.: Overlap amounts (shared area under the curve) for histograms shown in Figure 3.8.

3.1) have about a 50% false-negative rate [148, 147]; the PPI network AT covers only 2% of the interactome of *A. thaliana*. Compared with other PPI networks listed in Tabel 3.1, the five viral PPI networks (HSV-1, KSHV, VZV, EBV and mCMV) are substantially complete in the sense that all possible pairs of proteins were tested for interaction [26]. We compared these complete PPI networks to six random network models using the non-parametric test described above. Figure 3.8 depicts distributions of GDD agreement scores of data-vs-model comparisons and model-vs-model comparisons, and the overlap amounts are shown in Table 3.3. We find that the STICKY model is the best-fitting to these networks, as it has the greatest amount of overlap with four of the five viral PPI networks. The SF-GD model appears to be the second best, followed by GEO-GD, GEO, ER-DD (SF-RND) models (in that order). As expected, the ER model is the worst fitting model compared to the others since there is almost no overlap between the data-vs-model distribution and the model-vs-model distribution for the five viral PPI networks. To our knowledge, this is the first time that the structure of viral PPI networks has been modelled. Examining the biological reasons for the good fit of the STICKY model and why it is a less good fit for the PPI network KSHV is a subject of future research.

Similar plots for the other PPI networks listed in Table 3.3 can be found in Figure A.4. These plots also indicate that STICKY, SF-GD, and GEO-GD models are the best fitting models for these PPI networks.

3.4. Conclusions

In this chapter, we examined the use of graphlet-based measures for biological network comparison. By generating the empirical distributions of

GDD agreement scores, we identified the edge density regions in which the topology of model networks is ‘unstable’, and showed how graphlet-based measures correctly detect this topological instability. The ‘unstable’ regions do not affect the analysis of current PPI data, since current PPI data is dense enough to avoid these regions. We showed that data networks have high local densities thus graphlet-based measures are ‘stable’ in regions that are of interest in real networks.

We have also validated the use of GDD agreement for finding well-fitting random models for PPI networks. We showed for the first time that five viral PPI networks are well-fit by several network models (STICKY, SF-GD, and GEO-GD). We believe that this is a significant milestone in the modelling of biological networks. Though the biological significance of the fits is not immediately clear, the fact that we now have reasonably well-fitting models of viral PPI networks is a stepping stone towards understanding complex biological systems. Evaluating how and why these models are not perfect, examining why the STICKY model is often but not always the best fitting model, and improving current models, is a subject of future research.

3.5. Author’s contributions

In the work presented in this chapter, Kai Sun collected all the data, ran all the experiments and analysis, helped interpret the results, and prepared all figures.

4. Novel insight into disease-disease associations

Current knowledge of the relationship between diseases is mainly based on the similarity of clinical phenotypes. However, such knowledge could be greatly improved by considering the molecular causes and the underlying biological mechanisms of diseases. Further understanding of disease associations based on system-level biological data is expected to lead to improvements in disease diagnosis, prognosis and treatment.

In this chapter, we present three similarity measures for predicting disease associations. We take advantage of diverse biological data including disease-gene associations and a large-scale molecular network to gain novel insights into disease relationships. We analyse and compare four publicly available disease-gene association datasets, and apply three measures to estimate the similarity scores between diseases. We systematically evaluate disease associations obtained by our measures against a statistical measure of comorbidity which was derived from a large number of medical patient records. Our results show that the correlation between our similarity scores and comorbidity scores is substantially higher than expected at random, confirming that our similarity measures are able to recover comorbidity associations. We also demonstrate that our predicted disease associations correlate with disease associations generated from GWAS data significantly higher than expected at random. Furthermore, we evaluate our predicted disease associations via mining the literature on PubMed, and present case studies to demonstrate how these novel disease associations can be used to enhance our current knowledge of disease relationships.

The work presented in this chapter is in collaboration with Dr. Joana Gonçalves and Dr. Chris Larminie. The related manuscript is submitted.

4.1. Motivation

Correct diagnosis is critical for effective treatment and prevention of disease. As a result, disease classification has become a cornerstone of modern medicine. Disease may be classified by any one of a number of criteria: topographic, anatomic, pathological, physiological, etiological, juristic, epidemiological or statistical approaches. However, without considering the molecular mechanisms driving diseases, such knowledge is limited and can even be misleading. For example, a common phenotype can be caused by different underlying mechanisms, such as breast cancer, which can be divided into several subgroups that are characterized by distinct patterns of pathway activation [58]. Meanwhile, a common mechanism may lead to different phenotypes. For example, a mutation at the β -globin locus may lead to sickle-cell anemia with different phenotypes such as bony infarcts, acute chest syndrome and stroke [157].

During the past decade, a wealth of biological data has been generated from various large-scale genomic studies, prompting the scientific community to gain deeper insight into disease relationships based on their underlying biological mechanisms. Various types of biological data have been used to infer associations between diseases. One of the most commonly used biological data is disease-gene association. In a broad definition, a disease-gene association is a connection reported in the literature, which can be a genetic association (i.e., mutations in that gene may lead to that disease), or a connection inferred from other aspects. Disease-gene associations can be obtained from large-scale knowledge-bases such as the OMIM database [64]. Early studies used text mining to infer similarities between phenotypes contained in OMIM, and found those similarities were positively correlated with a number of measures of gene functions [65] and could be used to predict disease-causing genes [66]. Also by using OMIM, Goh *et al.* [67] constructed the human diseasome by connecting diseases that share a disease-causing gene. Other types of biological data such as biological pathways [57], gene expression data [81, 140], biomedical ontologies [141, 139], and GWAS data [137, 158, 138], have also been used to improve the current understanding of disease relationships from different aspects. Recently, networks have been used to model large-scale biological data, and network topology is beginning to provide insights into diseases and their associations [67, 144, 88, 42]. By

considering the interconnectivity of biomolecules in the cell, the topology of biological networks is expected to have various biological and clinical applications [14, 41].

However, early studies still have several limitations when inferring disease associations from biological data. First, some studies only considered several specific diseases, rather than giving a global comparison among all diseases. This is the case for GWAS-based studies, e.g., [137], since only a small number of GWAS studies have been completed to date in a relatively small proportion of the total disease population. Furthermore, most studies solely used OMIM as the source of disease-gene association data. OMIM is a catalogue of mendelian disorders and as a result most diseases are annotated with few genes in OMIM [159]. Limitations of using OMIM have also been discussed previously [160, 161]. Finally, most computationally predicted disease associations were not systematically evaluated due to the difficulty in identifying a suitable benchmark of known disease associations. In particular, most studies were only able to validate part of their results by comparing them with phenotypic similarities or mining the literature manually. A comparison of previous studies can be found in Table 4.1.

In the study presented in this chapter, we used diverse biological data from a number of repositories to gain novel insights into the relationship of all known human diseases by considering their underlying biological mechanisms. We used disease-gene associations obtained from four different sources to avoid the bias introduced by a single dataset. Moreover, we took advantage of the topology of a large-scale molecular network to examine its use for inferring disease associations. We applied three different disease similarity measures, namely annotation-based measure, function-based measure and topology-based measure, to estimate similarity scores between diseases. The disease associations obtained by the three measures were systematically evaluated against the ICD disease classification system, and a statistical measure of comorbidity derived from a large number of medical patient records. In addition, we evaluated our predicted disease associations by using disease associations generated from GWAS studies, which represent one of the most robust routes for identifying causal relationships between genes and diseases. To our knowledge, this is the first time that network topology has been used to infer disease-disease associations.

In the rest of this chapter, we will start with a description of the biological

Reference	Data	Size	Evaluation
van Driel <i>et al.</i> (2006) [65]	OMIM	5132 phenotypes in OMIM	Comparing results with genotypic similarities
Lage <i>et al.</i> (2007) [66]	OMIM	7000 OMIM record pairs	Evaluating results against the overlap of the OMIM record pairs
Goh <i>et al.</i> (2007) [67]	OMIM	1284 OMIM diseases	Analysing network topological properties
Huang <i>et al.</i> (2009) [137]	GWAS	7 diseases	Comparing results with phenotypic similarities
Li and Agarwal (2009) [57]	Pubmed abstracts, biological pathways	1028 diseases in MeSH	Comparing results with MeSH classification
Kim <i>et al.</i> (2009) [158]	GWAS	53 clinical traits related to severe asthma	Mining the literature manually
Hu and Agarwal (2009) [81]	Expression data	645 diseases in MeSH	Comparing results with MeSH classification
Suthram <i>et al.</i> (2010) [140]	Expression data, PPI	54 diseases	Evaluating results against genetic similarities
Lewis <i>et al.</i> (2011) [138]	GWAS	61 diseases	Comparing results with Huang <i>et al.</i> (2009) results
Mathur and Dinakarpanthian <i>et al.</i> (2007) [141]	DO annotation, GO annotation	36 diseases (for evaluation)	Evaluating results using 68 curated disease associations
Our study	Disease-gene associations, PPI, GO annotation	543 ICD-9 diseases	Evaluating results against ICD-9, comorbidity, and genetic similarities derived from GWAS data

Table 4.1.: Comparison of studies on inferring disease association. The comparison is based on the data used to derive associations (denoted by 'Data'), number of diseases evaluated (denoted by 'Size') and benchmarks used for evaluation (denoted by 'Evaluation'). The number of diseases evaluated in our study is computed as the union of diseases annotated in the four disease-gene association datasets we analysed, given in Figure 4.1.

data we analysed, followed by details of our methodology of measuring disease associations. Then we will show and discuss the evaluation of disease associations predicted by our similarity measures against known disease associations derived from ICD-9, comorbidity data and GWAS data. Finally, we will present case studies to demonstrate the ability of our similarity measures to predict novel disease associations.

4.2. Methods

4.2.1. Biological data

Three types of biological data were used in this study: PPIs, GO annotations and disease-gene associations.

PPI network

As described in Section 2.1.1, a PPI network models the physical interaction among proteins in the cell, in which a node represents a protein, and an undirected edge exists between a pair of nodes if their corresponding proteins can physically bind to each other. Currently available PPIs are mostly yielded from various high throughput proteomics experiments, such as yeast two-hybrid screening (e.g., [18]) and affinity capture mass spectrometry (e.g., [25]). We constructed a human PPI network using data obtained from BioGRID [30] version 3.1.93 (released in October 2012). All self-loops, duplicate interactions and inter species interactions were removed since we considered only simple, undirected graphs. We used the largest connected component of this PPI network for computation, which contains 11,261 nodes and 66,253 edges.

GO annotations

Genes are annotated with GO terms to represent their biological properties [162]. All GO terms are organised in three domains: cellular component, molecular function and biological process. We downloaded the ontology file and annotations of *Homo sapiens* from the Gene Ontology database¹ in November 2012. We removed annotations with evidence code ‘Inferred from

¹<http://www.geneontology.org>

Electronic Annotation' (IEAs), since IEAs are computationally inferred annotations which have not been reviewed by curators. In total, we collected 171,888 annotations between 13,166 genes and 10,787 GO terms.

Disease-gene associations.

Disease-gene associations can be modelled as a graph containing both known human diseases and disease-related genes in the human genome. The degree of a disease is the number of genes associated with that disease, while the degree of a gene is the number of diseases annotated with that gene. We used four disease-gene association datasets obtained from different sources: OMIM, CTD [68], Functional Disease Ontology annotations (FunDO) [163] and Human Genome Epidemiology Network (HuGENet) [164]. Among these datasets, OMIM, CTD, and FunDO contain curated associations, while HuGENet contains computationally inferred associations. Details of these disease-gene association datasets are described below.

- OMIM is considered to be the best curated resource of phenotype-genotype relationships, and it has been used in various disease-related studies (discussed in Section 4.1). We downloaded the OMIM database in November 2012. In total, it contains 3,537 diseases (annotated by OMIM IDs), 2,862 genes and 4,337 disease-gene associations.
- CTD provides scientific data describing relationships between chemicals, genes, and human diseases, with the goal of improving the understanding of environmental chemicals' effects on human health. It contains both curated and inferred disease-gene associations, but we only used curated associations as they have higher confidence than inferred associations. Disease-gene associations directly derived from OMIM were excluded to reduce the dependency between datasets. We downloaded the data from CTD in November 2012 and obtained 17,754 associations between 2,761 diseases (annotated by MeSH terms²) and 5,828 genes.
- FunDO contains disease-gene associations extracted from the NCBI Gene Reference Into Function (GeneRIF) database. A GeneRIF is a brief statement about the function of a gene, along with information

²<http://www.nlm.nih.gov/mesh/>

of its association with diseases. We downloaded the latest stable version of FunDO (released in October 2008) and obtained 1,854 diseases (annotated by Disease Ontology (DO) terms), 4,781 genes and 28,442 disease-gene associations.

- HuGENet is known as an integrated knowledge-base on human genome epidemiology. The Phenopedia collection of HuGENet contains disease-gene associations obtained by text-mining of abstracts on PubMed using machine learning techniques. Disease-gene association data were downloaded via HuGE Navigator in September 2012. We obtained 353,883 associations between 2,387 diseases (annotated by Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs)³) and 11,915 genes.

Since disease names or IDs used in these datasets are based on different labelling schemes, we mapped all disease names or IDs to ICD-9 codes, for the purpose of comparing these datasets and further evaluation (also see Section 4.3.2 for details). We used the mapping manually constructed by Goh *et al.* (2007) [67] and Park *et al.* (2009) [133] to convert OMIM IDs to ICD-9 codes, and used the corresponding mapping provided in DO version 3 (the latest stable version of DO, released in May 2007) to map DO IDs, MeSH terms and UMLS CUIs to ICD-9 codes. In total, 1,467 OMIM IDs in OMIM, 423 MeSH terms in CTD, 806 DO IDs in FunDO and 693 UMLS CUIs in HuGENet were mapped to ICD-9 codes.

4.2.2. Disease similarity measures

We applied three similarity measures to estimate similarity scores between diseases. These measures include standard methods (e.g., Jaccard index) and novel measures proposed in this study (e.g., graphlet-based measure). Considering the information used in calculation, the similarity score of a pair of diseases was measured in three different ways: annotation-based, function-based and topology-based.

³<http://www.nlm.nih.gov/research/umls/>

Annotation-based measure.

The annotation-based measure solely used the information obtained from disease-gene association data. We applied the Jaccard index, which is known as a standard method for comparing the similarity between two sets, to estimate the similarity score between diseases as follows. Let G_{D_i} be the set of genes associated with a disease D_i . We computed the annotation-based similarity score of two diseases D_i and D_j as the Jaccard index (or Jaccard similarity coefficient) of G_{D_i} and G_{D_j} :

$$Sim_{annotation}(D_i, D_j) = \frac{|G_{D_i} \cap G_{D_j}|}{|G_{D_i} \cup G_{D_j}|}. \quad (4.1)$$

Note that the F-measure, which corresponds to the Sørensen-Dice coefficient [165] in set theory, can also be applied to measure the similarity between G_{D_i} and G_{D_j} . Since it has been shown that the Sørensen-Dice coefficient is monotonically related to the Jaccard index [166], the choice between the two measures does not affect our results.

Function-based measure.

The function-based similarity measure used both GO term annotations and disease-gene associations to estimate the similarity score between a pair of diseases. For each disease D_i annotated in a specific disease-gene association dataset, we first identified the set of GO terms that were overrepresented within G_{D_i} , denoted by GO_{D_i} . The statistical significance (p -value) of the enrichment of a GO term was computed according to the hypergeometric distribution for sampling without replacement, and was corrected for multiple testing using the Benjamini-Hochberg test [167]. We used Benjamini-Hochberg correction rather than Bonferroni correction to control the false discovery rate since Bonferroni correction has been considered to be too conservative if the number of tests is large [168]. Therefore, only overrepresented GO terms from the ‘biological process’ domain of GO and having a p -value less than 0.05 were considered to be in GO_{D_i} . Then for a pair of diseases D_i and D_j , we computed the Jaccard index of GO_{D_i} and GO_{D_j} as their function-based similarity score, defined as:

$$Sim_{function}(D_i, D_j) = \frac{|GO_{D_i} \cap GO_{D_j}|}{|GO_{D_i} \cup GO_{D_j}|}. \quad (4.2)$$

Topology-based measure.

The topology of PPI networks has yielded insights into biological function and disease (e.g., [46]). Topological similarities of proteins in a PPI network are considered as a complementary information to sequence similarities [39]. Thus in this study, we took advantage of the topology of the human PPI network along with disease-gene association data to examine the use of network topology for uncovering novel disease associations. We hypothesised that the underlying biological mechanisms of diseases would be reflected by the topological property of the human PPI network. In particular, we proposed a measure to estimate the similarity score between a pair of diseases based on the topological similarity of their annotated genes.

We applied a graphlet-based method to assess the topological similarity of genes in the human PPI network. As described in Section 2.2.4, a *graphlet* is defined as a small, connected and induced subgraph of a larger network [5]. Topologically identical nodes within a graphlet are said to belong to the same *automorphism orbit* [6]. The *graphlet signature* of a node u is a 73-dimensional vector, whose i^{th} element u_i counts the number of times the node u is touched by the particular automorphism orbit i [46]. According to Milenković and Pržulj (2008) [46], the signature similarity of a pair of nodes u and v is defined as:

$$SigSim(u, v) = 1 - \frac{1}{\sum_{i=0}^{72} w_i} \left(\sum_{i=0}^{72} (w_i \times \frac{|\log(u_i + 1) - \log(v_i + 1)|}{\log(\max\{u_i, v_i\} + 2)}) \right) \quad (4.3)$$

where w_i is a weight assigned to orbit i defined as $1 - \log(o_i)/\log(73)$ (o_i is the dependency count of orbit i , see [46] as well as Section 2.2.4 for details).

We calculated the signature similarity of each pair of genes in the human PPI network. Note that the network has an edge density (the proportion of the number of edges to the maximum possible number of edges) of 0.001, which for its size (11,261 nodes and 66,253 edges) is dense enough to avoid low edge density regions in which the topology of networks is unstable (see [106] and Chapter 3 for details). We extended the use of graphlet-based method to measure disease similarities. We introduced two terms to quan-

tify the topology-based similarity score between diseases D_i and D_j . The first term, denoted by $AllSig$, is the maximum of the signature similarity between a gene in G_{D_i} and a gene in G_{D_j} :

$$AllSig(D_i, D_j) = \max_{\substack{g_m \in G_{D_i} \\ g_n \in G_{D_j}}} SigSim(g_m, g_n). \quad (4.4)$$

The second term, denoted by $ShareSig$, focuses on the topological similarity between genes shared between both diseases:

$$ShareSig(D_i, D_j) = \max_{\substack{g_m \neq g_n \\ g_m \in G_{D_i} \cap G_{D_j} \\ g_n \in G_{D_i} \cap G_{D_j}}} SigSim(g_m, g_n). \quad (4.5)$$

Finally we defined the topology-based similarity score between D_i and D_j as the average of these two terms:

$$Sim_{topology}(D_i, D_j) = \frac{1}{2} \times (ShareSig(D_i, D_j) + AllSig(D_i, D_j)). \quad (4.6)$$

4.2.3. Evaluation

Comorbidity associations of diseases.

The availability of electronic patient records facilitates studies into disease comorbidity, which indicates the potential for co-occurrence of two given diseases in the same individual. Comorbidity can be considered as a type of disease-disease association derived from electronic medical records, but the underlying drivers for comorbidity may be very different from one another. Comorbidity and its correlation with other types of disease associations such as genetic associations [145] and evolutionary associations [134] have previously been studied. Unlike these studies, we used comorbidity data to evaluate disease associations predicted by our similarity measures. Comorbidity associations were downloaded from the Human Disease Network (HuDiNe, [132]), which were obtained from the disease history of 32 million American patients. Diseases were annotated using ICD-9 codes in HuDiNe, and as many diseases in patient records were not specific enough to map to 4-digit or 5-digit codes, we used the comorbidity data annotated using 3-digit level ICD-9 codes for our analysis. The strength of comorbid-

ity association between a pair of diseases can be measured by the Relative Risk and ϕ -correlation [132]. Because comorbidity associations quantified by ϕ -correlation were reported to contain more connections across different ICD-9 categories [132], we chose ϕ -correlation as the measure of comoridity. The ϕ -correlation score between D_i and D_j was defined as the Pearson's correlation for binary variables, given by:

$$\phi(D_i, D_j) = \frac{C_{ij}N - P_iP_j}{\sqrt{P_iP_j(N - P_i)(N - P_j)}} \quad (4.7)$$

where C_{ij} is the number of individuals affected by both D_i and D_j , N is the total number of individuals in the population, P_i and P_j are the prevalences of D_i and D_j respectively. A ϕ -correlation higher than 0 indicates the co-occurrence of D_i and D_j is more frequently than expected by random. The statistical significance of ϕ -correlation was determined by using a t -test,

$$t = \frac{\phi\sqrt{n-2}}{\sqrt{1-\phi^2}} \quad (4.8)$$

where $n = \max(P_i, P_j)$ is the number of observations used to calculate ϕ . We used significant associations at 5% level ($t \geq 1.96$) for our analyses.

GWAS data.

GWAS is a powerful method for identifying genetic variations associated with diseases and is one of the most robust routes for identifying causal relationships between genes and diseases [70, 72]. As described in Section 2.1.3, GWAS studies examine the genome for SNPs that occur more frequently in people with a particular disease than in people without it. GWAS studies have enabled exploration of gene association in complex diseases in a systematic way on a genome scale. Whilst individual studies are extremely powerful, only a small number of diseases have been studied thus far using GWAS. Hence the GWAS database as a whole is only able to contribute a relatively small component to the overall knowledge base of general disease-gene associations. For this reason, we did not use GWAS data as a source of disease-gene association to measure disease similarity scores, but used them to evaluate our predicted disease associations. Note that for simplicity, we assume independency of SNPs in our analysis. It is possible that

some of the SNPs are highly correlated, i.e., in high linkage disequilibrium (LD). Despite several methods having been proposed to account for LD in GWAS data, such methods may increase the risk of including false positives [169]. In addition, our current knowledge of the dependencies of SNPs is still limited. For example, the interrelationships between driver SNPs (mutations that drive the development and progress of cancer) and passenger SNPs (mutations that occur as cancer progresses but have no effect on the disease) are still poorly understood [170]. For these reasons, although the dependencies between SNPs are typically present in GWAS data, we did not take them into account in our study.

We downloaded GWAS data from the NHGRI GWAS catalog [73] in May 2013. This resource collects significant associations between traits (or diseases) and SNPs from the literature. Similar to Sanseau *et al.* (2012) [84], we only considered highly confident associations with p -value lower than 10^{-7} . We also eliminated not replicated associations to minimize false-positives. For all disease-SNP associations, we used the corresponding disease-gene associations reported by the authors in the original publications in our analysis. After mapping diseases to ICD-9 codes, we obtained 1,756 genetic associations (from 478 publications) between 126 diseases and 1,298 genes.

4.3. Results and discussion

4.3.1. Comparison of disease-gene association datasets

We analysed four different disease-gene association datasets: three curated datasets, namely OMIM, CTD and FunDO, and one computationally predicted dataset, HuGENet. Although these datasets focus on different aspects of the connections between diseases and genes, they are not fully independent since information contained in these datasets is extracted from the literature. For example, disease-gene associations contained in CTD and FunDO were extracted from 9,269 and 48,436 publications respectively, and they have 799 publications in common. We mapped all disease names or IDs annotated in these datasets to ICD-9 codes for a correct comparison (see Section 4.2.1). If several diseases were mapped to a common ICD-9 code, we assigned the union of genes associated with those diseases to that ICD-9

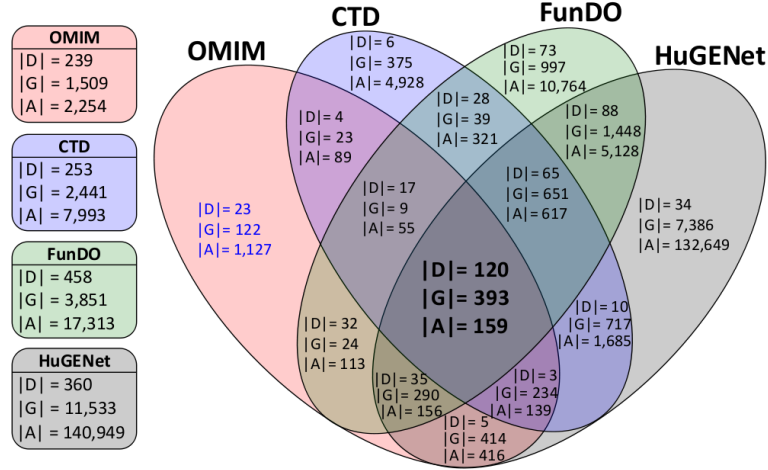


Figure 4.1.: The overlap of diseases (denoted by ‘D’), genes (denoted by ‘G’) and their associations (denoted by ‘A’) between the four disease-gene association datasets we analysed. Each ellipse in the Venn diagram represents a disease-gene association dataset. The interior of an ellipse symbolically represents the elements (i.e., diseases, genes and associations) annotated in the corresponding dataset, while the exterior represents elements that are not annotated in that dataset. For example, 23 diseases, 122 genes and 1127 disease-gene associations annotated in OMIM are not annotated in the other three datasets, as shown by blue numbers in the Venn diagram. Boxes on the left list the sizes of the four datasets. The size of the intersection of the datasets is marked in bold.

code. In order to evaluate our measures using comorbidity data, we further limited the ICD-9 codes to 3-digit level. We are aware that noise may be introduced when merging diseases into 3-digit level. Generally speaking, a 3-digit level ICD-9 code is always associated with more than one disease, thus the average degree of diseases increased after mapping.

Interestingly, the overlap among the four disease-gene association datasets is unexpectedly small, as shown in Figure 4.1. While a considerable number of diseases (120 diseases in total, that is, 50.21%, 47.43%, 26.20% and 33.33% of diseases annotated in OMIM, CTD, FunDO and HuGENet, respectively) have gene annotations in all four datasets, few disease-gene associations (159 associations in total, that is, 7.05%, 1.99%, 0.92% and 0.11% of associations in OMIM, CTD, FunDO and HuGENet, respectively) can

be found in all datasets. Figure A.7 further demonstrates the difference between these datasets according to the degree distribution of diseases. In general, these distributions follow power law distributions, indicating that most human diseases are associated with only a few disease genes, while a small number of diseases relate to many genes. However, this scale-free topology may also be an artifact of sampling: several diseases are better studied than others [52]. We notice that in OMIM, most diseases are associated with fewer genes compared with other datasets. The average number of genes associated with a disease in OMIM is 9.43, while in the two other curated datasets CTD and FunDO, these numbers are 31.59 and 37.80. On the other hand, on average a disease in HuGENet is annotated with more than 300 genes: HuGENet has a higher false positive rate compared to other datasets, since its associations were derived from computational predictions rather than manual curations.

The difference and inconsistency discussed above indicate that currently available disease-gene association datasets are still noisy and incomplete. The incompleteness may be due to the focus of the datasets and the nature of the curation process. For example, OMIM mainly focuses on mendelian diseases and traits. Meanwhile, many false positives may be introduced by text-mining the literature (e.g., HuGENet). However, there is no single standard and systematic method to assess the quality of these data. Therefore, to gain a more comprehensive view of human diseases and to test the robustness of our methods, we used all four disease-gene association datasets along with the intersection/union of the three curated datasets in further computation and evaluation.

4.3.2. Evaluation of similarity measures

Correlation with ICD-9.

The results obtained by these measures were first evaluated against the standard disease classification system ICD-9. We say that two diseases are associated according to ICD-9, if they are classified under the same ICD-9 category⁴. For example, diabetes mellitus (ICD-9 code: 250) and thyroiditis (ICD-9 code: 245) are classified under the same category ‘endocrine, nutritional and metabolic diseases, and immunity disorders’. To investigate

⁴<http://www.icd9data.com/2013/Volume1/default.htm>

Data	Group	Annotation-based	Function-based	Topology-based
OMIM	Same	0.0114 ± 0.0665	0.0355 ± 0.0892	0.4349 ± 0.1101
	Different	0.0010 ± 0.0139	0.0118 ± 0.0314	0.3996 ± 0.0760
	P-value	1.2785×10^{-13}	1.0423×10^{-52}	2.1257×10^{-54}
CTD	Same	0.0361 ± 0.1590	0.0728 ± 0.1754	0.4863 ± 0.1770
	Different	0.0050 ± 0.0274	0.0333 ± 0.0662	0.4408 ± 0.1368
	P-value	1.4887×10^{-23}	1.4040×10^{-9}	2.0240×10^{-25}
FunDO	Same	0.0418 ± 0.1344	0.0991 ± 0.1611	0.5560 ± 0.2214
	Different	0.0100 ± 0.0262	0.0549 ± 0.0830	0.4952 ± 0.1636
	P-value	1.7609×10^{-144}	9.6708×10^{-100}	2.7037×10^{-90}
HuGENet	Same	0.0931 ± 0.1798	0.2470 ± 0.2123	0.8031 ± 0.2248
	Different	0.0438 ± 0.0566	0.1881 ± 0.1522	0.7837 ± 0.2292
	P-value	1.4585×10^{-74}	9.9053×10^{-72}	4.5910×10^{-14}
Intersection	Same	0.0338 ± 0.1511	0.0593 ± 0.1907	0.3826 ± 0.1131
	Different	0.0024 ± 0.0329	0.0089 ± 0.0428	0.3496 ± 0.1020
	P-value	2.2667×10^{-2}	2.7448×10^{-4}	5.4716×10^{-4}
Union	Same	0.0350 ± 0.1179	0.0963 ± 0.1463	0.5680 ± 0.2226
	Different	0.0085 ± 0.0219	0.0583 ± 0.0818	0.5042 ± 0.1716
	P-value	1.3493×10^{-211}	7.1478×10^{-113}	4.1709×10^{-141}

Table 4.2.: Evaluation of our measures against ICD-9 classification. Numbers in the table are similarity scores between diseases from the same ICD-9 categories, compared with those from different ICD-9 categories. P -values are calculated by using the Mann–Whitney U test.

the correlation between our similarity measures and the ICD-9 classification, we tested whether a pair of diseases from the same ICD-9 category tends to have a higher similarity score than diseases from different ICD-9 categories (Table 4.2). Since similarity scores obtained by our measures are not normally distributed, we used a non-parametric test, namely the Mann–Whitney U test, to assess the statistical significance (p -value). Our results show that for all three similarity measures and all four disease-gene association datasets, similarity scores of diseases from the same ICD-9 category are significantly higher than those from different ICD-9 categories.

Correlation with comorbidity.

As the goal of our study is to uncover novel disease associations that may reflect common underlying mechanisms, we are more interested in the associations between diseases that belong to different ICD-9 categories. For this reason, we systematically evaluated our similarity measures against a statis-

tical measure of comorbidity. We say two diseases are associated according to comorbidity if they are reported to have a significant co-occurrence in the same individual. In particular, their ϕ -correlation score should be higher than a chosen threshold and statistically significant at 5% level. Figure A.8 shows the distribution of ϕ -correlation scores for all pairs of diseases we analysed.

To assess the ability of our measures to uncover highly confident comorbidity associations, we used Receiver Operating Characteristic (ROC) curves, in which we plotted the *True Positive Rate* (TPR, also known as *sensitivity*) versus the *False Positive Rate* (FPR, also known as $1 - \text{specificity}$) for different thresholds of similarity score. TPR is defined as the fraction of true positives (that is, all pairs of diseases having a similarity score higher than a chosen threshold and having comorbidity association) out of the positives (all pairs of diseases having comorbidity association), while FPR is defined as the fraction of false positives (all pairs of diseases having a similarity score higher than a chosen threshold but having no comorbidity association) out of the negatives (all pairs of diseases excluding those having comorbidity association). Figure 4.2, Figure A.5, Figure A.6 and Table 4.3 show the ROC curves and Area Under Curve (AUC) values obtained by the three disease similarity measures. The ϕ -correlation threshold was set to 0.06 (the same threshold was used by Hidalgo *et al.* (2009) [132]). For each combination of disease-gene association data and disease similarity measure, we evaluated diseases annotated with at least 1, 3, 5, 7, 10, 15 genes, shown by curves with different colours in each plot. For the intersection set, only two curves were shown in each plot since there were no comorbidity relationships between diseases annotated with more than 5 genes. To illustrate that our results cannot be obtained by chance, we assigned a randomised score which was drawn from the same distribution of the similarity scores to each pair of diseases, and evaluated associations derived from these randomised scores against comorbidity. We show that the correlation between our similarity measures and comorbidity scores is substantially higher than expected at random for all disease-gene association datasets we analysed. In particular, diseases yielding a high similarity score are very likely to have comorbidity associations, thus confirming that our measures are able to uncover known comorbidity relationships.

While varying the ϕ -correlation threshold, we obtained higher AUC values

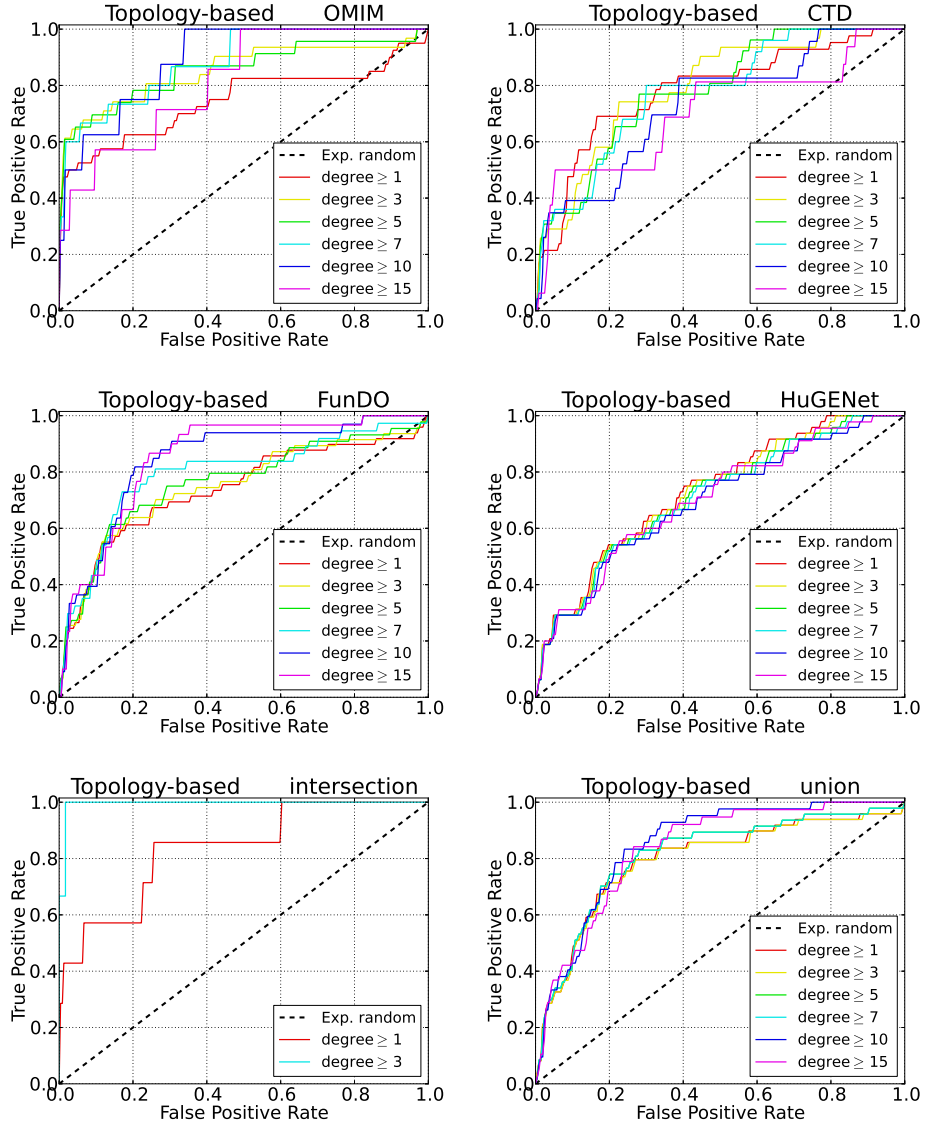


Figure 4.2.: ROC curves obtained by evaluating the topology-based measure against comorbidity. The ϕ -correlation threshold was set to 0.06. For each disease-gene association dataset, we evaluated diseases annotated with at least 1, 3, 5, 7, 10, 15 genes, shown by curves with different colours in each plot. For the intersection set, only two curves were shown in each plot since there were no comorbidity associations between diseases annotated with more than 5 genes.

Data	Annotation-based	Function-based	Topology-based
OMIM	0.8009 ± 0.0277 (0.57)	0.8694 ± 0.0073 (0.51)	0.8495 ± 0.0011 (0.50)
CTD	0.7849 ± 0.0164 (0.54)	0.7316 ± 0.0046 (0.50)	0.7949 ± 0.0042 (0.52)
FunDO	0.7426 ± 0.0088 (0.47)	0.7142 ± 0.0017 (0.49)	0.7497 ± 0.0016 (0.50)
HuGENet	0.7563 ± 0.0001 (0.51)	0.8185 ± 0.0001 (0.50)	0.7153 ± 0.0015 (0.49)
Intersection	0.9925 ± 0.0001 (0.60)	0.9802 ± 0.0001 (0.51)	0.9958 ± 0.0041 (0.47)
Union	0.8225 ± 0.0045 (0.47)	0.7491 ± 0.0001 (0.50)	0.7939 ± 0.0022 (0.50)
Average	0.8194 ± 0.0837 (0.53)	0.8106 ± 0.0930 (0.50)	0.8163 ± 0.0907 (0.50)

Table 4.3.: Evaluation of our measures against comorbidity. Numbers in the table are AUC values obtained by evaluating the three disease similarity measures against comorbidity associations. The ϕ -correlation threshold was set to 0.06, and all diseases annotated with least 3 genes were evaluated. Average AUC values obtained by using randomised scores are shown by numbers in brackets. Each evaluation test was run 30 times to compute the statistics reported in the table.

for higher thresholds. For example, when the ϕ -correlation threshold was set to 0.06, the AUC value was 0.7580 ± 0.0024 (using the topology-based measure and FunDO as the source of disease-gene associations). When the ϕ -correlation threshold was set to 0.08 and 0.10, the AUC value increased to 0.7669 ± 0.0027 and 0.7996 ± 0.0060 , respectively. This indicates our similarity measures tend to detect strong comorbidity associations with high ϕ -correlation. Meanwhile, when we decreased the number of false negatives in the comorbidity data by lowering the ϕ -correlation threshold from 0.06 to 0.02, the AUC values we obtained were still higher than expected at random. For example, when the ϕ -correlation threshold was set to 0.04 and 0.02 (namely increasing 0.90-fold and 5.41-fold the amount of comorbidity associations), the AUC values we obtained were 0.7064 ± 0.0019 and 0.6017 ± 0.0015 , respectively. These results suggest our similarity measures are robust to high false negatives in the comorbidity data. Better ROC curves can also be obtained by evaluating diseases annotated with higher numbers of genes (Figure 4.2, Figure A.5, Figure A.6). From Table 4.3, we observed that best performances of our similarity measures are achieved by using highly confident curated disease-gene associations (i. e., the intersection set of OMIM, CTD and FunDO), with AUC values higher than 0.98.

Data	Annotation-based	Function-based	Topology-based
F/G	0.7224 ± 0.0010 (0.49)	0.6781 ± 0.0001 (0.50)	0.6863 ± 0.0009 (0.50)
Common	0.7527 ± 0.0010 (0.49)	0.7147 ± 0.0001 (0.50)	0.7555 ± 0.0020 (0.40)

Table 4.4.: Evaluation of our measures against GWAS. Only diseases annotated with least 3 genes were evaluated. ‘F/G’ are diseases having associated genes in both FunDO and GWAS data (99 diseases in total). ‘Common’ are diseases having associated genes in all four disease-gene association datasets (given in Figure 4.1) and GWAS data (50 diseases in total). Average AUC values obtained by using randomised scores are shown by numbers in brackets. Each evaluation test was run 30 times to compute the statistics reported in the table.

Correlation with GWAS data.

We further examined the correlation between our predicted disease associations and currently available highly confident GWAS data (see the Methods section for details) to see whether our findings are supported by GWAS studies. A gene is said to be associated with a disease according to GWAS, if the occurrence of genetic variants (SNPs) within that gene is significantly higher in people with that disease than in people without it. We say that two diseases are associated according to GWAS if they share at least one gene in GWAS data. Since disease-gene associations collected in the four datasets we analysed were extracted from the literature, genetic associations reported in GWAS studies may also be collected in these datasets. To avoid bias in evaluation, we chose FunDO as the source of disease-genes associations, as it has few overlaps with GWAS data. In particular, since most GWAS data were published after FunDO’s last stable release (October 2008), only 42 out of 48,436 publications in FunDO were also found in GWAS data. We removed disease-gene associations collected from the common 42 publications before computing similarity scores between diseases using FunDO. Similar to our evaluation against comorbidity, we used ROC curve analysis to assess the ability of our similarity measures to recover disease associations derived from GWAS (Table 4.4). For each of the three measures, we found that the correlation between our similarity measures and GWAS data is substantially higher than expected at random. This result further confirms the validity of our methods.

4.3.3. Comparison of similarity measures

The three similarity measures, namely annotation-based measure, function-based measure, and topology-based measure, use different biological information to predict disease associations. For a pair of diseases, the annotation-based measure estimates their similarity score based on the overlap of their annotated genes, while the function-based measure estimates their similarity score based on the overlap of their associated biological functions derived from GO annotations. The topology-based measure makes use of the topology information derived from the underlying PPI network, and estimates disease similarity scores based on the topological similarity of their annotated genes. Based on our evaluation, the three similarity measures perform well in recovering known disease associations. Note that though the AUC values obtained by the three measures are similar, the predictions can differ from each other. Figure A.9 shows the overlap of disease associations predicted by the three measures. When considering the top 5% of the most associated disease pairs as our predicted disease associations, 14% ~ 38% of the predictions that are supported by all three similarity measures.

Our similarity measures are sensitive to the noise in disease-gene association data. We notice that prediction performances of our similarity measures generally decrease with the increase of noise level, thus using the intersection of curated disease-gene association datasets results in the best performance when predicting comorbidity associations (Table 4.3). Both the annotation-based measure and the topology-based measure have better performances when using curated disease-gene associations (i.e., OMIM, CTD and FunDO) than computationally predicted associations (i.e., HuGENet). However, the function-based measure obtains lower AUC values for curated datasets CTD and FunDO than the two other similarity measures, but higher AUC values for HuGENet. In this regard, the function-based measure may be more appropriate for analysing predicted datasets, while the annotation-based measure and topology-based measure may be more appropriate for analysing curated datasets.

The annotation-based measure is straightforward, but has relatively good performance according to our evaluation. However, as it only uses disease-gene associations to estimate similarity scores, for a pair of diseases sharing few genes, their annotation-based similarity score may be low, even if their

annotated genes are closely related. In particular, the annotation-based measure is highly affected by the occurrence of pleiotropic genes (genes that cause multiple phenotypes) in the dataset. We obtained the list of 802 pleiotropic genes from the OMIM Morbidmap by identifying genes that associated with more than one disease (a similar approach was used in [171]). To examine the influence of pleiotropic genes on our measures, we excluded these genes from OMIM and evaluated the performances of our similarity measures against comorbidity. Not surprisingly, the AUC value of annotation-based measure dropped to 0.5811, close to expected at random (0.5740, see Table 4.3). On the other hand, since both the function-based measure and the topology-based measure use additional data sources (GO annotations or network topology) to estimate similarity scores, they are less affected by pleiotropic genes. AUC values obtained by the function-based measure and the topology-based measure dropped to 0.7816 and 0.7199 respectively, after removing pleiotropic genes from OMIM. These results show the contribution of similarities between specific genes (genes associated with only one disease) to the prediction performances of our similarity measures.

Since disease-gene association datasets were obtained by different research groups and approaches, good performances for all datasets confirm the robustness of our similarity measures in predicting disease associations. In addition, the topology-based measure is also robust to the noise and incompleteness presented in PPI networks. We evaluated this by using PPI data obtained from different releases of BioGRID database (Table A.3). Generally speaking, the performance of the topology-based measure slightly decreases when using early PPI networks (Table A.4). However, AUC values obtained by using these early PPI networks are still substantially higher than expected at random. This result indicates that the ability of the topology-based measure to predict disease associations is independent of the noise and incompleteness level of PPI networks.

4.3.4. Case studies

To demonstrate how our similarity measures can be used for uncovering novel disease associations, we present a case study for diabetes mellitus (DM, ICD-9 code: 250). DM is a metabolic disease that affects the body's ability to produce or use insulin, a hormone for regulating carbohydrates.

It causes hyperglycemia and may lead to severe consequences such as brain damage, amputations and heart disease [172]. Table 4.5 lists the top 30 diseases associated with DM using the topology-based measure and FunDO as the source of disease-gene associations.

Among these 30 diseases, we found 6 of them are classified under the same ICD-9 catalogue (namely ‘Endocrine, nutritional and metabolic diseases, and immunity disorders’) with DM, e.g., ovarian dysfunction (ICD-9 code: 256). Meanwhile, the list also includes 5 diseases that have highly confident comorbidity associations with DM, e.g., essential hypertension (ϕ -correlation: 0.1275). Moreover, 10 diseases are associated with DM according to GWAS data, e.g., rheumatoid arthritis (number of shared genes: 8). Apart from these diseases known to be associated with DM (16 in total, as there are associations supported by multiple evidence), the remaining 14 diseases listed in the table are considered as novel associations predicted by our measure. We evaluated these associations via text mining the literature on PubMed⁵. We are able to confirm all of these associations, including surprising associations such as DM and ‘other cerebral degenerations’ (ICD-9 code: 331). This result highlights the power of our approaches to identify novel associations between diseases. Further exploration of potential underlying mechanisms shared by these diseases may lead to improvement in disease diagnosis, prognosis and treatment.

Table A.3 shows another case study for Parkinson’s disease (ICD-9 code: 332).

4.4. Conclusions

In this chapter, we gained novel insights into the relationship between human diseases by considering their molecular causes and underlying physical interactions. We used information derived from latest biological data, including disease-gene associations, gene functions and the topology of the human PPI network in our analysis. We applied three different measures to estimate the similarity score of diseases, and these measures were systematically evaluated against ICD-9 classification system, a statistical measure of comorbidity and GWAS data. Our results showed the correlation between associations predicted by our measures and known disease associations, and

⁵<http://www.ncbi.nlm.nih.gov/pubmed>

Rank	Code	Disease name	Reference
1	239	Neoplasms of unspecified nature	PMID: 23639840
2	155	Malignant neoplasm of liver and intra-hepatic bile ducts	GWAS
3	710	Diffuse diseases of connective tissue	GWAS
4	714	Rheumatoid arthritis and other inflammatory polyarthropathies	GWAS
5	256	Ovarian dysfunction	ICD-9, GWAS
5	278	Overweight, obesity and other hyperalimentation	ICD-9, comorbidity, GWAS
7	401	Essential hypertension	Comorbidity
8	295	Schizophrenic disorders	PMID: 17474808
9	282	Hereditary hemolytic anemias	GWAS
10	289	Other diseases of blood and blood-forming organs	PMID: 11727971
11	642	Hypertension complicating pregnancy childbirth and the puerperium	PMID: 18558027
12	365	Glaucoma	GWAS
13	135	Sarcoidosis	PMID: 23075651
13	414	Other forms of chronic ischemic heart disease	Comorbidity, GWAS
15	331	Other cerebral degenerations	PMID: 20837967
15	332	Parkinson's disease	PMID: 23335160
17	244	Acquired hypothyroidism	ICD-9
18	335	Anterior horn cell disease	PMID: 22017321
19	362	Other retinal disorders	Comorbidity
20	753	Congenital anomalies of urinary system	PMID: 22260488
21	277	Other and unspecified disorders of metabolism	ICD-9
22	286	Coagulation defects	PMID: 22460041
23	042	Human immunodeficiency virus [HIV] disease	PMID: 19419710
24	340	Multiple sclerosis	GWAS
24	579	Intestinal malabsorption	GWAS
26	272	Disorders of lipid metabolism	ICD-9
27	577	Diseases of pancreas	PMID: 22996690
28	287	Purpura and other hemorrhagic conditions	PMID: 21092704
28	290	Dementias	PMID: 23543134
28	581	Nephrotic syndrome	PMID: 16995591

Table 4.5.: List of the top 30 diseases associated with DM. The topology-based measure was used as the similarity measure, and FunDO was used as the source of disease-gene associations. Only diseases annotated in all four disease-gene association datasets are listed in the table. For a disease associated with DM according to ICD-9, comorbidity or GWAS, we added the supported evidence to the reference (the last column). The remaining disease associations were validated via mining the literature on PubMed, and for each disease only one reference (shown by PubMed ID) was listed in the table.

we also demonstrated the use of our measures in discovering novel disease associations and validated it via literature curation.

Novel disease associations uncovered in this study can be further used to improve our understanding of disease classification. For example, a human disease network that models the relationship of diseases can be constructed based on these similarity measures, and computational approaches, such as clustering, can be applied to detect communities in the disease network. This may provide the opportunity to redefine the current disease classification and further lead to improvements in disease diagnosis, prognosis and treatment.

4.5. Author’s contributions

In the work presented in this chapter, Kai Sun collected all the data, designed and implemented the similarity measures, ran all the experiments, and performed all the analysis.

5. Biological network integration

In this chapter, we further our results by integrating different types of large-scale biological data, including GWAS data, disease-chemical associations, biological pathways and GO annotations, with the aim of gaining more complete understanding of the relationships between diseases. In particular, we integrate these data into a heterogeneous network, where nodes are bio-entities (e.g., genes, diseases, chemicals, pathways and GO terms) and edges between nodes represent their associations. We extend our disease similarity measures to an integration-based measure to predict novel disease associations by mining the heterogeneous network. Our results are systematically evaluated against the MeSH classification and a statistical measure of disease co-occurrence in PubMed. In addition, we show that our similarity measure achieves better performance by using the integrated data than solely using a specific type of data. To illustrate the use of our approach to uncover novel diseases associations, we present a case study of Crohn's disease, in which we identify its top associated diseases and validate our predicted associations via mining the literature. We also give examples to demonstrate how our approach can be used for drug repositioning.

The work presented in this chapter is in collaboration with Dr. Chris Larminie and Dr. Natalie Buchan. The related manuscript is in preparation.

5.1. Motivation

Recent advances done in ‘omics’ technologies have lead to a wealth of biological data. These data, including genetic data (e.g., GWAS data [70, 173, 174, 71]), genomic data (e.g., microarray data [175, 176, 177]), proteomic data (e.g., PPI data [18, 19, 67, 14]), metabolomic data (e.g., metabolic pathways [178, 179]) and transcriptomic data (e.g., RNA-Sequencing data [180, 181, 182]), have been used to improve our understanding of the underlying mechanisms of human diseases. Since various types of biological data reside in diverse repositories, data integration has become crucial, as it not only provides a comprehensive view of data, but also enables further exploration of biological knowledge by efficiently mining and analysing these large-scale data.

One of the major tasks of biological data integration is to collect and combine heterogeneous data from different repositories. Repositories such as the OMIM database [64] for phenotype-genotype associations; the BioGRID database [30] for PPIs and genetic interactions; the NHGRI GWAS catalog [73] for GWAS data; KEGG [54] and Reactome [55] for biological pathways, are very valuable for biological knowledge discovery. Many studies in data integration collected information from these repositories and developed computational methods to gain novel biological insights into human diseases. Franke *et al.* (2006) [183] integrated information derived from PPIs, co-expression data, GO annotations and biological pathways, and developed a Bayesian framework to reconstruct a functional human gene network. They applied this network to rank positional candidate genes for various heritable disorders on the basis of the functional interactions of disease genes. Also by using a naïve Bayes classifier, Linghu *et al.* (2009) [184] constructed an integrated human functional linkage network from different datasets to prioritise disease candidate genes and infer associations between phenotypically dissimilar disease pairs. Ahn *et al.* (2011) [185] integrated PPIs, genetic interactions, gene regulation pathways and mRNA expression data to classify cancer-related genes. Specifically, they identified a prostate cancer-specific network of which interactions showed different behaviour between tumour and normal samples. Alcaraz *et al.* (2012) [186] combined a human PPI network with gene expression data and genome-scale DNA methylation profiles, and developed a software framework to discover

connected sub-networks which contained genes that were mainly dysregulated in most cases studied. There are also many other integration-based approaches that have been successfully applied to predict protein functions [187, 188, 189], to mine disease-specific pathways [190, 186], to infer disease-gene relationships [191, 183, 149, 192, 184, 193, 185] and to reposition drugs [194, 82, 83] from multiple heterogeneous data sources.

We propose to uncover novel associations between diseases via biological data integration. An accurate classification of human diseases is essential in consideration of reliable diagnosis and treatment. However, current classification of human diseases is mainly derived from the similarity of clinical phenotypes. It has been widely recognised that current disease classification lacks sensitivity in identifying preclinical diseases and specificity in defining disease unequivocally [157]. During the past decade, some studies have been proposed to improve our understanding of the relationship between diseases using large-scale biological data [65, 66, 67, 81, 132, 57, 140, 141]. Since most of these studies mainly focused on only one type of biological data (especially genetic data such as OMIM and GWAS data), we believe an integration-based approach that takes advantage of various types of biological data will provide further insights into disease relationships.

In this chapter, we integrated different types of biological data collected from diverse repositories, including disease-gene associations, disease-chemical associations, biological pathways and GO annotations, to gain further understanding of human diseases and their relationships. In particular, we integrated these data into a heterogeneous network, where nodes are bio-entities and edges between nodes represent their associations. To uncover novel disease associations, we developed a computational approach to measure the similarity between diseases, and constructed an integrated disease network. Our integration-based similarity measure was systematically evaluated against the hierarchical tree of the MeSH classification system¹ and a statistical measure of disease co-occurrence in the literature. Furthermore, we presented a case study of Crohn’s disease to illustrate how the integrated disease network could be used to improve our current understanding of disease relationships.

¹<http://www.nlm.nih.gov/mesh/>

5.2. Methods

5.2.1. Data collection

We integrated different types of large-scale biological data, including disease-gene associations, disease-chemical associations, biological pathways and GO annotations. These data were collected from a number of repositories. Since different repositories may use different disease naming schemes, we mapped all diseases annotated in these data to MeSH terms for a correct integration. Details of data sources and mapping approaches are described below.

Disease-gene associations

Disease-gene associations were collected from three repositories: CTD [68], FunDO [163] and the NHGRI GWAS catalog [73].

CTD provides curated disease-gene associations that were extracted from the published literature by CTD curators, or were derived from OMIM [64]. We downloaded disease-gene associations from CTD in July 2013, and collected 21,625 curated associations between 4,286 diseases (annotated by MeSH terms) and 6,940 genes.

FunDO contains DO² annotations of the human genome extracted from the NCBI GeneRIF database³. We used the latest stable version of FunDO (released in October 2008), which contained gene annotations for 1,854 diseases (annotated by DO IDs). We then mapped DO IDs to MeSH terms by using the mapping provided in DO version 3 (the latest stable version, released in May 2007). After mapping, the dataset contained 1,255 diseases (mapped from 1,528 DO IDs), 4,886 genes and 22,879 disease-gene associations.

GWAS catalog collects associations between traits (or diseases) and SNPs identified by GWAS studies. We downloaded GWAS catalog data in August 2013. Similar to Sanseau *et al.* (2012) [84], we eliminated unreplicated associations and associations with p -value higher than 10^{-7} to minimise false-positives. Diseases annotated in GWAS catalog were manually mapped to MeSH terms. In total, 6,430 associations between 370 diseases and 3,762

²<http://disease-ontology.org/>

³<http://www.ncbi.nlm.nih.gov/gene/about-generif>

genes were collected from the GWAS catalog.

The set of disease-gene associations used in this study was the union set of CTD, FunDO and GWAS catalog. It contained a total of 48,549 associations between 4,891 diseases and 11,160 genes.

Disease-chemical associations

Disease-chemical associations used in this study were downloaded from the CTD database in July 2013. CTD contains both curated and inferred disease-chemical associations, but we used only curated associations extracted by CTD curators. In total, we collected 80,225 associations between 3,029 diseases (annotated by MeSH terms) and 8,201 chemicals.

Biological pathways

We collected pathway data from four repositories: KEGG [54], Reactome [55], WikiPathways [56] and GeneGO⁴ (i.e., MetaCore from Thomson Reuters). All of these repositories contain manually curated pathway maps representing molecular reactions and interactions in a cell. We downloaded pathway data from the four repositories in September 2013 (pathway data from GeneGO were licensed). Only human pathways were considered in our study. In total, we collected 2,319 biological pathways (276 pathways from KEGG, 1,416 pathways from Reactome, 201 pathways from WikiPathways and 426 pathways from GeneGO) associated with 9,992 genes.

GO annotations

GO annotations of human were downloaded from the GO database [162] in September 2013. We removed all IEA annotations as they were computationally inferred annotations which had not been reviewed by curators. In total, we collected 123,743 annotations between 13,609 genes and 11,266 GO terms (1,034 cellular component terms, 7,480 biological process terms, and 2,752 molecular function terms).

⁴<https://portal.genego.com/>

5.2.2. Enrichment analysis

We inferred disease-pathway associations and disease-GO term associations by identifying pathways and GO terms that were significantly enriched within the set of genes associated with each disease. The statistical significance (p -value) of the enrichment was calculated according to the hypergeometric distribution for sampling without replacement (identical to one-tailed Fisher’s exact test). For a given disease D_i and a given pathway (or GO term) P_j , let $|G_{D_i}|$ be the number of genes associated with D_i , $|G_{P_j}|$ be the number of genes annotated with P_j , and $|G_P|$ be the total number of genes in the pathway (or GO annotation) data. If there were x genes annotated with pathway P_j that were associated with D_i (thus $x = |G_{D_i} \cap G_{P_j}|$), we computed the p -value of the enrichment (i.e., the probability of observing the same or higher enrichment purely by chance) as:

$$p - value = 1 - \sum_{k=0}^{x-1} \frac{\binom{|G_{P_j}|}{k} \binom{|G_P| - |G_{P_j}|}{|G_{D_i}| - k}}{\binom{|G_P|}{|G_{D_i}|}} \quad (5.1)$$

Since we used the union of three disease-gene association datasets in our analysis, we obtained a relatively larger set of annotated genes for each disease, compared to that obtained by solely using one disease-gene association dataset (which is the case for the study we presented in Chapter 4). It is known that a larger gene list may result in more significant p -values to slightly enriched pathways (or GO terms) [195]. For this reason, we used the Bonferroni correction, which is considered as a more stringent test compared to the others [168], to adjust p -values for multiple comparisons. A pathway or a GO term is said to be associated with a disease if the adjusted p -value is less than 0.05. Through this approach, we identified 29,432 disease-pathway associations between 1,128 diseases and 1,491 pathways, and 8,614 disease-GO term associations between 999 diseases and 1,714 GO terms.

5.2.3. Disease similarity measure

To uncover associations between diseases from our integrated data, we developed a novel measure to estimate disease similarity scores. The similarity measure is based on the *vector space model* [196, 197], which is frequently

used in the area of information retrieval. Four types of disease-related data were used for computing similarity scores, namely disease-gene associations, disease-chemical associations, disease-pathway associations and disease-GO term associations. Let E be the set of non-disease elements (i.e., genes, chemicals, pathways and GO terms) in the integrated dataset. For each disease D_i , we first constructed an $|E|$ -dimensional vector $\vec{D}_i$, called D_i 's *signature*, to represent information we gained for D_i :

$$\vec{D}_i = \left(w_{E_0} \times f(D_i, E_0), \dots, w_{E_k} \times f(D_i, E_k), \dots, w_{E_{|E|-1}} \times f(D_i, E_{|E|-1}) \right) \quad (5.2)$$

where $f(D_i, E_k)$ is a function that indicates the presence or absence of the association between the disease D_i and the non-disease element E_k :

$$f(D_i, E_k) = \begin{cases} 1 & \text{if } E_k \text{ is associated with } D_i \\ 0 & \text{else} \end{cases}$$

and w_{E_k} is a weight (also known as the *inverse document frequency* (*idf*) [198]) assigned to E_k . Let D be the set of diseases annotated in the data, and D_{E_k} be the set of diseases associated with E_k . The *idf* of E_k is defined as:

$$w_{E_k} = idf(E_k, D) = \log \frac{|D|}{|D_{E_k}|} \quad (5.3)$$

Assigning the *idf* as a weight to each non-disease element E_k allows us to increase the importance of E_k if E_k is rare across all diseases (i.e., E_k is associated with only a few diseases). Meanwhile, the logarithm in Equation 5.3 prevents our results from being dominated by the most rare non-disease elements, since the number of associated diseases may greatly differ from one to another. For a pair of diseases D_i and D_j , their similarity can be measured by the similarity between their corresponding signatures $\vec{D}_i$ and $\vec{D}_j$. Thus we computed the raw similarity score between D_i and D_j as the cosine of the angle θ_{ij} between $\vec{D}_i$ and $\vec{D}_j$:

$$\cos \theta_{ij} = \frac{\vec{D}_i \cdot \vec{D}_j}{\|\vec{D}_i\| \|\vec{D}_j\|} \quad (5.4)$$

If considering only the overlaps of non-disease elements (i.e., the number

of shared genes, chemicals, pathways and GO terms), a disease D_i may always have a higher similarity score with a complex disease (i.e., a disease that is caused by a combination of genetic, environmental, and lifestyle factors [199]) than a monogenic disease (i.e., a disease that is caused by a single gene). The cosine similarity reduces this bias: it compares disease signatures $\vec{D}_i$ and $\vec{D}_j$ on a normalised space, as it considers only their orientation and not magnitude. This enables that complex diseases do not always come out on top in computation of disease similarities.

To further ensure that our results can not be obtained by random chance, we computed a random similarity score for each pair of diseases D_i and D_j to adjusted their raw similarity score $\cos \theta_{ij}$. Random similarity scores were computed using a randomisation-based approach. We first generated random disease-gene association data by rewiring associations: for a disease-gene association (D_x, G_x) , we randomly chose another disease-gene association (D_y, G_y) , and replaced the two associations by (D_x, G_y) and (D_y, G_x) , if $D_x \neq D_y$, $G_x \notin G_{D_y}$ and $G_y \notin G_{D_x}$. This process was repeated until all disease-gene associations were rewired. We generated random disease-chemical associations by using a similar approach. Random disease-pathway associations and random disease-GO term associations were obtained by running enrichment analysis using the random disease-gene association data. We then used these random association data to compute a random similarity score for each pair of diseases. This process was repeated 100 times to generate the average random similarity score γ_{ij} for each pair of diseases D_i and D_j . Finally we computed the adjusted similarity score for diseases D_i and D_j as:

$$\text{Similarity}(D_i, D_j) = \max(\cos \theta_{ij} - \gamma_{ij}, 0) \quad (5.5)$$

5.3. Results and discussion

5.3.1. Integration of disease-related data

We constructed an Integrated Disease Network (IDN) in which nodes represent different bio-entities and edges represent the associations between them. In total the IDN consists 47,342 nodes and 432,660 edges, as shown in Table 5.1. The IDN was built by integrating large-scale biological data

Nodes		Edges	
Bio-entities	Number	Associations	Number
Disease	6,088	Disease-gene	48,549
Gene	18,812	Disease-chemical	80,225
Chemical	8,201	Gene-pathway	142,145
Pathway	2,975	Gene-GO term	123,743
GO term	11,266	Disease-pathway	29,432
		Disease-GO term	8,614
Total nodes	47,342	Total edges	432,660

Table 5.1.: Details of the statistics of the IDN.

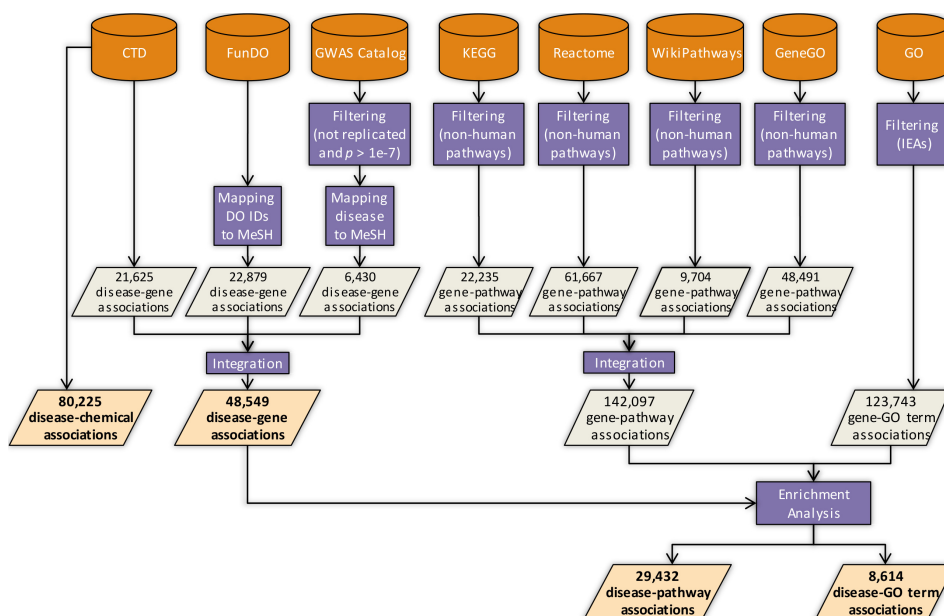


Figure 5.1.: Flow chart of data integration.

collected from different repositories, as demonstrated in Figure 5.1. All diseases annotated in the repositories were mapped to MeSH disease terms and all genes were mapped to Entrez gene IDs [200] for the purpose of integration.

Disease-gene associations were collected from three repositories: CTD, FunDO and GWAS Catalog. All of these repositories contain curated disease-gene associations extracted from the literature. The OMIM database is not included as CTD already contains disease-gene associations derived from OMIM. GWAS catalog has fewer associations compared to CTD and FunDO

because GWAS studies have been conducted on only a few diseases thus far. However, for a specific disease GWAS catalog may contain more information than the others (see Section 5.3.3). Since the three repositories focus on different aspects, the overlap among them is small: only 88 associations are supported by all three datasets (see Figure A.10 for details). Integrating the three datasets largely increases the coverage of disease-gene associations.

Disease-pathway and disease-GO term associations were inferred from disease-gene, gene-pathway, gene-GO term associations by using enrichment analysis. To investigate how the number of inferred associations deviates from random, we randomly rewired the associations between diseases and genes, while keeping the number of associations of each disease and gene unchanged (also see Section 5.2.3). The randomisation process was repeated 100 times. The average number of disease-pathway (disease-GO term) associations obtained from random disease-gene associations is $10,378 \pm 223$ ($2,395 \pm 57$), which is 2.84-fold (3.60-fold) lower than the actual number of significant disease-pathway (disease-GO term) associations (p -value < 0.01). These differences suggested that genes associated with the same disease are more likely to share common biological functions than those associated with different diseases.

On average, a disease node in the IDN is connected to 9.93 gene nodes, 26.49 chemical nodes, 26.09 pathway nodes and 8.62 GO term nodes; a gene node in the IDN is connected to 4.35 disease nodes, 14.22 pathway nodes and 9.09 GO term nodes. The largest connected component of the IDN contains 46,235 (97.66%) nodes, covering 5,541 (91.02%) diseases. In the connected network, the average distance between a pair of disease nodes is 4.06 (note that the minimum distance between two diseases is 2). These results indicate the interconnectedness of human diseases.

5.3.2. Evaluation of similarity measure

We proposed a novel integration-based similarity measure to infer disease-disease associations from the IDN. In particular, we assigned a score ranging from 0 to 1 to qualify the strength of the association between a pair of diseases in the IDN. A similarity score of 0 suggests that the two diseases are not directly associated: either they have no common neighbours in the IDN, or their association score is not higher than expected at random (see

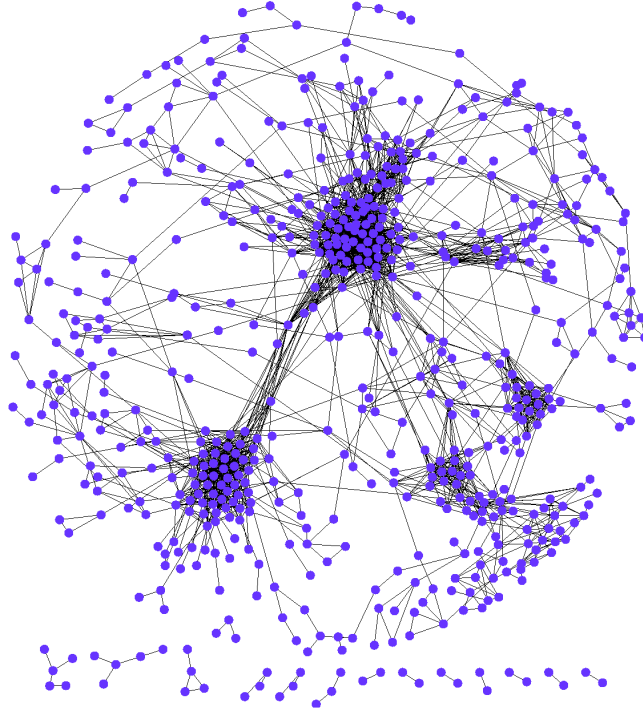


Figure 5.2.: The disease-disease network constructed based on the integration-based similarity scores. Only the top 1% disease associations were considered as the edges in the network. Isolated nodes were excluded from the network. Cytoscape 2.8.1 [48] was used for the visualisation. The spring embedding algorithm was chosen as the layout algorithm.

Section 5.2.3). We only considered the 711 diseases that are associated with at least one gene, one chemical, one pathway and one GO term in the evaluation. A disease-disease network can be constructed based on the similarity scores. When considering only the top 1% associations as the edges in the disease-disease network, the network contains 711 nodes (including 101 isolated nodes) and 2,525 edges (Figure 5.2). The average clustering coefficient of this network is 0.419, which is much higher than a random graph (e.g., for ER random graphs with the same size, the average clustering coefficient is 0.014 ± 0.017).

To assess the reliability of our similarity measure, we first evaluated our similarity measure against the MeSH tree classification (Table A.2) by comparing similarity scores of diseases from the same MeSH tree branch to

those from different MeSH branches. For example, Crohn’s disease and ulcerative colitis are both in the ‘Digestive System Diseases’ branch, while sarcoidosis is in the ‘Hemic and Lymphatic Diseases’ branch. Since similarity scores are not normal distributed, we used the Mann-Whitney U test to compute the statistical significance. We found that diseases from the same MeSH branch have significantly higher ($p\text{-value} < 4.9407 \times 10^{-324}$) similarity scores (0.0174 ± 0.0368) than diseases from different MeSH branch (0.0066 ± 0.0199).

We further evaluated our similarity measure against the co-occurrence of disease MeSH terms in the literature. To obtain co-occurrence scores for all pairs of diseases we analysed, we used a statistical approach proposed by Li and Agarwal (2009) [57]. In this approach, the co-occurrence score of a pair of diseases is measured by the statistical significance of their co-occurrence in the PubMed⁵ abstracts. We collected 15.29 million abstracts from PubMed covering articles published from January 1900 to October 2013, and identified disease MeSH terms that were associated with these abstracts as Major MeSH Headings. Statistical significance of disease co-occurrence was computed using one-tailed Fisher’s exact test and adjusted by Benjamini-Hochberg correction (see [57] for details). In total, we collected 8,459 significant co-occurrence associations between the 711 diseases we evaluated (corrected $p\text{-value} < 0.05$).

Similar to Section 4.3.2, we used ROC curves analysis to investigate the correlation between our similarity scores with disease co-occurrence scores. Figure 5.3 and Table A.6 show the ROC curves and AUC values obtained by evaluating our similarity measure. We found that the correlation between our integration-based similarity scores and disease co-occurrence scores is substantially higher than expected at random. In addition, our similarity measure tends to detect strong disease co-occurrence associations with small p -values (e.g., $p\text{-value} < 10^{-20}$). When setting the co-occurrence p -value threshold to 0.05 (i.e., all disease pairs having a co-occurrence p -value less than 0.05 are considered to be associated, 8,459 pairs in total), we obtained an AUC value of 0.7580 (Table A.6). This value increased to 0.7646 with a p -value threshold of 0.01 (7,713 co-occurrence associations in total), and 0.8282 with a p -value threshold of 10^{-20} (3,330 co-occurrence associations in total). These results demonstrates the ability of our approach to predict

⁵<http://www.ncbi.nlm.nih.gov/pubmed>

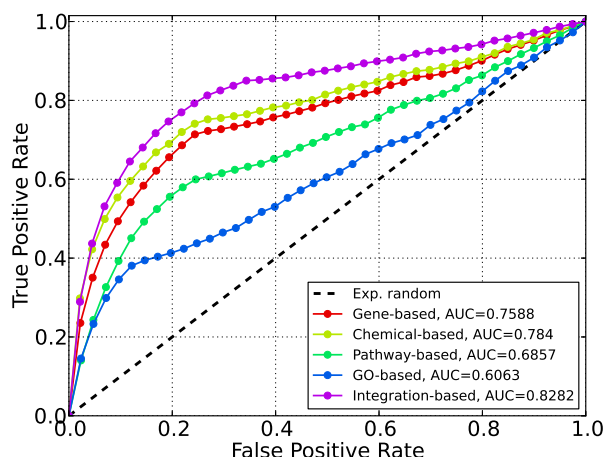


Figure 5.3.: Evaluation of our similarity measure against disease co-occurrence in the literature. ROC curves were obtained by using disease-gene associations (gene-based), disease-chemical associations (chemical-based), disease-pathway associations (pathway-based), disease-GO term associations (GO-based) or all four types of associations (integration-based). The p -value threshold for co-occurrence associations was set to 10^{-20} . AUC values obtained by using other p -value thresholds can be found in Table A.6.

co-occurrence associations.

More important, our results reveal the need of data integration for uncovering biological knowledge. We show that using the integration of disease-related data leads to better performance in recovering disease co-occurrence associations than solely using one type of data (Figure 5.3 and Table A.6). Among the four types of disease-related data, disease-gene association seems to be the most effective in predicting disease associations, followed by disease-chemical association, disease-pathway association and disease-GO term association. However, this result may also suggest that analysing shared biological pathways and shared GO annotations between diseases may offer the opportunity to uncover previously undiscovered disease associations.

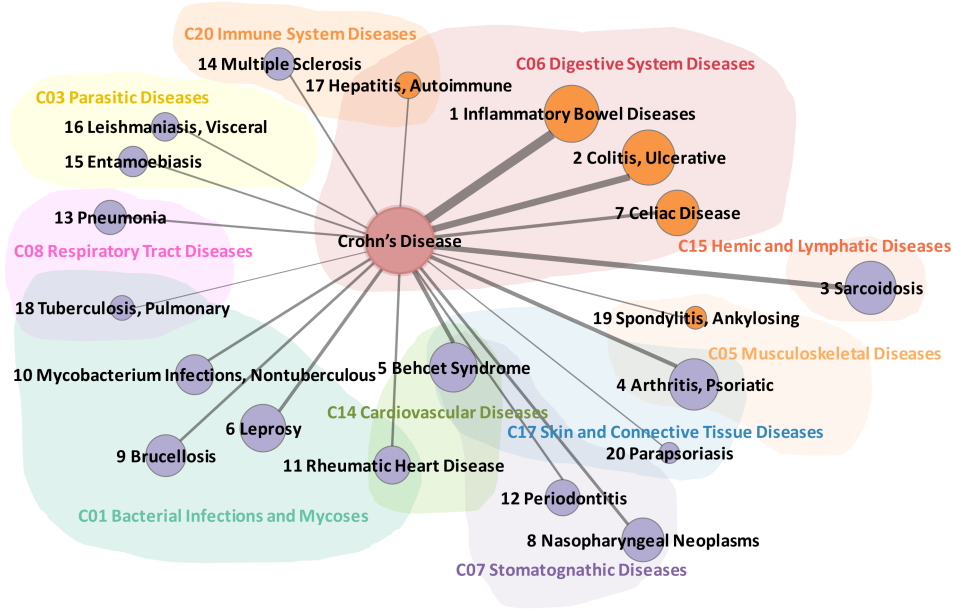


Figure 5.4.: Top 20 diseases associated with Crohn's disease. Diseases are grouped based on the MeSH tree classification (Table A.2). Numbers before disease names represent the ranking of association scores (see Table A.7 for the scores). The size of nodes is inversely proportional to the ranking. The width of edges is proportional to the association scores. Orange nodes represent diseases that are associated with Crohn's disease according to the literature co-occurrence measure (p -value < 0.05).

5.3.3. Crohn's disease: a case study

To illustrate how our approaches can be used to improve our current understanding of human diseases and their relationships, we present a case study of Crohn's disease (MeSH ID: D003424). Crohn's disease, also known as regional enteritis, is a chronic relapsing inflammatory disorder that can affect any part of the gastrointestinal tract. It can cause a wide variety of clinical symptoms, including abdominal pain, diarrhoea, and fever [201]. Recent studies suggested that Crohn's disease is an abnormal immune response to the bacteria in genetically susceptible individuals [202, 203]. However, the exact underlying cause of Crohn's disease is still unknown.

In the IDN, Crohn's disease is linked to 370 gene nodes (40 from CTD, 83 from FunDO, and 285 from GWAS Catalog). Among them only 5 genes are

reported in all three repositories, including tumor necrosis factor (TNF), interleukin 10 (IL10), nucleotide-binding oligomerization domain containing 2 (NOD2), immunity-related GTPase family, M (IRGM) and cyclin Y (CCNY). These 370 Crohn’s disease associated genes are significantly enriched in 120 biological pathways (39 from KEGG, 17 from Reactome, 11 from WikiPathways and 53 from GeneGO) and 47 GO terms (5 cellular component terms, 39 biological process terms, and 3 molecular function terms). There are also 25 chemical nodes that are connected to Crohn’s disease in the IDN.

Figure 5.4 shows the top 20 diseases associated with Crohn’s disease identified by our approach. 5 diseases among them are also associated with Crohn’s disease according to the literature co-occurrence measure (p -value < 0.05). The first two associated diseases are inflammatory bowel diseases (IBD) and ulcerative colitis (CU), which is not surprising since Crohn’s disease and CU are considered as two major types of IBD [204]. After grouping these 20 diseases based on the MeSH tree classification (Table A.2), we found that 4 out of them are under the same MeSH branch ‘Digestive System Diseases’ with Crohn’s disease. Meanwhile, our results show that Crohn’s disease is closely related to 5 diseases classified as ‘Bacterial Infections and Mycoses’ (e.g., leprosy) and 2 diseases classified as ‘Immune System Diseases’ (e.g., multiple sclerosis). This finding supports the most widely held hypothesis that Crohn’s disease is caused by the dysfunctional interaction between the intestinal immune system and a subset of commensal enteric bacteria [202, 203]. Apart from the 5 known disease-disease associations (i.e., diseases that are associated according to MeSH classification or co-occurrence measure), the remaining 15 associations are considered as novel associations inferred by our approach. High similarity scores and non-significant co-occurrence p -values of these novel associations suggest they may have been overlooked in the previous literature.

We further investigated the common underlying mechanisms of Crohn’s disease and its top 1% associated diseases, including IBD, CU, sarcoidosis (SA), psoriatic arthritis (AP), Behçet syndrome (BS), leprosy (LE) and celiac disease (CD). We generated an induced subgraph from the IDN representing the associations between the 8 diseases and their common annotated genes (Figure 5.5). TNF⁶, a gene encodes a multifunctional proinflamma-

⁶<http://www.ncbi.nlm.nih.gov/gene/7124>

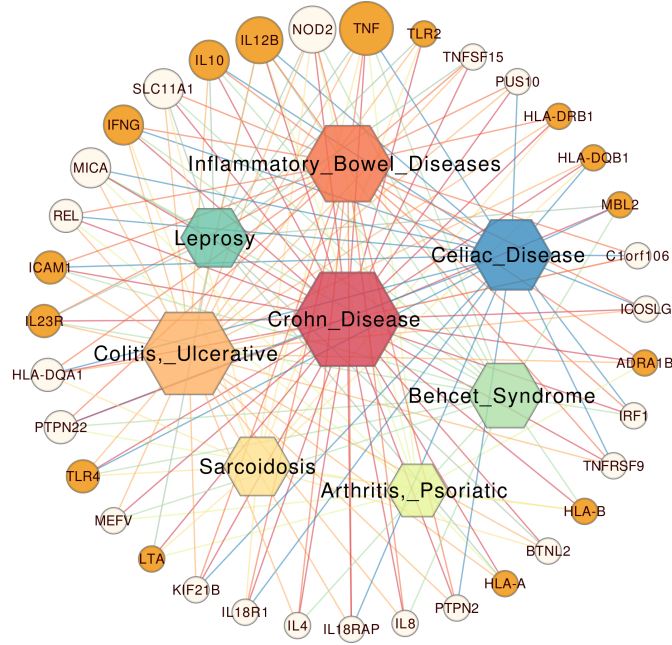


Figure 5.5.: The most common genes (represented by round nodes) shared between Crohn's disease and its top 7 associated diseases (represented by hexagon nodes). Only genes that associated with at least 4 diseases are shown in the figure. The size of a node is proportional to its degree (i.e., the number of diseases or genes it associated with). The colour of an edge corresponds to the disease it connects. A gene node is coloured in orange if its protein product is a drug target. Cytoscape 2.8.1 [48] was used for the visualisation.

tory cytokine involved in the regulation of many biological processes such as cell proliferation, differentiation, and apoptosis, is found to be associated with all the 8 diseases. It is known that TNF plays an important role in cancer and autoimmune diseases and it has been targeted for the treatment of many diseases including Crohn's disease [205]. Despite anti-TNF therapies have been very successful in ameliorating symptoms of autoimmune diseases, they might be harmful as TNF is still essential for the proper functioning of the immune system [206]. We also found that many of the genes shared among the 8 diseases are involved in the regulation of immune response, e.g., NOD2, interleukin 12B (IL12B), IL10, interferon gamma (IFNG) and intercellular adhesion molecule 1 (ICAM1). By investigating the biological

Drug	Target	Stage	Associated diseases
bertilimumab	CCL11	Preclinical	IBD, LE
ustekinumab	IL12B	Phase III Clinical Trial	IBD, CU, AP, BS, LE, CE
QAX-576	IL13	Phase II Clinical Trial	IBD, CU
clazakizumab	IL6	Phase II Clinical Trial	IBD, CU
PF-4236921	IL6	Phase II Clinical Trial	IBD, CU
olokizumab	IL6	Phase II Clinical Trial	IBD, CU
naltrexone	TLR4	Phase II Clinical Trial	CU, SA, BS, CE
VT-346	TNF	Preclinical	All top 7 diseases
ozoralizumab	TNF	Phase I Clinical Trial	All top 7 diseases
XPro-1595	TNF	Preclinical	All top 7 diseases
certolizumab pegol	TNF	Launched	All top 7 diseases
ESBA-105	TNF	Phase II Clinical Trial	All top 7 diseases
adalimumab	TNF	Launched	All top 7 diseases
infliximab	TNF	Launched	All top 7 diseases
PF-05230905	TNF	Phase I Clinical Trial	All top 7 diseases

Table 5.2.: Details of drugs developed for the treatment of Crohn’s disease, including the drug’s target, current development stage, and diseases that are associated with the target.

pathways shared between these diseases, we found that all of these diseases are associated with the ‘Cytokines and inflammatory response’ (Wikipathways), the ‘Immune response HSP60 and HSP70/TLR signaling pathway’ (GeneGO) and 7 KEGG disease pathways (Figure A.11 and Figure A.12).

Disease-disease associations uncovered by our approach can be further used to explore drug repositioning opportunities. Drug repositioning (i.e., identifying new disease indications for approved drugs) is considered as one of the most effective strategies in drug discovery. It can not only substantially reduce the costs and time of drug development, but also provide safer treatment for patients as it lowers the risk of unexpected toxicity and side effects [207]. Recently, many computational approaches have been developed for drug repositioning. These approaches can be roughly classified as ‘drug-based ’ or ‘disease-based’ [208]. ‘Drug-based’ approaches infer repositioning opportunities from chemical or pharmaceutical perspectives, such as the chemical properties [209, 210] and molecular activities [211, 212] of drug compounds. ‘Disease-based’ approaches infer repositioning opportunities from clinical perspectives, such as drug indications [81, 140, 213] and side effects [79, 214, 210, 215]. While most of these approaches have focused on either drug-target or drug-disease associations, only a few studies have provided an integrated approach to combine different types of associations for drug repositioning. Sanseau *et al.* (2012) [84] presented an analysis

that for the first time incorporated drug-target associations with GWAS data to identify potential drug repositioning opportunities. Another integrated approach for drug repositioning was proposed by Daminelli *et al.* (2012) [83]. They constructed a drug-target-disease network by integrating disease-chemical and drug-target associations, and mined the network for network motifs of bi-cliques in order to predict disease-drug associations. However, the major limitation of this approach is that the predictions may have a high false positive rate since they were inferred solely based on the network structure.

Here we present an integrated approach for computational drug repositioning. Four types of associations were used in this approach: disease-drug associations, disease-gene associations, drug-target associations and disease-disease associations. In particular, for a pair of associated diseases D_i and D_j , we checked their shared genes (if any) to see whether there existed a gene which had been targeted by a drug for the treatment of disease D_i . If the drug had not been developed for the treatment of disease D_j , we inferred an association between that drug and disease D_j . We obtained disease-drug-target associations from Drugbank 3.0 [76] and Informa Pipeline⁷ in November 2013 (data from Informa Pipeline were licensed). From all genes linked to Crohn’s disease in the IDN, we focused on the 190 genes that are associated with at least one of the other 7 diseases. 39 out of these genes encode proteins that are reported as drug targets in Drugbank or Informa Pipeline. We found that 6 of these druggable genes, including C-C motif chemokine 11 (CCL11), IL12B, interleukin 13 (IL13), interleukin 6 (IL6), toll-like receptor 4 (TLR4) and TNF, have been targeted for the treatment of Crohn’s disease (Table 5.2). The drugs developed for these targets may provide repositioning opportunities for diseases that are closely associated with Crohn’s disease. Meanwhile, drugs that have been developed for these associated diseases may be candidates for the treatment of Crohn’s diseases (Table 5.3). For example, thalidomide, an immunomodulatory drug that inhibits the production of TNF, is approved by the U.S. Food and Drug Administration (FDA) for the treatment of leprosy. Several recent clinical studies (e.g., [216, 217]) have shown the positive effect of thalidomide on the remission of Crohn’s disease, supporting the potential repositioning of thalidomide for Crohn’s disease.

⁷<http://sites.informahealthcare.com/pipeline/>

Drug	Target	Indication	Stage
CI-201	IL12B	CU	Phase II Clinical Trial
tralokinumab	IL13	CU	Phase II Clinical Trial
anrukinzumab	IL13	CU	Phase II Clinical Trial
VT-384	IL18	IBD	Preclinical
GSK-1070806	IL18	IBD	Phase I Clinical Trial
interleukin-2	IL2RA	LE	Launched
CR-3294	NOS2	IBD	Phase II Clinical Trial
DCAM-253	RIPK2	IBD	Preclinical
thalidomide	TNF	LE	Launched
golimumab	TNF	CU	Launched
golimumab	TNF	SA	Phase II Clinical Trial
AG-014	TNF	IBD	Preclinical
HL-036	TNF	IBD	Preclinical
etanercept	TNF	AP	Launched
AVX-470	TNF	CU	Phase I Clinical Trial

Table 5.3.: Details of potential drug repositioning opportunities for Crohn’s disease, including the drug’s target, disease indication, and current development stage.

5.4. Conclusions

In this chapter, we integrated different types of large-scale biological data collected from different repositories with the aim of gaining further understanding of how diseases associated with each other. We constructed a heterogeneous network to represent the integrated data, where nodes are bio-entities such as diseases, genes, chemicals, pathways or GO terms, and edges between nodes represent the associations between these bio-entities. In addition, we proposed a novel integration-based similarity measure to infer disease-disease associations from the heterogeneous network. Our similarity measure was systematically evaluated against the MeSH tree classification and a statistical measure of disease occurrence in the literature, confirming the ability of our approach to recover known disease-disease associations. We demonstrated how our approach can be used to provide insights into common underlying mechanisms of diseases and potential drug repositioning opportunities by presenting a case study of Crohn’s disease.

5.5. Author’s contributions

In the work presented in this chapter, Kai Sun collected all the data (excluding GeneGO pathways, PubMed abstracts and Informa Pipeline), con-

structed the IDN, designed and implemented the similarity measure, ran all the experiments, and performed all the analysis.

6. Conclusions

6.1. Summary of the dissertation

Gaining biological insights from system-level molecular data requires the development of efficient and biologically meaningful algorithms for their analysis. In this thesis, we presented network-based computational approaches for uncovering biological knowledge, especially disease-disease associations, from large-scale biological data, with the aim of providing novel insights into human diseases and their relationships. In particular, we introduced novel methods for solving three fundamental problems in biological network analysis, namely network modelling, comparison and integration. Our methods were applied to analyse diverse system-level biological data such as PPIs, disease-gene associations, drug-target associations, GO annotations and biological pathways, and proved to be useful in many biological tasks such as modelling PPI networks, predicting disease-disease associations and inferring drug repositioning opportunities.

We exploited systematic measures based on graphlets, which are defined as small, connected, induced subgraphs of large networks, to analyse the topology of biological networks. In Chapter 3, we examined the use of graphlet-based measures for biological network comparison and rebutted the claim that ‘graphlet-based measures were unstable in regions of low edge density’. By generating the empirical distributions of GDD agreement scores, we identified the edge density regions in which the topology of model networks is ‘unstable’, and showed how graphlet-based measures correctly detect this topological instability. Moreover, we analysed 18 PPI networks of different species and showed that these PPI networks had high edge densities thus the ‘unstable’ regions do not affect their analysis. These results demonstrated the suitability of graphlet-based measures for biological network comparison. We also validated the use of graphlet-based measures for finding well-fitting random models for PPI networks by using a recently de-

vised non-parametric statistical test. We showed for the first time that five viral species, possessing the latest and most complete PPI networks, were well-fit by several random models. Though the biological significance of the fits is not immediately clear, we believe the identification of well-fitting models for viral PPI networks is a stepping stone towards understanding biological systems.

In Chapter 4, we took advantage of diverse biological data collected from a number of repositories to gain novel insights into the relationship of human diseases by considering their underlying biological mechanisms. Unlike previous studies on disease-disease associations, we used disease-gene associations obtained from different sources to avoid the bias introduced by a single dataset. Moreover, we also used the topology of a large-scale molecular network to examine its use for inferring disease associations. We applied three disease similarity measures, including standard methods (e.g., annotation-based measure) and a novel measure (e.g., topology-based measure), to estimate similarity scores between diseases. Since most previous studies were only able to validate part of their predicted associations by manually mining the literature, we also provided a systematic approach to assess the quality of computationally predicted disease associations. In particular, we evaluated disease associations obtained by our measures against the ICD disease classification system, comorbidity data and GWAS data by using ROC curve analysis. The strong correlation between our predicted disease associations and known disease associations indicated the ability of our approaches to recover known disease associations and the potential to identify previously undiscovered disease associations.

We further integrated different types of large-scale biological data, including disease-gene associations derived from the literature and GWAS data, disease-chemical associations, biological pathway data and GO annotations, to gain more comprehensive understanding of the relationships between human diseases. In Chapter 5, we constructed a heterogeneous network in which nodes were bio-entities and edges between nodes represented their associations. We also extended our disease similarity measures to an integration-based measure to predict disease associations by mining the heterogeneous network. Our predictions were systematically evaluated against a statistical measure of the occurrence of disease MeSH terms in the literature. In addition, our results showed that using the integrated data led

to better performance on recovering known disease associations than solely using one type of data, revealing the need of biological data integration. The case study of Crohn’s disease well demonstrated the ability our approach to identify previously undiscovered disease associations, investigate common underlying mechanisms driving diseases, and infer drug repositioning opportunities. Our approach may also be a very effective knowledge mining tool, as it was able to survey and summarise a large molecular knowledge space and still provide direct access to the underlying evidence, which is particularly important for biological knowledge discovery.

Overall, we believe the methodologies and results presented in this thesis can provide novel insights into human diseases and their relationships. Disease associations inferred by our measures have the potential to improve our current knowledge of disease diagnosis, prognosis and treatment.

6.2. Future work

There exist several open questions closely related to the work presented in this thesis, as described below.

6.2.1. Modelling biological networks

Finding a model that represents the structural features of biological networks is fundamental to understanding complex biological systems. In Chapter 3, we showed that current PPI networks of many species were better modelled by the GEO-GD, SF-GD and STICKY models than other random network models. The STICKY model appeared to be the best fitting model for the five viral PPI networks we analysed except KSHV (i.e., the PPI network of Kaposi sarcoma-associated herpes virus). Meanwhile, PPI networks of *Arabidopsis thaliana*, nematode worm and fruit fly were also best fit by the STICKY model, while the best fitting model of the human PPI network was the GEO-GD model. Investigating the biological reasons for these observations is a subject of future research and may have important implications for understanding the similarity and diversity in cellular mechanisms of different species.

Despite these advances in modelling PPI networks, we are still far away from finding the ‘right’ model where the biological data are generated from.

When comparing the current yeast and human PPI networks with the six random network models discussed in Chapter 3, none of these models perfectly fit the data since there is no overlap between the data-vs-model and model-vs-model distributions. Further refinement of random models is expected to better describe the structure of biological networks and could be used for assessing confidence levels of existing data and predicting new interactions to guide biological experiments in a time- and cost-efficient way.

The approach we used to model and compare PPI networks can be easily applied to any types of networks such as metabolic networks, gene regulation networks, social networks and economic networks to study the structure of other complex systems.

6.2.2. Integration of molecular interaction networks

In Chapter 4, we introduced a measure to estimate the similarity score between a pair of diseases based on the topological similarity of their annotated genes in a human PPI network. A potential extension of our topology-based measure would be to use topological similarities of disease genes in other molecular intersection networks along with those in the PPI network to measure disease similarities. For example, we could construct a human metabolic network, genetic interaction network, cell signalling network, gene regulation network and PPI network and compute the signature similarities of disease genes in these networks separately. The topological similarities between a pair of genes could be measured by the average or maximum of their signature similarities in different networks. Including topological information of other molecular interaction networks would provide a more comprehensive view of the structure of human interactome and might yield to better understanding of the underlying mechanisms of diseases.

Molecular interaction networks could also be integrated into the heterogeneous network discussed in Chapter 5. New algorithms and tools designed for mining the integrated network which contains different types of molecular interaction data along with disease- and drug-related data might have many potential uses such as disease gene prioritisation, disease association predication and drug repositioning.

6.2.3. Constructing and analysing weighted biological networks

In the work presented in this thesis, we modelled most of the biological data as unweighted networks in which an edge between two nodes represents the presence or absence of the interaction (or association) between the corresponding bio-entities. For some of these networks, it is possible to assign weights to their edges based on confidence levels. For example, a weighted disease-gene association network can be constructed from the GWAS data by considering the statistical significance of the associations. How to construct these weighted biological networks, how to apply graphlet-based measures to analyse their structures, and how to integrate them with other unweighed networks such as PPI networks, are open research problems that would be addressed in future research.

6.2.4. Combining similarity measures

We introduced three disease similarity measures to estimate the strength of disease-disease associations in Chapter 4. The three measures compared diseases from different aspects: shared gene annotations (annotation-based measure), common biological processes (function-based measure) and topological similarity of annotated genes (topology-based). A new measure that combines these three similarity measures may provide a more comprehensive view on disease-disease associations. One potential approach is to use supervised learning algorithms, such as Naïve Bayes classifier or support vector machine, to assign the best weights to the similarity scores computed by using the three similarity measures. Known disease-disease associations derived from standard disease classifications, comorbidity data, GWAS studies and literature co-occurrence data could be used as training examples to guide the learning process. The new measure might achieve better performances on recovering known disease associations than the three measures discussed in Chapter 4.

6.2.5. Disease reclassification

Disease-disease associations uncovered in our studies would be further used to redefine disease classification. For example, a disease-disease association network would be constructed according to the disease similarity scores,

and clustering algorithms would be applied to detect the communities in the network. Diseases that belong to the same cluster might have common underlying mechanisms and would be grouped under the same category. It is also possible to use bi-clustering algorithms to discover bi-clusters in the integrated disease network presented in Chapter 5. Bi-clustering, also known as co-clustering, is a distinct class of clustering algorithms which is able to classify the objects and features of a dataset (i.e., rows and columns of a data matrix) simultaneously [218]. Since the problem of finding a maximum size bi-cluster has been proven to be NP-complete [218], many algorithms use heuristic approaches to identify bi-clusters. For example, BiCluE [219] and Bi-Force [220] algorithms identify bi-clusters by solving the weighted bi-cluster editing problem, which transforms a given bipartite graph into a union of disjoint bi-cliques by edge insertions and deletions with minimal costs for these modifications. In the studies presented in Chapter 4 and Chapter 5, we mainly focused on quantifying the strength of disease-disease associations rather than grouping diseases, therefore we predicted disease associations by using disease similarity measures but not bi-clustering algorithms. In addition, since most bi-clustering algorithms were developed for analysing bipartite graphs, extending these algorithms for analysing heterogeneous networks which contain different types of nodes and edges is still a subject of future research.

Disease groups identified by clustering algorithms might lead to a novel integration-based disease classification. Investigating and understanding the similarity and difference between such classification and existing disease classifications such as ICD, MeSH and DO, validating the novel classification via biological experiments and using it to improve current disease diagnosis, might lead to exciting discoveries in biology and medicine.

Bibliography

- [1] W. H. Organization, *ICD-10: International statistical classification of diseases and related health problems*. World Health Organization, 2004.
- [2] F. Rogers, “Medical subject headings,” *Bulletin of the Medical Library Association*, vol. 51, p. 114, 1963.
- [3] R. A. Cote and S. Robboy, “Progress in medical information management,” *JAMA: the Journal of the American Medical Association*, vol. 243, no. 8, pp. 756–762, 1980.
- [4] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, “The Unified Medical Language System,” *Methods of Information in Medicine*, vol. 32, no. 4, pp. 281–291, 1993.
- [5] N. Pržulj, D. G. Corneil, and I. Jurisica, “Modeling interactome: scale-free or geometric?,” *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.
- [6] N. Pržulj, “Biological network comparison using graphlet degree distribution,” *Bioinformatics*, vol. 23, no. 2, pp. e177–e183, 2007.
- [7] N. Pržulj and D. J. Higham, “Modelling protein-protein interaction networks via a stickiness index,” *Journal of the Royal Society Interface*, vol. 3, no. 10, pp. 711–716, 2006.
- [8] N. Pržulj, O. Kuchaiev, A. Stevanovic, and W. Hayes, “Geometric Evolutionary Dynamics of Protein Interaction Networks,” in *Pacific Symposium on Biocomputing*, vol. 2009, pp. 178–189, World Scientific, 2010.
- [9] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, “Topological network alignment uncovers biological function and

- phylogeny,” *Journal of the Royal Society Interface*, vol. 7, no. 50, pp. 1341–1354, 2010.
- [10] T. Milenković, W. L. Ng, W. Hayes, and N. Pržulj, “Optimal network alignment with graphlet degree vectors,” *Cancer Informatics*, vol. 9, p. 121, 2010.
 - [11] O. Kuchaiev and N. Pržulj, “Integrative network alignment reveals large regions of global network similarity in yeast and human,” *Bioinformatics*, vol. 27, no. 10, pp. 1390–1396, 2011.
 - [12] V. Memišević and N. Pržulj, “C-GRAAL: Common-neighbors-based global GRAPh ALIGNment of biological networks,” *Integrative Biology*, vol. 4, no. 7, pp. 734–743, 2012.
 - [13] T. Rito, Z. Wang, C. M. Deane, and G. Reinert, “How threshold behaviour affects the use of subgraphs for network comparison,” *Bioinformatics*, vol. 26, no. 18, pp. i611–i617, 2010.
 - [14] T. Ideker and R. Sharan, “Protein networks in disease,” *Genome Research*, vol. 18, no. 4, pp. 644–652, 2008.
 - [15] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki, “Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 3, pp. 1143–1147, 2000.
 - [16] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, *et al.*, “A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*,” *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.
 - [17] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, “A comprehensive two-hybrid analysis to explore the yeast protein interactome,” *Proceedings of the National Academy of Sciences*, vol. 98, no. 8, pp. 4569–4574, 2001.
 - [18] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen,

- et al.*, “A human protein-protein interaction network: a resource for annotating the proteome,” *Cell*, vol. 122, no. 6, pp. 957–968, 2005.
- [19] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, *et al.*, “Towards a proteome-scale map of the human protein–protein interaction network,” *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.
 - [20] N. Simonis, J.-F. Rual, A.-R. Carvunis, M. Tasan, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, J. M. Sahalie, K. Venkatesan, F. Gebreab, *et al.*, “Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network,” *Nature Methods*, vol. 6, no. 1, pp. 47–54, 2008.
 - [21] M. Dreze, A.-R. Carvunis, B. Charlotiaux, M. Galli, S. J. Pevzner, M. Tasan, Y.-Y. Ahn, P. Balumuri, A.-L. Barabási, V. Bautista, *et al.*, “Evidence for network evolution in an *Arabidopsis* interactome map,” *Science*, vol. 333, no. 6042, pp. 601–607, 2011.
 - [22] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin, “A generic protein purification method for protein complex characterization and proteome exploration,” *Nature Biotechnology*, vol. 17, no. 10, pp. 1030–1032, 1999.
 - [23] A.-C. Gavin, M. Bösch, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, *et al.*, “Functional organization of the yeast proteome by systematic analysis of protein complexes,” *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.
 - [24] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dimpelfeld, *et al.*, “Proteome survey reveals modularity of the yeast cell machinery,” *Nature*, vol. 440, no. 7084, pp. 631–636, 2006.
 - [25] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, *et al.*, “Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*,” *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

- [26] E. Fossum, C. C. Friedel, S. V. Rajagopala, B. Titz, A. Baiker, T. Schmidt, T. Kraus, T. Stellberger, C. Rutenberg, S. Suthram, *et al.*, “Evolutionarily conserved herpesviral protein interaction networks,” *PLoS Pathogens*, vol. 5, no. 9, p. e1000570, 2009.
- [27] J. M. Peregrín-Alvarez, X. Xiong, C. Su, and J. Parkinson, “The modular organization of protein interactions in *Escherichia coli*,” *PLoS Computational Biology*, vol. 5, no. 10, p. e1000523, 2009.
- [28] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. J. Han, A. Chesneau, T. Hao, *et al.*, “A map of the interactome network of the metazoan *C. elegans*,” *Science*, vol. 303, no. 5657, pp. 540–543, 2004.
- [29] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. Hao, C. Ooi, B. Godwin, E. Vitols, *et al.*, “A protein interaction map of *Drosophila melanogaster*,” *Science*, vol. 302, no. 5651, pp. 1727–1736, 2003.
- [30] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “BioGRID: a general repository for interaction datasets,” *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.
- [31] T. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, *et al.*, “Human protein reference database—2009 update,” *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D767–D772, 2009.
- [32] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.-W. Mewes, *et al.*, “The MIPS mammalian protein–protein interaction database,” *Bioinformatics*, vol. 21, no. 6, pp. 832–834, 2005.
- [33] A. Chatr-Aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni, “MINT: the Molecular INTeraction database,” *Nucleic Acids Research*, vol. 35, no. suppl 1, pp. D572–D574, 2007.

- [34] K. R. Brown and I. Jurisica, “Online predicted human interaction database,” *Bioinformatics*, vol. 21, no. 9, pp. 2076–2082, 2005.
- [35] K. R. Brown and I. Jurisica, “Unequal evolutionary conservation of human protein interactions in interologous networks,” *Genome Biology*, vol. 8, no. 5, p. R95, 2007.
- [36] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, “The database of interacting proteins: 2004 update,” *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D449–D451, 2004.
- [37] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del Toro, *et al.*, “The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases,” *Nucleic Acids Research*, p. gkt1115, 2013.
- [38] M. P. Stumpf, W. P. Kelly, T. Thorne, and C. Wiuf, “Evolution at the system level: the natural history of protein interaction networks,” *Trends in Ecology & Evolution*, vol. 22, no. 7, pp. 366–373, 2007.
- [39] V. Memišević, T. Milenković, and N. Pržulj, “Complementarity of network and sequence structure in homologous proteins,” *Journal of Integrative Bioinformatics*, vol. 7, no. 3, p. 135, 2010.
- [40] T. Milenković, V. Memišević, A. K. Ganesan, and N. Pržulj, “Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data,” *Journal of The Royal Society Interface*, vol. 7, no. 44, pp. 423–437, 2010.
- [41] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, “Network medicine: a network-based approach to human disease,” *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.
- [42] V. Janjić and N. Pržulj, “The core diseasome,” *Molecular BioSystems*, vol. 8, no. 10, pp. 2614–2625, 2012.
- [43] N. Pržulj, D. A. Wigle, and I. Jurisica, “Functional topology in a network of protein interactions,” *Bioinformatics*, vol. 20, no. 3, pp. 340–348, 2004.

- [44] H. N. Chua, W.-K. Sung, and L. Wong, “Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions,” *Bioinformatics*, vol. 22, no. 13, pp. 1623–1630, 2006.
- [45] R. Sharan, I. Ulitsky, and R. Shamir, “Network-based prediction of protein function,” *Molecular Systems Biology*, vol. 3, no. 1, 2007.
- [46] T. Milenković and N. Pržulj, “Uncovering biological network function via graphlet degree signatures,” *Cancer Informatics*, vol. 6, p. 257, 2008.
- [47] N. Pržulj, *Biological networks uncover evolution , disease , and gene functions*, pp. 1–31. 2011.
- [48] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, “Cytoscape 2.8: new features for data integration and network visualization,” *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2011.
- [49] R. Mrowka, A. Patzak, and H. Herzog, “Is there a bias in proteome research?,” *Genome Research*, vol. 11, no. 12, pp. 1971–1973, 2001.
- [50] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein, “Bridging structural biology and genomics: assessing protein interaction data with known complexes,” *TRENDS in Genetics*, vol. 18, no. 10, pp. 529–536, 2002.
- [51] H. Huang, B. M. Jedynek, and J. S. Bader, “Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps,” *PLoS Computational Biology*, vol. 3, no. 11, p. e214, 2007.
- [52] J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal, “Effect of sampling on topology predictions of protein-protein interaction networks,” *Nature Biotechnology*, vol. 23, no. 7, pp. 839–844, 2005.
- [53] L. Hakes, J. W. Pinney, D. L. Robertson, and S. C. Lovell, “Protein-protein interaction networks and biology—what’s the connection?,” *Nature Biotechnology*, vol. 26, no. 1, pp. 69–72, 2008.

- [54] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, “KEGG for integration and interpretation of large-scale molecular data sets,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D109–D114, 2012.
- [55] D. Croft, G. O’Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, *et al.*, “Reactome: a database of reactions, pathways and biological processes,” *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D691–D697, 2011.
- [56] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico, “WikiPathways: building research communities on biological pathways,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D1301–D1307, 2012.
- [57] Y. Li and P. Agarwal, “A pathway-based view of human diseases and disease relationships,” *PloS One*, vol. 4, no. 2, p. e4346, 2009.
- [58] M. L. Gatz, J. E. Lucas, W. T. Barry, J. W. Kim, Q. Wang, M. D. Crawford, M. B. Datto, M. Kelley, B. Mathey-Prevot, A. Potti, *et al.*, “A pathway-based classification of human breast cancer,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 15, pp. 6994–6999, 2010.
- [59] J. Zhao, H. Yu, J. Luo, Z. Cao, and Y. Li, “Complex networks theory for analyzing metabolic networks,” *Chinese Science Bulletin*, vol. 51, no. 13, pp. 1529–1537, 2006.
- [60] H. Ma and A.-P. Zeng, “Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms,” *Bioinformatics*, vol. 19, no. 2, pp. 270–277, 2003.
- [61] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: simple building blocks of complex networks,” *Science*, vol. 298, no. 5594, pp. 824–827, 2002.
- [62] R. Rottger, U. Ruckert, J. Taubert, and J. Baumbach, “How little do we actually know? On the size of gene regulatory networks,” *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 9, no. 5, pp. 1293–1300, 2012.

- [63] M. Gage, D. Wattendorf, and L. Henry, “Translational advances regarding hereditary breast cancer syndromes,” *Journal of Surgical Oncology*, vol. 105, no. 5, pp. 444–451, 2012.
- [64] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders,” *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D514–D517, 2005.
- [65] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen, “A text-mining analysis of the human phenome,” *European Journal of Human Genetics*, vol. 14, no. 5, pp. 535–542, 2006.
- [66] K. Lage, E. O. Karlberg, Z. M. Størling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, *et al.*, “A human phenome-interactome network of protein complexes implicated in genetic disorders,” *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.
- [67] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi, “The human disease network,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.
- [68] A. P. Davis, B. L. King, S. Mockus, C. G. Murphy, C. Saraceni-Richards, M. Rosenstein, T. Wieggers, and C. J. Mattingly, “The comparative toxicogenomics database: update 2011,” *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D1067–D1072, 2011.
- [69] M. Whirl-Carrillo, E. McDonagh, J. Hebert, L. Gong, K. Sangkuhl, C. Thorn, R. Altman, and T. E. Klein, “Pharmacogenomics knowledge for personalized medicine,” *Clinical Pharmacology & Therapeutics*, vol. 92, no. 4, pp. 414–417, 2012.
- [70] J. N. Hirschhorn and M. J. Daly, “Genome-wide association studies for common diseases and complex traits,” *Nature Reviews Genetics*, vol. 6, no. 2, pp. 95–108, 2005.
- [71] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, *et al.*, “Genome-wide association study of 14,000 cases

- of seven common diseases and 3,000 shared controls,” *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.
- [72] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn, “Genome-wide association studies for complex traits: consensus, uncertainty and challenges,” *Nature Reviews Genetics*, vol. 9, no. 5, pp. 356–369, 2008.
 - [73] L. A. Hindorff, H. A. Junkins, J. Mehta, T. Manolio, *et al.*, “A catalog of published genome-wide association studies,” 2011.
 - [74] M. J. Li, P. Wang, X. Liu, E. L. Lim, Z. Wang, M. Yeager, M. P. Wong, P. C. Sham, S. J. Chanock, and J. Wang, “GWASdb: a database for human genetic variants identified by genome-wide association studies,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D1047–D1054, 2012.
 - [75] A. L. Hopkins and C. R. Groom, “The druggable genome,” *Nature Reviews Drug discovery*, vol. 1, no. 9, pp. 727–730, 2002.
 - [76] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, *et al.*, “DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs,” *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D1035–D1041, 2011.
 - [77] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, *et al.*, “ChEMBL: a large-scale bioactivity database for drug discovery,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D1100–D1107, 2012.
 - [78] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, “Drug—target network,” *Nature Biotechnology*, vol. 25, no. 10, p. 1119, 2007.
 - [79] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, “Drug target identification using side-effect similarity,” *Science*, vol. 321, no. 5886, pp. 263–266, 2008.
 - [80] X. Chen, M.-X. Liu, and G.-Y. Yan, “Drug-target interaction prediction by random walk on the heterogeneous network,” *Molecular BioSystems*, vol. 8, no. 7, pp. 1970–1978, 2012.

- [81] G. Hu and P. Agarwal, “Human disease-drug network based on genomic expression profiles,” *PLoS One*, vol. 4, no. 8, p. e6536, 2009.
- [82] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang, “Prediction of drug-target interactions and drug repositioning via network-based inference,” *PLoS Computational Biology*, vol. 8, no. 5, p. e1002503, 2012.
- [83] S. Daminelli, V. J. Haupt, M. Reimann, and M. Schroeder, “Drug repositioning through incomplete bi-cliques in an integrated drug–target–disease network,” *Integrative Biology*, vol. 4, no. 7, pp. 778–788, 2012.
- [84] P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon, and V. Mooser, “Use of genome-wide association studies for drug repositioning,” *Nature Biotechnology*, vol. 30, no. 4, pp. 317–320, 2012.
- [85] L. Brouwers, M. Iskar, G. Zeller, V. van Noort, and P. Bork, “Network neighbors of drug targets contribute to drug side-effect similarity,” *PLoS One*, vol. 6, no. 7, p. e22187, 2011.
- [86] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, *et al.*, “Large-scale prediction and testing of drug activity on side-effect targets,” *Nature*, vol. 486, no. 7403, pp. 361–367, 2012.
- [87] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, “Network motifs in the transcriptional regulation network of *Escherichia coli*,” *Nature Genetics*, vol. 31, no. 1, pp. 64–68, 2002.
- [88] T. Milenković, V. Memišević, A. Bonato, and N. Pržulj, “Dominating biological networks,” *PloS One*, vol. 6, no. 8, p. e23016, 2011.
- [89] S. A. Cook, “The complexity of theorem-proving procedures,” in *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, pp. 151–158, ACM, 1971.
- [90] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

- [91] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [92] A. Clauset, C. R. Shalizi, and M. E. Newman, “Power-law distributions in empirical data,” *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.
- [93] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, “Superfamilies of evolved and designed networks,” *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004.
- [94] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone, “Comment on” Network motifs: simple building blocks of complex networks” and” Superfamilies of evolved and designed networks”,” *Science*, vol. 305, no. 5687, pp. 1107–1107, 2004.
- [95] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, and U. Alon, “Response to Comment on ”Network Motifs: Simple Building Blocks of Complex Networks” and ”Superfamilies of Evolved and Designed Networks”,” *Science*, vol. 305, no. 5687, p. 1107, 2004.
- [96] P. Erdős and A. Rényi, “On random graphs,” *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.
- [97] P. Erdős and A. Rényi, “On the evolution of random graphs,” *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, vol. 5, pp. 17–61, 1960.
- [98] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, “Modeling of protein interaction networks,” *Complexus*, vol. 1, no. 1, pp. 38–44, 2002.
- [99] M. Penrose, *Random geometric graphs*, vol. 5. Oxford University Press Oxford, 2003.
- [100] M. Lappe and L. Holm, “Unraveling protein interaction networks with near-optimal efficiency,” *Nature Biotechnology*, vol. 22, no. 1, pp. 98–103, 2003.
- [101] O. Kuchaiev, M. Rašajski, D. J. Higham, and N. Pržulj, “Geometric de-noising of protein-protein interaction networks,” *PLoS Computational Biology*, vol. 5, no. 8, p. e1000454, 2009.

- [102] B. Bollobás, *Random Graphs*, vol. 73. Cambridge university press, 2001.
- [103] M. E. Newman, “The structure and function of complex networks,” *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.
- [104] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, “Lethality and centrality in protein networks,” *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.
- [105] S. Maslov and K. Sneppen, “Specificity and stability in topology of protein networks,” *Science*, vol. 296, no. 5569, pp. 910–913, 2002.
- [106] W. Hayes, K. Sun, and N. Pržulj, “Graphlet-based measures are suitable for biological network comparison,” *Bioinformatics*, vol. 29, no. 4, pp. 483–491, 2013.
- [107] A. Sarajlić, V. Janjić, N. Stojković, D. Radak, and N. Pržulj, “Network Topology Reveals Key Cardiovascular Disease Genes,” *PloS One*, vol. 8, no. 8, p. e71537, 2013.
- [108] N. Malod-Dognin and N. Pržulj, “GR-Align: fast and flexible alignment of protein 3D structures using graphlet degree similarity,” *Bioinformatics*, vol. 30, no. 9, pp. 1259–1265, 2014.
- [109] B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker, “PathBLAST: a tool for alignment of protein interaction networks,” *Nucleic Acids Research*, vol. 32, no. suppl 2, pp. W83–W88, 2004.
- [110] M. Kalaev, M. Smoot, T. Ideker, and R. Sharan, “NetworkBLAST: comparative analysis of protein networks,” *Bioinformatics*, vol. 24, no. 4, pp. 594–596, 2008.
- [111] M. Kalaev, V. Bafna, and R. Sharan, “Fast and accurate alignment of multiple protein networks,” *Journal of Computational Biology*, vol. 16, no. 8, pp. 989–999, 2009.
- [112] M. Koyutürk, Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama, “Pairwise alignment of protein interaction networks,” *Journal of Computational Biology*, vol. 13, no. 2, pp. 182–199, 2006.

- [113] Z. Liang, M. Xu, M. Teng, and L. Niu, “NetAlign: a web-based tool for comparison of protein interaction networks,” *Bioinformatics*, vol. 22, no. 17, pp. 2175–2177, 2006.
- [114] J. Flannick, A. Novak, B. S. Srinivasan, H. H. McAdams, and S. Batzoglou, “Graemlin: general and robust alignment of multiple large interaction networks,” *Genome Research*, vol. 16, no. 9, pp. 1169–1181, 2006.
- [115] R. Singh, J. Xu, and B. Berger, “Pairwise global alignment of protein interaction networks by matching neighborhood topology,” in *Research in Computational Molecular Biology*, pp. 16–31, Springer, 2007.
- [116] R. Singh, J. Xu, and B. Berger, “Global alignment of multiple protein interaction networks,” in *Pacific Symposium on Biocomputing*, vol. 13, pp. 303–314, Citeseer, 2008.
- [117] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, “IsoRankN: spectral methods for global alignment of multiple protein networks,” *Bioinformatics*, vol. 25, no. 12, pp. i253–i258, 2009.
- [118] J. Flannick, A. Novak, C. B. Do, B. S. Srinivasan, and S. Batzoglou, “Automatic parameter learning for multiple network alignment,” in *Research in Computational Molecular Biology*, pp. 214–231, Springer, 2008.
- [119] M. Zaslavskiy, F. Bach, and J.-P. Vert, “Global alignment of protein–protein interaction networks by graph matching methods,” *Bioinformatics*, vol. 25, no. 12, pp. i259–i267, 2009.
- [120] W. Tian and N. F. Samatova, “Pairwise alignment of interaction networks by fast identification of maximal conserved patterns,” in *Pacific Symposium on Biocomputing*, vol. 14, pp. 99–110, Citeseer, 2009.
- [121] G. W. Klau, “A new graph-based method for pairwise global network alignment,” *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S59, 2009.
- [122] M. El-Kebir, J. Heringa, and G. W. Klau, “Lagrangian relaxation applied to sparse global network alignment,” in *Pattern Recognition in Bioinformatics*, pp. 225–236, Springer, 2011.

- [123] L. Chindelevitch, C.-S. Liao, and B. Berger, “Local optimization for global alignment of protein interaction networks,” in *Pacific Symposium on Biocomputing*, vol. 15, pp. 123–132, World Scientific, 2010.
- [124] R. Ibragimov, M. Malek, J. Guo, and J. Baumbach, “GEDEVO: an evolutionary graph edit distance algorithm for biological network alignment,” in *German Conference on Bioinformatics 2013*, vol. 34, pp. 68–79, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013.
- [125] A. E. Aladağ and C. Erten, “SPINAL: scalable protein interaction network alignment,” *Bioinformatics*, vol. 29, no. 7, pp. 917–924, 2013.
- [126] B. Neyshabur, A. Khadem, S. Hashemifar, and S. S. Arab, “NETAL: a new graph-based method for global alignment of protein–protein interaction networks,” *Bioinformatics*, vol. 29, no. 13, pp. 1654–1662, 2013.
- [127] T. Milenković, J. Lai, and N. Pržulj, “GraphCrunch: a tool for large network analyses,” *BMC Bioinformatics*, vol. 9, no. 1, p. 70, 2008.
- [128] O. Kuchaiev, A. Stevanović, W. Hayes, and N. Pržulj, “GraphCrunch 2: Software tool for network modeling, alignment and clustering,” *BMC Bioinformatics*, vol. 12, no. 1, p. 24, 2011.
- [129] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe, “Disease Ontology: a backbone for disease semantic integration,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D940–D946, 2012.
- [130] X. Jiang, B. Liu, J. Jiang, H. Zhao, M. Fan, J. Zhang, Z. Fan, and T. Jiang, “Modularity in the genetic disease-phenotype network,” *FEBS Letters*, vol. 582, no. 17, pp. 2549–2554, 2008.
- [131] R. Xu, L. Li, and Q. Wang, “Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature,” *Bioinformatics*, vol. 29, no. 17, pp. 2186–2194, 2013.
- [132] C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis, “A dynamic network approach for the study of human phenotypes,” *PLoS Computational Biology*, vol. 5, no. 4, p. e1000353, 2009.

- [133] J. Park, D.-S. Lee, N. A. Christakis, and A.-L. Barabási, “The impact of cellular networks on disease comorbidity,” *Molecular Systems Biology*, vol. 5, no. 1, 2009.
- [134] S. Park, J.-S. Yang, J. Kim, Y.-E. Shin, J. Hwang, J. Park, S. K. Jang, and S. Kim, “Evolutionary history of human disease genes reveals phenotypic connections and comorbidity among genetic diseases,” *Scientific Reports*, vol. 2, 2012.
- [135] A. Rzhetsky, D. Wajngurt, N. Park, and T. Zheng, “Probing genetic overlap among complex human phenotypes,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 28, pp. 11694–11699, 2007.
- [136] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søbey, S. Bredkjær, A. Juul, T. Werge, *et al.*, “Using electronic patient records to discover disease correlations and stratify patient cohorts,” *PLoS Computational Biology*, vol. 7, no. 8, p. e1002141, 2011.
- [137] W. Huang, P. Wang, Z. Liu, and L. Zhang, “Identifying disease associations via genome-wide association studies,” *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S68, 2009.
- [138] S. N. Lewis, E. Nsoesie, C. Weeks, D. Qiao, and L. Zhang, “Prediction of disease and phenotype associations from genome-wide association studies,” *PloS one*, vol. 6, no. 11, p. e27175, 2011.
- [139] M. Žitnik, V. Janjić, C. Larminie, B. Zupan, and N. Pržulj, “Discovering disease-disease associations by fusing systems-level molecular data,” *Scientific Reports*, vol. 3, 2013.
- [140] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte, “Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets,” *PLoS Computational Biology*, vol. 6, no. 2, p. e1000662, 2010.
- [141] S. Mathur and D. Dinakarpanthian, “Finding disease similarity based on implicit semantic similarity,” *Journal of Biomedical Informatics*, vol. 45, no. 2, pp. 363–371, 2012.

- [142] J. Li, B. Gong, X. Chen, T. Liu, C. Wu, F. Zhang, C. Li, X. Li, S. Rao, and X. Li, “DOSim: An R package for similarity between diseases based on Disease Ontology,” *BMC Bioinformatics*, vol. 12, no. 1, p. 266, 2011.
- [143] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, “The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease,” *The American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008.
- [144] D.-S. Lee, J. Park, K. Kay, N. Christakis, Z. Oltvai, and A.-L. Barabási, “The implications of human metabolic network topology for disease comorbidity,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 29, pp. 9880–9885, 2008.
- [145] D. A. Davis and N. V. Chawla, “Exploring and exploiting disease interactions from multi-relational gene and phenotype networks,” *PloS One*, vol. 6, no. 7, p. e22670, 2011.
- [146] J. R. Parrish, J. Yu, G. Liu, J. A. Hines, J. E. Chan, B. A. Mangiola, H. Zhang, S. Pacifico, F. Fotouhi, V. J. DiRita, *et al.*, “A proteome-wide protein interaction map for *Campylobacter jejuni*,” *Genome Biology*, vol. 8, no. 7, p. R130, 2007.
- [147] Y. Shimoda, S. Shinpo, M. Kohara, Y. Nakamura, S. Tabata, and S. Sato, “A large scale analysis of protein–protein interactions in the nitrogen-fixing bacterium *Mesorhizobium loti*,” *DNA Research*, vol. 15, no. 1, pp. 13–23, 2008.
- [148] S. Sato, Y. Shimoda, A. Muraki, M. Kohara, Y. Nakamura, and S. Tabata, “A large-scale protein–protein interaction analysis in *Synechocystis* sp. PCC6803,” *DNA Research*, vol. 14, no. 5, pp. 207–216, 2007.
- [149] P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle, and S. D. Mooney, “An integrated approach to inferring gene-disease associations in humans,” *Proteins: Structure, Function, and Bioinformatics*, vol. 72, no. 3, pp. 1030–1037, 2008.

- [150] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. Gandhi, K. Chandrika, N. Deshpande, S. Suresh, *et al.*, “Human protein reference database as a discovery resource for proteomics,” *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D497–D501, 2004.
- [151] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, “Comparative assessment of large-scale data sets of protein-protein interactions,” *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.
- [152] M. P. Stumpf, C. Wiuf, and R. M. May, “Subnets of scale-free networks are not scale-free: sampling properties of networks,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 12, pp. 4221–4224, 2005.
- [153] S. R. Collins, P. Kemmeren, X.-C. Zhao, J. F. Greenblatt, F. Spencer, F. C. Holstege, J. S. Weissman, and N. J. Krogan, “Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*,” *Molecular & Cellular Proteomics*, vol. 6, no. 3, pp. 439–450, 2007.
- [154] R. Colak, F. Hormozdiari, F. Moser, A. Schönhuth, J. Holman, M. Ester, and S. C. Sahinalp, “Dense graphlet statistics of protein interaction and random networks,” in *Pacific Symposium on Biocomputing*, vol. 548, pp. 178–189, Citeseer, 2009.
- [155] T. A. Gibson and D. S. Goldberg, “Improving evolutionary models of protein interaction networks,” *Bioinformatics*, vol. 27, no. 3, pp. 376–382, 2011.
- [156] E. Hartuv and R. Shamir, “A clustering algorithm based on graph connectivity,” *Information Processing Letters*, vol. 76, no. 4, pp. 175–181, 2000.
- [157] J. Loscalzo, I. Kohane, and A.-L. Barabasi, “Human disease classification in the postgenomic era: a complex systems approach to human pathobiology,” *Molecular Systems Biology*, vol. 3, no. 1, 2007.

- [158] S. Kim, K.-A. Sohn, and E. P. Xing, “A multivariate regression approach to association analysis of a quantitative trait network,” *Bioinformatics*, vol. 25, no. 12, pp. i204–i212, 2009.
- [159] A. Bauer-Mehren, M. Bundschuh, M. Rautschka, M. A. Mayer, F. Sanz, and L. I. Furlong, “Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases,” *PloS One*, vol. 6, no. 6, p. e20284, 2011.
- [160] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, “The genetic association database,” *Nature Genetics*, vol. 36, no. 5, pp. 431–432, 2004.
- [161] M. Oti, M. A. Huynen, and H. G. Brunner, “Phenome connections,” *Trends in Genetics*, vol. 24, no. 3, pp. 103–106, 2008.
- [162] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, “Gene Ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [163] J. Osborne, J. Flatow, M. Holko, S. Lin, W. Kibbe, L. Zhu, M. Danila, G. Feng, and R. Chisholm, “Annotating the human genome with Disease Ontology,” *BMC Genomics*, vol. 10, no. Suppl 1, p. S6, 2009.
- [164] W. Yu, M. Clyne, M. J. Khoury, and M. Gwinn, “Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations,” *Bioinformatics*, vol. 26, no. 1, pp. 145–146, 2010.
- [165] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
- [166] D. M. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
- [167] Y. Hochberg and Y. Benjamini, “More powerful procedures for multiple significance testing,” *Statistics in Medicine*, vol. 9, no. 7, pp. 811–818, 1990.

- [168] R. Bender and S. Lange, “Multiple test procedures other than Bonferroni deserve wider use,” *BMJ: British Medical Journal*, vol. 318, no. 7183, p. 600, 1999.
- [169] A. Ziegler, I. R. König, and J. R. Thompson, “Biostatistical aspects of genome-wide association studies,” *Biometrical Journal*, vol. 50, no. 1, pp. 8–28, 2008.
- [170] M. R. Stratton, P. J. Campbell, and P. A. Futreal, “The cancer genome,” *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.
- [171] S. Chavali, F. Barrenas, K. Kanduri, and M. Benson, “Network properties of human disease genes with pleiotropic effects,” *BMC Systems Biology*, vol. 4, no. 1, p. 78, 2010.
- [172] R. Samreen, “Diabetes mellitus: A review,” *Scientific Research and Essays*, vol. 4, no. 5, pp. 367–373, 2009.
- [173] R. H. Duerr, K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhart, C. Abraham, M. Regueiro, A. Griffiths, *et al.*, “A genome-wide association study identifies IL23R as an inflammatory bowel disease gene,” *Science*, vol. 314, no. 5804, pp. 1461–1463, 2006.
- [174] R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, *et al.*, “A genome-wide association study identifies novel risk loci for type 2 diabetes,” *Nature*, vol. 445, no. 7130, pp. 881–885, 2007.
- [175] J. Quackenbush, “Computational analysis of microarray data,” *Nature Reviews Genetics*, vol. 2, no. 6, pp. 418–427, 2001.
- [176] A. Butte, “The use and analysis of microarray data,” *Nature Reviews Drug Discovery*, vol. 1, no. 12, pp. 951–960, 2002.
- [177] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin, “GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways,” *Nature Genetics*, vol. 31, no. 1, pp. 19–20, 2002.

- [178] S. Schuster, D. A. Fell, and T. Dandekar, “A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks,” *Nature Biotechnology*, vol. 18, no. 3, pp. 326–332, 2000.
- [179] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, *et al.*, “KEGG for linking genomes to life and the environment,” *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D480–D484, 2008.
- [180] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays,” *Genome Research*, vol. 18, no. 9, pp. 1509–1517, 2008.
- [181] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.
- [182] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.
- [183] L. Franke, H. v. Bakel, L. Fokkens, E. D. De Jong, M. Egmont-Petersen, and C. Wijmenga, “Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes,” *The American Journal of Human Genetics*, vol. 78, no. 6, pp. 1011–1025, 2006.
- [184] B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia, C. DeLisi, *et al.*, “Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network,” *Genome Biology*, vol. 10, no. 9, p. R91, 2009.
- [185] J. Ahn, Y. Yoon, C. Park, E. Shin, and S. Park, “Integrative gene network construction for predicting a set of complementary prostate cancer genes,” *Bioinformatics*, vol. 27, no. 13, pp. 1846–1853, 2011.
- [186] N. Alcaraz, T. Friedrich, T. Kötzing, A. Krohmer, J. Müller, J. Pauling, and J. Baumbach, “Efficient key pathway mining: combining net-

- works and OMICS data,” *Integrative Biology*, vol. 4, no. 7, pp. 756–764, 2012.
- [187] A. C. Haugen, R. Kelley, J. B. Collins, C. J. Tucker, C. Deng, C. A. Afshari, J. M. Brown, T. Ideker, and B. Van Houten, “Integrating phenotypic and expression profiles to map arsenic-response networks,” *Genome Biology*, vol. 5, no. 12, p. R95, 2004.
 - [188] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir, “Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2981–2986, 2004.
 - [189] R. Kelley and T. Ideker, “Systematic interpretation of genetic interactions using protein networks,” *Nature Biotechnology*, vol. 23, no. 5, pp. 561–566, 2005.
 - [190] N. Alcaraz, H. Küçük, J. Weile, A. Wipat, and J. Baumbach, “Key-PathwayMiner: detecting case-specific biological pathways using expression data,” *Internet Mathematics*, vol. 7, no. 4, pp. 299–313, 2011.
 - [191] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, *et al.*, “Gene prioritization through genomic data fusion,” *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.
 - [192] C. Huttenhower, E. M. Haley, M. A. Hibbs, V. Dumeaux, D. R. Barrett, H. A. Coller, and O. G. Troyanskaya, “Exploring the human genome with functional maps,” *Genome Research*, vol. 19, no. 6, pp. 1093–1106, 2009.
 - [193] Y. Li and J. C. Patra, “Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network,” *Bioinformatics*, vol. 26, no. 9, pp. 1219–1224, 2010.
 - [194] J. von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne, and R. Preissner, “PROMISCUOUS: a database for network-based drug-repositioning,” *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D1060–D1066, 2011.

- [195] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.
- [196] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
- [197] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.
- [198] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.
- [199] J. Craig *et al.*, "Complex diseases: Research and applications," *Nature Education*, vol. 1, no. 1, 2008.
- [200] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D54–D58, 2005.
- [201] D. C. Baumgart and W. J. Sandborn, "Crohn's disease," *The Lancet*, 2012.
- [202] R. B. Sartor, "Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis," *Nature Clinical Practice Gastroenterology & Hepatology*, vol. 3, no. 7, pp. 390–407, 2006.
- [203] J. H. Cho and S. R. Brant, "Recent insights into the genetics of inflammatory bowel disease," *Gastroenterology*, vol. 140, no. 6, pp. 1704–1712, 2011.
- [204] D. C. Baumgart and S. R. Carding, "Inflammatory bowel disease: cause and immunobiology," *The Lancet*, vol. 369, no. 9573, pp. 1627–1640, 2007.

- [205] M. Feldmann and R. N. Maini, “TNF defined as a therapeutic target for rheumatoid arthritis and other autoimmune diseases,” *Nature Medicine*, vol. 9, no. 10, pp. 1245–1250, 2003.
- [206] G. Sethi, B. Sung, A. B. Kunnumakkara, and B. B. Aggarwal, “Targeting TNF for treatment of cancer and autoimmunity,” in *Therapeutic Targets of the TNF Superfamily*, pp. 37–51, Springer, 2009.
- [207] V. J. Haupt and M. Schroeder, “Old friends in new guise: repositioning of known drugs with structural bioinformatics,” *Briefings in Bioinformatics*, vol. 12, no. 4, pp. 312–326, 2011.
- [208] J. T. Dudley, T. Deshpande, and A. J. Butte, “Exploiting drug-disease relationships for computational drug repositioning,” *Briefings in Bioinformatics*, vol. 12, no. 4, pp. 303–311, 2011.
- [209] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijter, R. C. Matos, T. B. Tran, *et al.*, “Predicting new molecular targets for known drugs,” *Nature*, vol. 462, no. 7270, pp. 175–181, 2009.
- [210] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, “PREDICT: a method for inferring novel drug indications with application to personalized medicine,” *Molecular Systems Biology*, vol. 7, no. 1, 2011.
- [211] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, *et al.*, “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease,” *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.
- [212] Y. Pan, T. Cheng, Y. Wang, and S. H. Bryant, “Pathway analysis for drug repositioning based on public database mining,” *Journal of Chemical Information and Modeling*, 2014.
- [213] J. Jung and D. Lee, “Inferring disease association using clinical factors in a combinatorial manner and their use in drug repositioning,” *Bioinformatics*, vol. 29, no. 16, pp. 2017–2023, 2013.
- [214] L. Yang and P. Agarwal, “Systematic drug repositioning based on clinical side-effects,” *PloS One*, vol. 6, no. 12, p. e28025, 2011.

- [215] H. Ye, Q. Liu, and J. Wei, “Construction of drug network based on side effects and its application for drug repositioning,” *PloS One*, vol. 9, no. 2, p. e87864, 2014.
- [216] L. M. Felipez, R. Gokhale, M. P. Tierney, and B. S. Kirschner, “Thalidomide use and outcomes in pediatric patients with Crohn disease refractory to infliximab and adalimumab,” *Journal of Pediatric Gastroenterology and Nutrition*, vol. 54, no. 1, pp. 28–33, 2012.
- [217] M. Lazzerini, S. Martelossi, G. Magazzù, S. Pellegrino, M. C. Lucanto, A. Barabino, A. Calvi, S. Arrigo, P. Lionetti, M. Lorusso, *et al.*, “Effect of Thalidomide on Clinical Remission in Children and Adolescents With Refractory Crohn DiseaseA Randomized Clinical TrialThalidomide in Pediatric Crohn DiseaseThalidomide in Pediatric Crohn Disease,” *JAMA*, vol. 310, no. 20, pp. 2164–2173, 2013.
- [218] S. C. Madeira and A. L. Oliveira, “Biclustering algorithms for biological data analysis: a survey,” *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 1, no. 1, pp. 24–45, 2004.
- [219] P. Sun, J. Guo, and J. Baumbach, “BiCluE - Exact and heuristic algorithms for weighted bi-cluster editing of biomedical data,” in *BMC Proceedings*, vol. 7, pp. 1–9, Springer, 2013.
- [220] P. Sun, N. K. Speicher, R. Röttger, J. Guo, and J. Baumbach, “Bi-Force: large-scale bicluster editing and its application to gene expression data biclustering,” *Nucleic Acids Research*, vol. 42, no. 9, p. e78, 2014.

A. Appendix

A.1. Supplementary information for Chapter 2

ICD-9 codes	Diseases
001-139	Infectious and parasitic diseases
140-239	Neoplasms
240-279	Endocrine, nutritional and metabolic diseases, and immunity disorders
280-289	Diseases of the blood and blood-forming organs
290-319	Mental disorders
320-389	Diseases of the nervous system and sense organs
390-459	Diseases of the circulatory system
460-519	Diseases of the respiratory system
520-579	Diseases of the digestive system
580-629	Diseases of the genitourinary system
630-679	Complications of pregnancy, childbirth, and the puerperium
680-709	Diseases of the skin and subcutaneous tissue
710-739	Diseases of the musculoskeletal system and connective tissue
740-759	Congenital anomalies
760-779	Certain conditions originating in the perinatal period
780-799	Symptoms, signs, and ill-defined conditions
800-999	Injury and poisoning
V01-V91	Supplementary classification of factors influencing health status and contact with health services
E000-E999	Supplementary classification of external causes of injury and poisoning

Table A.1.: Three-digit level ICD-9 classification¹.

¹<http://www.icd9data.com/2014/Volume1/default.htm>

MeSH tree ID	Diseases
C01	Bacterial infections and mycoses
C02	Virus diseases
C03	Parasitic diseases
C04	Neoplasms
C05	Musculoskeletal diseases
C06	Digestive system diseases
C07	Stomatognathic diseases
C08	Respiratory Tract diseases
C09	Otorhinolaryngologic diseases
C10	Nervous system diseases
C11	Eye diseases
C12	Male urogenital diseases
C13	Female urogenital diseases and pregnancy complications
C14	Cardiovascular diseases
C15	Hemic and lymphatic diseases
C16	Congenital, hereditary, and neonatal diseases and abnormalities
C17	Skin and connective tissue diseases
C18	Nutritional and metabolic diseases
C19	Endocrine system diseases
C20	Immune system diseases
C21	Disorders of environmental origin
C22	Animal diseases
C23	Pathological conditions, signs and symptoms
C24	Occupational diseases
C25	Chemically-Induced disorders
C26	Wounds and injuries
F03	Mental disorders

Table A.2.: Top level MeSH tree categories². Only MeSH categories that are related to human diseases are listed above.

²<http://www.nlm.nih.gov/mesh/trees.html>

A.2. Supplementary information for Chapter 3

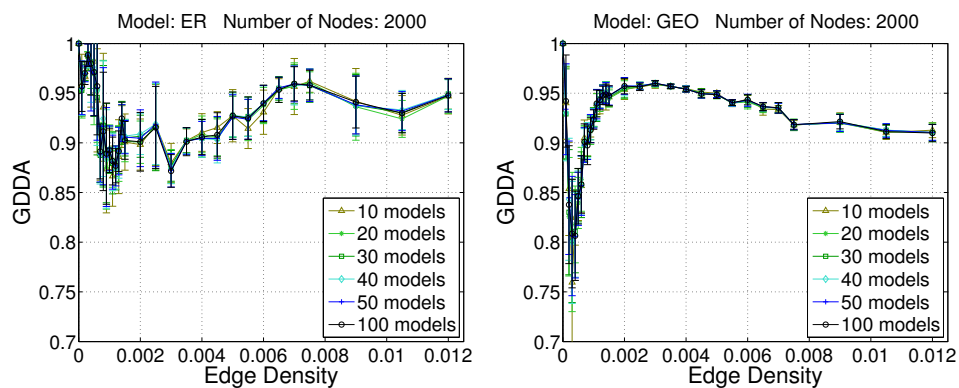


Figure A.1.: GDD agreement vs. edge density when comparing 10, 20, 30, 40, 50 and 100 random model networks to each other with 2000 nodes for ER models (left) and GEO models (right). Error bars are standard deviations of GDD agreement scores.

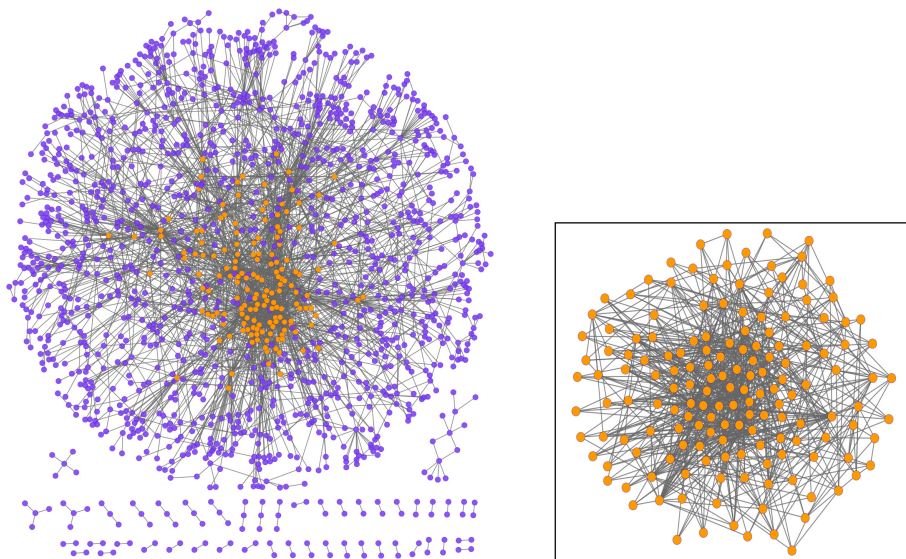


Figure A.2.: Visualisation of the human PPI network HS (left) and one of its highly connected subgraphs (right).

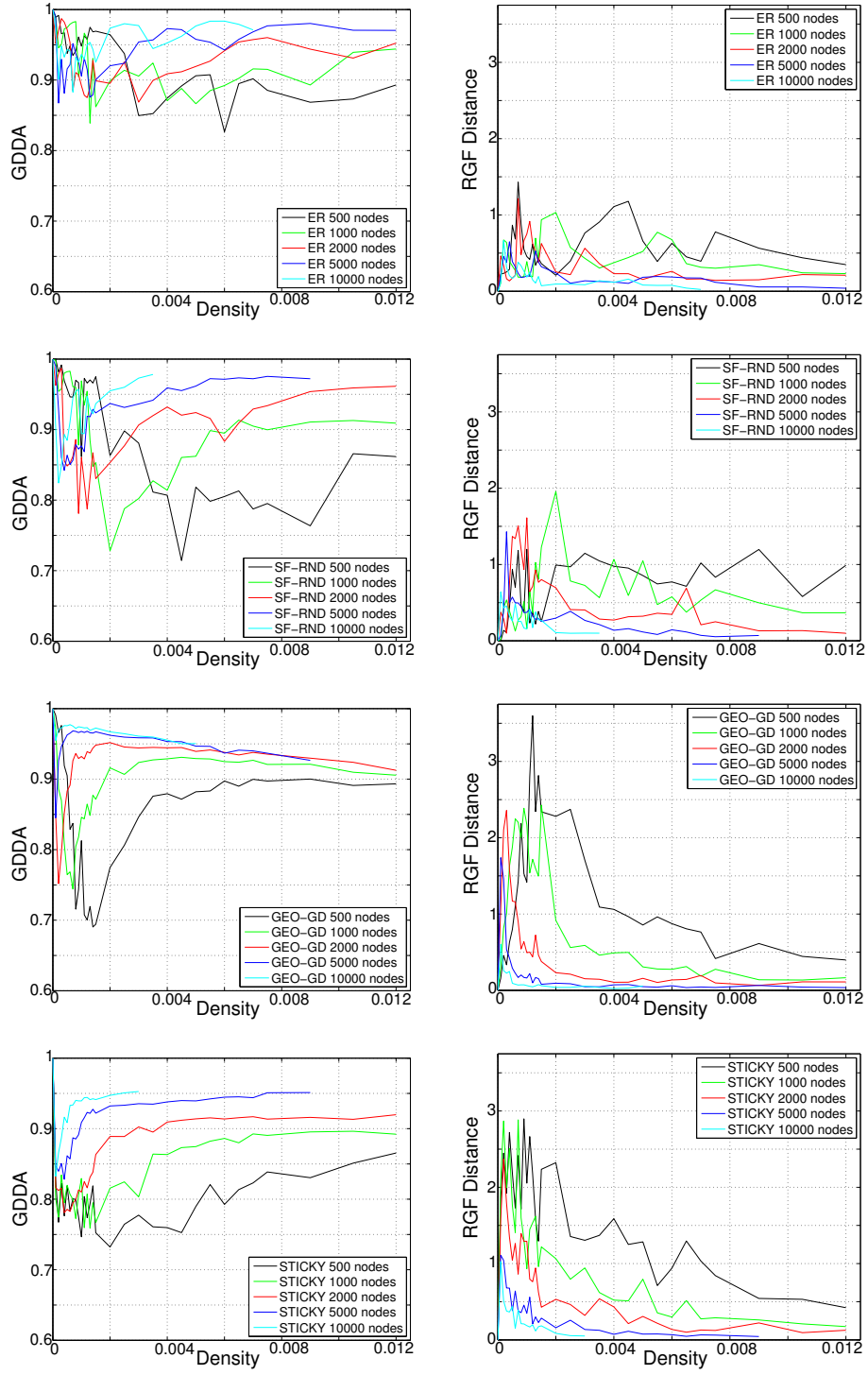


Figure A.3.: GDD agreement (left) and RGF distance (right) vs. edge density when comparing 30 random model networks with each other with 500-10000 nodes for ER, ER-DD (SF-RND), GEO-GD and STICKY models.

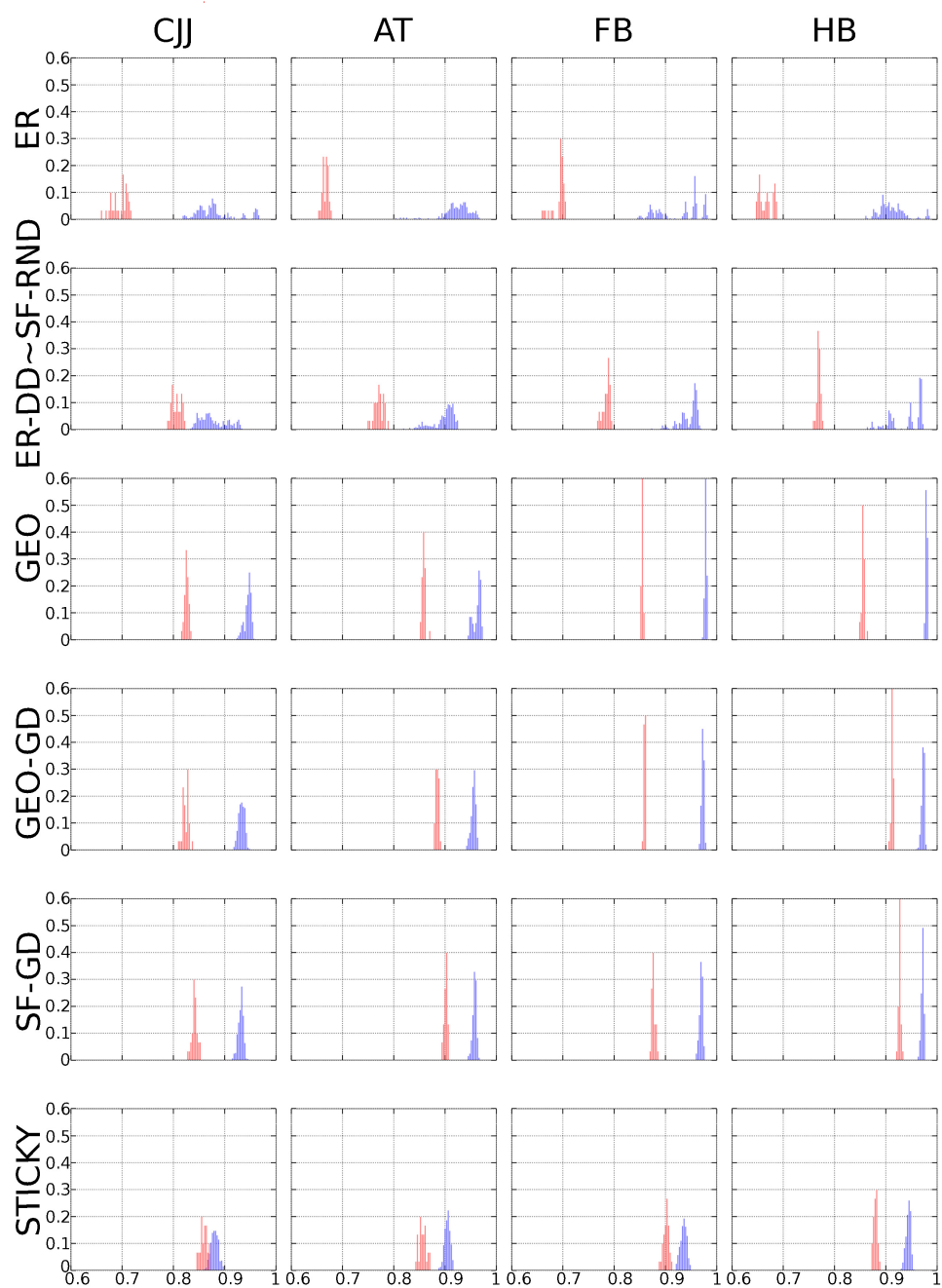


Figure A.4.: Comparing six random network models with PPI networks of different species.

A.3. Supplementary information for Chapter 4

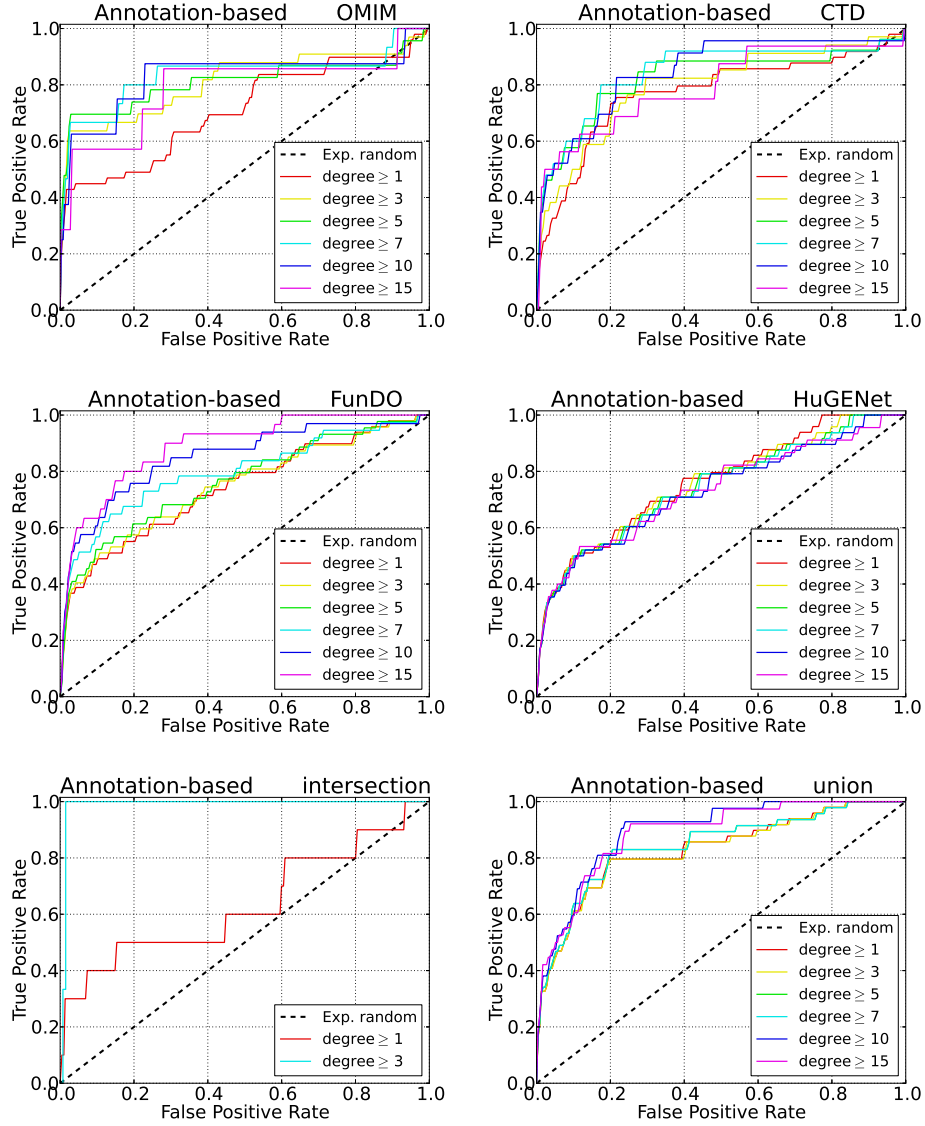


Figure A.5.: ROC curves obtained by evaluating the annotation-based measure against comorbidity. The ϕ -correlation threshold was set to 0.06. For each disease-gene association dataset, we evaluated diseases annotated with at least 1, 3, 5, 7, 10, 15 genes, showing by curves with different colours in each plot. For the intersection set, only two curves were shown in each plot since there were no comorbidity associations between diseases annotated with more than 5 genes.

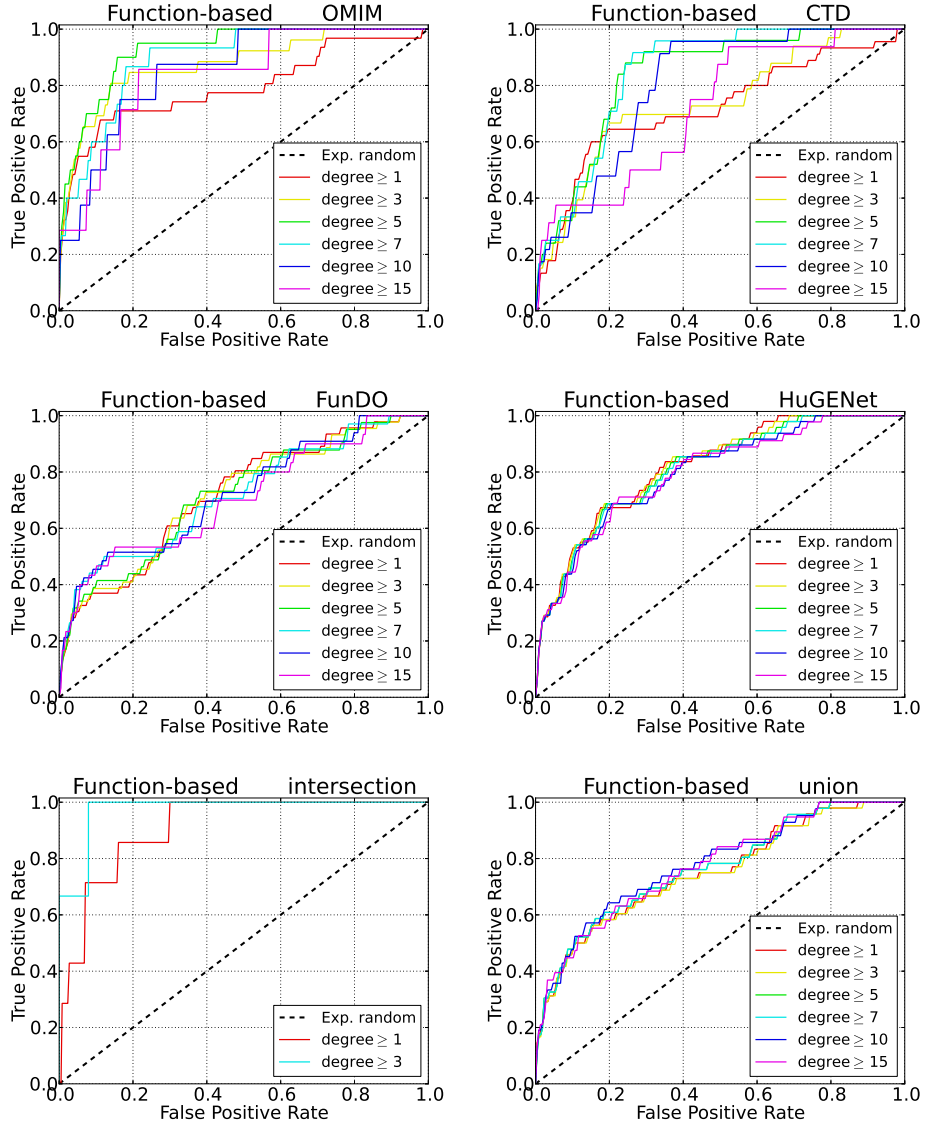


Figure A.6.: ROC curves obtained by evaluating the function-based measure against comorbidity. The ϕ -correlation threshold was set to 0.06. For each disease-gene association dataset, we evaluated diseases annotated with at least 1, 3, 5, 7, 10, 15 genes, showing by curves with different colours in each plot. For the intersection set, only two curves were shown in each plot since there were no comorbidity associations between diseases annotated with more than 5 genes.

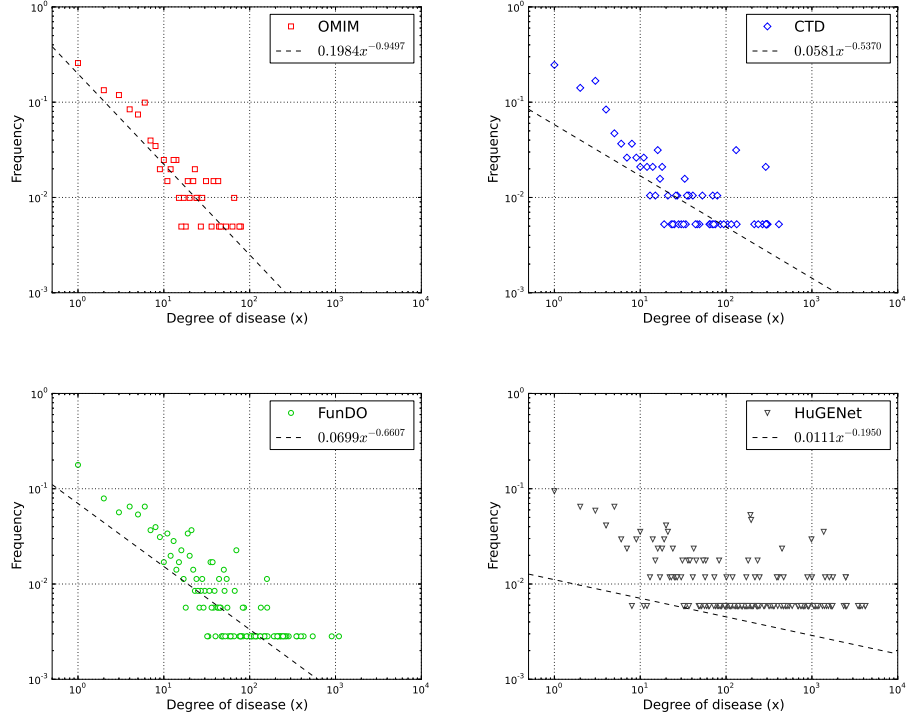


Figure A.7.: Degree distributions of diseases of the four disease-gene association datasets we analysed. Matlab curve fitting tools were used to estimate parameters of the best-fitting power-law distributions to the data.

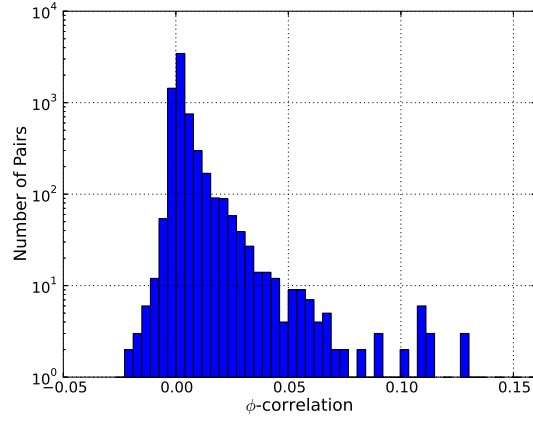


Figure A.8.: The distribution of ϕ -correlation scores.

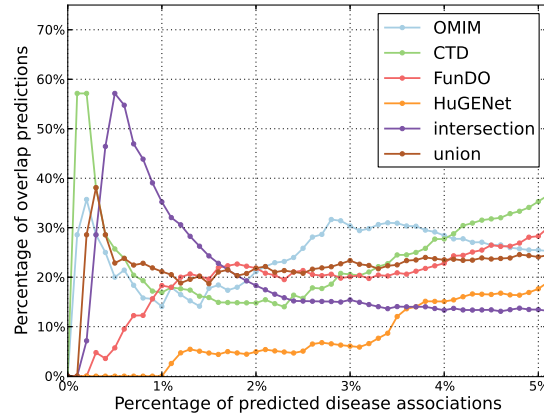


Figure A.9.: The overlap of predictions. The x-axis shows the percentage of predicted associations in all disease pairs we analysed, and the y-axis shows the percentage of overlap among associations predicted by the three similarity measures.

PPI Data	Nodes	Edges
BioGRID version 2.0.56 (released in September 2009)	6,098	18,744
BioGRID version 3.0.68 (released in September 2010)	8,433	29,971
BioGRID version 3.1.80 (released in September 2011)	9,056	37,640
BioGRID version 3.1.93 (released in October 2012)	11,261	66,253

Table A.3.: Details of PPI data obtained from different versions of the BioGRID database.

A.4. Supplementary information for Chapter 5

Data	v. 2.0.56	v. 3.0.68	v. 3.1.80	v. 3.1.93
OMIM	0.7222	0.7431	0.7262	0.8495
CTD	0.7669	0.7524	0.7569	0.7949
FunDO	0.7451	0.7401	0.7274	0.7497
HuGENet	0.6853	0.6965	0.7099	0.7153
Intersection	0.9993	0.9965	0.9962	0.9958
Union	0.7605	0.7685	0.7688	0.7939

Table A.4.: AUC values obtained by evaluating the topology-based similarity measure against comorbidity, using PPI data obtained from different versions of BioGRID database. The ϕ -correlation threshold was set to 0.06 and all diseases annotated with least 3 genes were evaluated. Each evaluation test was run 30 times to compute the statistics reported in the table.

Rank	Code	Disease name	Reference
1	239	Neoplasms of unspecified nature	PMID: 22278152
2	331	Other cerebral degenerations	ICD-9, Comorbidity
3	250	Diabetes mellitus	PMID: 23335160
3	256	Ovarian dysfunction	PMID: 15351195
5	714	Rheumatoid arthritis and other inflammatory polyarthropathies	PMID: 18525447
5	155	Malignant neoplasm of liver and intrahepatic bile ducts	PMID: 24148818
7	290	Dementias	Comorbidity
8	202	Other malignant neoplasms of lymphoid and histiocytic tissue	GWAS
9	335	Anterior horn cell disease	ICD-9
9	758	Chromosomal anomalies	PMID: 23162423
11	216	Benign neoplasm of skin	–
11	401	Essential hypertension	PMID: 9403584
11	642	Hypertension complicating pregnancy childbirth and the puerperium	–
14	710	Diffuse diseases of connective tissue	GWAS
15	340	Multiple sclerosis	ICD-9, GWAS
16	300	Anxiety, dissociative and somatoform disorders	PMID: 20479358
16	301	Personality disorders	PMID: 22083431
16	305	Nondependent abuse of drugs	PMID: 20443774
16	307	Special symptoms or syndromes not elsewhere classified	PMID: 18181204
20	277	Other and unspecified disorders of metabolism	PMID: 21645034
21	365	Glaucoma	ICD-9
22	295	Schizophrenic disorders	GWAS
22	333	Other extrapyramidal disease and abnormal movement disorders	ICD-9, Comorbidity, GWAS
24	362	Other retinal disorders	ICD-9
25	577	Diseases of pancreas	PMID: 22745701
26	278	Overweight, obesity and other hyperalimentation	PMID: 23175195
26	783	Symptoms concerning nutrition metabolism and development	PMID: 17131227
28	414	Other forms of chronic ischemic heart disease	GWAS
28	733	Other disorders of bone and cartilage	PMID: 23000281
30	042	Human immunodeficiency virus [HIV] disease	PMID: 19748551

Table A.5.: List of the top 30 diseases associated with Parkinson’s disease (PD). The topology-based measure was used as the similarity measure, and FunDO was used as the source of disease-gene associations.

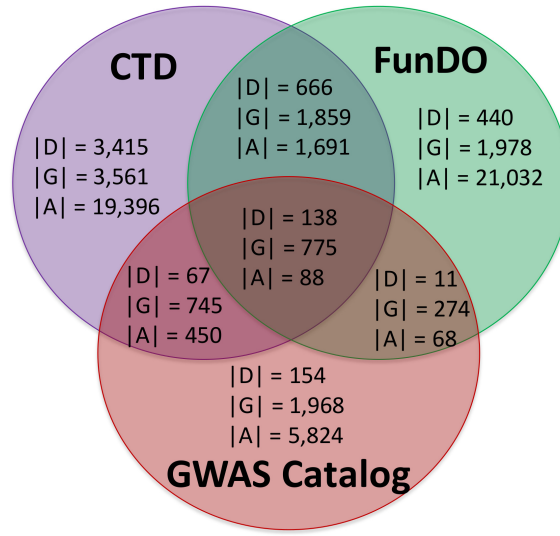


Figure A.10.: The overlap of diseases (denoted by ‘D’), genes (denoted by ‘G’) and their associations (denoted by ‘A’) between the three disease-gene association datasets we integrated.

P-value threshold	5×10^{-2}	10^{-2}	10^{-3}	10^{-5}	10^{-10}	10^{-20}
Gene-based	0.6914	0.6981	0.6926	0.7171	0.7388	0.7588
Chemical-based	0.7250	0.7293	0.7260	0.7463	0.7647	0.7840
Pathway-based	0.6329	0.6382	0.6335	0.6540	0.6704	0.6857
GO-based	0.5748	0.5768	0.5708	0.5879	0.5971	0.6063
Integration-based	0.7580	0.7646	0.7607	0.7846	0.8064	0.8282

Table A.6.: AUC values obtained by evaluating the integration-based similarity measure against disease co-occurrence in the literature. AUC values obtained by solely using disease-gene associations (gene-based), disease-chemical associations (chemical-based), disease-pathway associations (pathway-based) or disease-GO term associations (GO-based) are also listed in the table.

	Disease	Score	Gene	Chemical	Pathway	GO term
1	IBD	0.2006	0.2124 (1)	0.1879 (2)	0.2039 (72)	0.1249 (40)
2	CU	0.1640	0.1580 (2)	0.1727 (4)	0.2686 (33)	0.1399 (33)
3	SA	0.1429	0.0584 (7)	0.0925 (16)	0.3845 (4)	0.2942 (4)
4	AP	0.1197	0.0643 (6)	0.0434 (76)	0.3005 (18)	0.2962 (3)
5	BS	0.1186	0.0689 (5)	0.0924 (17)	0.3095 (16)	0.2221 (12)
6	LE	0.1145	0.1069 (3)	0.0467 (63)	0.0890 (154)	0.2501 (10)
7	CD	0.1067	0.0777 (4)	0.1829 (3)	0.3424 (8)	0.0360 (144)
8	NN	0.0999	0.0372 (18)	0.0000 (487)	0.3292 (12)	0.2603 (9)
9	BC	0.0970	0.0467 (13)	0.0000 (300)	0.2988 (19)	0.1565 (26)
10	MIN	0.0948	0.0292 (36)	0.0000 (480)	0.4271 (1)	0.0652 (95)
11	RHD	0.0916	0.0172 (73)	0.0000 (548)	0.3856 (3)	0.1192 (47)
12	PE	0.0912	0.0306 (30)	0.0201 (151)	0.2983 (20)	0.2654 (7)
13	PN	0.0908	0.0322 (26)	0.0428 (77)	0.2982 (21)	0.3369 (1)
14	MS	0.0900	0.0522 (10)	0.0514 (53)	0.2847 (25)	0.1458 (30)
15	EN	0.0869	0.0031 (179)	0.0457 (70)	0.3518 (7)	0.0374 (139)
16	LV	0.0859	0.0319 (28)	0.0000 (448)	0.4200 (2)	0.0707 (89)
17	HA	0.0849	0.0205 (60)	0.1099 (11)	0.3121 (15)	0.1928 (17)
18	TP	0.0836	0.0185 (68)	0.0000 (583)	0.3745 (5)	0.1720 (21)
19	SP	0.0835	0.0460 (14)	0.0425 (78)	0.2123 (67)	0.3192 (2)
20	PA	0.0796	0.0043 (168)	0.0000 (515)	0.2296 (55)	0.2396 (11)

Table A.7.: Similarity scores between Crohn’s disease and its top 20 associated diseases. Similarity scores were obtained by using the integrated data (the 3rd column, denoted by ‘Score’), or by solely using one type of disease-related data, namely disease-gene associations (the 4th column, denoted by ‘Gene’), disease-chemical associations (the 5th column, denoted by ‘Chemical’), disease-pathway associations (the 6th column, denoted by ‘Pathway’) or disease-GO term associations (the 7th column, denoted by ‘GO term’). Numbers in brackets are ranks of scores. MeSH terms of diseases are denoted by abbreviations: ‘Inflammatory Bowel Diseases’ (IBD), ‘Colitis, Ulcerative’ (CU), ‘Sarcoidosis’ (SA), ‘Arthritis, Psoriatic’ (AP), ‘Behcet Syndrome’ (BS), ‘Leprosy’ (LE), ‘Celiac Disease’ (CD), ‘Nasopharyngeal Neoplasms’ (NN), ‘Brucellosis’ (BC), ‘Mycobacterium Infections, Nontuberculous’ (MIN), ‘Rheumatic Heart Disease’ (RHD), ‘Periodontitis’ (PE), ‘Pneumonia’ (PN), ‘Multiple Sclerosis’ (MS), ‘Entamoebiasis’ (EN), ‘Leishmaniasis, Visceral’ (LV), ‘Hepatitis, Autoimmune’ (HA), ‘Tuberculosis, Pulmonary’ (TP), ‘Spondylitis, Ankylosing’ (SP), ‘Parapsoriasis’ (PA).

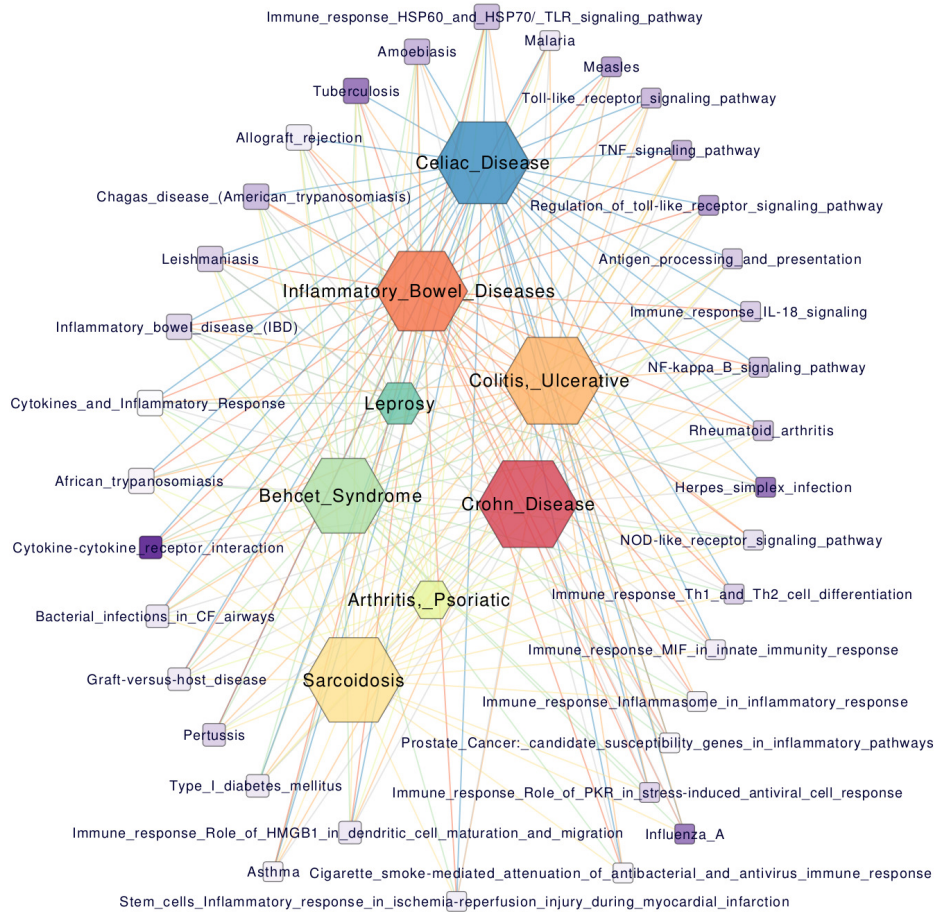


Figure A.11.: The most common biological pathways (represented by rounded rectangle nodes) shared between Crohn's disease and its top 7 associated diseases (represented by hexagon nodes). The size of nodes is proportional to its degree. The color of an edge corresponds to the disease it connects. The color of a pathway node is darker if there are more genes associated with that pathway. Only pathways that associated with at least 6 diseases are shown in the figure. Cytoscape 2.8.1 [48] was used for the visualisation.

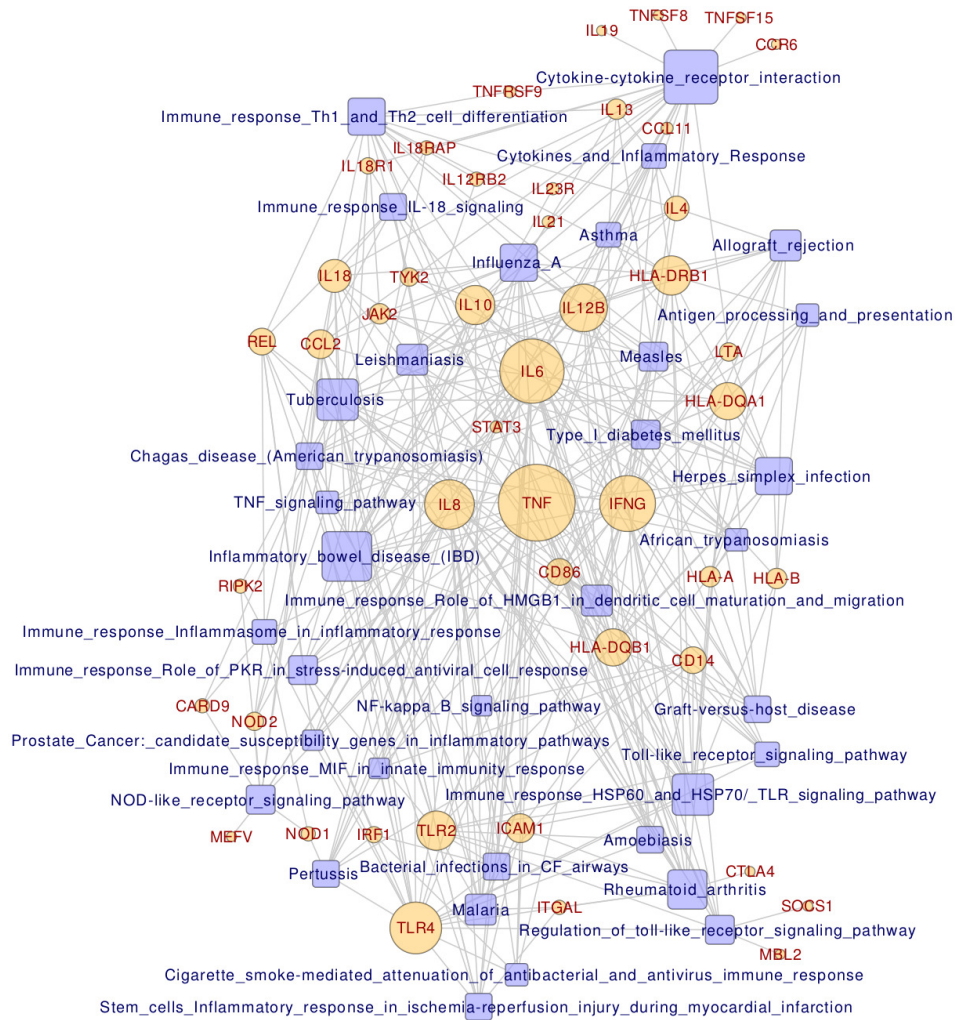


Figure A.12.: Associations between the most common genes (represented by round nodes) and the most common biological pathways (represented by rounded rectangle nodes). The size of nodes is proportional to its degree. Only pathways that associated with at least 6 diseases are shown in the figure. Cytoscape 2.8.1 [48] was used for the visualisation.