UNIVERSITÄT
DES
SAARLANDES

max planck institut
informatik

Network Biology Methods for Functional Characterization and Integrative Prioritization of Disease Genes and Proteins

Nadezhda T. Doncheva

Network Biology Methods for Functional Characterization and Integrative Prioritization of Disease Genes and Proteins

Dissertation

zur Erlangung des Grades des
Doktors der Naturwissenschaften der
Naturwissenschaftlich-Technischen Fakultäten der
Universität des Saarlandes

vorgelegt von
Nadezhda T. Doncheva

Saarbrücken, Juli 2016

Tag des Kolloquiums:	6. Juni 2016
Dekan der Fakultät:	Univ.-Prof. Dr. Frank-Olaf Schreyer
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Thorsten Herfet
Berichterstatter:	Prof. Dr. Dr. Thomas Lengauer
	Prof. Dr. Hans-Peter Lenhof
	John H. Morris, Ph.D.
Akademischer Beisitzer:	Dr. Markus List
Betreuer:	Dr. Mario Albrecht

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, 29. Juli 2016

(Nadezhda T. Doncheva)

Abstract

Nowadays, large amounts of experimental data have been produced by high-throughput techniques, in order to provide more insight into complex phenotypes and cellular processes. The development of a variety of computational and, in particular, network-based approaches to analyze these data have already shed light on previously unknown mechanisms. However, we are still far from a comprehensive understanding of human diseases and their causes as well as appropriate preventive measures and successful therapies.

This thesis describes the development of methods and user-friendly software tools for the integrative analysis and interactive visualization of biological networks as well as their application to biomedical data for understanding diseases. We design an integrative phenotype-specific framework for prioritizing candidate disease genes and functionally characterizing similar phenotypes. It is applied to the identification of several disease-relevant genes and processes for inflammatory bowel diseases and primary sclerosing cholangitis as well as for Parkinson's disease.

Since finding the causative disease genes does often not suffice to understand diseases, we also concentrate on the molecular characterization of sequence mutations and their effect on protein structure and function. We develop a software suite to support the interactive, multi-layered visual analysis of molecular interaction mechanisms such as protein binding, allostery and drug resistance. To capture the dynamic nature of proteins, we also devise an approach to visualizing and analyzing ensembles of protein structures as, for example, generated by molecular dynamics simulations.

Kurzfassung

In den letzten Jahren wurde mittels Hochdurchsatzverfahren eine große Menge experimenteller Daten generiert, um einen Einblick in komplexe Phänotypen und zelluläre Prozesse zu ermöglichen. Die Entwicklung von verschiedenen bioinformatischen und insbesondere netzwerkbasierten Ansätzen zur Analyse dieser Daten konnte bereits Aufschluss über bisher unbekannte Mechanismen geben. Dennoch sind wir weit entfernt von einem umfassenden Verständnis menschlicher Krankheiten und ihrer Ursachen sowie geeigneter präventiver Maßnahmen und erfolgreicher Therapien.

Diese Dissertation beschreibt die Entwicklung von Methoden und benutzerfreundlichen Softwarewerkzeugen für die integrative Analyse und interaktive Visualisierung biologischer Netzwerke sowie ihre Anwendung auf biomedizinische Daten zum Verständnis von Krankheiten. Wir entwerfen ein integratives, phänotypspezifisches Framework für die Priorisierung potentiell krankheitserregender Gene und die funktionelle Charakterisierung ähnlicher Phänotypen. Es wird angewandt, um mehrere krankheitsspezifische Gene und Prozesse von chronisch-entzündlichen Darmerkrankungen und primär sklerosierender Cholangitis sowie von Parkinson zu bestimmen.

Da es für das Verständnis von Krankheiten oft nicht genügt, die krankheitserregenden Gene zu entdecken, konzentrieren wir uns auch auf die molekulare Charakterisierung von Sequenzmutationen und ihren Effekt auf die Proteinstruktur und -funktion. Wir entwickeln eine Software, um die interaktive, vielschichtige visuelle Analyse von molekularen Mechanismen wie Proteinfaltung, Allosterie und Arzneimittelresistenz zu unterstützen. Um den dynamischen Charakter von Proteinen zu erfassen, ersinnen wir auch eine Methode für die Visualisierung und Analyse von Proteinstrukturen, welche sich zum Beispiel während Molekulardynamiksimulationen ergeben.

Acknowledgments

First and foremost, I wish to thank Mario Albrecht, who shared with me his great expertise and deep aspiration for solving scientific questions and patiently mentored me during my studies. I would like to express my gratitude to Thomas Lengauer for his continuous support and insightful feedback on my work. My special thanks also go to John "Scooter" Morris, who welcomed me at UCSF, showed me once again what a pleasure it is to do scientific research and so kindly offered to review my thesis. I am thankful to Hans-Peter Lenhof for introducing me to the field of bioinformatics all those years ago and for willing to review my thesis. I also wish to thank Thorsten Herfet and Markus List for being part of my defense committee.

I am especially grateful to my current and former colleagues from D3 for the nice atmosphere at work, the fruitful discussions and the pleasant coffee breaks. In particular, these include Yassen Assenov, Hagen Blankenburg, Andreas Schlicker, Dorothea Emig, Jan Hoinka, Matthias Bieg, Tim Kacprowski, Tomas Bastys, Olga Kalinina, Olga Voitenko, Glenn Lawyer, Bastian Beggel, Prabhav Kalaghatgi, Matthias Dietzen, and Nora Speicher. Special thanks go to my best office mate Alejandro Pironti, our great secretary Ruth Schneppen-Christmann, and our irreplaceable system admins Georg Friedrich and Joachim Büch.

I also owe my gratitude to my collaboration partners Karsten Klein, Christoph Welsch, and Francisco Domingues for our inspiring work on residue interaction networks. I am particularly thankful to Thomas Ferrin and the whole UCSF Chimera team for hosting me at UCSF and for the wonderful time spent working with them. I would also like to thank Tim Kacprowski, David Buezas, Andre Franke, Eva Ellinghaus, David Ellinghaus, Tom Karlsen, Johannes Roksund Hov, Trine Folseraas, Ole Andreassen, Jimmy Liu, Carl Anderson, and Francisco Domingues for our fruitful collaboration on disease gene prioritization. Last but not least, I am very grateful to the dedicated team that organizes and teaches the EMBO practical course on computational analysis of protein-protein interactions for providing such a great teaching experience every year.

Finally, I am indebted to my friends and family for their encouragement, patience, and support throughout the years. In particular, I would like to thank Javor Kalojanov, Elitza Kalojanova, Iva Baykova, Ekaterina Krasteva, Elora Krancheva, Violeta Ivanova, Goran Doychev, Zwetelina Iliewa, and Maria Slavtscheva. I cannot express how grateful I am to my parents, Rayna Doncheva and Tzanko Donchev, for raising me to be the person I am today and ultimately making all of this possible.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

This chapter addresses the importance of studying human diseases, the challenges involved in the era of high-throughput data, and the arising need of novel methods and tools for the functional characterization and integrative prioritization of disease genes and proteins. Then, the research contributions and the structure of the thesis are outlined.

1.1 Motivation

Human diseases are still one of the biggest global burdens of our population. The term disease is broadly used and may refer to any abnormal condition (disorder) that affects an organism and, in particular for humans, causes pain, dysfunction, distress or death. A number of external and internal factors such as the environment or our own genetic code may play a crucial role in the manifestation and development of a disease. For example, diseases that result from a single mutation in a single gene are referred to as monogenic, while complex diseases are caused by the intricate interplay of several genes and external factors. Often enough, however, such factors might also be the key to preventing or curing a disease. This has motivated physicians and scientists to dedicate their time to the study and understanding of diseases throughout the centuries until today. Currently, a major goal of biomedical research is to identify and characterize the genes that predispose to, are causative of, or modify the respective diseases.

The sequencing of the human genome in 2001 was a big step toward partially accomplishing this goal (International Human Genome Sequencing Consortium *et al.*, 2001). Since then, sequencing technologies have advanced tremendously and, nowadays, the sequences of more than thousand human genomes are available to the medical research community (Sudmant *et al.*, 2015). Although these vast amounts of data have already enabled a better understanding of monogenic diseases, there is still a long way to go in the case of complex and infectious diseases. Experimental techniques such as genome-wide association studies (Franke *et al.*, 2008; Manolio, 2010; Hirschhorn and Gajdos, 2011) and large-scale RNA interference screens (Boutros and Ahringer, 2008; Hawkins *et al.*, 2010a; Reiss *et al.*, 2011) have been specifically developed for the discovery of genes associated with complex diseases such as inflammatory bowel diseases. The downside of these approaches is that they yield long lists of candidate genes that need to be experimentally validated in time-consuming and costly follow-up studies. In other areas, such as transcriptomics, proteomics, and metabolomics, great volumes of experimental data have also been produced in an effort to increase our knowledge of cellular and disease mechanisms (The ENCODE Project Consortium, 2012; Rolland *et al.*, 2014; Moignard *et al.*, 2015; Huttlin *et al.*, 2015; Sahni *et al.*, 2015).

Making sense out of this wealth of generated data is a big challenge nowadays. Therefore, new bioinformatics approaches and software tools are needed to integrate, analyze, prioritize, and visualize such large-scale datasets. For instance, the field of disease gene prioritization has grown substantially over the last years. Many prioritization approaches make use of functional annotations, protein interaction networks, or integrate multiple data sources using network-based representations or statistical learning techniques (Doncheva *et al.*, 2012b). Network-based methods are very effective for representing complex relationships between interacting molecules and rely on well-known methods from the graph theory field to gain more insight into complex disease mechanisms. However, there is still room for methodological improvements, in particular, with regard to the sources of biomedical knowledge and how they are exploited, integrated and evaluated. For example, prioritization approaches should be tailored to specific phenotypes or groups of phenotypes, especially for less studied diseases.

Recent studies have revealed that combining systems and structural biology could be very beneficial for both fields, but there is still a gap between them in terms of available methods and software tools (Fraser *et al.*, 2013). For example, understanding how disease-associated genes alter a given phenotype usually involves analyzing the effect of mutations on the structure and function of the respective gene products. So far, most computational methods focus on identifying whether a mutation will affect a protein or not, but there are few tools that can aid the interpretation of the resulting changes. Structural biologists have started using network representations to study the interactions of residues in protein structures for understanding complex protein structure-function relationships (Csermely, 2008; Vishveshwara *et al.*, 2009; Doncheva *et al.*, 2011). Furthermore, the field of visual analytics has gained more attention in the biological community addressing the importance of interactive visualization of big data sets (Ray *et al.*, 2014). Thus, combining network biology, structural biology and visual analytics presents a promising way of tackling some of the challenges involved in studying complex diseases.

1.2 Overview

In this thesis, we explore two avenues of research. On the one hand, we focus on the development and application of network-based approaches for understanding the molecular mechanisms of human diseases. In close collaboration with biological and medical experts, we design an integrative phenotype-specific framework for prioritizing candidate disease genes and functionally characterizing similar phenotypes, such as immune-related diseases. In particular, we combine known protein interactions and strong functional similarities between proteins into integrative networks and apply appropriate network analysis and visualization techniques. Our framework is an indispensable part of three independent functional studies of inflammatory bowel diseases, primary sclerosing cholangitis, and Parkinson's disease.

On the other hand, we develop novel methods and user-friendly tools to bridge the gap between network and structural biology with a special focus on visual analytics. By enhancing molecular networks with structural information and further providing a network representation of residue interactions, we provide an interactive software suite that supports the multi-layered visual analysis of molecular interaction mechanisms such as protein binding, allostery, and drug resistance. To specifically analyze the impact of sequence mutations on protein structure and function, we devise an integrative visual approach that combines different structure and network visualizations and enriches them with external biophysical knowledge. To account for the dynamic nature of proteins, we extend our methodology to the analysis and visualization of ensembles of protein structures as generated from molecular dynamics simulations. We use dynamic, weighted residue interaction networks to capture the different protein conformations within the ensemble. We apply our approach to identify important residues and interactions on the interface of bound molecules and to better characterize the effect of protein mutations.

To a large extent, the research described in this thesis is based on 10 co-authored publications that appeared in peer-reviewed journals in the last few years (see Appendix C for a full list). The publications in the field of disease gene prioritization result from the collaboration with experimental partners at the Christian-Albrechts-University of Kiel, the Oslo University Hospital, the Wellcome Trust Sanger Institute, and the European Academy of Bozen/Bolzano. The structural bioinformatics projects were conducted together with experts from the Max Planck Institute for Informatics, the University of California, San Francisco, the Monash University, the European Academy of Bozen/Bolzano, and the University Hospital Frankfurt. Altogether, this work was performed in the context of several bioinformatics research projects and was financially supported by the Max Planck Society, the German National Genome Research Network, the DFG-funded Cluster of Excellence on Multimodal Computing and Interaction, and a Boehringer Ingelheim Fonds travel grant for a three-month stay at the University of California, San Francisco.

1.3 Outline

The remainder of this thesis is divided into five chapters followed by a bibliography and three appendices. The contents of each chapter are briefly described below.

Chapter 2 introduces the field of network biology and bioinformatics. First, we define the term biological network and describe several representative types of biological networks. Then, we give an overview of state-of-the-art analytical network approaches and visualization techniques. Finally, we highlight some of the successful applications of network biology and the accompanying software tools.

Chapter 3 describes our methodological contributions to the field of candidate disease gene prioritization and their application to three distinct phenotypes. After introducing the state-of-the-art of the field, our network-based prioritization framework is presented in detail. We then report our findings from the computational analysis of data associated with inflammatory bowel diseases. The application of our framework to prioritize candidate genes for the less studied disease primary sclerosing cholangitis and to characterize its functional overlap with inflammatory bowel diseases is presented thereafter. At last, we describe our contribution to the prioritization of candidate proteins for Parkinson's disease.

Chapter 4 addresses the problem of characterizing the effect of amino acid mutations on protein structure and function. We introduce our visual analytics approach that integrates different biological views of protein sequence, structure, and residue interaction networks with external biophysical data and detail the major implementation tasks involved in the development and realization. The effectiveness of the approach is demonstrated in a proof-of-concept study of the functionally defective protein provided by the BioVis 2013 Data Analysis contest. Another more systematic analysis investigates the physico-chemical, structural and topological properties of drug resistance mutations in the HCV NS3 protease.

Chapter 5 presents a novel approach for the analysis and visualization of ensembles of protein structures. After discussing the application of network biology to study protein dynamics, we give details of our dynamic residue interaction networks, the used methodology and provided software tools. The first application of our approach is the visual exploratory analysis of data from molecular dynamics simulations with focus on characterizing the effect of sequence mutations. The second is the identification of frequent interface residues and interactions in ensembles of docking structures.

Chapter 6 summarizes and evaluates the main contributions of this thesis and closes with a discussion of prospective work.

Appendix A is a Nature protocol giving detailed instructions for performing three exemplary network analysis and visualization workflows. Appendix B contains additional figures omitted from the main text of the thesis for brevity and clarity. Appendix C lists own publications.

CHAPTER 2

Biological networks

The field of network biology has already become an indispensable part of bioinformatics despite its rather recent origin. In this chapter, we introduce the most common types of biological networks and outline the most important analysis and visualization techniques. Then we discuss prominent examples of the successful application of network biology in cell biology, network medicine, and structural biology.

2.1 The origins of network biology

Network science is a broad interdisciplinary research field on the interface between mathematics, physics, computer science, sociology, and, only recently, the life sciences. It originates from the foundation of graph theory in the early 18th century by Leonhard Euler. The principles of graph theory have played a particularly central role in the development of network science. Nevertheless, it has also made use of statistical mechanics, data mining and information visualization, statistical inference and social structures, just to name a few. The main paradigm of network science is the representation of complex systems as networks of interacting elements with the ultimate goal of understanding and modeling these systems using network analysis and visualization techniques.

For a long time, network analysis was mostly employed in the social sciences for studying the structure of relationships between social entities, for example, the spread of news, rumors, or diseases among a group of people. However, in the late 1990s, two groundbreaking papers gained considerable publicity by revealing some fundamental properties shared among many large real-world networks. On the one hand, Watts and Strogatz (1998) discovered that the neural network of the worm *Caenorhabditis elegans*, the power grid of the western United States, and the collaboration graph of film actors are highly clustered networks with surprisingly small characteristic path length. They called these networks small-world networks. On the other hand, Barabási and Albert (1999) introduced the term scale-free networks for networks with scale-free power-law node degree distribution, i.e., having a few highly connected nodes and many nodes with just a few edges. The representatives include well-known networks such as the Internet, the World Wide Web, power grids, transportation networks, citation networks, some social networks, and lately some biological networks. The discovery that many real-world networks share similar architecture boosted the network science research field. It expanded into an even more interdisciplinary direction, namely, to biology and medicine.

Generally, a biological network is any network that represents a biological process or system. A well-known representative that dates a while back are pathway diagrams like the tricarboxylic acid (TCA) cycle. However, with the constantly growing amount of new experimental data and the need to understand biology at a systems level, research has shifted from the traditional reductionist approach, which focuses on one element at a time and studies it in detail, to a holistic approach, where the biological system is considered as a whole and its properties, structure and dynamics are further investigated (Kitano, 2002). The latter has emerged as part of the new interdisciplinary field called systems biology. With that, from being used for purely illustrative and didactic goals, biological networks have become a central player in the study of cellular systems. Nowadays, networks are commonly used to represent the relationships between biological entities, such as genes, proteins, residues, and network analysis is crucial for understanding these relationships and formulating new hypotheses about the biological function of the involved molecules.

Furthermore, biological networks are well suited for integrating large-scale datasets generated from high-throughput experimental techniques such as next generation sequencing or mass spectrometry. Such data can either be interpreted as additional relationships between the involved entities and represented as edges in an integrative network or superimposed upon the network using different visualization cues and layouts. An appropriate network visualization can often reduce the complexity of the data and reveal interesting patterns and characteristics that cannot be detected otherwise. Finally, many analytical techniques from graph theory can be transferred to biological networks and used to detect key elements or whole subnetworks to explain complex cellular mechanisms and interactions. In the following sections, we will introduce some typical representatives of biological networks, present common analysis and visualization techniques, and discuss how these are combined to gain new insights into cell biology.

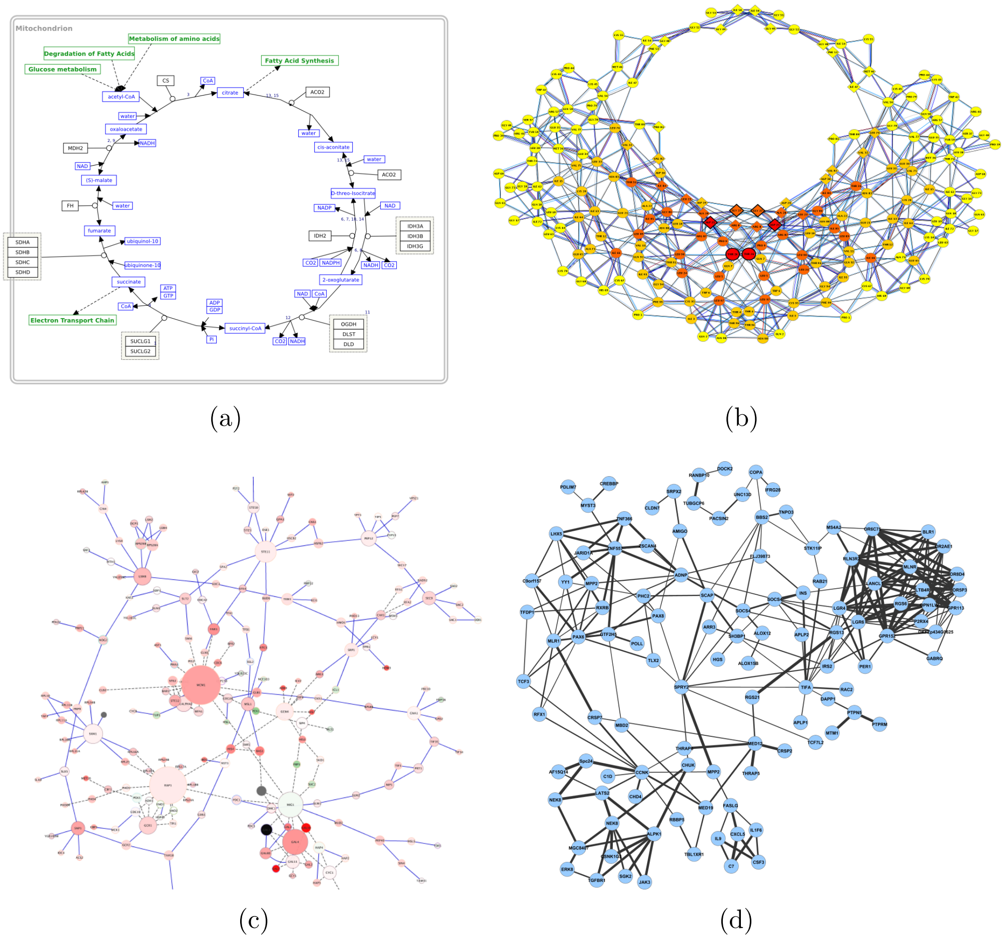
2.2 Types of biological networks

The term biological network is quite general and can refer to many different types of networks. We can distinguish between many levels of detail starting at the molecular level where DNA, RNA, proteins and metabolites interact with each other, going through cells, tissues and organs, and ending at an ecosystem, which is formed by the relationships between organisms (Junker and Schreiber, 2008). On the other hand, we can divide networks into three types based on their structure and properties: pathways, interaction networks and similarity networks (Morris *et al.*, 2015a). Typically, each of these groups is visualized and analyzed using different techniques.

Pathways are probably the most familiar as they are usually visualized by hand-curated diagrams and used to represent signaling, metabolic or regulatory pathways for educational purposes, even at school. Another common characteristic of biological pathways is that they usually describe a sequence of directional events such as signaling cascades, metabolic reactions, and gene activation or deactivation. Pathways can be found in textbooks as well as in several online databases, for instance, KEGG (Kanehisa and Goto, 2000), Reactome (Croft *et al.*, 2011), WikiPathways (Pico *et al.*, 2008), SignaLink 2 (Fazekas *et al.*, 2013), and ConsensusPathDB (Kamburov *et al.*, 2009). Phylogenetic trees represent the evolutionary relationships between organisms and can also be considered part of network science (Felsenstein, 1985). Although pathways are static illustrations, they can be combined with other information related to the involved entities such as gene expression data.

In contrast, interaction networks are the most common type of biological networks and illustrate the interactions between biological entities such as genes, proteins, metabolites, nucleic or amino-acid residues, small molecules, diseases, etc. Typical biological representatives in this category are protein interaction networks, gene regulation networks, signal transduction networks, and metabolic networks. Gene or transcriptional regulation networks are usually directed graphs, in which the edges indicate events of gene expression control between genes and their products. By extending these networks with protein-protein interactions and phosphorylation events, we can create signal transduction networks. In contrast, metabolic networks represent biochemical reactions, for instance, the conversion of metabolites into each other in reactions catalyzed by enzymes. Social networks are also an example for interaction networks and they have laid the ground for most techniques used for the topological analysis of biological networks (see Section 2.3.1 for details). Interaction networks are especially useful for bringing together the multivariate omics datasets available nowadays.

Protein interaction networks (PINs) represent physical interactions between proteins in a complex or a cell, which are usually detected by experimental techniques such as yeast two-hybrid (Y2H) (Fields and Song, 1989), co-immunoprecipitation (Co-IP) (Auerbach *et al.*, 2002), and tandem affinity purification (TAP) (Roque and Lowe, 2008), recently also coupled to mass spectrometry (Dunham *et al.*, 2012; Altelaar *et al.*, 2013; Morris *et al.*, 2014), or computational prediction approaches (Skrabanek *et al.*, 2008; Frishman *et al.*, 2008; Papanikolaou *et al.*, 2015). In the last years, several large-scale experiments were performed with the ultimate goal to create systematic proteome-wide maps for different organisms (Rual *et al.*, 2005; Yu *et al.*, 2008, 2011a; Rolland *et al.*, 2014; Kim *et al.*, 2014; Wilhelm *et al.*, 2014).



(a) (b)

(c) (d)

Figure 2.1: Examples for biological networks: (a) pathway representation of the TCA cycle from WikiPathways (Stobbe *et al.*, 2013); (b) residue interaction network of HIV-1 protease (Doncheva *et al.*, 2012a); (c) interaction network of proteins associated with galactose metabolism in yeast (Ideker *et al.*, 2001); (d) network of strong functional similarities between selected human genes based on their Gene Ontology annotations (Liu *et al.*, 2013).

For example, Rolland *et al.* (2014) presented the so far largest high-quality human dataset consisting of 14 000 protein interactions and demonstrated its usefulness for studying genotype-phenotype relationships. Overall, PINs have played a key role for understanding complex biological systems, both their natural and disrupted states in human diseases (Kann, 2007; Bader *et al.*, 2008; Ideker and Sharan, 2008; Barabási *et al.*, 2011; Vidal *et al.*, 2011; Ideker and Krogan, 2012; Meyniel-Schicklin *et al.*, 2012; De Las Rivas and Fontanillo, 2012; Jia and Zhao, 2014; Gustafsson *et al.*, 2014). Some of the widely used online resources that integrate protein interaction data from different primary sources are iRefIndex (Razick *et al.*, 2008), STRING (Franceschini *et al.*, 2013), PSICQUIC (Aranda *et al.*, 2011).

Of particular interest in this work are residue interaction networks (RINs), which represent the non-covalent interactions between amino-acid residues in a 3D protein structure as network edges. The first approaches for generation of RINs considered the spatial proximity of atoms with a distance cutoff between 5 and 8 Å (Vishveshwara *et al.*, 2009; Yan *et al.*, 2014). Recently, we released several software tools that support the interactive generation, visualization and analysis of RINs, where edges are defined as different types of non-covalent residue interactions such as hydrogen bonds and van der Waals interactions (Doncheva *et al.*, 2011). In the last years, RINs have been successfully applied for the analysis of protein structure-function relationships (Csermely, 2008; Greene, 2012; Di Paola *et al.*, 2013; Hu *et al.*, 2013; Doncheva *et al.*, 2014) and protein dynamics (Vishveshwara *et al.*, 2009; Sethi *et al.*, 2013; Xue *et al.*, 2012; Bromley *et al.*, 2013; Tiberti *et al.*, 2014; Seeber *et al.*, 2014).

The third network type are similarity networks where the nodes represent biological entities, which are connected by edges based on some similarity measure. One common example are gene or transcript correlation networks generated from expression data (Bergmann *et al.*, 2004) or metabolic correlation networks from profiling data (Weckwerth *et al.*, 2004). We also encounter other gene or protein networks based on sequence or structural similarity such as BLAST values or RMSD (Atkinson *et al.*, 2009; Holm and Sander, 1996). Another interesting representative are networks of small molecules, such as drugs, which are created according to the similarity of their chemical fingerprints to find small molecules with similar structural characteristics (Maggiora *et al.*, 2014). In Section 3.2.1, we will focus on functional similarity networks (FSNs), in which an edge corresponds to strong functional similarity of two genes based on their annotated Gene Ontology (GO) terms. In particular, FSNs have advanced the prioritization of candidate genes and the functional characterization of diseases (Liu *et al.*, 2013; Jiang *et al.*, 2011).

2.3 Analysis and visualization techniques

A key to the usefulness of networks in biology and other fields is the availability of appropriate analysis and visualization techniques that allow us to characterize these networks. In the following sections, we will give an overview of graph theory concepts and how they are applied to analyze networks. We will also present common visualization techniques and explain how they enhance our ability to interpret biological networks.

2.3.1 Analytical approaches

The origin of standard analytical approaches in network biology is the wide and well-known field of graph theory as well as the more interdisciplinary social network analysis field (Bondy, 1976; Brandes and Erlebach, 2005; Diestel, 2012). Mathematically, networks are defined as graphs and analyzed using various graph algorithms. Social sciences researchers have already discovered an appropriate set of graph algorithms to describe the structure and characteristics of social networks (Freeman, 2006; Abraham *et al.*, 2010; Borgatti *et al.*, 2013). Complex networks have different levels of organization, as shown in Figure 2.2, that can be used to breakdown the hairball that arises when we usually visualize a large network. First, we can look at single nodes and their local properties, such as the node degree. These nodes are then linked to form motifs, small subnetworks of three or more nodes. Motifs are combined to form communities or modules and finally, communities are joined into the entire network. The hierarchy of the network describes how the various structural elements are combined. Building upon this, the network biology field has adopted some of these analysis tasks and also developed new ones as reviewed by Junker and Schreiber (2008); Yamada and Bork (2009); Pavlopoulos *et al.* (2011); Przulj (2011).

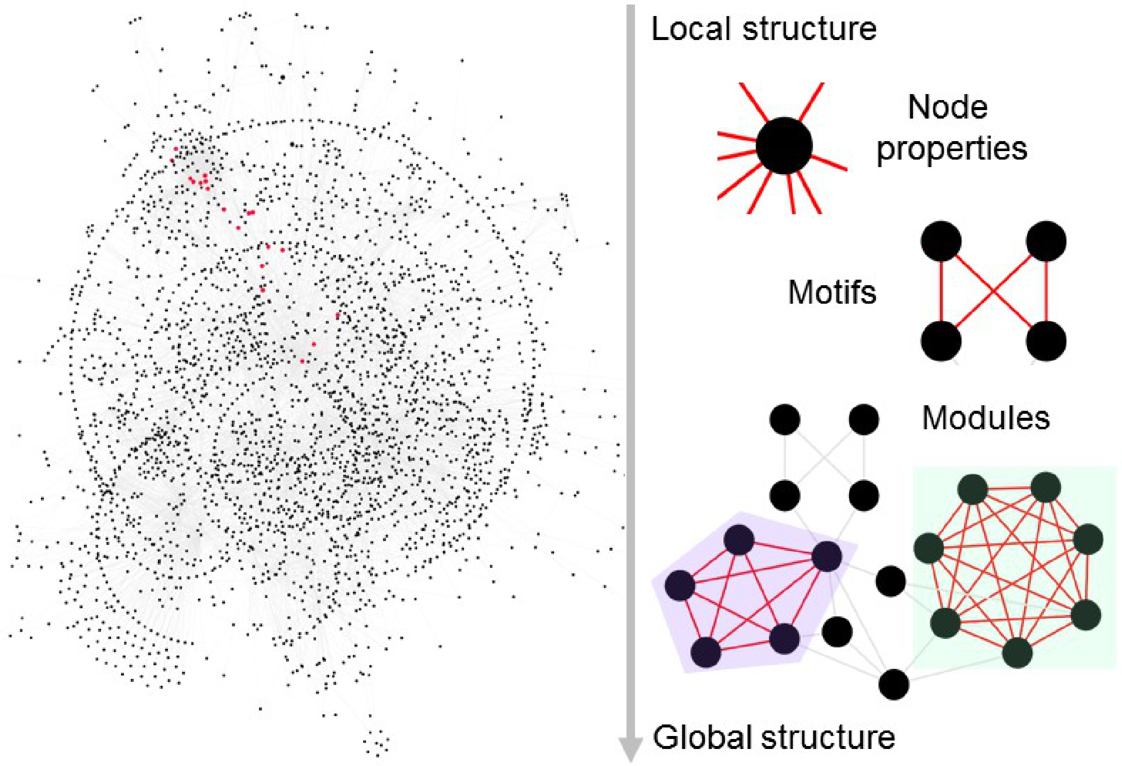


Figure 2.2: The levels of organization in complex networks (inspired by Gulbahce and Lehmann (2008)).

To characterize the structure of a biological network, we can compute a set of network topology statistics such as degree distribution of nodes, clustering coefficient, node centrality, shortest path between nodes, and robustness of the network to the random removal of single nodes (Jeong *et al.*, 2001; Ravasz *et al.*, 2002; Barabási and Oltvai, 2004). Another important characteristic of a network is its modularity, i.e., the presence or absence of subnetworks of interconnected nodes that might represent molecules, which are physically or functionally linked and work coordinately to achieve a specific function (Ravasz *et al.*, 2002; Barabási and Oltvai, 2004). Furthermore, motif analysis is used to identify small network patterns that are over-represented compared with a randomized version of the same network (Ciriello and Guerra, 2008; Masoudi-Nejad *et al.*, 2012). Discrete biological processes such as regulatory elements are often composed of such motifs (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002). Finally, network alignment and comparison tools can be applied to determine similarities between networks and have been used to study evolutionary relationships between protein networks of organisms (Kuchaiev *et al.*, 2010; Atias and Sharan, 2012; Clark and Kalita, 2014; Panni and Rombo, 2015).

2.3.2 Network topology

Next, we will describe a number of topological parameters that are commonly used in network biology and are also addressed throughout this work, in particular, in the analysis workflows in Appendix A. This text in this section has been adapted and extended from Doncheva *et al.* (2012a).

Graph definition In mathematical terms, a biological network is represented as a graph. Formally defined, a graph G is a pair of two sets $G = (V, E)$, where V is the set of nodes, E the set of edges. Each edge $e \in E$ connects the nodes $u \in V$ and $v \in V$ and is denoted as an unordered pair $e = \{u, v\}$ in undirected networks and an ordered pair $e = (u, v)$ in directed networks. Additionally, different attributes such as text, numerical values, types, colors, etc are associated with the nodes and edges of a network. For example, edge weights that represent the stoichiometry of reactions in metabolic networks can be expressed as a function $\omega : E \rightarrow \Re$, which assigns each edge $e \in E$ a weight $\omega(e)$. There are also other graph models, such as mixed (undirected and directed edges), multi-graphs (multiple edges between two nodes), hyper-graphs (an edge connects more than two elements), bipartite graphs (two distinct sets of nodes such that each edge connects them), and trees (undirected connected acyclic graphs). Depending on the specific biological network at hand, any of these models might be used.

Connected components. In undirected networks, two nodes are connected if there is a path of edges between them. All nodes that are pairwise connected form a *connected component*. The number of connected components in a network is an indicator of the global connectivity of a network. A low number of connected components relates to strong network connectivity because many nodes are connected to form few connected components of large node size.

Degree distributions. In undirected networks, the *node degree* of a node n is the number of edges linked to n (Barabási and Oltvai, 2004). A self-loop of a node is counted like two edges for the node degree (Diestel, 2012). A node with a high degree is referred to as *hub*. The *node degree distribution* gives the number of nodes with degree k for $k = 0, 1, \ldots$.

Scale-free property. A network is called *scale-free* if its degree distribution approximates a power law $k^{-\alpha}$ with the degree exponent α (Barabási and Oltvai, 2004). The topological role of network hubs depends on the α value. For $\alpha > 3$ the hubs are not relevant, for $3 > \alpha > 2$ the hubs are organized in a hierarchy, and for $\alpha = 2$ a hub-and-spoke model emerges, in which the largest hub is in contact with a large fraction of all nodes. For most biological networks, it has been observed that $2 < \alpha < 3$. Barabási and Albert (1999) used this network property to distinguish between random (as defined by Erdös and Rényi (1959, 1960)) and scale-free network topologies (Barabási and Oltvai, 2004). There are also continued discussions about the observed power law and scale-freeness (Barabási, 2009; Lima-Mendez and van Helden, 2009). Since it is usually difficult to accurately fit a power law to a distribution, for biological problems it is often enough to note that the distribution is inhomogeneous and long-tailed (Junker and Schreiber, 2008).

Neighborhood-related parameters. The *neighborhood* of a node n is the set of its neighbors. The *connectivity* k_n is the size of the neighborhood of n and should not be confused with the degree k since cases exist for which $k \neq k_n$ (Dong and Horvath, 2007). The *average number of neighbors* is an indicator for the average connectivity of the nodes in the network. The *neighborhood connectivity* of a node n is the average connectivity of all neighbors of n (Maslov and Sneppen, 2002). The *neighborhood connectivity distribution* gives the average of the neighborhood connectivities of all nodes n with k neighbors for $k = 0, 1, \ldots$. If the neighborhood connectivity distribution is a decreasing function in k, edges between low connected and highly connected nodes prevail in the network (Maslov and Sneppen, 2002). $P(n, m)$ is the number of *shared neighbors* between the nodes n and m, that is, the interaction partners that are neighbors of both n and m (Assenov *et al.*, 2008).

Clustering coefficients. The *clustering coefficient* C_n of a node n is defined as $C_n = 2e_n/(k_n(k_n - 1))$, where k_n is the number of neighbors of n and e_n the number of edges between all neighbors of n (Barabási and Oltvai, 2004; Watts and Strogatz, 1998). The clustering coefficient constitutes a ratio N/M, where N is the number of edges between the neighbors of n, and M the maximum number of edges that could possibly exist between the neighbors of n. The clustering coefficient of a node is always a number between 0 and 1. The *network clustering coefficient* is the average of the clustering coefficients of all nodes in the network and relates to the local cohesiveness and the tendency of the nodes to form clusters. The *average clustering coefficient distribution* gives the average of the clustering coefficients for all nodes n with k neighbors for $k = 2, \ldots$ and was used to suggest a modular organization of metabolic networks (Ravasz *et al.*, 2002).

Shortest paths. The length of a path is the number of edges forming it. The *length* of the *shortest path*, the *distance*, between two nodes n and m is denoted by $L(n, m)$. The *shortest path length distribution* gives the number of node pairs (n, m) with $L(n, m) = k$ for $k = 1, 2, \ldots$ and may indicate *small-world* properties of a network (Watts and Strogatz, 1998). The *eccentricity* of a node n is the maximum non-infinite length of a shortest path between n and another node in the network. The *network diameter* is the maximum node eccentricity. In contrast, the *network radius* is the minimum of the non-zero eccentricities of the nodes in the network. The *average shortest path length* L, also known as the *characteristic path length*, indicates the expected distance between two connected nodes. If the average shortest path length is much smaller than the number of nodes N in the network ($L \propto \log N$ for $N \to \infty$), it is referred to as a small-world network (Watts and Strogatz, 1998). Barabási and Oltvai (2004) showed that this is the case in interaction networks.

Shortest path centralities. The *degree centrality* of a node n is defined as its degree k_n and is sometimes also normalized by the number of nodes in the network. Several studies have shown that the degree centrality correlates well with the importance of a node for the network (Albert *et al.*, 2000) and that the removal of proteins with high degree from protein interaction networks is related to lethality (Jeong *et al.*, 2001).

The *betweenness centrality* $C_b(n)$ of a node n is defined as $C_b(n) = \sum_{s \neq n \neq t} \sigma_{st}(n)/\sigma_{st}$ (Brandes, 2001). Here, σ_{st} denotes the number of shortest paths from s to t, $\sigma_{st}(n)$ is the number of shortest paths from s to t that n lies on, and s and t are nodes in the network different from n. The betweenness centrality for each node n is normalized to a value between 0 and 1 by dividing with the number of node pairs excluding n: $(N-1)(N-2)/2$, where N is the total number of nodes in the connected component that n belongs to. The betweenness centrality of a node reflects the amount of control that this node exerts over the interactions of other nodes in the network (Yoon *et al.*, 2006). The *stress centrality* of a node n is the number of shortest paths passing through n (Brandes, 2001; Shimbel, 1953).

The *closeness centrality* $C_c(n)$ of a node n is the reciprocal of the average shortest path length from n to any other node in the network (Freeman, 1979). It measures how quickly information spreads from a given node to other reachable nodes in the network (Freeman, 1979).

Current flow centralities. In contrast to shortest path centralities, where distance is measured by the length of the shortest path between two nodes, here the distance between two nodes is computed as the *effective electric resistance* between them. This is defined as the difference of the potentials of two nodes required for generating one unit of electrical current between them (Newman, 2005). *Current flow closeness* is the inverted sum of the effective resistances between a node n and all other nodes. *Current flow betweenness* is the amount of current that passes through a node n, when a current unit flows from a source to a target node, over all source-target node pairs in the network.

Random walk centralities. Here, the distance between two nodes is measured by the *hitting time*, i.e., the expected number of steps needed by a random walk from one node to the other. *Random walk closeness* is the mean hitting time over all random walks starting at any node in the network and ending at the node n. *Random walk betweenness* is the expected number of visits to a node by a random walk between each pair of root set nodes relative to the hitting time of the random walk. The computation of hitting time and the expected number of visits is based on the relationship between random walks and the distribution of electrical current through the network (Tetali, 1991).

Clustering and modularity. Network *clustering* can be described as the process of finding subsets of nodes that satisfy some pre-defined property, for instance, the nodes within the cluster are densely connected with each other and sparsely connected to nodes outside their cluster (Girvan and Newman, 2002). Unfortunately, there is no generally accepted definition of a network cluster. It can also be referred to as a *module* or a *community*. Similarly, *modularity* is a general term that is used to describe whether a network tends to have modules or not, i.e., networks with high modularity have many distinct modules. In a biological context, it might refer to a functional module or a group of nodes that work together to perform a particular cellular function (Hartwell *et al.*, 1999; Barabási and Oltvai, 2004). Similar to other clustering techniques, network clustering methods can be either bottom-up or top-down and use any type of similarity measure between the nodes, such as edge weight or shortest path (Junker and Schreiber, 2008). In biological networks, network clustering is guided by the assumption of a modular organization of biological functions (Hartwell *et al.*, 1999) and is used to classify and reduce the underlying complex data. It aids the identification of natural clusters of evolutionary or functionally related entities (e.g. protein complexes in protein interaction networks) as well the understanding of the functional organization of networks (e.g. different metabolic functions in metabolic networks) (Ravasz *et al.*, 2002; Barabási and Oltvai, 2004; Costanzo *et al.*, 2010; Mitra *et al.*, 2013). However, real networks rarely present a unique clustering, but rather several alternative solutions that need to be evaluated and interpreted. In addition, many different models and methods exist, which makes it difficult to select the right one and usually results in the necessity to try out several different ones.

Motifs. A *network motif* is small subgraph of linked nodes (a connectivity pattern) that occurs more frequently than might be expected for randomly connected nodes. Typical representatives are feed-forward loops (Mangan and Alon, 2003) and feedback loops (Glossop *et al.*, 1999), but the definition is not restricted to subgraphs with a fixed number of nodes and can account for more complex topological structures, such as multi-input motifs (Lee *et al.*, 2002; Shen-Orr *et al.*, 2002). Network motifs are often referred to as the building blocks of complex networks since, in particular, transcriptional and signaling networks are often composed of highly overrepresented motifs (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002; Lee *et al.*, 2002; Milo *et al.*, 2004).

Robustness. *Robustness* describes the behavior of a system in response to random or intentional attacks. In the context of complex networks, this can be understood as the persistence of topological network properties, such as characteristic path length or number of connected components, upon removal of nodes or edges (Albert and Barabási, 2002; Albert *et al.*, 2000; Callaway *et al.*, 2000). For example, while random networks with homogeneous degree distribution are equally vulnerable to both types of attacks, scale-free networks are very robust to random attacks, which are expected to affect nodes with low degree, and fragile against selective intentional removal of nodes, in particular, with high degree (Albert *et al.*, 2000; Albert and Barabási, 2002). In general, networks can also be optimized to be robust against targeted attacks, but then they are often vulnerable to unanticipated perturbations (Junker and Schreiber, 2008). Robustness is also a common feature of most biological networks and reflects the ability of biological systems to accommodate fluctuations and perturbations without losing their functionality (Barabási and Oltvai, 2004; Kitano, 2004; Stelling *et al.*, 2004).

Random network models. The topological structure and characteristics of a network can be determined using a combination of the parameters described so far. However, some of these properties are informative of the network only if they are compared to a similar *null model*. Since it is difficult to develop an appropriate probability model for a complex network, several *random network* models have been suggested and each of them focuses on distinct features. The *Erdös-Rényi* network model is a graph, where each pair of nodes is connected by an edge with an equal probability (Erdös and Rényi, 1959, 1960). This network is often called flat because its degree distribution approximates a binomial distribution, most nodes have a similar degree, and there is no local structure or cohesiveness. In order to model a random network that exhibits small-world properties, i.e., has both short average path lengths and high clustering, Watts and Strogatz (1998) proposed another approach. First, a regular ring lattice network is generated such that each node is connected to its $\langle k \rangle /2$ nearest neighbors, where $\langle k \rangle$ is the average node degree. In the next step, each edge is rewired with a probability between 0 and 1 in order to connect distant nodes and decrease the average shortest path length. The *Watts-Strogatz* network model has high clustering coefficients and low average path lengths. Since none of these models captures the properties of scale-free networks, Barabási and Albert (1999) introduced another approach that generates a scale-free random network. In contrast to the other methods, the *Barabási-Albert* model starts with a small network of unconnected nodes, which is then grown larger by a preferential attachment rule. This means that, at each step, a new node n is added to the network and it is connected to an existing node m with a probability proportional to the degree of m. In this way, hubs continue to get new edges, while nodes with fewer edges remain less connected.

2.3.3 Visualization techniques

In addition to the topological analysis of complex networks, their visualization is also an important tool for exploratory analysis and visual interpretation of large datasets (Kelder *et al.*, 2010; Merico *et al.*, 2009). A good network visualization can facilitate the study of complex relationships between the elements of a system, the identification of dense clusters of interacting entities, and ultimately the generation of new hypotheses and insights (Barabási and Oltvai, 2004; Barabási *et al.*, 2011; Chuang *et al.*, 2011). There are different design principles and techniques employed for the visualization of pathways, interaction and similarity networks, and their implementation differs in the various software tools available nowadays (Suderman and Hallett, 2007; Pavlopoulos *et al.*, 2008; Gehlenborg *et al.*, 2010; Fung *et al.*, 2012; Agapito *et al.*, 2013; Villaveces *et al.*, 2015). The most straight-forward depiction of networks is as dots (circles) for the nodes and lines for the edges. The positions of the nodes and edges are usually not fixed or pre-defined but rather suggested by an automatic layout algorithms that helps to convey the relationships between the nodes. The exception are pathway diagrams, which are manually curated and arranged in the best possible way for educational purposes. Other visual features, such as shape, size, color, labels, etc., are flexible and are used to enrich the visualization with additional data.

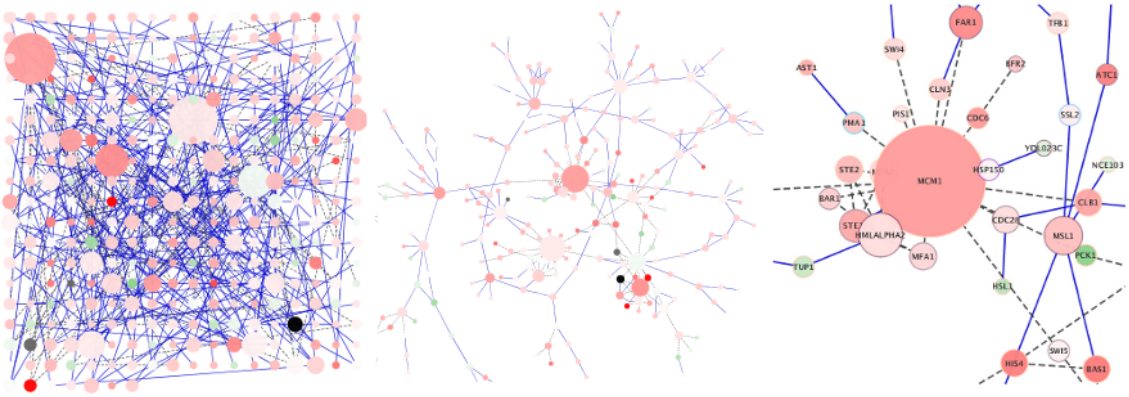


Figure 2.3: Examples for network visualization.

Layouts. Many layout algorithms have been developed for different purposes and they are usually made available as part of the respective tools for network visualization. Force-directed layout and spring-embedded layout are two very common layouts that simulate the edges as springs such that nodes repel each other and the edges pull them together. These layouts result in clusters of highly connected nodes, and nodes with low degree are placed in the periphery. An example for a force-directed layout applied on a protein interaction network is shown in the middle of Figure 2.3. Tree-like networks are best visualized using a hierarchical layout, which defines ordered layers of nodes in a tree structure. The circular (radial) layout places nodes on an imaginary circumference of a circle and can arrange them either in arbitrary order or use some attribute like the name or degree to govern the order. Especially for large networks, one layout is often not enough to achieve the right visualization. In such cases, suitable solution might be to tune the settings of the layout algorithm to the specific network or to apply different algorithms to different parts of the network.

Data mapping and visual features. The nodes and edges in a network can be associated with any data value, annotation or attribute. These might be simple text annotations such as gene symbols, protein identifiers, Gene Ontology or pathway annotations as well as numerical values such as gene expression data, degree, closeness centrality or any other topological value. To gain more insight about the whole system and the interacting entities, the additional information is usually mapped onto the network using different visual properties. The set of visual properties available to the users strongly depends on the software tool used for network visualization. For example, Cytoscape provides a huge amount of editable visual properties including fill color, shape, width and height, border color and width, opacity, and label for nodes as well as line type, color, width, arrow type, size and color for edges (Shannon *et al.*, 2003). An example is shown in the right part of Figure 2.3, where two different interaction types (protein-protein and protein-DNA) are distinguished by solid blue lines and dashed black lines, the node size is based on the node degree and the node color on expression values (red for down-regulated and green for up-regulated).

2.4 Network biology applications

Network-based approaches can aid in the structural and functional characterization of complex biological mechanisms. Thus, network biology is very interdisciplinary and its applications stretch across several disciplines. Here, we will first give an overview on the general usage of networks in cell biology and the resulting findings. Then we will discuss in more detail applications in network medicine and structural biology. More examples can be found in Barabási and Oltvai (2004); Zhang *et al.* (2007); Junker and Schreiber (2008); Yamada and Bork (2009); Ideker and Krogan (2012).

2.4.1 Cell biology

The first applications of network theory to biological data were focused on uncovering the generic organizational principles of cellular networks such as protein interaction, gene regulatory and metabolic networks (Barabási and Oltvai, 2004). In particular, some metabolic and interaction networks are small-world and approximate scale-free topology (Jeong *et al.*, 2000; Wagner, 2001; Jeong *et al.*, 2001; Yook *et al.*, 2004). One of the hypotheses for the evolutionary origin of these properties in cellular networks is gene duplication, i.e., duplicated genes encode for proteins interacting with the same partners (Barabási and Oltvai, 2004). An interesting implication of the scale-free property is that such networks are robust against random attacks and vulnerable against attacks targeting hub nodes (Albert *et al.*, 2000; Albert and Barabási, 2002). Robustness is also an important characteristic of biological systems, which need to accommodate fluctuations and perturbations without losing their functionality, but at the same time contain essential molecules (hubs) that play a crucial role in cell growth and survival (Barabási and Oltvai, 2004; Kitano, 2004; Stelling *et al.*, 2004).

Furthermore, there has been strong evidence for the modularity and hierarchical organization of biological networks (Hartwell *et al.*, 1999; Ravasz *et al.*, 2002; Barabási and Oltvai, 2004). Therefore, the next important step was the usage of network theory to identify groups of nodes such as motifs and modules. Several groups revealed that clusters of nodes in molecular interaction networks represent protein complexes (Bader and Hogue, 2003; Girvan and Newman, 2002; Rives and Galitski, 2003; Spirin and Mirny, 2003; Krogan *et al.*, 2006; Bandyopadhyay *et al.*, 2008). On the other hand, transcriptional and signaling networks were found to contain highly overrepresented motifs, such as feedback loops (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002; Lee *et al.*, 2002; Milo *et al.*, 2004). The finding that interacting proteins are often involved in the same biological process and thus are likely to have the same function, lead to a new group of methods for gene function prediction based on the interaction network neighborhood (Ideker *et al.*, 2002; Letovsky and Kasif, 2003; Deng *et al.*, 2003; Vazquez *et al.*, 2003; Nabieva *et al.*, 2005; Sharan *et al.*, 2007).

Another area of research involved the comparison of interaction networks from different species and the identification of common network structures such as pathways and clusters (Sharan *et al.*, 2005; Pinter *et al.*, 2005; Sharan and Ideker, 2006). As a result, a number of methods and tools for network alignment and subgraph matching in biological networks were developed (Sharan *et al.*, 2005; Pinter *et al.*, 2005; Flannick *et al.*, 2006; Shlomi *et al.*, 2006; Kuchaiev and Przulj, 2011). A recent trend in the network biology field is the construction and analysis of dynamic networks, i.e., networks that represent biological systems at different times or conditions (Przytycka *et al.*, 2010; Ideker and Krogan, 2012).

2.4.2 Network medicine

Network medicine is a recent interdisciplinary field that employs the tools of network science to study diseases and to discover new drugs or drug targets. The term 'network medicine' was introduced by Barabási (2007) in his review paper, where he points out the need of a map of all cellular processes involved in diseases instead of just lists of disease genes. In the last 10 years, several studies have shown that network models and analysis techniques are very suitable for identifying new disease genes, drug targets and biomarkers for complex diseases, as described in more detail in Section 3.1. In particular, molecular networks of protein-protein and regulatory interactions as well as metabolic pathways and reactions play a very important role. In addition, the construction of networks of all diseases and their known associations (Goh *et al.*, 2007), or of the available drugs and their known targets

(Yildirim *et al.*, 2007) has provided key insights into human diseases and therapy. Finally, epidemiology has been advanced by the application of network models and the integration of knowledge about social and transportation networks to study the spreading of diseases in populations (Pastor-Satorras and Vespignani, 2001; Keeling and Eames, 2005).

A number of approaches for candidate disease gene prioritization make use of networks to represent and integrate the available biomedical knowledge and of network analysis techniques to identify promising disease genes and modules (Kann, 2007; Ideker and Sharan, 2008; Baudot *et al.*, 2009; Vidal *et al.*, 2011; Wang *et al.*, 2011; Aranda *et al.*, 2011). In particular, shortest path centrality measures and random walks in protein interaction networks were successfully applied to find novel relationships between genes and diseases (Lage *et al.*, 2007; Köhler *et al.*, 2008; Dezso *et al.*, 2009; Chen *et al.*, 2009; Navlakha and Kingsford, 2010; Guney *et al.*, 2014). Comprehensive network topology analysis also revealed a number of characteristic properties for disease genes that distinguish them from other genes (Jonsson and Bates, 2006; Goh *et al.*, 2007; Yu *et al.*, 2007; Collins, 2015). In this context, we developed the versatile software tool NetworkPrioritizer that supports the integrative network-based prioritization of candidate disease genes using a number of centrality measures and rank aggregation algorithms (Kacprowski *et al.*, 2013).

The integration of protein interaction networks with gene expression data enabled the discovery of disease subnetworks with differential transcriptional profile that suggest key pathways involved in disease progression (Ideker *et al.*, 2002; Chuang *et al.*, 2007; Dittrich *et al.*, 2008; Dobrin *et al.*, 2009; Alcaraz *et al.*, 2012). Other recent approaches have also focused on the discovery of disease modules (Rossin *et al.*, 2011; Jia *et al.*, 2011; Hwang *et al.*, 2012; Guala *et al.*, 2014; Menche *et al.*, 2015; Tasan *et al.*, 2015). In particular, we developed a framework for the functional characterization and integrative prioritization of candidate disease genes and identified a number of promising disease subnetworks for inflammatory bowel diseases (Ellinghaus *et al.*, 2013b), primary sclerosing cholangitis (Liu *et al.*, 2013), and Parkinson's disease (Zanon *et al.*, 2013) (see Chapter 3 for more details).

Furthermore, the application of network biology in drug discovery and development has become more popular in the last years (Hopkins, 2008; Arrell and Terzic, 2010; Barabási *et al.*, 2011; Csermely *et al.*, 2013). Molecular networks are used to identify novel targets for already known drugs (aka drug repositioning) as well as to predict unwanted side effects through network neighborhoods (Yildirim *et al.*, 2007; Keiser *et al.*, 2009; Iorio *et al.*, 2010; Lounkine *et al.*, 2012). In addition, several proteins and even whole modules were detected through network topology analysis as promising biomarkers for disease classification and therapy prediction (Chuang *et al.*, 2007).

2.4.3 Structural biology

Structural biologists have also used network representations to study the interactions of residues in protein structures towards the understanding of complex protein structure-function relationships (Csermely, 2008; Vishveshwara *et al.*, 2009; Krishnan *et al.*, 2008; Csermely *et al.*, 2013; Greene, 2012; Yan *et al.*, 2014). In particular, we presented a novel approach to investigating protein structure-function relationships based on the interactive visual analysis of residue interaction networks (RINs) (Doncheva *et al.*, 2011). These networks are derived from the 3D protein structure and can be analyzed by graph-theoretic methods to identify residues crucial for structure and function and their long-range interactions as also shown by Vendruscolo *et al.* (2001); Amitai *et al.* (2004); Swint-Kruse (2004); del Sol *et al.* (2006); Welsch *et al.* (2008); Susser *et al.* (2009). To support the automatic generation, visualization and analysis of RINs, we developed the software tools RINalyzer and RINerator, which are described in more detail in Chapter 4 and Appendix A.

Recently, several groups have focused on the application of networks for analyzing dynamic protein processes such as (un)folding, allosteric interactions, protein and ligand binding as well as drug resistance in viral proteins (Vishveshwara *et al.*, 2009; Bhattacharyya *et al.*, 2013; Xue *et al.*, 2012; Seeber *et al.*, 2011; Pasi *et al.*, 2012; Sethi *et al.*, 2009; Eargle and Luthey-Schulten, 2012). Thereby, molecular dynamics simulations are represented as networks of interacting residues and are further characterized using topological analysis. More examples as well as our new method for visualizing and analyzing ensembles of protein structures are described in Chapter 5.

To understand complex molecular mechanisms, it is crucial to bridge the gap between systems biology and structural biology (Fraser *et al.*, 2013). Two initiatives aimed at this common goal were linking the visualization of biological networks to the visualization and analysis of protein structures (Morris *et al.*, 2007; Nepomnyachiy *et al.*, 2015) and annotating protein interaction networks with structural information (Mosca *et al.*, 2013). To enhance molecular networks with sequence and structure information and, at the same time, provide a complementary network representation of residue interactions, we recently released a software suite that facilitates a novel interactive, multi-layered analysis of protein interactions and their molecular function in protein binding, allosteric effects, drug resistance and other mechanisms (see Figure 2.4).

2.4.4 Software tools

A number of software tools are available for the visual exploration and computational analysis of networks (Gehlenborg *et al.*, 2010; Chuang *et al.*, 2011; Agapito *et al.*, 2013; Suderman and Hallett, 2007; Pavlopoulos *et al.*, 2008; Villaveces *et al.*, 2015). General software libraries for network analysis are the Java framework JUNG (O'Madadhain *et al.*, 2003), the C++ library LEDA (Mehlhorn and Näher, 1999), the dot-based software Graphviz (Gansner and North, 2000), the Python package NetworkX (Hagberg *et al.*, 2008), and R packages such as igraph (Csárdi and Nepusz, 2006), statnet (Handcock *et al.*, 2008), sna (Butts, 2008), tnet (Opsahl *et al.*, 2010), WGCNA (Langfelder and Horvath, 2008), and QuACN (Mueller *et al.*, 2011). However, they cannot be applied by users without programming expertise.

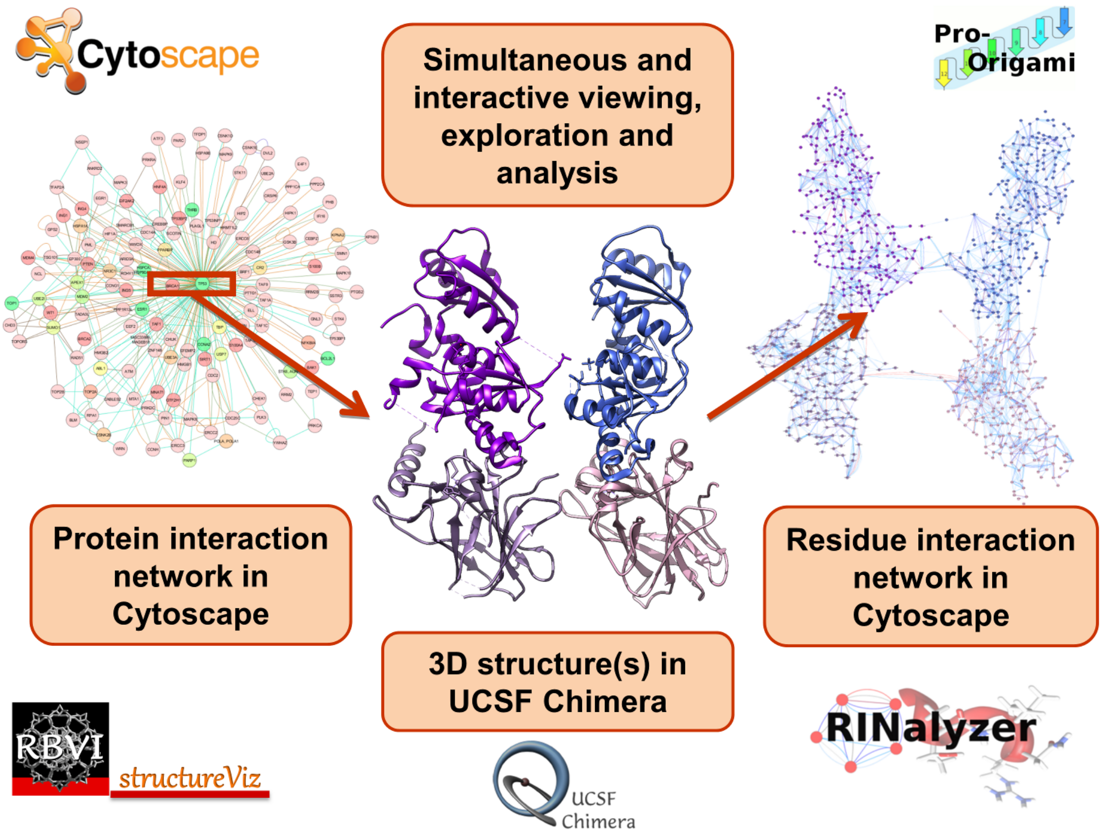

Cytoscape
Simultaneous and interactive viewing, exploration and analysis
Pro-Origami
Protein interaction network in Cytoscape
Residue interaction network in Cytoscape
3D structure(s) in UCSF Chimera
RBVI
structureViz
UCSF Chimera
RINalyzer

Figure 2.4: Bridging the gap between network and structural biology by linking the network analysis and visualization capabilities of Cytoscape and the molecular structure visualizer UCSF Chimera with the help of the Cytoscape apps structureViz and RINalyzer.

In contrast, sophisticated free software platforms such as Pajek (Batagelj and Mrvar, 1998), Osprey (Breitkreutz *et al.*, 2003), VisANT (Hu *et al.*, 2005), ONDEX (Köhler *et al.*, 2006), Gephi (Bastian *et al.*, 2009), NAViGaTOR (Brown *et al.*, 2009), BioLayout Express(3D) (Theocharidis *et al.*, 2009), BIANA (Garcia-Garcia *et al.*, 2010), and BiNA (Gerasch *et al.*, 2014) provide graphical user interfaces and versatile functionality for the analysis and visualization of, in particular, biological networks.

In the last years, the free and stand-alone Cytoscape platform has gained considerable interest because of its open-source code development and its rapidly growing community of users and developers (Shannon *et al.*, 2003). In particular, its functionality is easily extendable by additional apps (previously called plugins) that offer complementary features for the analysis of biological networks (Saito *et al.*, 2012; Pico *et al.*, 2014). Recently, the Cytoscape app store was launched to provide an overview of all available apps and their important features (Lotia *et al.*, 2013). For instance, NetworkAnalyzer (Assenov *et al.*, 2008) performs a comprehensive analysis of network topologies without requiring advanced knowledge in graph theory or programming expertise and has become part of the Cytoscape core due to its wide application by users. While our app RINalyzer (Doncheva *et al.*, 2011) complements NetworkAnalyzer on the particular task of analyzing and visualizing residue interaction networks (RINs) interactively, our other tool NetworkPrioritizer (Kacprowski *et al.*, 2013) supports the network-based prioritization of candidate disease genes.

Some of the most frequently downloaded and used plugins are ClusterMaker (Morris *et al.*, 2010), MCODE (Bader and Hogue, 2003) and jActiveModules (Cline *et al.*, 2007), which can be used to identify and visualize clusters in networks, as well as BiNGO (Maere *et al.*, 2005), ClueGO (Bindea *et al.*, 2009) and CluePedia (Bindea *et al.*, 2013), which provide functional enrichment analysis and facilitate the biological interpretation of sets of genes or proteins. More specialized plugins are, for example, DomainGraph, which has a special focus on the visual analysis of the effect of alternative splicing on gene and protein networks (Emig *et al.*, 2011), and structureViz (Morris *et al.*, 2007), which links biological networks with protein structures visualization and analysis features provided by UCSF Chimera. Another Cytoscape plugin with functionality related to NetworkAnalyzer and RINalyzer is CentiScaPe, which also computes network centrality measures (Scardoni *et al.*, 2009). However, among other differences, it does not provide global measures of network topology as NetworkAnalyzer does, and it does not support weighted networks as RINalyzer does.

In addition to the online tutorials and extensive Cytoscape documentation, several software protocols are already available for network exploration with Cytoscape 2.8.3 (Millán, 2013) and Cytoscape 3 (Su *et al.*, 2014), for the integration of interaction networks with gene expression data (Cline *et al.*, 2007), for cluster analysis with the TransClust and ClusterExplorer plugins (Wittkop *et al.*, 2011), and for the integration of physical and genetic interactions into module maps with the PanGIA plugin (Srivas *et al.*, 2011).

We have recently demonstrated how to apply two of our Cytoscape plugins, NetworkAnalyzer (Assenov *et al.*, 2008) and RINalyzer (Doncheva *et al.*, 2011), for the standard and advanced analysis of network topologies (Doncheva *et al.*, 2012a) as outlined in Appendix A. The first workflow uses NetworkAnalyzer and shows how to conduct a typical topology analysis of biological networks such as protein interaction networks or RINs (Appendix A Step 2A). The second workflow covers various aspects related to the use of RINalyzer for the visual exploration of RINs, the study of protein binding interfaces and the network centrality analysis (Appendix A Step 2B). The third workflow details how to combine NetworkAnalyzer and RINalyzer for the comparison of multiple RINs (Appendix A Step 2C).

CHAPTER 3

Network-based prioritization and functional characterization of candidate disease genes

The field of disease gene prioritization has grown substantially over the last years, and a common feature of all approaches is that they rely on the currently available biomedical knowledge. Besides network-based prioritization approaches, many others have been presented that, for instance, make use of functional annotations or integrate multiple data sources using statistical learning techniques. A distinct advantage of network-based methods is their effectiveness for representing complex relationships between interacting molecules using only a single data source such as molecular interactions or by integrating different data sources in an-easy-to-interpret manner. Many graph theory methods can be applied to integrative gene or protein networks to gain more insight into complex disease mechanisms by characterizing the topological network structure as well as global and local interaction properties.

In this chapter, we first give an overview of recent computational approaches to the identification of the most promising candidates for experimental follow-up validation in line with our recent review (Doncheva *et al.*, 2012b). Furthermore, we describe our efforts towards improving existing methods and tools for candidate disease gene prioritization with focus on network-based approaches and integration of biomedical data. In addition to our methodological contributions to bioinformatics research, an important part of our work is dedicated to the direct application of newly developed or established methods to answer biomedical research questions. We designed a phenotype-specific framework for prioritization and functional characterization of candidate genes that makes use of existing databases and tools developed by colleagues in our group or by collaborators, in particular FunSimMat (Schlicker and Albrecht, 2008), Cytoscape (Shannon *et al.*, 2003), and the Cytoscape plugins NetworkPrioritizer (Kacprowski *et al.*, 2013) and ClusterOne (Nepusz *et al.*, 2012). In

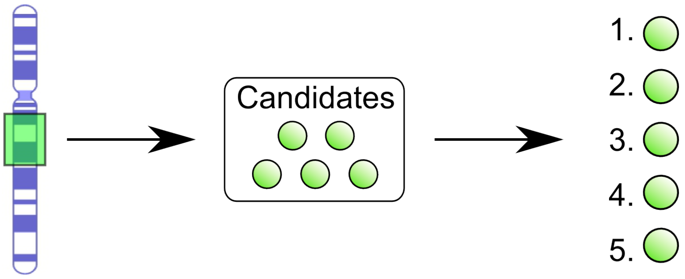


Figure 3.1: Schematic representation of the process of disease gene prioritization

close collaboration with biologists and physicians, we applied it to three different phenotypes: inflammatory bowel diseases (Ellinghaus *et al.*, 2013b), primary sclerosing cholangitis (Liu *et al.*, 2013), and Parkinson's disease (Zanon *et al.*, 2013).

3.1 Introduction

Many common diseases are complex and polygenic, involving dozens of human genes that might predispose to, be causative of, or modify the respective disease phenotype (Schreiber *et al.*, 2005; Hirschhorn and Gajdos, 2011; Raychaudhuri, 2011). This intricate interplay of disease genotypes and phenotypes still complicates the identification of all relevant disease genes (Frazer *et al.*, 2009; Mackay *et al.*, 2009; Hawkins *et al.*, 2010a). Therefore, a number of techniques exist to discover disease genes. In particular, high-throughput methods such as genome-wide association studies (Franke *et al.*, 2008; Manolio, 2010; Hirschhorn and Gajdos, 2011) and large-scale RNA interference screens (Boutros and Ahringer, 2008; Hawkins *et al.*, 2010a; Reiss *et al.*, 2011) yield lists of up to hundreds of candidate disease genes. As validating the actual disease relevance of candidate genes in experimental follow-up studies is a time-consuming and expensive task, many methods and web services for the computational prioritization of candidate disease genes have already been developed and recently reviewed in Vidal *et al.* (2011); Barabási *et al.* (2011); Tranchevent *et al.* (2011); Wang *et al.* (2011); Piro and Di Cunto (2012); Doncheva *et al.* (2012b); Moreau and Tranchevent (2012); Lehner (2013); Bromberg (2013); Gustafsson *et al.* (2014); Collins (2015).

The concrete problem of candidate gene prioritization can be formulated as follows: Given a disease (or a specific phenotype) of interest and some list of candidate genes, identify potential gene-disease associations by ranking the candidate genes in decreasing order of their relevance to the disease phenotype (Figure 3.1). When abstracting from the methodological details, the vast majority of computational approaches to this prioritization problem work in a similar manner. Most of them rely on the biological information already available for the disease of interest, the known, already verified, disease genes, and the newly suggested candidate genes.

These data then serve as input for statistical learning methods or are integrated into network representations, which are further analyzed by network scoring algorithms. Although individual data sources such as functional annotations or protein interactions provide quite powerful information for prioritizing candidate genes, the integration of multiple data sources has been reported to increase the performance even more. However, a generally accepted and consistent benchmarking strategy for all the diverse prioritization methods has not emerged yet, which complicates performance evaluation and comparison.

3.1.1 Our methodological contributions

One of the main goals of current methodological development is the investigation of less frequently studied phenotypes with unknown causative genes. A suitable platform for such studies is the recently developed Immunochip, a custom Illumina array that allows the dense genotyping of poorly understood immune-mediated diseases (Cortes and Brown, 2011). Using this array, our cooperation partners Liu *et al.* (2013) performed a study on primary sclerosing cholangitis (PSC), a severe liver disease of unknown etiology, and suggested several previously unknown disease associated loci. In order to prioritize and functionally characterize the PSC loci, we generated networks of strong functional similarities between the candidate genes based on their Gene Ontology annotations. We also devised a novel method for assessing the connectivity of the candidate genes without knowledge of known causative genes. Through network topology analysis, we selected one disease-relevant gene per locus and built a disease-specific network. Furthermore, we extended our method to the analysis and comparison of the functional similarities between genes associated with both diseases and revealed a large functional overlap of PSC with inflammatory bowel disease (IBD). Our computational framework is described in more detail in Section 3.2 and the analysis of PSC and IBD loci as included in the publication by Liu *et al.* (2013) is presented in Section 3.4.

We also applied our computational framework on recently published GWAS data for IBD, a chronic inflammatory disorder of the gastrointestinal tract (Franke *et al.*, 2010; Anderson *et al.*, 2011). In particular, we assessed the functional overlap between the two main IBD subtypes, Crohn's disease (CD) and ulcerative colitis (UC). We also generated an integrative network of publicly available physical protein interactions and strong functional similarities between candidate and known IBD genes. Based on node connectivity and relevance in this network, we prioritized the IBD candidates and generated a disease-specific network with the top-ranked gene per locus. In addition to our already established framework, we made use of the NetworkPrioritizer plugin developed by Tim Kacprowski for performing rank aggregation of the different measures (Kacprowski *et al.*, 2013). In a second study, we combined exome sequencing data of CD patients and healthy individuals, provided to us by our collaboration partners, with a network of genes associated with CD to suggest candidate genes with rare sequence variants. In addition, we explored the relationships between a gene identified in the exome analysis and genes known to be associated with CD or to be involved in autophagy. The latter network analysis was included in the publication by Ellinghaus *et al.* (2013b). For more details, refer to Section 3.3.

For our cooperation project with Francisco S. Domingues, we adapted our computational analysis framework to the prioritization of proteins associated with Parkinson's disease (PD), a progressive neurodegenerative disorder of the central nervous system (Zanon *et al.*, 2013). Our contribution to this project was to rank the PD candidate proteins based on their shortest path to known PD proteins in a network of human protein-protein interactions and to identify groups of functionally similar PD candidate and known proteins in a network of strong functional similarities (see Section 3.5 for more details). These and other selection criteria were finally combined by our collaboration partners to suggest a list of best-ranking predictions for further experimental validation (Zanon *et al.*, 2013).

3.1.2 Related work

This section is an adapted and modified version of text contained in Doncheva *et al.* (2012b). In our review (Doncheva *et al.*, 2012b), we categorize the various prioritization methods according to the biological data and their representation that are primarily considered when scoring and ranking candidate disease genes: gene and protein characteristics, network information on molecular interactions, and integrated biomedical knowledge. In the following, we summarize and update each of these categories with focus on network-based approaches. We also discuss different benchmarking strategies and the need of standardized procedures for performance measurement.

Many of the earlier disease gene prioritization methods exploited discriminative gene and protein properties, assuming that candidate genes, which satisfy properties derived from known disease genes and proteins, are more likely to be relevant to the disease. López-Bigas and Ouzounis (2004) and Adie *et al.* (2005) identified several sequence features that discriminate disease genes from non-disease genes. After Jimenez-Sanchez *et al.* (2001) demonstrated that gene and protein function strongly correlates with disease features, such as age of onset, many prioritization approaches exploited the functional annotations of known disease genes and successfully ranked candidate genes based on their functional similarity to the disease of interest or its associated genes (Perez-Iratxeta *et al.*, 2002; Freudenberg and Propping, 2002; Turner *et al.*, 2003; Schlicker *et al.*, 2010; Ramírez *et al.*, 2012). These studies also confirmed the assumption that phenotypically similar diseases often involve common molecular mechanisms and thus functionally related genes.

In the last decade, molecular interaction networks have become an indispensable tool and a valuable information source in the study of human diseases. Thus, many prioritization methods use protein interaction data as a powerful source for finding relationships between gene products of candidate genes and disease genes (Kann, 2007; Ideker and Sharan, 2008; Baudot *et al.*, 2009; Vidal *et al.*, 2011; Wang *et al.*, 2011; Aranda *et al.*, 2011). Disease genes and their products have discriminatory interaction network properties that allow their distinction from non-disease genes. Furthermore, the application of the *guilt-by-association* principle is straightforward in protein interaction networks.

Since disease proteins tend to cluster and interact with each other (Jonsson and Bates, 2006; Lim *et al.*, 2006; Goh *et al.*, 2007; Feldman *et al.*, 2008; Yao *et al.*, 2011; Menche *et al.*, 2015), early prioritization methods focused on local network information such as the topological neighborhood of a node representing a candidate gene or protein (Krauthammer *et al.*, 2004; Karni *et al.*, 2009; Oti *et al.*, 2006; Xu and Li, 2006; Lage *et al.*, 2007). Examples for local topology measures are node degree (number of edges linked to node n) and shortest path length (minimum number of edges between two nodes). Oti *et al.* (2006) proposed a very simple method for a genome-wide prediction of disease genes based on their disease-associated chromosomal location and direct protein interaction with a disease protein. In contrast, Xu and Li (2006) trained a k-nearest-neighbor classifier on multiple topological properties and three different molecular networks. The approach by Lage *et al.* (2007) assigns a high score to those candidates that have protein interactions neighbors associated with phenotypes similar to the disease of interest.

However, local measures are less sensitive to the overall network topology and ignore potential network-mediated effects from distant nodes. Thus, global network measures were introduced in the methods for candidate disease gene prioritization and considerably improved their performance (Köhler *et al.*, 2008; Suthram *et al.*, 2008; Dezso *et al.*, 2009; Chen *et al.*, 2009; Navlakha and Kingsford, 2010; Bottomly *et al.*, 2013; Guney *et al.*, 2014). Global network information relates to the overall network topology and is retrieved by measures that characterize the role of a node with respect to the whole network. Commonly used centrality measures are shortest path closeness and betweenness as well as random-walk related properties such as hitting time and visit frequency (see Section 2.3.1 for details). The most famous and frequently re-used approach based on global network information was introduced by Köhler *et al.* (2008). They used a random walk with restart (from known disease genes) on a protein interaction network to rank the candidate gene products.

Especially for the study of complex diseases, network topology analysis is more useful than approaches based on individual gene or protein properties as it can provide more insight into the functionality and interplay of disease genes by revealing the alternative paths of interactions of their gene products. The performance of network-based prioritization approaches depends heavily on the quality of the data. Protein interaction networks are well known to be biased towards extensively studied proteins and subject to inherent noise (Xu and Li, 2006; Ramírez *et al.*, 2007; Cusick *et al.*, 2009).

Further performance gain can be achieved by comprehensive knowledge integration from multiple data sources. This procedure is able to reduce the noise in the integrated data and to provide additional information that is not captured (yet) by other approaches (Moreau and Tranchevent, 2012). Two distinct approaches to disease gene prioritization that exploit multiple data sources are exemplarily highlighted in Figure 3.2. The first approach combines heterogeneous datasets in a network representation, and subsequently applies specific analysis techniques, such as network topology measures, to score and rank candidates with regard to their

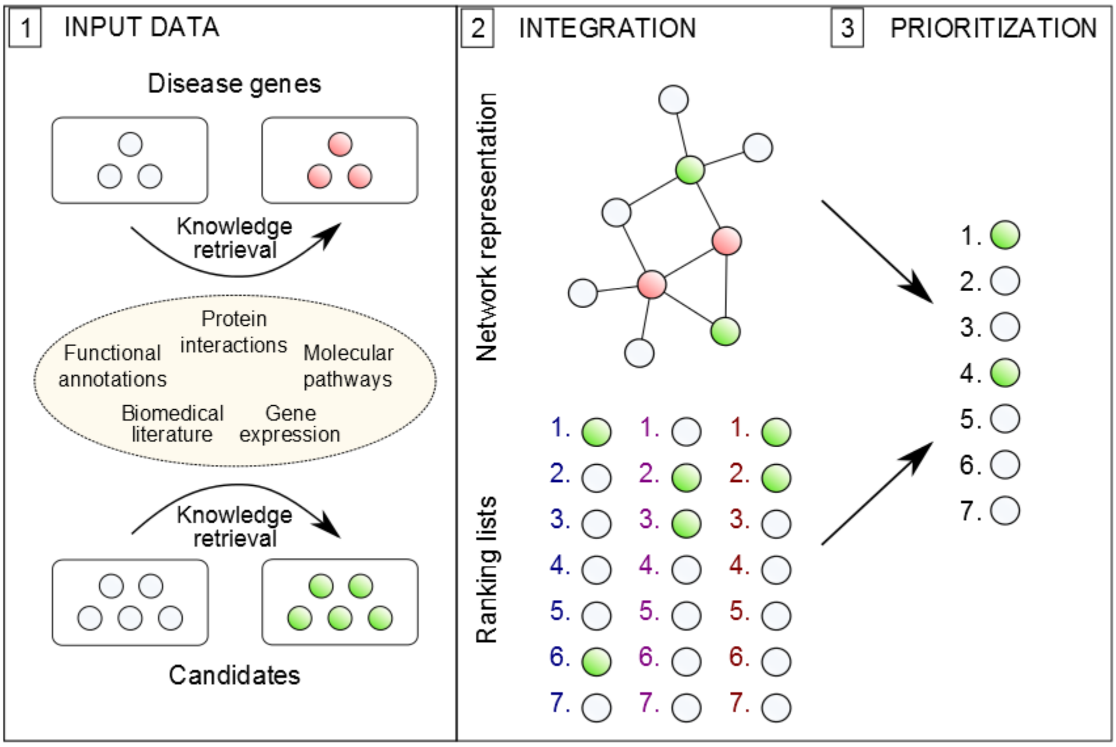


Figure 3.2: Integrative approaches to disease gene prioritization. The typical workflow of integrative prioritization approaches based on multiple data sources consists of three major steps. The first step involves preparing the input data consisting of two different sets of genes, the known disease genes and the candidate genes. For each gene, further biomedical knowledge is retrieved from various data sources such as functional annotations from the Gene Ontology and molecular pathways from the KEGG database. In the second step, the collected information is integrated using a network representation (top) or evaluated individually for each data source, resulting in different ranking lists (bottom). The third step computes a final ranking list of candidate genes based on network measures or rank aggregation. The candidate genes are thus prioritized by their relevance to the disease of interest. Figure first published in Doncheva *et al.* (2012b).

network proximity to nodes representing known disease genes (Franke *et al.*, 2006; Lage *et al.*, 2007; Linghu *et al.*, 2009; Huttenhower *et al.*, 2009; Li and Patra, 2010a; Lee *et al.*, 2011; Liekens *et al.*, 2011; Hoehndorf *et al.*, 2011; Wu *et al.*, 2008; Yao *et al.*, 2011; Chen *et al.*, 2011a; Guo *et al.*, 2011; Vanunu *et al.*, 2010; Yang *et al.*, 2011; Hwang *et al.*, 2011). However, the usefulness and biological relevance of the individual data sources are usually not evaluated by such approaches. It is also beneficial to first analyze each data source separately using the most suitable techniques and then combine the resulting ranking lists using sophisticated rank aggregation algorithms (Aerts *et al.*, 2006; Li and Patra, 2010b; Pers *et al.*, 2011; De Bie *et al.*, 2007; Radivojac *et al.*, 2008; Costa *et al.*, 2010; Yu *et al.*, 2011b; Chen *et al.*, 2011b; Mordelet and Vert, 2011). This procedure also facilitates backtracking the origin of the most relevant information.

One of the earliest and still well-recognized integrative approaches is Endeavour (Aerts *et al.*, 2006; Tranchevent *et al.*, 2008; Schuierer *et al.*, 2010). It utilizes more than twenty data sources, such as ontologies and functional annotations, protein-protein interactions, cis-regulatory information, gene expression data, sequence information, and text-mining results. For each data source, candidate genes are first ranked separately based on their similarity to a profile derived from known disease genes. Afterwards, all individual candidate rankings are merged into a final overall ranking using rank order statistics. MetaRanker is a similar approach that combines many heterogeneous data sources and is particularly suited to uncover associations in complex polygenic diseases (Pers *et al.*, 2011).

An alternative way of integrating information from multiple data sources is the application of machine learning techniques. Each data source can be represented as one or more individual features and used as input for the training of supervised learning methods. In particular, support vector machines (De Bie *et al.*, 2007; Radivojac *et al.*, 2008; Yu *et al.*, 2011b), decision tree based classifiers (Costa *et al.*, 2010), and PU learning (machine learning from positive and unlabeled examples) (Mordelet and Vert, 2011) have been applied to prioritize candidate disease genes using multiple data sources.

Prioritizer was one of the first approaches to integrate information from multiple data sources into a network representation and rank the candidate genes according to the length of the shortest paths between them (Franke *et al.*, 2006). Building upon Prioritizer, several research groups have assembled different types of integrated networks as biological evidence for candidate disease gene prioritization and applied local or, preferably, global network topology measures (Li and Patra, 2010a; Linghu *et al.*, 2009; Huttenhower *et al.*, 2009; Lee *et al.*, 2011). A more general view of the relationships between phenotypes and genes is introduced by BioGraph, a heterogeneous network containing diverse biomedical entities and relations between them (Liekens *et al.*, 2011).

After several studies indicated that similar phenotypes often share underlying genes or even pathways (Limviphuvadh *et al.*, 2007; Oti and Brunner, 2007; Van Driel *et al.*, 2006), phenotypic similarity has become another major data source exploited by computational methods for prioritization of candidate disease genes. Such phenotypic knowledge can be very useful to discover new potential disease genes by transferring known gene-phenotype associations to similar diseases and phenotypes (Lage *et al.*, 2007; Hoehndorf *et al.*, 2011). Several methods are based on a two-layered heterogeneous data network such that the phenome layer consists of connections between similar phenotypes, the interactome layer of protein-protein interactions, and known gene-phenotype associations link both layers (Wu *et al.*, 2008; Li and Patra, 2010a; Yao *et al.*, 2011; Chen *et al.*, 2011a; Guo *et al.*, 2011; Vanunu *et al.*, 2010; Yang *et al.*, 2011; Hwang *et al.*, 2011). These approaches usually apply measures based on random walks to find causal genes among the candidates.

Another useful source of information is gene expression data, which is usually mapped onto protein interaction networks to identify groups of differentially expressed candidate genes (Nitsch *et al.*, 2009, 2010; Zhao *et al.*, 2011; Nitsch *et al.*, 2011; Wu *et al.*, 2012; Poirel *et al.*, 2013; Wang *et al.*, 2014). Furthermore, gene expression data was used to construct tissue-specific protein interaction networks that aid in prioritizing disease causing genes (Magger *et al.*, 2012) and in understanding the local manifestation of hereditary diseases (Barshir *et al.*, 2014).

Recently, the application of local network topology, in particular, the identification of network modules or clusters, has also gained more attention (Menche *et al.*, 2015; Tasan *et al.*, 2015; Rossin *et al.*, 2011; Jia *et al.*, 2011; Guala *et al.*, 2014; Hwang *et al.*, 2012). In particular, Menche *et al.* (2015) uncovered disease modules in protein interaction networks that consist of proteins associated with the same disease as well as overlapping modules that confirm disease similarity, while Rossin *et al.* (2011) showed that genes associated with GWA loci are more densely connected by protein interactions than expected by chance. Overall, the integration of GWAS results into prioritization approaches has been very popular recently (Azencott *et al.*, 2013; Hou *et al.*, 2014; Wang *et al.*, 2015). Tasan *et al.* (2015) presented an interesting approach that identifies subnetworks of mutually functionally related genes that span multiple GWA loci. On top of GWAS, several recent studies successfully combined biological networks with exome sequencing data to identify novel disease genes (Dand *et al.*, 2013; Smedley *et al.*, 2014).

3.1.3 Evaluation and benchmarking

This section has been adapted from Doncheva *et al.* (2012b). To show the biological applicability and scientific value of disease gene prioritization methods, their authors are normally expected to conduct an extensive performance evaluation and, if possible, a thorough comparison with other methods. To this end, many authors usually benchmark disease phenotypes from OMIM. Depending on the requirements of their method, only phenotypes with at least 2 or 3 known disease genes may be suitable. Hence, the number of evaluated diseases can vary from tens to hundreds with hundreds to thousands corresponding genes. The range of disease phenotypes and genes, for which a given method is applicable, depends on the data used by the method. For instance, only about 20 % of all possible human protein-protein interactions have been described so far and only about 10 % of all human genes have at least one known disease association (Menche *et al.*, 2015; Barabási *et al.*, 2011; Amberger *et al.*, 2009). In addition, only about every second gene or protein is functionally annotated (Huntley *et al.*, 2015).

Here, we briefly describe frequently used measures for evaluating the performance of disease gene prioritization methods. Leave-one-out cross-validation is a widely used and generally accepted test for how a method might perform on previously unseen data. In each run, one of the known disease genes, the so-called target disease gene, is removed from the training data. The remaining disease genes are used to identify the omitted gene from a test set of genes that are not known to be associated with the disease of interest. In the best case, the top rank should be assigned to the target disease gene and lower ranks to the other test genes. Since cross-validation is a standard performance test, a number of suitable measures of predictive power exist, for example, sensitivity and specificity, receiver-operating characteristic (ROC) curve, precision and recall, enrichment and mean rank ratio. They are calculated using the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) at a specific rank or score cut-off that discriminates predicted from not predicted ones.

In this specific scenario, positives are disease genes, while negatives are candidate genes without a disease association. For instance, sensitivity *(TP/(TP + FN))* is the percentage of correctly identified disease genes among all genes above the cut-off, while specificity *(TN/(TN + FP))* is the percentage of correctly dismissed candidate genes among all genes below the cut-off. Plotting sensitivity versus specificity while varying the cut-off yields a ROC curve. The area-under-the-ROC curve (AUC) is a standard measure for the overall performance of binary classification methods (here, disease genes vs. others). The AUC is 100 % in case of perfect prioritization and 50 % if the disease genes were ranked randomly.

In some cases, the authors of prioritization methods give the percentage of disease genes ranked in the top 1 % and 5 % of all genes, which corresponds to reporting the sensitivity at 99 % or 95 % specificity, respectively. The percentage of correctly prioritized disease genes among all disease genes is defined as precision (*TP/(TP + FP)*), while recall is equal to sensitivity. Thus, the plot of a precision-recall curve can also be used to evaluate method performance (Navlakha and Kingsford, 2010). An additional, rather simple measure is the mean rank ratio defined as the average of rank ratios for all tested disease genes (Zhang *et al.*, 2011). Finally, authors refer to the n/m-fold enrichment on average if disease genes are ranked in the top m % of all genes in n % of the linkage intervals (Wu *et al.*, 2008).

Since all of these measures evaluate the performance of a method in a slightly different way, none of them is considered as default and authors rarely report all of them. This complicates the comparison between different methods of disease gene prioritization. In this particular case, it would be useful to also report the performance for the top-ranked candidate genes, e.g., the first ten or twenty genes because only a few candidates can usually be considered for further validation experiments.

Another important aspect of the benchmarking strategy is the choice of genes in the test set, that is, the candidate genes that are prioritized together with the target disease gene. One usual input for prioritization methods is a set of susceptibility loci as determined by GWA studies. These loci typically contain up to several hundreds of possible disease genes. Therefore, different strategies have been followed to derive useful test sets, that is, the definition of artificial gene loci, the random selection of genes, the use of the whole genome, and the small-scale choice of genes. Endeavour (Aerts *et al.*, 2006) and several other methods, which were either related or compared with it, were evaluated with a test set containing 99 candidate genes chosen at random from the whole genome in addition to the target disease gene.

However, since similar genes tend to cluster in chromosomal neighborhoods (Lee and Sonnhammer, 2003), another, presumably, more difficult setting for performance benchmarking and especially more relevant for GWAS is the definition of artificial linkage intervals with genes that surround the disease gene on the chromosome. The size of such intervals, as found in the relevant literature, ranges from the 100 nearest genes to 300 genes on average if a 10 Mb genomic neighborhood is considered (Schlicker *et al.*, 2010). The average gene number of linkage intervals associated with diseases according to OMIM is estimated to be 108 (Lage *et al.*, 2007). The third option for assembling a test set is the use of all genes in the genome except for the known disease genes in the training set and can be chosen only by the few methods that are capable of performing genome-wide disease gene prioritization. Finally, prioritization methods that consider, for instance, gene expression data are evaluated only on a smaller scale because there is not enough data for a comprehensive benchmarking over many disease phenotypes. Therefore, the authors commonly choose only few diseases that have, for example, the required experimental data available.

Nevertheless, experimental validation remains the most important and valued evaluation of the outcome of prioritization methods. This is, however, very difficult to accomplish as most computational biology labs do not have an in-house wet lab or even a direct cooperation with biologists or physicians, who are currently studying the disease in question. Therefore, we made an effort to work closely with researchers, who are also either physicians themselves or work with physicians. During this work, we successfully cooperated with Andre Franke from the Christian-Albrechts-University of Kiel (Ellinghaus *et al.*, 2013b), Tom H. Karlsen from Oslo University Hospital (Liu *et al.*, 2013), and Francisco S. Domingues from EURAC research (Zanon *et al.*, 2013).

3.2 Network-based prioritization framework

Based on our review (Doncheva *et al.*, 2012a) and more recently published papers, we concluded that there is still room for methodological improvements, in particular, with regard to the sources of biomedical knowledge and how they are exploited, integrated and evaluated. Since phenotypes can strongly differ in their genetic characteristics as well as in the amount of research dedicated to them, prioritization approaches should be tailored to specific phenotypes or groups of phenotypes.

The available sources of biological information can be used not only to prioritize candidate genes, but also to generate disease-specific networks that provide more insight into the functional characteristics of the underlying phenotypes. Last but not least, there is a demand for more user-friendly and interactive web interfaces for disease gene prioritization as well as for stand-alone tools that support the input of user-specific data and provide further visualization and analysis functionality. In this section, we will describe our efforts towards the aforementioned challenges.

3.2.1 Functional similarity networks

The Gene Ontology (GO) is a comprehensive resource for functional gene and protein annotations represented as three structured controlled vocabularies: biological process (BP), molecular function (MF) and cellular component (CC) (Ashburner *et al.*, 2000; Gene Ontology Consortium, 2015). Two of the widely used applications of GO are gene set enrichment analysis (Subramanian *et al.*, 2005; Huang *et al.*, 2009a; Hung *et al.*, 2012) and semantic similarity analyses (Pesquita *et al.*, 2009; Du Plessis *et al.*, 2011; Guzzi *et al.*, 2012). In this section, we will discuss how GO annotations have advanced candidate disease gene prioritization and suggest a new way of representing functional relationships between genes and proteins based on GO annotations.

Motivation

After Jimenez-Sanchez *et al.* (2001) demonstrated a strong correlation between disease features, such as age of onset, and the function of genes and proteins, the first prioritization approaches that exploit functional annotations of known disease genes for ranking candidates were presented (Perez-Iratxeta *et al.*, 2002; Freudenberg and Propping, 2002; Turner *et al.*, 2003; Schlicker *et al.*, 2010; Ramírez *et al.*, 2012; Li *et al.*, 2013). Perez-Iratxeta *et al.* (2002) mined the biomedical literature using GO and Medical Subject Headings (MeSH) terms (Lowe and Barnett, 1994) to relate disease phenotypes with functional annotations. In a similar fashion, Freudenberg and Propping (2002) identified candidate genes based on their annotated GO terms that are shared with groups of known disease genes associated with similar phenotypes. In contrast, the approach POCUS assesses the shared over-representation of functional annotation terms between genes in different loci for the same disease (Turner *et al.*, 2003).

In order to facilitate the usage of phenotypic data in a similar way as the GO functional annotations, Robinson *et al.* (2008) developed the Human Phenotype Ontology (HPO). Currently, it consists of 10,088 terms describing 7,278 human hereditary syndromes and 13,326 relations between the terms (Köhler *et al.*, 2014). One of the first applications of the HPO was a method developed by Köhler *et al.* (2009) and called Phenomizer. They made use of the structured data in HPO and applied a set of similarity measures to rank candidate genes and refine differential clinical diagnosis by suggesting clinical features that differentiate among several candidate diagnoses. Several other computational approaches successfully employed the semantic phenotype similarities from HPO to prioritize candidate disease genes (Robinson *et al.*, 2014; Singleton *et al.*, 2014; Javed *et al.*, 2014; Yang *et al.*, 2015).

Recently, Schlicker *et al.* (2010) developed a prioritization method based on the similarity between the functional annotations of disease genes and candidates. In contrast to the approaches that consider solely identical functional annotations or GO term enrichments, MedSim automatically derives functional profiles for each disease phenotype from the GO term annotation of known disease genes. Candidate genes are then scored and ranked according to the functional similarity of their annotation profiles to a disease profile. In addition, Ramírez *et al.* (2012) introduced the BioSim method for discovering biological relationships between genes or proteins. While MedSim is based only on GO term annotations, BioSim quantifies functional gene and protein similarity according to multiple data sources of functional annotations and can also be applied to rank candidate genes based on their functional similarity to known disease genes.

In a case study of Hepatitis C virus siRNA screen, Reiss *et al.* (2011) successfully applied a network-based method for the analysis of host factor candidates by annotating them with molecular interaction data and functional protein and gene annotations. The method was originally developed by Nora Speicher under the joint supervision of Hagen Blankenburg and Mario Albrecht. Among others, Speicher (2010) showed that pairs of human host factors identified for different viral infections in the literature are significantly more similar to each other than randomly selected protein pairs.

The success of the presented studies also indicates that diseases with similar phenotype often involve common molecular mechanisms and thus functionally related genes. This also explains the frequent use of functional annotations as important biological evidence in integrative prioritization approaches. Notably, the information value of functional annotations can be further increased by improved scoring of functional similarity, reaching the performance of complex integrative methods based on multiple data sources (Schlicker *et al.*, 2010).

Definition

Motivated by the wide and successful application of GO annotations and, in particular, functional similarity values, we constructed functional similarity networks (FSNs) as a complement to protein interaction networks and designed an integrative prioritization framework based on them. In a functional similarity network, a node corresponds to a gene or gene product and two nodes are connected by an edge if their pair-wise functional similarity as derived from their GO annotations is above a user-defined cut-off (Figure 3.3). Some of the advantages of FSNs are their higher coverage compared to PINs as well as their ability to relate genes or proteins with respect to their cellular function even if they are not physically interacting (Hawkins *et al.*, 2010b; Jiang *et al.*, 2011).

To construct an FSN, we retrieve functional similarities from the FunSimMat web

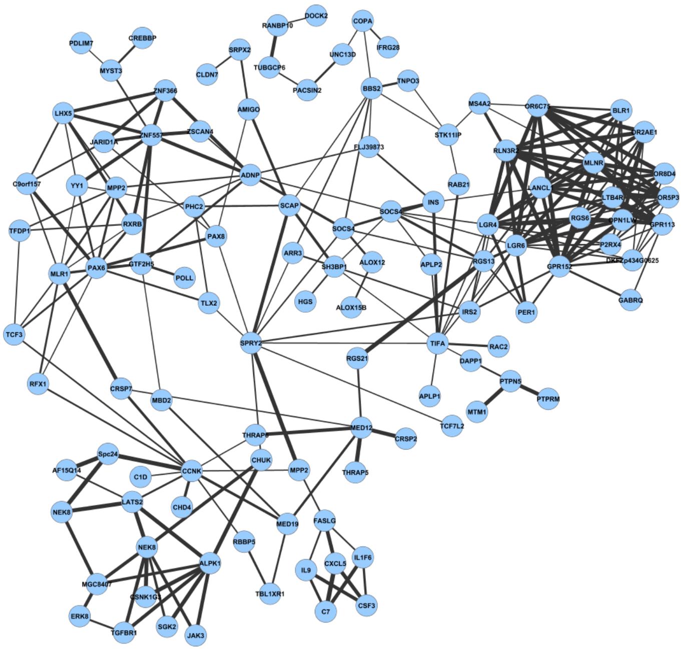


Figure 3.3: Example for a functional similarity network. Nodes represent genes and edges indicate strong functional similarity (between 0.8 and 1.0) based on the GO annotations of the genes. The edge lines thickness reflects the functional similarity.

service (Schlicker and Albrecht, 2008, 2010) for a given set of genes or gene products using the Python package developed by Speicher (2010). For most of our analyses, we decided on the *rfunSim* measure (Schlicker *et al.*, 2006) as it accounts for both the Biological Process and Molecular Function domains of GO and the established cut-off of 0.8 for strong functional similarity (Liu *et al.*, 2013). Alternatively, any other functional similarity measure or a different cut-off might be chosen. We explain the most common semantic and functional similarity measures in the following section. We visualize and analyze the resulting networks using Cytoscape (Shannon *et al.*, 2003) and the Cytoscape apps NetworkAnalyzer (Doncheva *et al.*, 2012a), NetworkPrioritizer (Kacprowski *et al.*, 2013), SubnetworkGenerator, setsApp (Morris *et al.*, 2015b), and ClusterONE (Nepusz *et al.*, 2012).

Previous analysis of genome-wide FSNs for *Escherichia coli*, *Saccharomyces cerevisiae*, and *Plasmodium falciparum* (malaria) revealed that they have different network topology than protein interaction networks, in particular, higher modularity (Hawkins *et al.*, 2010b). Furthermore, FSNs constructed with the *funSim* measure exhibited a tendency to be hierarchical compared to FSNs accounting only for one of the GO domains (*BPscore*, *MFscore* or *CCscore*) (Hawkins *et al.*, 2010b). By constructing an FSN for all human genes, Jiang *et al.* (2011) demonstrated that it covers 50 % more genes than a widely used PIN, the Human Protein Reference Database network (Keshava Prasad *et al.*, 2009), and that the functional similarity between two genes correlates with the proximity of their gene products in the PIN. Moreover, the authors showed that genes associated with similar diseases are functionally more similar than randomly selected genes and that FSNs are at least as good or even better data source for network-based candidate disease gene prioritization using a random walk with restart algorithm (Jiang *et al.*, 2011).

Semantic and functional similarity

There are different ways to assess the functional similarity between two biological entities and here we focus on the application of the Gene Ontology. In the last years, many methods were developed for comparing sets of GO terms with respect to their information content and the GO graph structure (Pesquita *et al.*, 2009; Du Plessis *et al.*, 2011; Guzzi *et al.*, 2012). They are called semantic similarity measures. In order to compare two genes or proteins annotated with a set of GO terms, functional similarity measures were introduced. They are based on the pair-wise semantic similarity of the annotated terms. Two of the most well-known semantic similarity measures are Lin's and Resnik's measure.

Resnik's measure is based on the concept of information content (IC) (Resnik, 1995, 1998). It is defined as

$$sim_{Resnik}(c_1, c_2) = \max_{c \in S(c_1,c_2)} (-\log p(c))$$

where $S(c_1, c_2)$ is the set of common ancestors of terms c_1 and c_2. The information content (IC) of term c can be expressed as $IC(c) = -\log_{10} p(c)$, where the probability of a term is its relative frequency of occurrence in the whole ontology. The more information two terms share, the higher is their similarity. Thus, the lowest common ancestor (LCA) is $\max_{c \in S(c_1,c_2)} IC(c)$, i.e., the most informative term.

Lin's measure is represented as the ratio of the commonality of two GO terms (their common ancestors) and the information needed to fully describe them (the sum of their information) (Lin, 1998). It is formally defined as

$$sim_{Lin}(c_1, c_2) = \max_{c \in S(c_1,c_2)} \left(\frac{2 \log p(c)}{\log p(c_1) + \log p(c_2)} \right)$$

In contrast to Resnik's measure, which ranges from zero to infinity, Lin's measure is bounded by 0 and 1.

In 2006, Schlicker *et al.* (2006) presented a new semantic and functional similarity measure called *simRel* and *funSim*, respectively. As a combination of Lin's and Resnik's semantic similarity measures, the relevance similarity (sim_{Rel}) captures how close two terms are to their LCA as well as how informative the LCA is. It is defined as

$$sim_{Rel}(c_1, c_2) = \max_{c \in S(c_1,c_2)} \left(\frac{2 \log p(c)}{\log p(c_1) + \log p(c_2)} (1 - p(c)) \right)$$

and the values range between 0 and 1.

As already mentioned before, a functional similarity measure needs to account for the comparison of two sets of GO terms. The *funSim* measure proposed by Schlicker *et al.* (2006) is defined in several steps. Given two sets of GO terms GO^A and

GO^B of size N and M annotated to proteins A and protein B, respectively, a matrix of the pairwise semantic similarity is computed such that each entry $s_{ij} = sim(GO_i^A, GO_j^B), \forall i \in 1, \ldots, N, \forall j \in 1, \ldots, M$ is the semantic similarity between GO term i annotated to protein A and GO term j annotated to protein B. At this point, any semantic similarity measure can be chosen to construct the matrix.

Since the sets of GO terms are different, the resulting matrix is not necessarily square or symmetric. More importantly, the rows and columns represent the comparison of A to B and B to A, respectively, and the maxima of each row/column are the best matching pairs of GO terms. To get a final score for the comparison of proteins A and B, a *GOscore* is defined as

$$GOscore = \max\{rowScore, columnScore\}$$

where the *rowScore* and *columnScore* are the averages over the row maxima and the column maxima:

$$rowScore = \frac{1}{N}\sum_{i=1}^{N} \max_{1 \leq j \leq M} s_{ij} \text{ and } columnScore = \frac{1}{M}\sum_{j=1}^{M} \max_{1 \leq i \leq M} s_{ij}.$$

Alternatively, the *GOscore* can also be calculated as the average of the row and column scores. Since there are three different GO domains (BP, MF, CC), a *GOscore* can be computed for each of them and called accordingly, i.e., *BPscore*, *MFscore* and *CCscore*.

Schlicker *et al.* (2006) suggested a functional similarity score *funSim* as a combination of the BP similarity and the MF similarity between two proteins. Formally, it is defined as

$$funSim = \frac{1}{2}\left[\left(\frac{BPscore}{\max(BPscore)}\right) + \left(\frac{MFscore}{\max(MFscore)}\right)\right]$$

where $\max(BPscore)$ and $\max(MFscore)$ are the maximum possible scores for the whole biological process and molecular function domains, respectively. In particular, this score favors proteins that are very similar based on one GO domain to proteins that are less similar according to both GO domains. However, the *funSim* score is lower than the average of *BPscore* and *MFscore*. Thus, the more intuitive *rfunSim* score can be calculated as the square root of the *funSim* score. It ranges from 0 for no functional similarity to 1 for maximal functional similarity. Compared to the *funSim* score, *rfunSim* was better suited for classifying protein pairs without sequence similarity and orthologous protein pairs (Schlicker *et al.*, 2007b).

In contrast to methods that only consider the overlap of identical GO terms, the *funSim* measure can be used to compare proteins with partial functional annotations or multi-functional proteins. Several studies have shown that this functional similarity is very useful for the prediction and validation of protein-protein and domain-domain interactions (Suthram *et al.*, 2006; Ramírez *et al.*, 2007; Schlicker *et al.*, 2007a) as well as for the prioritization of putative disease genes (Schlicker *et al.*, 2010). A generally accepted cut-off of 0.8 has been established as an indicator for strong functional similarity between two proteins (Schlicker *et al.*, 2007a, 2010). Furthermore, an evaluation of different measures for semantic similarity indicated that the *simRel* measure is best suited for computing functional gene and protein similarities between host factors based on the available GO annotations (Speicher, 2010).

Since the computation of semantic and functional similarity for a large set of proteins could be computationally intensive and time-consuming, the comprehensive web resource FunSimMat was provided (Schlicker and Albrecht, 2008, 2010). It contains pre-computed values for four different semantic similarity measures and over 10 functional similarity measures for all UniProt proteins (The UniProt Consortium, 2014) as well as all Pfam (Finn *et al.*, 2014) and SMART families (Letunic *et al.*, 2015). FunSimMat also allows ranking disease candidate proteins for OMIM diseases using the MedSim method (Schlicker *et al.*, 2010).

3.2.2 Network-based analyses

We designed an analytical framework that uses FSNs for a phenotype-specific prioritization of candidate disease genes and the functional characterization of diseases. In the following subsection, we will present the individual analysis steps that we proposed. Their application to biomedical data is described in the next sections.

Functional phenotype overlap

Understanding the interplay between different diseases, how much they do and do not have in common with respect to their genetic and phenotypic profiles is still a challenge (Lehner, 2013; Parkes *et al.*, 2013). Motivated by the findings that similar diseases share functional annotations (Goh *et al.*, 2007; Van Driel *et al.*, 2006; Oti and Brunner, 2007; Suthram *et al.*, 2010), we proposed to assess the functional phenotype overlap based on the number of functional similarity edges in a FSN of the respective phenotypes.

For this purpose, we constructed an overview network that provides a quantitative visual representation of the phenotype overlap as determined by the number of functional similarity edges between the associated genes. In this network, each node represents a phenotype and is labeled by the number of genes associated with this phenotype, while each edge indicates a strong functional similarity between at least one pair of genes from the connected phenotypes and is labeled by the number of such pairs. The functional similarity edges connecting genes in the same group are represented as self-loops. Based on these counts, we can compute a ratio of functional overlap as a quotient of the number of edges within the phenotype and to the other phenotype. Examples are shown in Figure 3.6 for the two IBD subtypes and in Figure 3.14 for PSC and IBD.

Functional cluster identification

Several studies have indicated that biological systems have a modular organization (Hartwell *et al.*, 1999), which is reflected by the topological structure of biological networks (Hartwell *et al.*, 1999). They contain groups (clusters) of densely connected entities that usually represent functionally or evolutionary related genes, proteins or metabolites (Ravasz *et al.*, 2002; Barabási and Oltvai, 2004; Costanzo *et al.*, 2010; Mitra *et al.*, 2013). For example, Ulitsky and Shamir (2007) developed an approach for the accurate identification of functional modules in similarity networks. Since the edges in a FSN represent strong functional similarity between the connected genes, we expect to find biologically meaningful clusters of genes that are, for instance, enriched for particular GO terms. Thus, we decided to apply the ClusterONE algorithm (Nepusz *et al.*, 2012) to identify overlapping clusters of highly connected nodes.

To detect possibly overlapping protein complexes from weighted protein interaction networks, Nepusz *et al.* (2012) introduced a measure called cohesiveness that accounts for two important structural properties expected for a densely connected subgraph, namely, the nodes within the subgraph are connected by many reliable interactions with each other and at the same time they are well-separated from other nodes in the network. The clustering algorithm uses the cohesiveness to guide a greedy procedure that searches for groups of densely connected nodes and is repeated for several different seed nodes. In the next step of the algorithm, highly overlapping groups are merged and finally, the clusters are filtered based on user-defined thresholds, such as minimal cluster size or density.

We applied this algorithm to find clusters of functionally related disease gene and protein candidates in the IBD FSN (Section 3.3.2) as well as in a Parkinson's disease FSN (Section 3.5). These clusters and their members are likely participating in processes that are disturbed in the respective diseases and should be considered for further experimental validation.

Network-based prioritization

Many network-based prioritization approaches rely on network topology measures, such as degree, shortest path centralities or random walk, to assess the importance of candidate genes or proteins in a protein interaction network (Navlakha and Kingsford, 2010; Wang *et al.*, 2011; Ideker and Sharan, 2008) or in different types of integrative networks (Franke *et al.*, 2006; Linghu *et al.*, 2009; Lee *et al.*, 2011; Hoehndorf *et al.*, 2011). In the special case of PINs, it has been demonstrated several times and, thus, it is generally accepted that global topology measures perform better than local ones (Köhler *et al.*, 2008; Chen *et al.*, 2009; Navlakha and Kingsford, 2010). Thus, Jiang *et al.* (2011) applied a random walk with restart (Köhler *et al.*, 2008) on a FSN to prioritize candidate genes and achieved similar performance as in a protein interaction network. Apart from their work, no comprehensive study has been published that compares the performance of different topology measures for prioritization of candidate genes in other types of networks, such as the integrative networks or FSNs. Another limitation of these widely used network-based approaches is that they rely on known disease genes as seed nodes for the computation of centrality measures in the network. We address some of these issues as follows.

First, we suggested the combination of centrality measures for a consensus score. Thus, in his master's thesis, Kacprowski (2011) explored the performance of different centrality measures and rank aggregation algorithms to prioritize candidate genes in FSNs and PINs. A comprehensive evaluation procedure on the benchmark set used by Schlicker *et al.* (2010) demonstrated good performance of the individual centrality measures (area under the ROC curve between 80 % and 90 %) with an expected increase for global and random-walk based measures compared to local shortest path based, respectively. An aggregation of the centrality-based rankings pushed the performance over 90 % AUC with best AUC for the MaxRank Fuse algorithm. This approach was implemented as a user-friendly Cytoscape plugin NetworkPrioritizer (Kacprowski *et al.*, 2013). It enables the user to choose which data sources to integrate and how to combine them into a network as well as which centrality measures to compute and how to aggregate the resulting rankings into a final score for the candidates.

Second, we developed a method for the prioritization of candidate genes in the absence of known disease genes (Section 3.3.2 and Liu *et al.* (2013)). Thereby, we assumed that the candidate genes originate from a genome-wide association study of a complex disease and thus, are associated with different loci found to be significant for this disease. The method is composed of three steps. First, a FSN (or an integrative FSN and PIN) is constructed for the candidate genes and each gene is assigned to a group based on its association with a particular locus. Second, selected global centrality measures are computed for all nodes in the network and the genes are ranked according to the centrality values. Third, the individual rankings are aggregated into a final ranking using the MaxRank Fuse algorithm, i.e., each gene is assigned the highest rank it achieved. Finally, for each locus, the best-ranked gene is selected (in an iterative fashion). Naturally, this approach can also be applied to networks that contain already known disease genes by integrating this information into the computation of the centrality measures (Section 3.3).

Third, we constructed phenotype-specific networks and subnetworks using different strategies. Since we perform our analyses on a FSN, we expect that functionally more related genes are located in densely connected regions, such as clusters, or at least in the same connected component, i.e., every node can be reached by any other node. Thus, we constructed a FSN for subsets of genes like the top candidates per locus or the top-N ranked genes. The number of nodes and edges in these networks are an indication for the functional coherence of the given set of genes. An example for IBD is shown in Section 3.3.2 and for PSC in Section 3.4.3.

An alternative strategy for extracting a subnetwork from large human protein interaction networks is to find the nodes that, for example, connect two proteins of interest either by a direct connection or through a shortest path. Therefore, we have also created shortest path (SPNs) and common neighbors networks (CNNs) for two sets of proteins. Both networks are defined for two sets of nodes, whereas the first set are the source nodes and the second are the target nodes. Given these two sets, we find all nodes that lie on all shortest paths between each source and each target node and include them in the resulting SPN. For CNN, we identify all nodes that are shared between the source and target nodes, e.g. have a direct edge to a source and a target node at the same time. An example for the application of SPN is presented in Section 3.5.

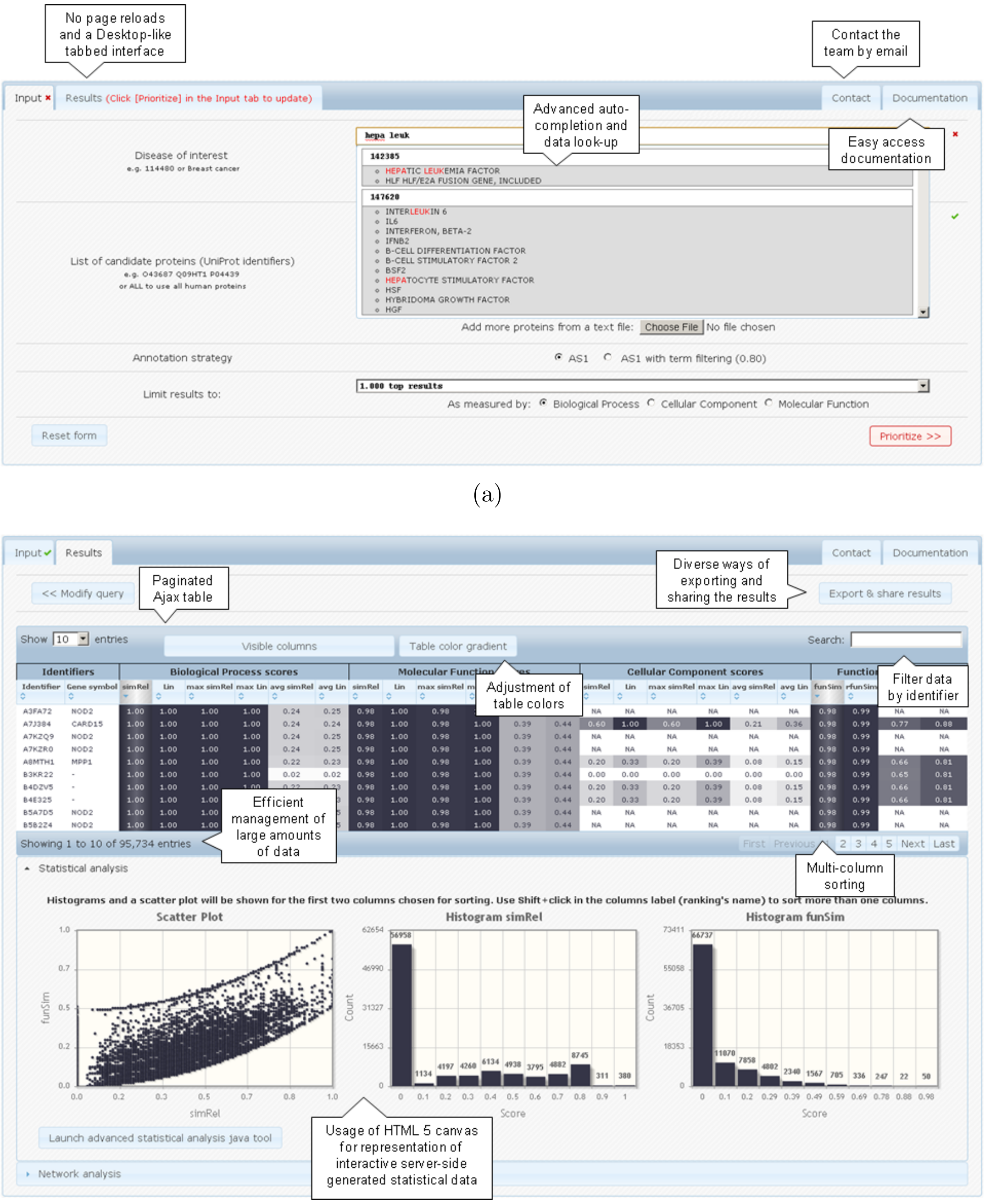
3.2.3 Accompanying software

After reviewing several methods for candidate disease gene prioritization, we realized that there is still room for improvement with regard to their realization as software tools or web services. For example, while it might be convenient to indirectly provide the user with the biological knowledge used by the approach, the user usually does not have access to the data itself and cannot judge its quality and up-to-dateness. Because of the complexity related to data integration, especially when different identifiers are involved, many approaches do not update their underlying biological data sources frequently enough. In addition, it is also difficult to prevent circularity, i.e., ranking a candidate gene on top because it is annotated as a true gene by one of the data sources. Finally, many approaches still provide very static and unhandy web interfaces. In the course of this thesis, we have initiated the development of a few Cytoscape plugins and apps that support the input of user-specific data and provide further visualization and analysis functionality as well as of a user-friendly and interactive web interface for disease gene prioritization using functional similarities.

MedSimX

The method MedSim developed by Schlicker *et al.* (2010) was implemented as part of the FunSimMat web service, which is a resource for pre-computed semantic and functional similarity values. However, this means that the MedSim implementation was not tailored to the specifications and requirements of an advanced web service for candidate disease gene prioritization. Thus, we developed the MedSimX interface with David Buezas.

MedSimX is designed to be interactive and user-friendly as well as to provide more options for a fast and comfortable analysis. Among others, its functionality includes an auto-complete option for disease search by name or OMIM identifier, real-time color spell check-like validation of the candidate proteins, remaining time dialog, and easy access to documentation (Figure 3.4(a)). The results page is easy to bookmark, share and export due to an URL that reflects the state of the page (Figure 3.4(b)). It also provides in-place data analysis functionality such as HTML5 histograms and scatter plots, a paginated color-coded results table with a hierarchical folder-like filter for visible columns (e.g. different similarity scores), and row filtering by



No page reloads and a Desktop-like tabbed interface
Contact the team by email
Advanced auto-completion and data look-up
Easy access documentation
Input ✖
Results (Click [Prioritize] in the Input tab to update)
Contact
Documentation

Disease of interest
e.g. 114480 or Breast cancer

hepa leuk

142385
- HEPATIC LEUKEMIA FACTOR
- HLF HLF/E2A FUSION GENE, INCLUDED

147620
- INTERLEUKIN 6
- IL6
- INTERFERON, BETA-2
- IFNB2
- B-CELL DIFFERENTIATION FACTOR
- B-CELL STIMULATORY FACTOR 2
- BSF2
- HEPATOCYTE STIMULATORY FACTOR
- HSF
- HYBRIDOMA GROWTH FACTOR
- HGF

List of candidate proteins (UniProt identifiers)
e.g. O43687 Q09HT1 P04439
or ALL to use all human proteins

Add more proteins from a text file: Choose File No file chosen

Annotation strategy ● AS1 ○ AS1 with term filtering (0.80)

Limit results to: 1.000 top results

As measured by: ● Biological Process ○ Cellular Component ○ Molecular Function

(a)

Show 10 entries

Search:

Identifiers		Biological Process scores						Molecular Function						Cellular Component scores						Function			
Identifier	Gene symbol	simRel	Lin	max simRel	max Lin	avg simRel	avg Lin	simRel	Lin	max simRel	m			simRel	Lin	max simRel	max Lin	avg simRel	avg Lin	funSim	rfunSim		
A3FA72	NOD2	1.00	1.00	1.00	1.00	0.24	0.25	0.98	1.00	0.98	[illegible]	[illegible]	[illegible]	NA	NA	NA	NA	NA	NA	0.98	0.99	NA	NA
A7J384	CARD15	1.00	1.00	1.00	1.00	0.24	0.24	0.98	1.00	0.98	1.00	0.39	0.44	0.60	1.00	0.60	1.00	0.21	0.36	0.98	0.99	0.77	0.88
A7KZQ9	NOD2	1.00	1.00	1.00	1.00	0.24	0.25	0.98	1.00	0.98	1.00	0.39	0.44	NA	NA	NA	NA	NA	NA	0.98	0.99	NA	NA
A7KZR0	NOD2	1.00	1.00	1.00	1.00	0.24	0.25	0.98	1.00	0.98	1.00	0.39	0.44	NA	NA	NA	NA	NA	NA	0.98	0.99	NA	NA
A8MTH1	MPP1	1.00	1.00	1.00	1.00	0.22	0.23	0.98	1.00	0.98	1.00	0.39	0.44	0.20	0.33	0.20	0.39	0.08	0.15	0.98	0.99	0.66	0.81
B3KR22	-	1.00	1.00	1.00	1.00	0.02	0.02	0.98	1.00	0.98	1.00	0.39	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.99	0.65	0.81
B4DZV5	-	1.00	1.00	1.00	1.00	0.22	0.23	0.98	1.00	0.98	1.00	0.39	0.44	0.20	0.33	0.20	0.39	0.08	0.15	0.98	0.99	0.66	0.81
B4E325	-	1.00	1.00	1.00	[illegible]	[illegible]	[illegible]	0.98	1.00	0.98	1.00	0.39	0.44	0.20	0.33	0.20	0.39	0.08	0.15	0.98	0.99	0.66	0.81
B5A7D5	NOD2	1.00	1.00	1.00	[illegible]	[illegible]	[illegible]	0.98	1.00	0.98	1.00	0.39	0.44	NA	NA	NA	NA	NA	NA	0.98	0.99	NA	NA
B5B2Z4	NOD2	1.00	1.00	1.00	[illegible]	[illegible]	[illegible]	0.98	1.00	0.98	1.00	0.39	0.44	NA	NA	NA	NA	NA	NA	0.98	0.99	NA	NA

Showing 1 to 10 of 95,734 entries

Statistical analysis

Histograms and a scatter plot will be shown for the first two columns chosen for sorting. Use Shift+click in the columns label (ranking's name) to sort more than one columns.

(b)

Figure 3.4: MedSimX web interface: (a) data input page and (b) results page.

protein/gene name. One very important new feature of MedSimX is that it allows users to export the top-ranked candidates to BioMyn (Ramírez *et al.*, 2012), a data warehouse integrating 22 publicly available annotation sources for human genes and proteins. Once having the list loaded in BioMyn, the user can look for enrichment of KEGG pathways, GO terms, etc. The implementation followed good programming practices by separating code in libraries and modules as well separating HTML, PHP, Java script and CSS code. The use of JQuery UI libraries made it possible to create an interface that is appealing to the eye, easy to modify and compatible with all main browsers (Chrome / Firefox / IE 8+ / Opera).

The further development of MedSimX will focus on the integration of FSNs into the results of a prioritization run, such that the user can visually explore the FSN of the top-ranked genes. The network interface will be rather simple and only provide basic functionality using Cytoscape Web (Lopes *et al.*, 2010), but it will also support the export of network data into the stand-alone Cytoscape application, where more complex analysis can be performed.

Cytoscape plugins

As part of his master's thesis, Tim Kacprowski implemented the Cytoscape plugin NetworkPrioritizer for the integrative network-based prioritization of candidate genes and proteins (Kacprowski *et al.*, 2013). Given a network selected by the user and a set of known disease genes, the plugin estimates the importance of the candidate genes for the network connectivity using a number of centrality measures such as shortest path betweenness, shortest path closeness, random walk betweenness, random walk receiver closeness and random walk transmitter closeness (Borgatti, 2005). NetworkPrioritizer works on both unweighted and weighted networks and allows users to adjust the effect of the edge weights on the computed centrality measures (Opsahl *et al.*, 2010).

The plugin also facilitates the user-guided aggregation and comparison of multiple node rankings derived from the different centrality measures. The rank aggregation methods include Weighted Borda fuse (WBF), Weighted AddScore fuse (WASF), and MaxRank Fuse (MRF). While WBF ranks a node based on the sum of its sores computed as the number of nodes ranked lower in the respective rankings (Saari, 1999), WASF assigns the highest rank to the node with the largest weighted sum of scores in its rankings and MRF ranks a node according to the highest rank it achieved in any singal ranking. Furthermore, NetworkPrioritizer provides two common measures of ranking distance. The Spearman footrule distance computes the sum of the difference between the ranks of a node, while the Kendall tau distance is the number of nodes with different ranks (Dwork *et al.*, 2001).

We also implemented a simple plugin for subnetwork generation from an initial network for our project Parkinson's disease (Zanon *et al.*, 2013). SubnetworkGenerator can create a shortest path network (SPN) and a common neighbors network (CNN). The user interface of SubnetworkGenerator is very simple and allows the user to set the source and target nodes sets as well as to choose which network should be created. In addition, the user might also include the direct neighbors of the source or target nodes in the final subnetwork. A year later, Lemetre *et al.* (2013) also released a Java application that supports the generation of subnetworks using shortest paths and other graph-based algorithms.

3.3 Inflammatory bowel diseases

In addition to the development of methods and tools for network-based prioritization, we are also dedicated to the application of newly developed or established methods to open biomedical research questions. Over the last years, we have cooperated with several research groups on the disease-specific application of integrative network-based methods for candidate gene prioritization.

In cooperation with the group of Andre Franke at the Christian-Albrechts-University of Kiel, we have performed extensive data analysis for IBD. Among others, we constructed a phenotype-specific network of all known GWAS candidates for IBD and suggested putative disease genes for follow-up analysis by an integrative network approach. Furthermore, we combined next generation sequencing data from IBD patients and healthy individuals with publicly available interaction data to derive subnetworks of genes associated with the disease.

3.3.1 Medical background

IBD is a chronic relapsing-remitting inflammatory disorder that affects the gastrointestinal tract (Ellinghaus *et al.*, 2015). The two major IBD subtypes are ulcerative colitis (UC) and Crohn's disease (CD), and they affect over 2.5 million people of European ancestry (Molodecky *et al.*, 2012). In North America and Europe, IBD has the highest prevalence ranging from 21 to 246 per 100 000 for UC and 8 to 214 per 100 000 for CD, and the rates are increasing in other populations (Ellinghaus *et al.*, 2015). While ulcerative colitis is characterized by inflammation that spreads in continuous fashion and is limited to the mucosal layer of the colon, Crohn's disease can involve any part of the gastrointestinal tract and is often associated with complications such as strictures, abscesses and fistulas (Khor *et al.*, 2011). Since there is no known cure yet, the therapy of IBD symptoms involves a combination of immune-suppressing medications and dietary changes, in more than half of the patients surgery (Liu and Anderson, 2014; Conrad *et al.*, 2014). The differences in the localization, endoscopic appearance, histology and behavior of CD and UC suggest differences in the underlying pathophysiology despite some shared clinical and pathological features (Ellinghaus *et al.*, 2015).

The etiology of IBD is so far understood as a complex interplay between individual genetic predisposition and environmental factors, such as the gut microbiome (Jostins *et al.*, 2012; Ellinghaus *et al.*, 2015). Family history is a risk factor for developing IBD with a 26-fold and 9-fold increased risk when another sibling already has CD or UC, respectively (Bengtson *et al.*, 2009). Meta-analyses of genome-wide association studies (GWAS) have established a total of 163 IBD susceptibility loci (Barrett *et al.*, 2008; Franke *et al.*, 2010; Anderson *et al.*, 2011; Jostins *et al.*, 2012), which explain only 13.6 % of the disease variance for CD and 7.5 % for UC (Jostins *et al.*, 2012). CD and UC share 110 of these loci, which might explain the clinical similarities between both diseases and, at the same time, it points to a complex, heterogeneous and difficult to diagnose disease spectrum (Ellinghaus *et al.*, 2015).

Although the majority of the genetic contribution to disease risk is still unknown, the identified loci and the candidate genes associated with them have deepened our understanding of the pathogenesis of IBD. Most loci appear to exert their effect by influencing the regulation of gene transcription rather than disrupting the coding sequence (Ellinghaus *et al.*, 2013b). Enrichment analysis has revealed relevant disease pathways, such as autophagy, ER stress response, innate immunity, and epithelial barrier dysfunction among others (Khor *et al.*, 2011). Recent evidence points to an essential role for host defense against infection in IBD, in particular, the interaction between the host mucosal immune system and microbes, both at the epithelial cell surface and within the gut lumen (Jostins *et al.*, 2012). However, the biological mechanisms that underlie IBD risk loci are yet to be identified by fine-mapping studies, direct experimental work, and functional interrogation (Liu and Anderson, 2014; Ellinghaus *et al.*, 2015). In particular, next-generation sequencing is expected to reveal rare and low-frequency variants that contribute to disease risk.

3.3.2 Phenotype-specific network analysis

Although the two IBD subtypes, CD and UC, have similar phenotypes, their genotypes are distinct from each other (Ellinghaus *et al.*, 2015). Nevertheless, genome-wide association studies have identified many significant loci associated with either CD, UC, or both. We have combined data from two recent GWAS on CD and UC performed by our cooperation partners from Kiel. Overall, 625 genes were identified for the 71 CD risk loci (Franke *et al.*, 2010) and 377 for the 47 UC risk loci (Anderson *et al.*, 2011), resulting in 276 overlapping genes and 726 genes in total.

We constructed an integrative IBD candidate network (Figure 3.5) by combining protein interactions from the iRefIndex database (Razick *et al.*, 2008) and strong functional similarity links based on the Gene Ontology from the FunSimMat database (Schlicker and Albrecht, 2008). In the resulting network, 480 out of the 726 genes were connected by 2875 edges, including 183 protein interactions and 2692 functional similarity edges. Since we were interested in the functional relationships between genes in different loci, we distinguished between edges connecting genes located in different loci or within the same locus. From the 2875 edges, only 7.3 % are within the same loci and were not considered in the following network analysis.

First, we applied the graph clustering algorithm provided by the Cytoscape plugin ClusterONE (Nepusz *et al.*, 2012) on the IBD network to identify groups of functionally related genes. From the 9 significant clusters ($P \leq 0.05$), the top 5 are highlighted in Figure 3.5. The genes corresponding to each cluster were uploaded to BioMyn (Ramírez *et al.*, 2012) for functional annotation, and the top-enriched biological process terms (adjusted p-value < 0.05) are listed:

- Cluster 1 (56 genes, P-value = 0.00): cell differentiation, reproduction
- Cluster 2 (37 genes, P-value = 5.77E-9): signal transduction, immune system process, response to stress

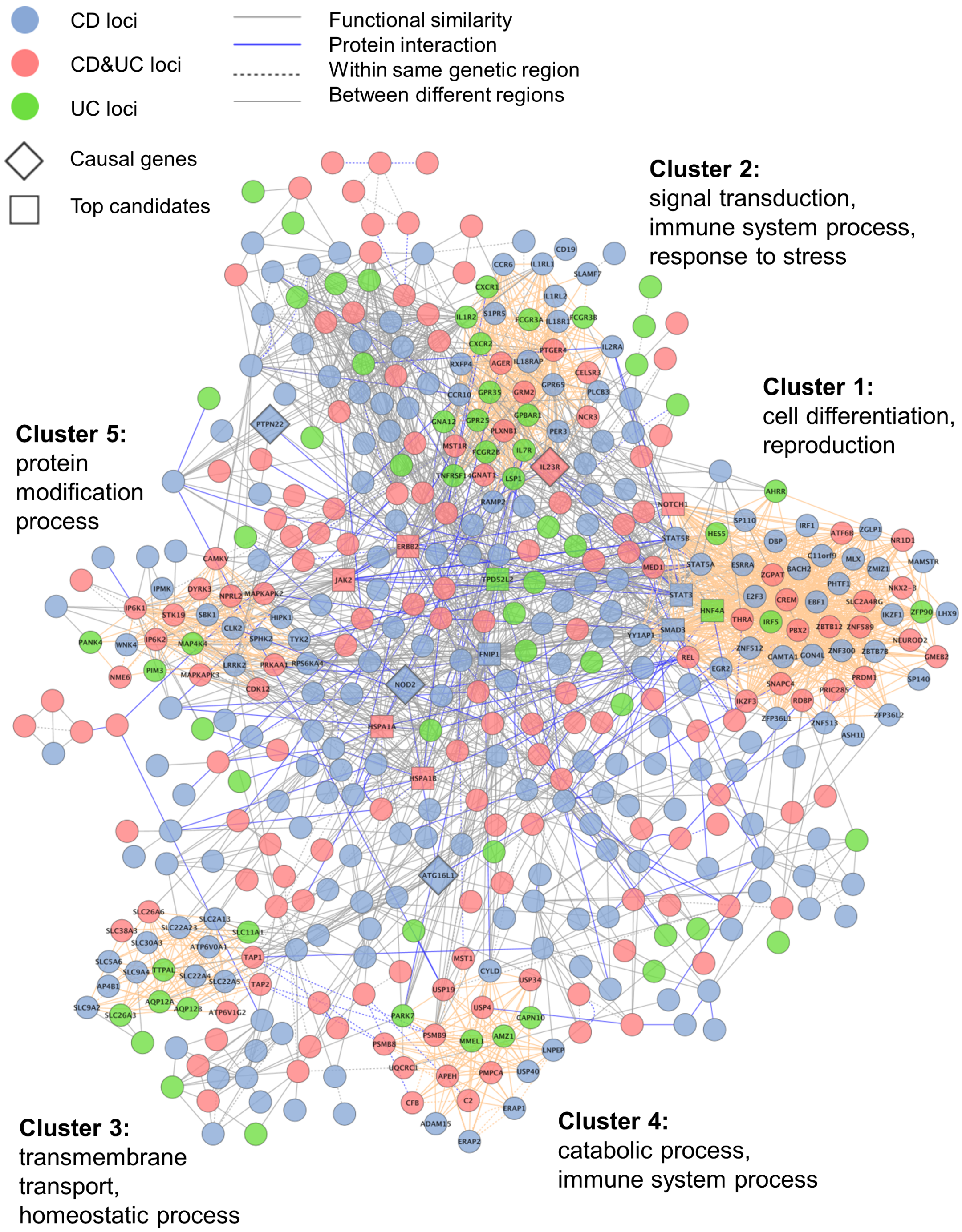
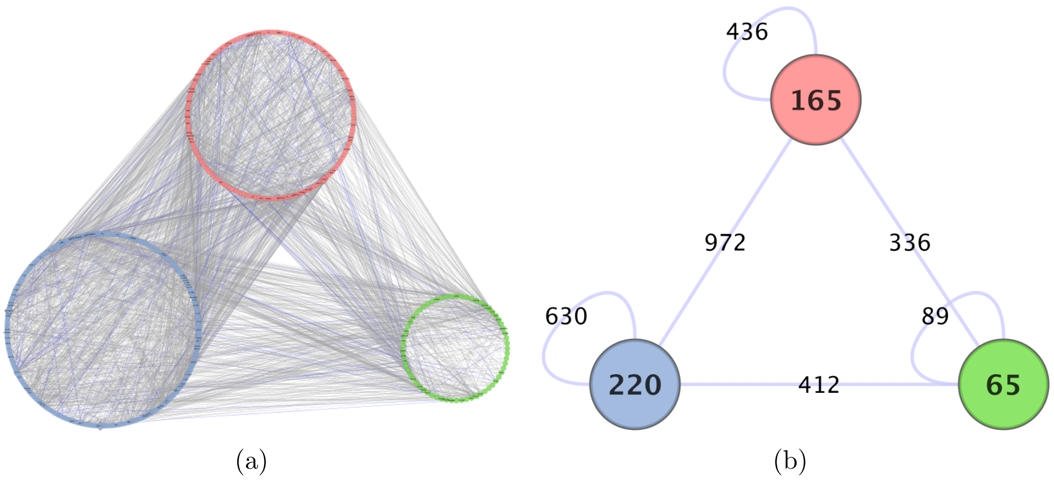


Figure 3.5: Network representation for genes associated with IBD. In this network, blue nodes correspond to genes in CD loci, green nodes to genes in UC loci, and pink nodes to genes in both CD and UC loci. Gray and blue edge lines represent strong functional similarity based on Gene Ontology annotations and direct protein interactions, respectively.

- Cluster 3 (20 genes, P-value = 1.32E-7): transmembrane transport, homeostatic process

- Cluster 4 (21 genes, P-value = 2.73E-7): catabolic process, immune system process
- Cluster 5 (23 genes, P-value = 1.66E-4): protein modification process

Overall, we did not identify a subtype-specific cluster, e.g. one that only contains CD or UC candidates. Cluster 2 stands out as it is enriched with UC associated genes compared to the other clusters and is the only cluster that contains a known IBD gene (IL23R). This observation considered together with the functional annotations renders the cluster an interesting target for further investigation of UC. In addition, cluster 1 contains 4 out of the 10 top-ranked IBD candidates, including two genes associated with CD, one with UC, and one with both subtypes (see details on the ranking below).



(a) (b)

Figure 3.6: Functional phenotype overlap between Crohn's disease (CD) and ulcerative colitis (UC). The network in (a) is the same network shown in Figure 3.5 grouped by phenotype, e.g., CD genes are represented by blue nodes, UC genes by green nodes, and CD & UC genes by pink nodes, while the gray and blue edges correspond to strong functional similarity and protein interactions, respectively. Figure (b) shows an overview of the same network, where each node represents one of the phenotypes and is labeled with the number of associated genes. An edge indicates that the phenotypes are connected and is labeled by the numbed of connections between the corresponding genes. The edges connecting genes in the same group are represented as self-loops.

Since we could not identify any UC or CD specific clusters, we analyzed the phenotypic overlap between the two IBD subtypes by creating an overview network (Figure 3.6). Here, the nodes represent the phenotypes (CD only, UC only, CD & UC) and are labeled by the number of genes associated with the respective phenotype. The edge labels correspond to the number of functional similarity; protein interaction edges connecting the genes and edges within the same phenotype are represented as self-loops. This simplified view of the network of IBD-associated genes confirmed our observation that the CD-associated genes are more functionally similar within their phenotype than are the UC candidates (edge-to-node ratio

Table 3.1: Top-ranked IBD candidate genes using MaxRank Fuse aggregation.

Symbol	Band	Status	GDC	GRWB	GRWRC	GRWTC	GSPB	GSPC	BF rank
ERBB2	17q12	IBD	49	0.312	7.25E-04	1.27E-03	0.261	0.667	1
MTMR3	22q12	CD	5	0.094	2.07E-03	1.96E-04	0.002	0.364	164
FNIP1	5q31	CD	80	0.475	6.55E-04	1.70E-03	0.173	0.571	4
STAT3	17q21	CD	53	0.188	5.70E-04	2.30E-03	0.050	0.571	7
TPD52L2	20q13	UC	51	0.336	7.60E-04	1.13E-03	0.072	0.571	2
NDUFAF3	3p21	IBD	20	0.369	7.64E-04	7.77E-04	0.065	0.571	17
GPBAR1	2q35	UC	37	0.075	5.49E-04	2.01E-03	0.005	0.444	43
SMAD3	15q22	CD	75	0.241	8.01E-04	1.11E-03	0.032	0.444	5
PPM1G	2p23	CD	2	0.010	1.38E-03	1.12E-04	0.000	0.308	315
RAMP2	17q21	CD	13	0.164	7.78E-04	6.41E-04	0.037	0.571	35

of 2.9 and 1.4, respectively). The CD & UC group is located in-between with a ratio of 2.6 and is more functionally similar to UC (ratio of 5.2) than to CD (ratio of 4.4). Furthermore, most of the UC-associated genes are connected to CD genes (edge-to-node ratio of 6.3), while this is not the case for the CD group (ratio of 1.9). Although there might be a bias toward CD because there are more loci identified for it, we expect the edge-to-node ratios to be less affected by this and thus to reveal relevant subtype relationships that indicate some functional overlap between CD and UC.

Furthermore, we ranked the candidate genes based on their importance for the connectivity of known IBD genes included in this network. There are four such genes (NOD2, IL23R, ATG16L1, PTPN22) and we refer to them as seed nodes. We computed seven centrality measures to rank the candidate genes with respect to the seed genes using the RINalyzer plugin for Cytoscape (Doncheva *et al.*, 2012a) and merged the rankings using the MaxRank Fuse (MRF) and Weighted Borda Fuse (WBF) rank aggregation algorithms using the NetworkPrioritizer plugin (Kacprowski *et al.*, 2013). The 10 top-ranked genes and their respective values are listed in Table 3.1 and 3.2. The 10 top-ranked genes according to WBF are also represented as rectangles in the network shown in Figure 3.5.

Since the WBF algorithm computes the weighted mean rank of a node in all primary rankings without considering its actual scores, a node is ranked high in the aggregated ranking if it is ranked high in many of the primary rankings. The MRF assigns each node the highest rank, which it has achieved in any primary ranking. In this case, a node has a high rank in the aggregated ranking if it is important according to one centrality measure and not to all of them. As can be expected, these two aggregation strategies deliver different results (see Tables 3.1 and 3.2). However, there is still an overlap of 5 genes on the top 10 positions. While the remaining top genes according to WBF are also ranked rather high by MRF (up to rank 23), this is not the case for the remaining MRF top genes, e.g., two of them have WBF ranks of 164 and 315.

Although most of the top candidates are spread throughout the network (Figure 3.5), they are also physically interacting or functionally similar to each other.

Table 3.2: Top-ranked IBD candidate genes using Weighted Borda Fuse rank aggregation.

Symbol	Band	Status	GDC	GRWB	GRWRC	GRWTC	GSPB	GSPC	MRF rank
ERBB2	17q12	IBD	49	0.312	7.25E-04	1.27E-03	0.261	0.667	1
TPD52L2	20q13	UC	51	0.336	7.60E-04	1.13E-03	0.072	0.571	5
JAK2	9p24	IBD	45	0.217	7.61E-04	1.07E-03	0.051	0.571	15
FNIP1	5q31	CD	80	0.475	6.55E-04	1.70E-03	0.173	0.571	3
SMAD3	15q22	CD	75	0.241	8.01E-04	1.11E-03	0.032	0.444	8
HNF4A	20q13	UC	49	0.099	7.30E-04	1.15E-03	0.024	0.500	23
STAT3	17q21	CD	53	0.188	5.70E-04	2.30E-03	0.050	0.571	4
NOTCH1	9q34	IBD	52	0.135	7.84E-04	1.07E-03	0.018	0.444	20
HSPA1A	6p21	IBD	31	0.281	7.04E-04	1.08E-03	0.032	0.571	18
HSPA1B	6p21	IBD	31	0.281	6.97E-04	1.09E-03	0.030	0.571	13

This can be observed in the two subnetworks shown in Figure 3.7(a) and 3.7(b). They include the 4 known IBD genes and the 10 top-ranked candidates using MRF and WBF aggregation, respectively. We also checked for enriched functional annotations using DAVID (Huang *et al.*, 2009b). For the WBF top genes, we obtained the following representative BP terms (Benjamin corrected p-value ≤ 0.05): response to wounding, regulation of cell proliferation, regulation of apoptosis. In contrast, while five of the top 10 MRF genes were annotated with signal transduction, neither this BP term nor other GO terms were significantly enriched for the top 10 MRF genes.

We performed a second round of prioritization by selecting the top-ranked gene for each locus. From 65 loci represented in the IBD network, we constructed a subnetwork of 51 genes connected by 218 functional similarity or protein interaction edges (Figure 3.7(c)). In particular, this network contains the 5 overlapping top-ranked genes as well as 4 additional genes ranked in the top 10 positions by MRF. GO enrichment analysis using DAVID (Huang *et al.*, 2009b) resulted in several enriched BP terms, including regulation of transcription, metabolic process and cell proliferation as well as T cell activation and immune response. In addition, we observed that 20 of the locus candidates are physically or functionally connected to the four already known IBD genes.

3.3.3 Exome sequencing in Crohn's disease

Increasingly, GWAS studies have been criticized because their findings account only for a low proportion of overall heritability across common complex diseases so far (Manolio *et al.*, 2009). Possible explanations include overestimation of disease heritability, epigenetic effects, and a major contribution of low-frequency and rare variants. After the successes of whole-exome sequencing to detect rare coding variants in Mendelian disorders (Bamshad *et al.*, 2011), its application for identifying the contribution of rare and low-frequency coding variants in complex phenotypes is a logical step. Therefore, our collaboration partner Andre Franke and his colleagues

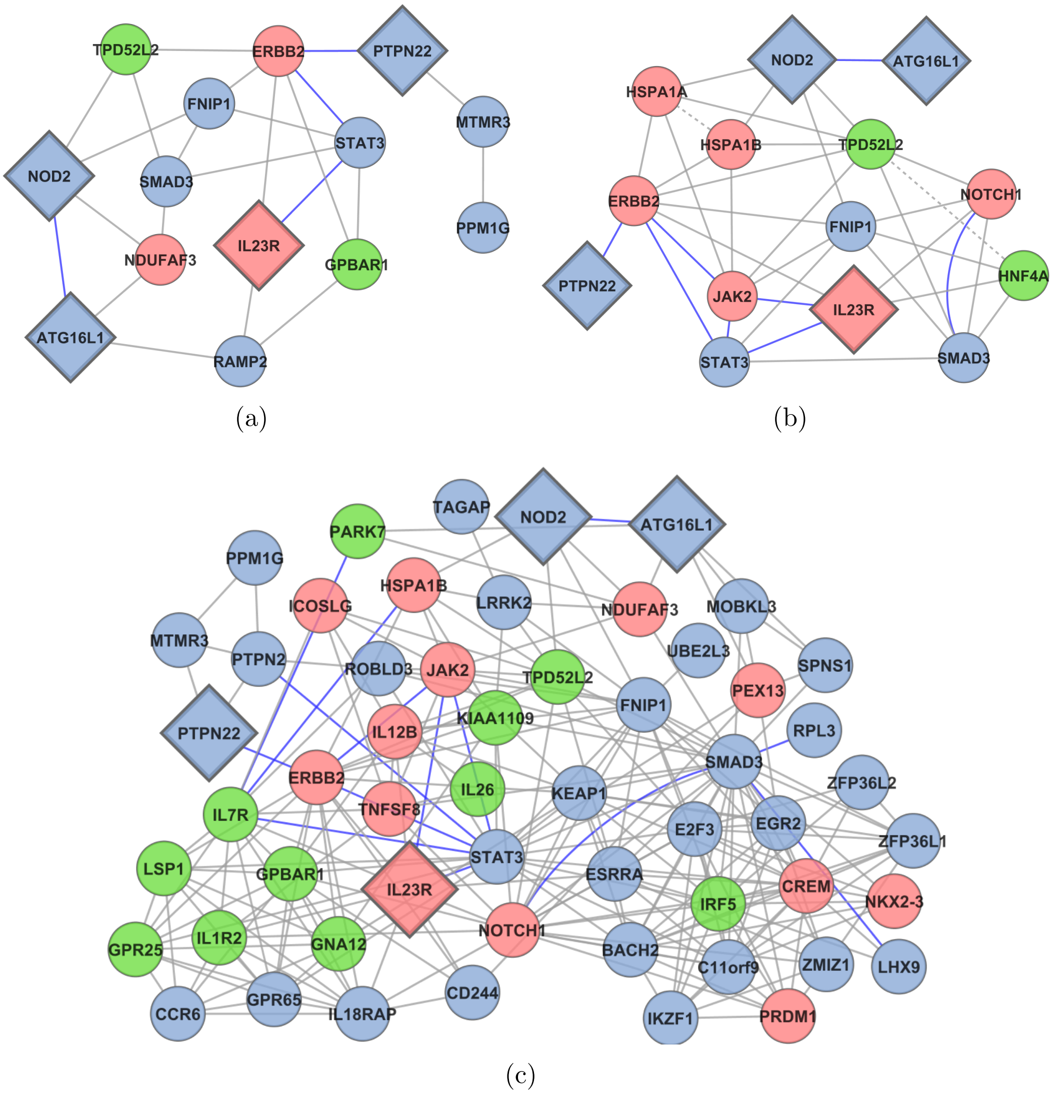


Figure 3.7: Network of top-ranked IBD candidates: (a) using MaxRank Fuse aggregation; (b) using Weighted Borda Fuse aggregation; (c) selecting top-ranked gene per locus. In all networks, blue nodes correspond to genes in CD loci, green nodes to genes in UC loci, and pink nodes to genes in both CD and UC loci. Gray and blue edge lines represent strong functional similarity based on Gene Ontology annotations and direct protein interactions, respectively.

initiated a study to identify novel CD variants by combining functional studies with exome sequence data from patents (Ellinghaus *et al.*, 2013b).

After sequencing the whole exomes of 42 unrelated subjects with CD and 5 healthy subjects (controls), 117 957 SNVs were identified, including 59 076 coding and splice site SNVs. In order to identify functionally relevant variants, the large set of SNVs was annotated and filtered using a two-fold strategy. First, based on the results of tools such as SNAP (Bromberg and Rost, 2007), SIFT (Sim *et al.*, 2012), and Polyphen2 (Adzhubei *et al.*, 2010), non-synonymous and splice site SNVs with at least one in silico prediction of a protein-altering effect or a known disease mutation annotation in the HGMD database (Stenson *et al.*, 2008) were considered. Then, the *Targeted* approach filtered SNVs from the established 71 CD associated regions (Franke *et al.*, 2010), while in the more hypothesis-free *Whole-exome* approach, the whole exome data set was scanned for IBD candidate susceptibility SNVs (association signal of $P < 10^{-4}$ within a region of ±250 kb around the variant from the large GWAS meta-analysis on CD (Franke *et al.*, 2010)). After a conventional Sanger re-sequencing to correct for sequence errors, 93 and 159 SNVs were verified for approaches one and two, respectively.

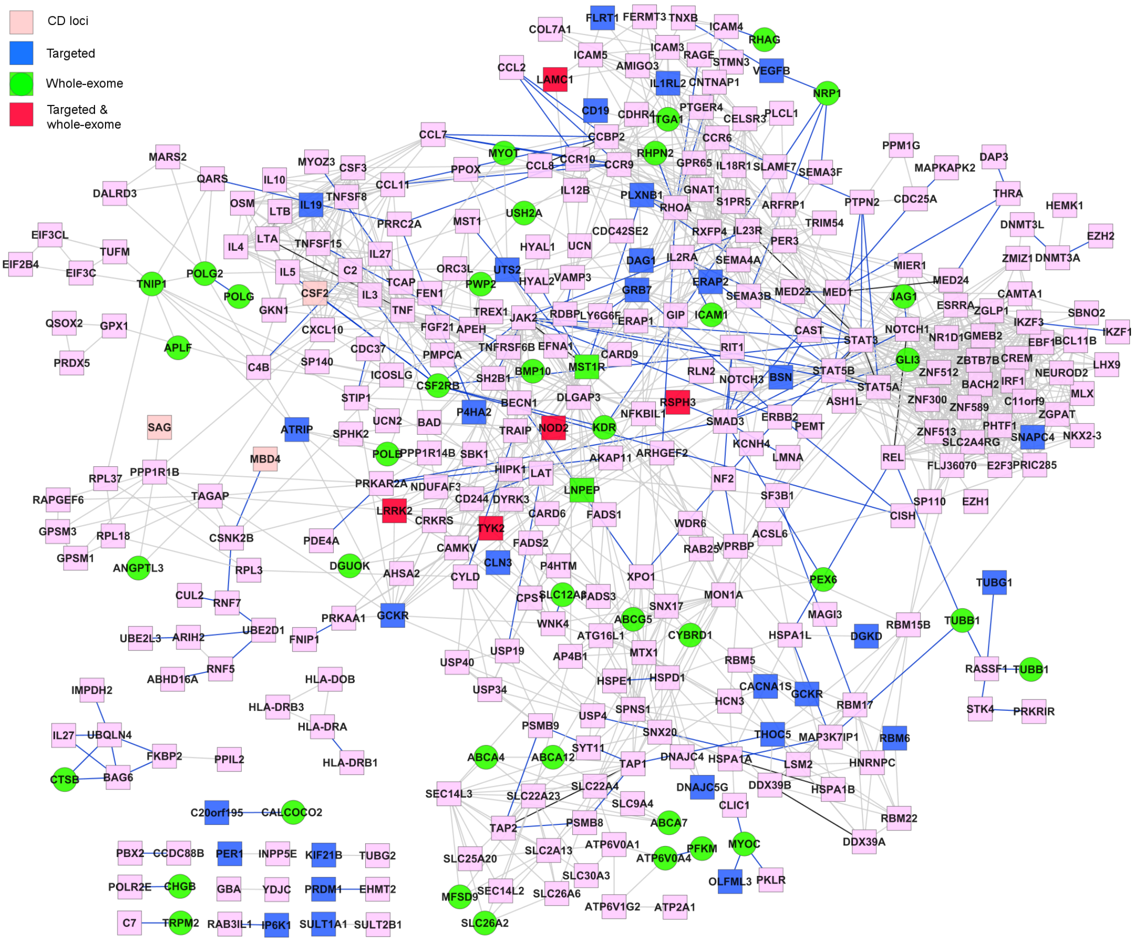


Figure 3.8: Network representation of genes associated with CD or implicated by exome sequencing. In this network, blue and green nodes represent genes with SNVs selected by the *Targeted* or *Whole-exome* strategy, respectively, while red nodes refer to genes with SNVs selected by both strategies. Pink nodes or nodes with a squared shape indicate genes located in the previously associated CD loci. Blue network edges represent direct protein-protein interactions of the gene products, while gray edges indicate strong functional similarity based on the Gene Ontology annotations. If both a protein-protein interaction and a strong functional similarity are evident for a pair of genes, the corresponding edge is colored in black. Genes and their nodes that are not connected to any other node in the network are omitted.

In order to investigate the functional relationships between genes associated with CD (663 genes) and the selected SNVs, we extracted all genes located close to these SNVs. This resulted in 77 genes for the *Targeted* approach and 151 for the *Whole-exome* approach with an overlap of 8 genes (RSPH3, LAMC1, LAMB2, LMNB2, NOD2, LRRK2, FBXW12, and TYK2). As expected from the design of the studies, all 77 *Targeted* genes coincided with genes previously associated with CD. Interestingly, beside the 8 genes with SNVs from both strategies, four additional genes (LNPEP, MST1R, CCDC71, and LRRK2) with whole-exome SNVs were located in the CD loci. In particular, NOD2 and LRRK2 are known IBD-associated autophagy genes (Franke *et al.*, 2010; Anderson *et al.*, 2011).

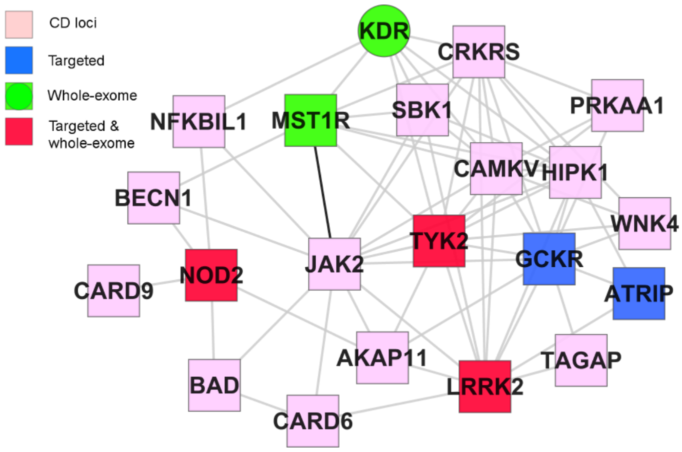


Figure 3.9: Network of selected genes implicated by exome sequencing. In this network, blue and green nodes represent genes with SNVs selected by the *Targeted* or *Whole-exome* strategy, respectively, while red nodes refer to genes with SNVs selected by both strategies. Pink nodes or nodes with a squared shape indicate genes located in the previously associated CD loci. Blue network edges represent direct protein-protein interactions of the gene products, gray edges correspond to strong functional similarity based on the Gene Ontology annotations, and black edges indicate that both interaction types are evident.

Then, we constructed an integrative network of all genes with SNVs selected by the *Targeted* or *Whole-exome* strategy as well as genes located in the 71 CD associated loci (Franke *et al.*, 2010) by combining known protein interactions and strong functional similarities based on Gene Ontology annotations (Figure 3.8). Overall, 358 genes are connected by 1293 edges, including 30 representatives from the *Targeted* and 40 from the *Whole-exome* approach as well as the 5 genes selected by both strategies (RSPH3, LAMC1, NOD2, LRRK2, and TYK2). Of the network edges, 134 represent physical protein interactions, 1138 indicate strong functional similarity, and 13 gene pairs are connected by both types. Visual inspection of the network revealed that the genes with SNVs do not cluster together and they are spread throughout the network of CD associated genes. In particular, GRB7 is located very central in the network (high degree, closeness and betweenness centrality values) and connects several genes associated with CD or closely located to an SNV. Furthermore, three of the genes implicated by both strategies and their direct neighbors form a small subnetwork (Figure 3.9) enriched in regulation of programmed cell death, phosphorylation, and intracellular signaling according to DAVID (Huang *et al.*, 2009b).

The 93 *Targeted* and 159 *Whole-exome* SNVs were further genotyped in an independent case-control cohort and 147 SNVs were found to be polymorphic in the German population. From the 40 SNVs selected for follow-up genotyping, 6 SNVs were found to be associated with CD at the 0.05 level in a small German panel. They were subjected to further genotyping in 8 independent case-control sets comprising 9348 subjects with CD, 2868 subjects with ulcerative colitis (UC), and 14 567 healthy control subjects of European ancestry. Using this approach, two missense SNPs, chr6:106659789 (Ser354Asn) and chr6:106660076 (Leu450Phe) at PRDM1 on chromosome 6q21, were significantly associated with CD and UC, respectively. The missense SNP rs2303015 (Val248Ala) in the novel CD candidate gene NDP52 (also known as CALCOCO2) was associated with CD.

PRDM1 encodes PR domain containing 1, a zinc finger-containing transcriptional repressor that regulates terminal B- and T-cell differentiation (Crotty *et al.*, 2010). Functional studies performed by Ellinghaus and colleagues demonstrated that the Ser354Asn mutation led to increased peripheral blood lymphocyte (PBL) expression of the adhesion molecule L-selectin, which is critical for PBL migration to the sites of intestinal inflammation, and increased CD4+ and CD8+ T-cell proliferation, IFN-g secretion, and up-regulation of activation markers on stimulation (Ellinghaus *et al.*, 2013b). Furthermore, eQTL analysis revealed that an adjacent CD risk allele (Franke *et al.*, 2010) correlated with reduced expression of PRDM1 in ileal biopsy specimens and peripheral blood mononuclear cells. These are all factors that may contribute to its pathogenic role in CD.

NDP52 was initially described as a 52 kilodaltons subunit of nuclear domain 10 bodies (Korioth *et al.*, 1995). Nowadays, NDP52 is rather known as CALCOCO2 (calcium binding and coiled-coil domain 2), a cytosolic protein with a crucial role in immunity and as an adaptor for selective autophagy (Morriswood *et al.*, 2007; Thurston *et al.*, 2009; Ivanov and Roy, 2009; von Muhlinen *et al.*, 2010). A combination of functional characterization and structural analysis of the identified missense variant Val248Ala indicated that it affects the inhibitory role of NDP52 for nuclear factor κB activation of genes involved in inflammation as well as the stability of proteins in Toll-like receptor pathways (Ellinghaus *et al.*, 2013b; Till *et al.*, 2013).

The autophagy adaptor gene NDP52 can be considered as a new member to the group of IBD risk factors contributing to autophagy regulation (ATG16L1, BAD, BECN1, CUL2, IRGM, KEAP1, LRRK2, NOD2, PARK7, PRKAA1, PTPN2, SMURF1, ULK1, VAMP3). To investigate the interactions between NDP52 (labeled as CALCOCO2) and other IBD genes, we created a phenotype-specific subnetwork by extracting all direct protein interactions for the IBD risk genes from the iRefIndex database (release 9.0) (Razick *et al.*, 2008) and manually including regulatory interactions for FOXO3 and IRGM (Jegga *et al.*, 2011). We reduced the obtained dataset from 2328 to 381 interactions to emphasize the role of the autophagy pathway for pathogenesis of CD. The resulting interaction network shown in Figure 3.10 contains the 14 IBD-associated autophagy genes (Franke *et al.*, 2010; Anderson *et al.*, 2011), their direct interactors involved in autophagy (Xie and Klionsky, 2007; Yang and Klionsky, 2010; Wong *et al.*, 2011), the interactors shared

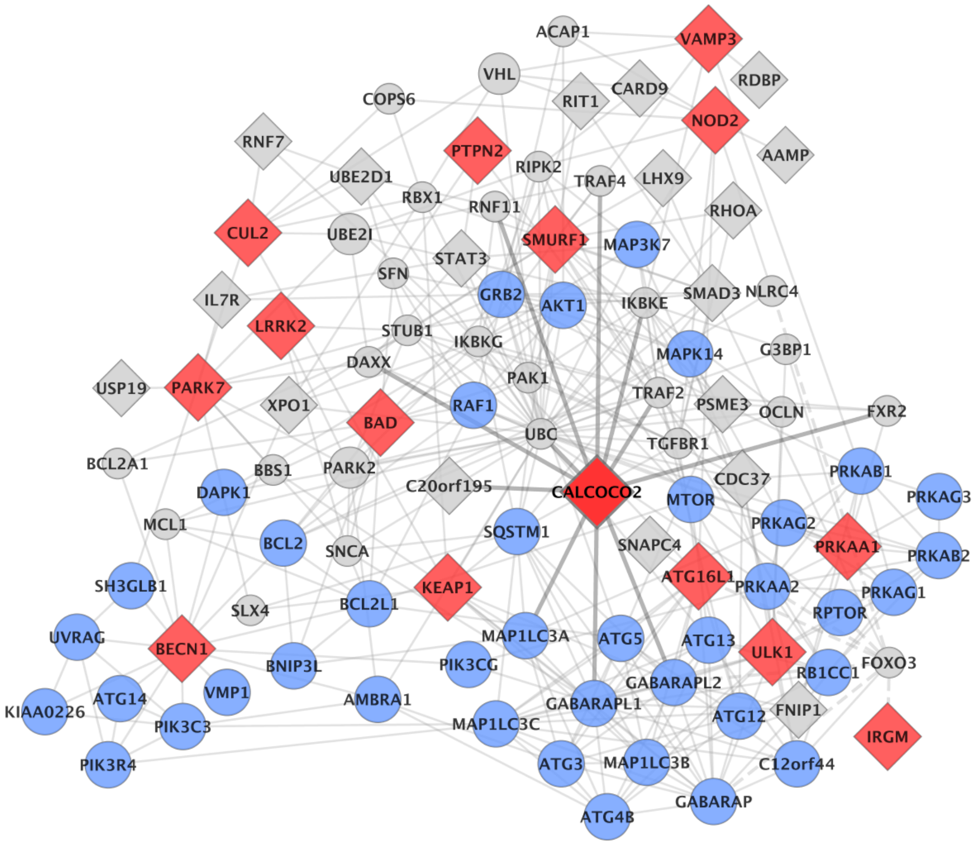


Figure 3.10: Interaction network of CD-associated autophagy loci. In this network, direct protein interactions are shown as solid gray edges, while manually included regulatory interactions are represented by dashed gray edges. The network contains the 14 IBD-associated autophagy genes (red diamonds), their direct interactors involved in autophagy (blue circles), the interactors shared between at least two IBD risk factors (small gray circles), and the direct interactors associated with IBD but not contributing to autophagy (gray diamonds). The interactions of the newly identified gene CALCOCO2 (also known as NDP52) are highlighted by thicker edge lines. Figure first published in Ellinghaus *et al.* (2013b).

between at least two IBD risk factors, and the direct interactors associated with IBD but not contributing to autophagy (Franke *et al.*, 2010; Anderson *et al.*, 2011). The newly identified gene CALCOCO2 interacts directly with three IBD interactors involved in autophagy and is connected to the remaining IBD autophagy genes through eight interactions with other nodes in the network.

3.3.4 Summary and discussion

Currently, 163 susceptibility loci have been associated with IBD through GWA studies and they explain 13.6 % and 7.5 % of the disease variance for CD and UC, respectively (Jostins *et al.*, 2012). These findings have been instrumental for advancing our knowledge of IBD pathogenesis and relevant disease pathways. However, many questions still remain unanswered (Ellinghaus *et al.*, 2015). For instance, can the overlapping loci between CD and UC explain their clinical similarities and which associations are really subtype-specific or shared? Which of these candidate genes and risk variants are really relevant for the disease? How do genetic and environmental factors combine together to lead to the development of IBD?

In order to address these questions, we performed a network-based phenotype-specific analysis of known IBD susceptibility loci. We constructed a network of strong functional similarities and known physical protein interactions for the gene products of genes associated with IBD loci. Based on topological analysis of this network, we concluded that there are no subtype specific groups (clusters) of genes, although the CD genes are more functionally similar among each other than the UC genes. We also identified meaningful groups of functionally related genes associated with both CD and UC that point to relevant disease pathways.

Furthermore, we developed and applied two different strategies for network-based prioritization of the IBD candidate genes. Although the resulting predictions do not overlap completely, in all cases, we were able to construct a subnetwork of functionally similar or physically interacting candidate genes originating from different loci. Such disease-specific subnetworks might be the first step towards understanding the interdependence between the genetic factors of complex diseases such as IBD. In addition, we identified five candidate genes (ERBB2, TPD52L2, FNIP1, SMAD3, STAT3) as common for our prioritization strategies.

Finally, we showed that exome sequencing data can be combined with network data to suggest candidate genes for IBD that would not be identified otherwise. We were also able to put the newly identified CALCOCO2 gene in a functional context with known IBD genes and autophagy-related genes. In this way, we supported the findings of our collaboration partners, who performed a whole-exome sequencing followed by rare variant analysis and detailed expression and functional studies to extend the genetic insights beyond those derived from GWAS alone.

With the foreseeable improvements of network data quality and quantity as well as disease association mapping we will be able to perform a more complete and insightful network analysis. Still, while the network of candidate genes advances our understanding, functional studies will be needed to prove the disease relevance of particular candidate genes or risk variants and how it affects the molecular mechanisms of the involved pathways. The next challenge in IBD research will be the profiling of large cohorts of patients in terms of longitudinal data retrieval, whole-genome sequencing, gene expression data generation and the study of epigenetic factors (Ellinghaus *et al.*, 2015).

3.4 Primary sclerosing cholangitis

The field of disease gene prioritization has made substantial progress over the last years. However, many approaches still neglect the fact that phenotypes can strongly differ in their genetic characteristics as well as in the amount of research dedicated to them. Therefore, the investigation of less well studied phenotypes or groups of phenotypes with unknown causative genes needs to be approached in an alternative way and stands in the focus of current methodological developments. Thus, we have developed and applied a method for the generation of a disease-specific network of primary sclerosing cholangitis (PSC) to provide more insight into the functional characteristics of the underlying phenotype and to identify candidate disease genes from multiple loci without previous knowledge of known disease genes.

3.4.1 Medical background

PSC is a progressive chronic inflammatory condition caused by inflammation and subsequent obstruction of the intrahepatic and extrahepatic bile ducts that, in most cases, progresses to cirrhosis of the liver and end-stage liver disease (Hirschfield *et al.*, 2013). Clinically, it is characterized by abdominal pain, chills, diarrhea, fatigue, fever, itchiness, weight loss, yellowing of eyes and skin. Most PSC patients are diagnosed relatively young, at a median age of 30-40 years, and approximately 2/3 of them are male (Karlsen *et al.*, 2010b). Although PSC has a relatively low prevalence (1 in 10,000), it remains a leading indicator for liver transplantation in northern Europe and the United States due to the lack of effective medical therapy (Karlsen *et al.*, 2010b; Williamson and Chapman, 2015).

A characteristic feature of PSC is the presence of comorbidity with autoimmune diseases. Most PSC patients suffer from an increased frequency of IBD (60-80 %), mostly ulcerative colitis (Karlsen *et al.*, 2010b; Karlsen and Kaser, 2011). In contrast, only around 5 % of patients with IBD develop PSC (Karlsen *et al.*, 2010b; Karlsen and Kaser, 2011). A variety of other autoimmune diseases were reported at an increased frequency (25 %) in PSC (Saarinen *et al.*, 2000). The strong HLA associations and the clinical occurrence of PSC with immune-mediated diseases suggest that autoimmunity has a role in pathogenesis (Karlsen *et al.*, 2010b).

The main challenges in PSC research remain in learning more about the etiology and pathogenesis of the disease. Several theories have been proposed to explain the development of PSC (Karlsen and Boberg, 2013; Hirschfield *et al.*, 2013; Folseraas *et al.*, 2014). For instance, there are ongoing discussions whether the bile duct injury in PSC might be caused by immune mediated mechanisms or biochemical aspects related to bile physiology or potentially a combination of these two aspects (Karlsen and Boberg, 2013). Sibling relative risk of 9- to 39-fold indicates a strong genetic component to PSC risk (Bergquist *et al.*, 2008). In addition to multiple strong associations within the HLA complex, association studies before 2012 have identified genome-wide significant loci at 1p36 (MMEL1-TNFRSF14), 2q13 (BCL2L11), 2q37 (GPR35), 3p21 (MST1), 10p15 (IL2RA) and 18q21 (TCF4) (Karlsen *et al.*, 2010a; Melum *et al.*, 2011; Srivastava *et al.*, 2012; Folseraas *et al.*, 2012; Ellinghaus *et al.*, 2013a). In the following, we will present a recent study by our collaboration partners (Liu *et al.*, 2013) that revealed nine novel loci associated with PSC using the Immunochip (Cortes and Brown, 2011) and the accompanying network analysis that we performed to prioritize the candidate genes.

3.4.2 Experimental candidate genes identification

Recently, the Wellcome Trust Case-Control Consortium designed the Immunochip, a custom-made Illumina Infinium genotyping chip that includes approx. 200 000 SNPs relevant to major immune-mediated diseases (Cortes and Brown, 2011). It has a dense marker coverage across 186 known disease loci from 12 autoimmune and seronegative diseases (rheumatoid arthritis, ankylosing spondylitis, systemic lupus erythematosus, type 1 diabetes, autoimmune thyroid disease, celiac disease, multiple sclerosis, ulcerative colitis, Crohn's disease, and psoriasis) as well as thousands of SNPs of intermediate significance from multiple meta-analyses of immune-mediated diseases. The chip represents a cost-effective way of fine-mapping known disease loci and more thoroughly surveying those that have been associated with related autoimmune diseases. Some of the SNPs were included for deep replication of statistically weaker signals from GWAS studies and large meta-analyses. Since immune-mediated diseases are genetically related, the Immunochip can also provide many insights into the shared genetic susceptibility between them. For instance, Parkes *et al.* (2013) analyzed Immunochip data for several major immune-mediated diseases and revealed that, although multiple loci are shared, the most associated variants from the same locus or those with the largest effect sizes usually differ.

To further characterize the genetic etiology of PSC, our collaborators, Tom H. Karlsen (Oslo University Hospital, Norway) and Carl A. Andersson (Wellcome Trust Sanger Institute, UK), designed a study (Liu *et al.*, 2013), in which 3 789 PSC cases of European ancestry were compared to 25,079 population controls across 130,422 SNPs genotyped using the Immunochip. 12 genome-wide significant ($P < 5 \times 10^{-8}$) associations were identified, 9 of which were novel. Three out of the four known non-HLA PSC risk loci present on the Immunochip were also confirmed. For the further functional network analysis and prioritization, we considered all genes within 0.1cM of the 12 non-HLA genome-wide significant PSC loci.

Besides the loci that reach a stringent significance threshold of 5×10^{-8}, it is likely that there are additional true associations among the SNPs with weaker associations. Therefore, Liu *et al.* (2013) applied an alternative approach to exploit the known pleiotropy between seven related immune-mediated traits (Crohn's disease, celiac disease, psoriasis, rheumatoid arthritis, sarcoidosis; type 1 diabetes and ulcerative colitis) (Zhernakova *et al.*, 2009). Additional 33 non-HLA loci were discovered and all of them showed suggestive levels of significance ($5 \times 10^{-8} < P < 5 \times 10^{-5}$) in the standard association analysis (Figure 3.11). We also integrated the genes within 0.1cM of these loci into the functional network analysis to highlight PSC susceptibility genes.

3.4.3 Phenotype-specific network analysis

In order to identify putative disease genes, we performed several network-based analyses with the genes in the PSC associated loci. We created functional similarity networks for the genes within 0.1cM of the 12 non-HLA genome-wide significant

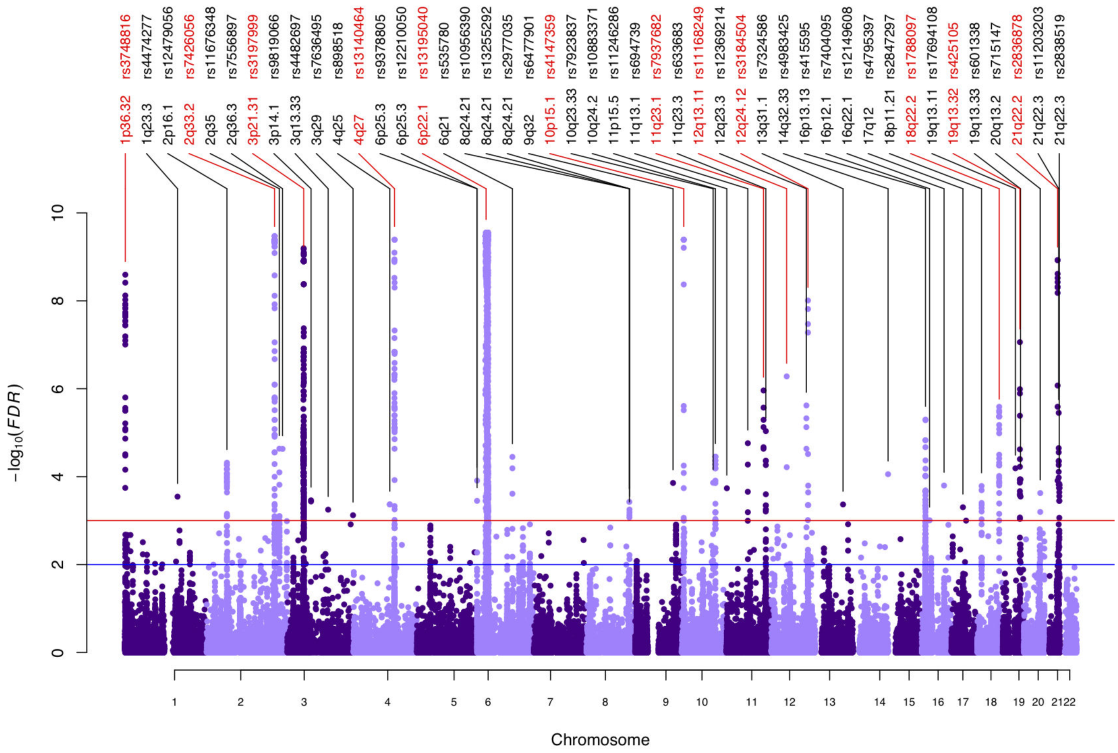
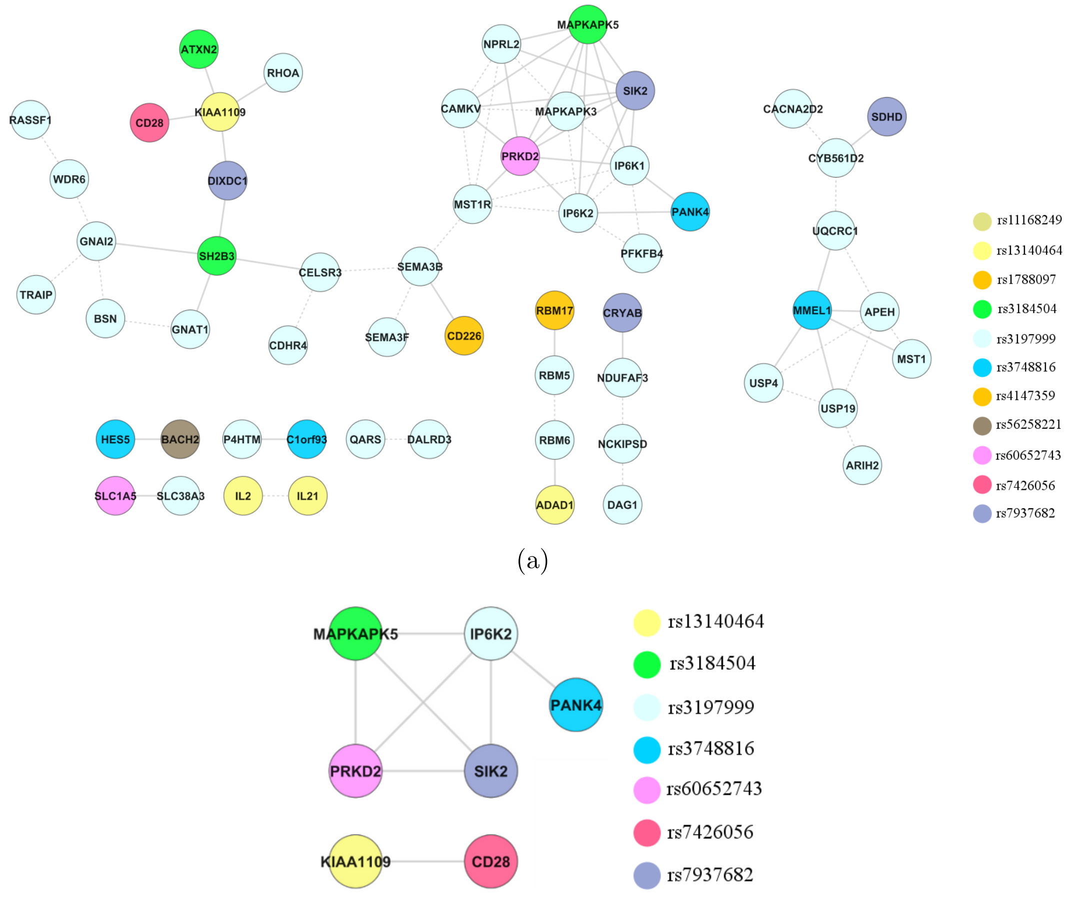


Figure 3.11: Manhattan plot of conditional associations in PSC calculated on the basis of the results of the present PSC analysis and genetic associations previously reported in seven immune-mediated diseases. SNPs in red represent genome-wide significant findings from the main association analysis, and SNPs in black are significantly associated with PSC conditional on their pleiotropic effects across the related immune-mediated diseases. The horizontal red and blue line represent a significance threshold of 0.001 and 0.01, respectively. Figure first published in Liu *et al.* (2013).

PSC loci and the 33 pleiotropic loci. In these networks, each edge represents strong functional similarity of two genes based on annotated Gene Ontology (GO) terms (Ashburner *et al.*, 2000) as determined by the functional similarity measure rfunSim (Schlicker *et al.*, 2006). rfunSim similarity values above the recommended cutoff of 0.8 were retrieved using the FunSimMat web service (Schlicker and Albrecht, 2008, 2010). We did not construct a protein interaction network since we could not retrieve enough information on physical protein interactions for the PSC gene products from the iRefIndex database (Razick *et al.*, 2008).

Figure 3.12(a) shows the functional similarity network of the genes in the genome-wide significant PSC loci, where the nodes are colored by their SNP association. The network contains 56 out of the 148 genes (representative for 10 out of the 12 loci) and these are connected by 76 functional similarity edges. Due to linkage disequilibrium the number of genes in one locus can vary from one to several dozens and this difference is emphasized in the network visualization. For example, there are 83 genes in the genomic region of SNP rs3197999 and at most 2 genes for 5 of the other loci. However, we expect only a few (at most one or two) genes from each locus to be actually causative. Thus, we are interested in associations between



(a)

(b)

Figure 3.12: Figure 3.12(a) shows the functional similarity network of the genes within 0.1cM of the 12 non-HLA genome-wide significant PSC loci and the corresponding nodes are colored by SNP association. The network contains 56 out of the 148 genes in PSC loci and these are connected by 76 edges (34 within the same locus and 42 between different loci). The network in Figure 3.12(b) consists of seven PSC candidate genes as suggested by topological analysis on the network of high functional similarities between individual genes in the associated loci (Figure 3.12(a)), one gene in each of the seven out of 12 non-HLA loci. In both networks, gray edges indicate strong functional similarity based on Gene Ontology annotations and connect genes either from different loci (solid edge lines) or within the same locus (dashed edge lines). Genes and their nodes that are not connected to any other node in the network are omitted. Figure first published in Liu *et al.* (2013).

genes from different loci, and we distinguish between edges within the same locus (34 edges) and between different loci (42 edges). Although there is a representative pair of genes connected by functional similarity edges for each pair of loci, we can observe that many genes within the same locus are also functionally similar.

To prioritize the PSC candidate genes in the absence of known disease genes such that only one gene per locus remains, we assessed the importance of each gene for the network structure and the connection between different loci. First, we computed three different topology measures to assess the connectivity of the candidate genes: degree (number of direct edges to other nodes), shortest path closeness (inverted average shortest path distance to other nodes), and shortest path betweenness (fraction of shortest paths passing through the node). Similarity edges between genes in the same locus and gene nodes that were not contained in the resulting largest connected subnetworks were ignored. The genes were ranked according to each measure and were then assigned the best of the three ranks. Third, we iteratively constructed a PSC-specific network from the top ranked genes in their respective loci such that each gene is connected to a gene in another locus. This resulted in a network that contains 7 genes that represent 7 out of the 10 loci considered in the analysis (Figure 3.12(b)). We performed the analysis and visualization in Cytoscape (Shannon *et al.*, 2003) with the help of the NetworkAnalyzer plugin for the computation of the topology measures (Doncheva *et al.*, 2012a).

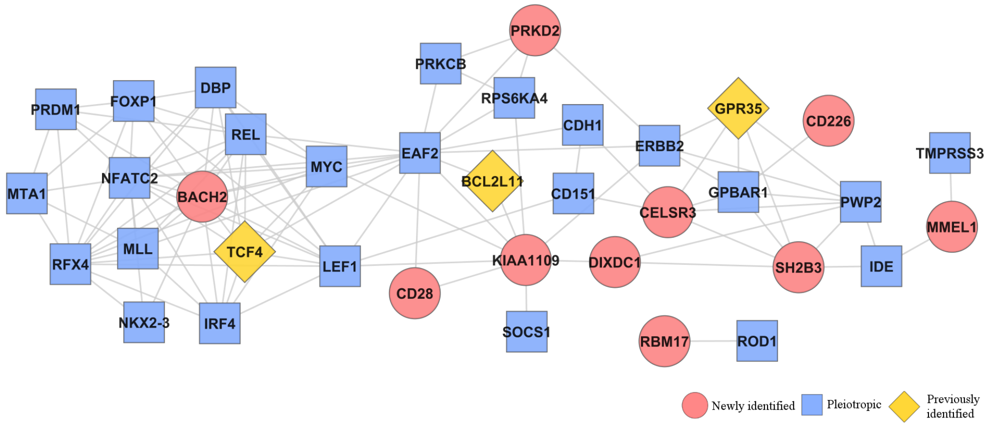


Figure 3.13: Functional similarity network of candidate genes as suggested by topological analysis on the network of functional similarities between individual genes in all loci associated with PSC (Figure B.1), one gene in each of the loci. Grey edge lines indicate strong functional similarity based on Gene Ontology annotations. Genes and their nodes that are not connected to any other node in the network are omitted. Figure first published in Liu *et al.* (2013).

Furthermore, we constructed a larger PSC candidate network (*FSN-all*) for the genes within 0.1cM of the 12 non-HLA identified PSC susceptibility regions, the 33 pleiotropic PSC loci, the 3 previously identified loci and a representative of the HLA locus (Figure B.1). It contains 177 out of the 341 considered genes and these are connected by 511 edges (51 within the same locus and 460 between different loci). We applied the same prioritization approach on this network and constructed the PSC-specific network shown in Figure 3.13. This network contains 35 genes connected by 109 strong functional similarity edges and they represent 35 out of the 42 considered loci.

Besides our prioritization approach based on FSNs, our collaboration partners applied several other methods on the significant PSC association data to prioritize candidate genes within these loci. The functional consequences of the most associated SNPs or the SNPs in high LD with these were evaluated by identifying

Table 3.3: Candidate genes among the genome-wide significant loci implicated by eQTL, missense mutation, GRAIL, DAPPLE, or FSN. *eQTL* and *Missense* are SNPs in high LD with the most significantly associated SNP in the locus that are either known eQTLs or missense mutations. Genes are implicated by GRAIL or DAPPLE if they show nominal significant number of connections. FSN genes are suggested by topological analysis on the functional similarity network of the genes in the 12 identified loci only (*FSN*) or in all PSC associated loci, including the pleiotropic and previously known (*FSN-all*). Adapted from (Liu *et al.*, 2013).

Locus	SNP	Candidate genes						consensus
		eQTL	Missense	GRAIL	DAPPLE	FSN	FSN-all	gene
1p36	rs3748816	MMEL1	MMEL1			PANK4	MMEL1	MMEL1
2q33	rs7426056			CD28		CD28	CD28	CD28
3p21	rs3197999	USP4	BSN MST1	GPX1 MST1		IP6K2	CELSR3	MST1
4q27	rs13140464			IL2		KIAA1109	KIAA1109	(KIAA1109)
6q15	rs56258221			BACH2			BACH2	BACH2
10p15	rs4147359			IL2RA				
11q23	rs7937682			CRYAB HSPB2	SIK2	SIK2	DIXDC1	SIK2
12q13	rs11168249			VDR		–	–	
12q24	rs3184504		SH2B3	SH2B3 TRAFD1	C12orf51	MAPKAPK5	SH2B3	SH2B3
18q22	rs1788097		CD226	CD226			CD226	CD226
19q13	rs60652743					PRKD2	PRKD2	(PRKD2)
21q22	rs2836883			ETS2		–	–	

missense SNPs using PolyPhen (Adzhubei *et al.*, 2010) and SIFT (Sim *et al.*, 2012) and by retrieving data on expression quantitative trait loci (eQTLs) from the University of Chicago eQTL browser (Gilad and Pritchard, 2010). Five SNPs that are in high linkage disequilibrium ($r^2 > 0.8$) with PSC associated loci were predicted as *benign*, *damaging* or *not tolerated* with respect to their effect on the respective protein structure and function. An eQTL represents a SNP where different genotypes are associated with variation in the expression levels of nearby genes. 10 eQTLs that are in high linkage disequilibrium with the most strongly associated SNP at two out of the 12 significant PSC loci were retrieved.

Furthermore, a GRAIL pathway analysis was performed in order to assess the functional relationship among the PSC risk regions. The GRAIL software is a statistical tool that uses text mining of published abstracts in the PubMed database to identify and quantify functional similarity among genes within disease-associated regions (Raychaudhuri *et al.*, 2009). In total, 13 genes in high linkage disequilibrium with PSC associated SNPs received a significant score ($P_{text} < 0.05$). Finally, the DAPPLE tool (Rossin *et al.*, 2011) was used for the construction of a network of known protein-protein interactions between the products of the genes in the 12 genome-wide significant PSC loci, the 33 pleiotropic loci, and the 3 previously known loci. Then, the gene connectivity was assessed based on the number of direct and indirect (via other proteins) connections between them. The 2 genes with a permuted p-value above 0.05 were listed as causal candidates.

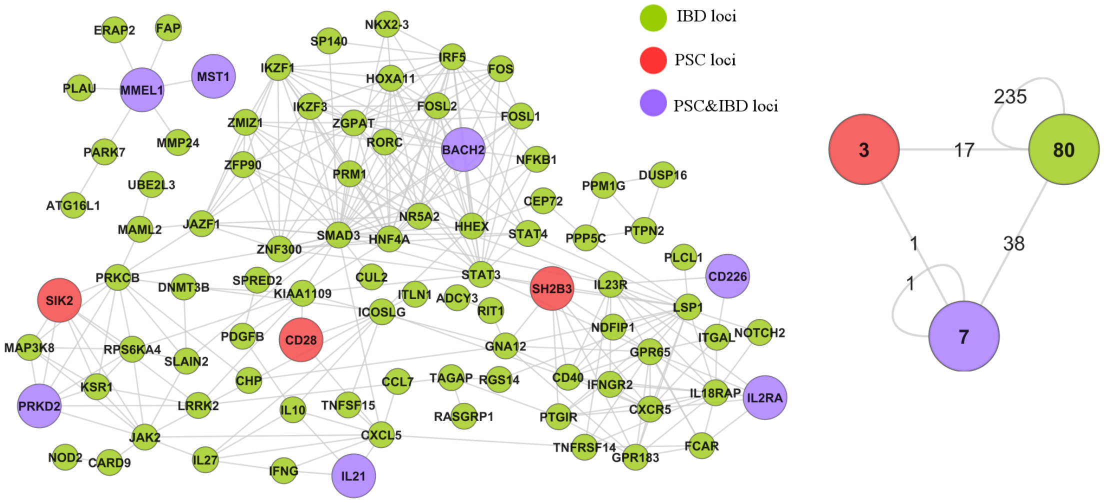
Table 3.3 summarizes the results of the aforementioned prioritization approaches. For 7 of the 12 genome-wide significant loci, the same gene was annotated by more than 1 method (here both FSN-based methods are considered as one), suggesting these genes (MMEL1, CD28, MST1, BACH2, SH2B3, CD226, and SIK2) as good candidates for further investigation at the given loci. For two additional loci, our two FSN-based approaches (with and without the pleiotropic loci) suggested the same genes (KIAA1109 and PRKD2) and they agreed on another one (CD28). Apart from this, the *FSN-all* approach selected 5 out of the 7 consensus genes, while the other FSN approach suggested only 2 of them, which might be interpreted as an indication that more biological knowledge improves the prioritization procedure.

3.4.4 Overlap with inflammatory bowel diseases

Although 72 % of the PSC patients in the Immunochip study have a diagnosis of concomitant IBD, only half of the genome-wide significant loci were associated with IBD in the recent International IBD Genetics Consortium (IIBDGC) Immunochip analysis (Jostins *et al.*, 2012), despite the greater sample size of that study (25,683 cases and 15.977 controls). Further analysis revealed that PSC is genetically more similar to ulcerative colitis than to Crohn's disease. This is consistent with clinical observations of greater comorbidity of PSC with ulcerative colitis than with Crohn's disease (Broomé and Bergquist, 2006).

To further compare the genetic profiles of PSC and IBD, we combined the genome-wide significant PSC-associated loci with the 163 confirmed IBD-associated loci (Jostins *et al.*, 2012) in a functional similarity network. Figure 3.14(a) shows the network of the protein-coding genes closest to the most associated SNP (compact PSC&IBD-FSN), while the FSN in Figure B.2 contains all genes within 0.1cM of the associated loci (PSC&IBD-FSN). The PSC&IBD-FSN contains 34 % of all PSC and 42 % of all IBD genes. In both cases, no PSC specific clusters could be determined within the FSNs, indicating that genes within PSC associated loci are distributed across the entire IBD network rather than localized to a particular cluster.

Furthermore, we investigated how the network structure changes when adding PSC genes to the IBD network and vice versa. We identified 41 PSC nodes that are new in the PSC&IBD-FSN compared to the PSC only FSN. Apparently, the number of connected genes increases because several PSC genes are connected through intermediate IBD nodes and vice versa. In contrast, only 3 IBD nodes are new in the PSC&IBD-FSN compared to the IBD only FSN. In addition, we determined that all PSC loci genes are functionally similar to (60 % of the) IBD loci genes in the PSC&IBD-FSN (Figure B.3(a)). Although rather small, the number of edges is considerable for the compact PSC&IBD-FSN (Figure 3.14(b)) and quite impressive for the network that additionally includes the genes within 0.1cM of the pleiotropic PSC loci (Figure B.3(b)). The increase of the number of connected genes after combining PSC and IBD loci in one network as well as the very high number of PSC nodes connected to IBD nodes indicate that there is high functional overlap of PSC with IBD. However, this statement is based only on the set of annotated



(a) (b)

Figure 3.14: Functional similarity network of PSC and IBD associated loci. The protein-coding genes closest to the most associated SNP in the 12 non-HLA genome-wide significant PSC loci and the 163 confirmed IBD loci were used to construct a functional similarity network (Figure 3.14(a)). The network contains 90 gene nodes that are connected by 292 similarity edges. Genes associated with only PSC are represented by large red nodes, with only IBD by small green nodes, and with both PSC and IBD by large violet nodes. Grey edge lines indicate strong functional similarity between the connected genes based on their Gene Ontology annotations. Genes and their nodes that are not connected to any other node in the network are omitted from the Figure. Figure 3.14(b) shows an overview of the network from Figure 3.14(a). The genes are grouped based on their association with PSC (red), IBD (green), and both PSC and IBD (violet) and each node is labeled with the number of associated genes. An edge indicates that the corresponding genes are connected by similarity edges and is labeled by the numbed of such connections. The functional similarity edges connecting genes in the same group are represented as self-loops. Figure first published in Liu *et al.* (2013).

genes and further analyses have to be performed to investigate if the remaining genes are not part of the PSC&IBD networks because of actual differences between the phenotypes or due to missing annotations.

3.4.5 Summary and discussion

Although PSC is the leading indicator for liver transplantation in Northern Europe and the United States (Karlsen *et al.*, 2010a), its etiology and pathogenesis are still not well understood. So far, 16 risk genes were identified by GWAS and they account for 7.3 % of the overall PSC susceptibility (Folseraas *et al.*, 2014). However, it is firmly established that genetics have an important role in the development of PSC. Another important feature of PSC is its comorbidity with autoimmune diseases, in particular IBD (60-80 %).

More than half of the known 16 PSC risk genes were identified in a GWA study with the custom-made Immunochip performed by our collaboration partners (Liu *et al.*, 2013). In addition to the 9 novel and 3 confirmed PSC loci, they also identified 33 pleiotropic loci based on data for seven related diseases. We constructed a functional similarity network of all genes located within 0.1cM of the 12 non-HLA genome-wide significant PSC loci as well as one additionally including the genes in the pleiotropic loci and the three previously identified PSC loci. Applying a network-based algorithm specifically designed for this scenario, we suggested 7 and 35 candidate genes for further experimental validation, respectively. Six of these genes were annotated as causative by at least one additional prioritization approach.

In order to characterize the genetic and phenotypic overlap of PSC and IBD, we constructed a functional similarity network of genes in close genetic proximity to the identified PSC and IBD risk loci. We observed that the PSC loci are distributed throughout the IBD loci, suggesting that there is no particular functional cluster of IBD susceptibility genes associated with PSC and vice versa. However, we estimated a considerable functional overlap between the two phenotypes based on the high number of similarity edges between the respective genes. This approach could also be applied to investigate the functional similarity of PSC with other related diseases, such as the 12 autoimmune diseases on the Immunochip, as well as their relationships among each other.

The described Immunochip study of PSC (Liu *et al.*, 2013) is the first of its kind for this disease and its findings make it possible to perform the first analyses on purely PSC associated data as compared to previously, when only similar diseases were used. These are the first promising steps in the study of the mechanisms underlying PSC and its comorbidity with other autoimmune diseases. Future efforts are expected to identify further PSC risk factors as well as the exact genes and pathways involved in the disease development. As with other complex diseases, it is very likely that several environmental factors interplay with specific genetic predisposition.

3.5 Parkinson's disease

3.5.1 Medical background

Parkinson's disease (PD) is the most common neurodegenerative disorder after Alzheimer's disease and has a prevalence of approximately 12 % in persons over 60 years of age (Nussbaum and Ellis, 2003). Clinically, it is characterized by motor abnormalities (tremor, rigidity, slowness, balance problems), autonomic disturbances, psychiatric sequelae (usually depression), and cognitive impairment (Hoehn and Yahr, 1967; Greenamyre and Hastings, 2004). The neuropathological characteristics include loss of neurons in the substantia nigra and the presence of α-synuclein positive inclusions in the cytoplasm of neurons, referred to as Lewy bodies or Lewy neurites depending on their structure (Forno, 1996; Spillantini *et al.*, 1998).

So far the underlying pathogenetic mechanisms of PD are still largely unknown. The first important insights were provided from the study of early-onset parkinsonism (< 10 % of all cases) and the discovery of several monogenic mutations that could cause mitochondrial impairment, oxidative stress, and protein mishandling (Greenamyre and Hastings, 2004). These common mechanisms are believed to play a central role in the pathogenesis of PD and may be induced by non-genetic factors in sporadic PD cases (De Lau and Breteler, 2006). The most common mutations known so far affect Parkin (Hedrich *et al.*, 2004), a ubiquitin E3 ligase that is responsible for the transfer of activated ubiquitin molecules to a protein substrate (Shimura *et al.*, 2000) and thus directly affects several cellular processes such as protein degradation, regulation of receptor trafficking, cell cycle progression, gene transcription, DNA repair, and immune responses (Ikeda and Dikic, 2008). The identification the cytoplasmic and mitochondrial interaction partners of Parkin could provide further biological insights into its complex role for the PD pathogenesis and elucidate novel therapeutic targets.

Therefore, Francisco S. Domingues and his colleagues at the EURAC research institute performed several Tandem Affinity Purification(TAP)/mass spectrometry (MS) interaction screens and identified 203 unique candidate Parkin-binding proteins (Zanon *et al.*, 2013). Approximately 50 % of them were detected in the mitochondrial fractions and 50 % in the cytosolic fractions of two different cell lines, with an overlap of 49 proteins between the fractions. Then the set of Parkin-binding candidates was further computationally analyzed for involvement in PD.

In addition, a set of proteins known to be related to genetic parkinsonism was assembled. It includes 9 proteins encoded by genes implicated in monogenic forms of parkinsonism (MonogenicPD) (Marras *et al.*, 2012) and 77 proteins known to interact with Parkin as retrieved from the iRefIndex database (Razick *et al.*, 2008).

3.5.2 Phenotype-specific network analysis

We contributed to this project with the design of a network-based prioritization pipeline involving topological analysis of protein-protein interaction networks (PINs) and functional similarity networks (FSNs). We implemented the accompanying software and performed the initial network analysis. The final biological analysis of the resulting high-ranked candidates was performed by Hagen Blankenburg; figures presented in this section were created by our collaboration partners (Zanon *et al.*, 2013).

To build a human PIN, we used the iRefIndex database (version 9.0), which combines protein interaction data from multiple primary resources (Razick *et al.*, 2008). Both binary and complex interactions were considered, whereas protein complexes were expanded using the matrix expansion model so that pairwise interactions are assumed between all interactors within a complex. The interaction data were filtered to exclude predicted interactions such as interactions for which the detection method contained 'predicted', 'interologs mapping' or 'confirmational text mining'. The resulting network contained 12,013 proteins connected by 350,917 interactions.

Out of the 203 ParkinTAP candidates, 194 are in the same large connected component of the human PIN. Furthermore, six of the nine MonogenicPD proteins are direct Parkin interactors, two more (UCHL1 and FBXO7) interact with Parkin interactors, and only one (ATP13A2) is not connected to any of them.

In order to create a PD-specific PIN, we constructed a shortest path network by selecting all proteins and interactions from the human PIN that are on the shortest paths between the ParkinTAP candidates and the PD-related proteins. We also included all direct interactors of the MonogenicPD proteins. The resulting network contains 4,009 proteins and 290,496 interactions, most of which are expanded complexes (268,484).

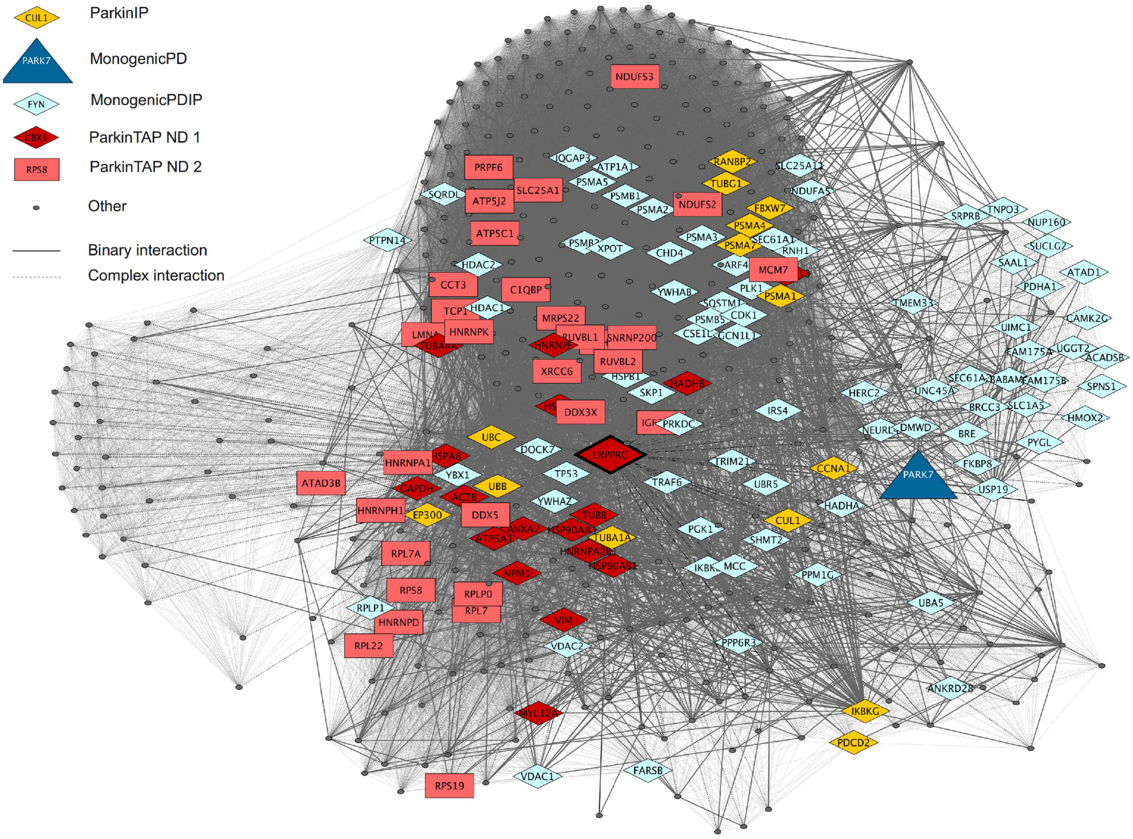


Figure 3.15: Direct protein interactions of the ParkinTAP candidate LRPPRC. Proteins are represented as nodes, binary interactions as solid lines and complex interactions as dashed lines. Binary interactions to the selected candidates are represented by thicker edges. ParkinTAP ND 1 or 2 are ParkinTAP candidates at network distance 1 or 2 of MonogenicPD, i.e., ParkinTAP ND 1 are direct MonogenicPD interactors. There are many interactors (461 nodes) of LRPPRC in iRefIndex, resulting in a dense network of complex interactions (34,732 edges). LRPPRC interacts with MonogenicPD PARK7, as well as with 48 other ParkinTAP candidates, and the network includes 14 Parkin and 77 MonogenicPD protein interactors. Figure first published in Zanon *et al.* (2013).

In this network, we computed the shortest path network distance to Parkin and the minimum network distance to MonogenicPD proteins for all ParkinTAP candidates as well as the number of MonogenicPD proteins at a given network distance.

The results for these two topological measures were later included in the candidate selection procedure designed by our collaborators. Out of the 203 ParkinTAP candidates, three have a network distance of 1, e.g., they are known Parkin-binding proteins (DNAJA1, HSPA1A, HSPA8) (Imai *et al.*, 2002), and 164 interact with Parkin through one intermediate protein (network distance of 2). In total, 40 candidates are direct interactors of MonogenicPD proteins and six of them interact with two different MonogenicPD proteins. An example network for the ParkinTAP candidate LRPPRC is shown in Figure 3.15.

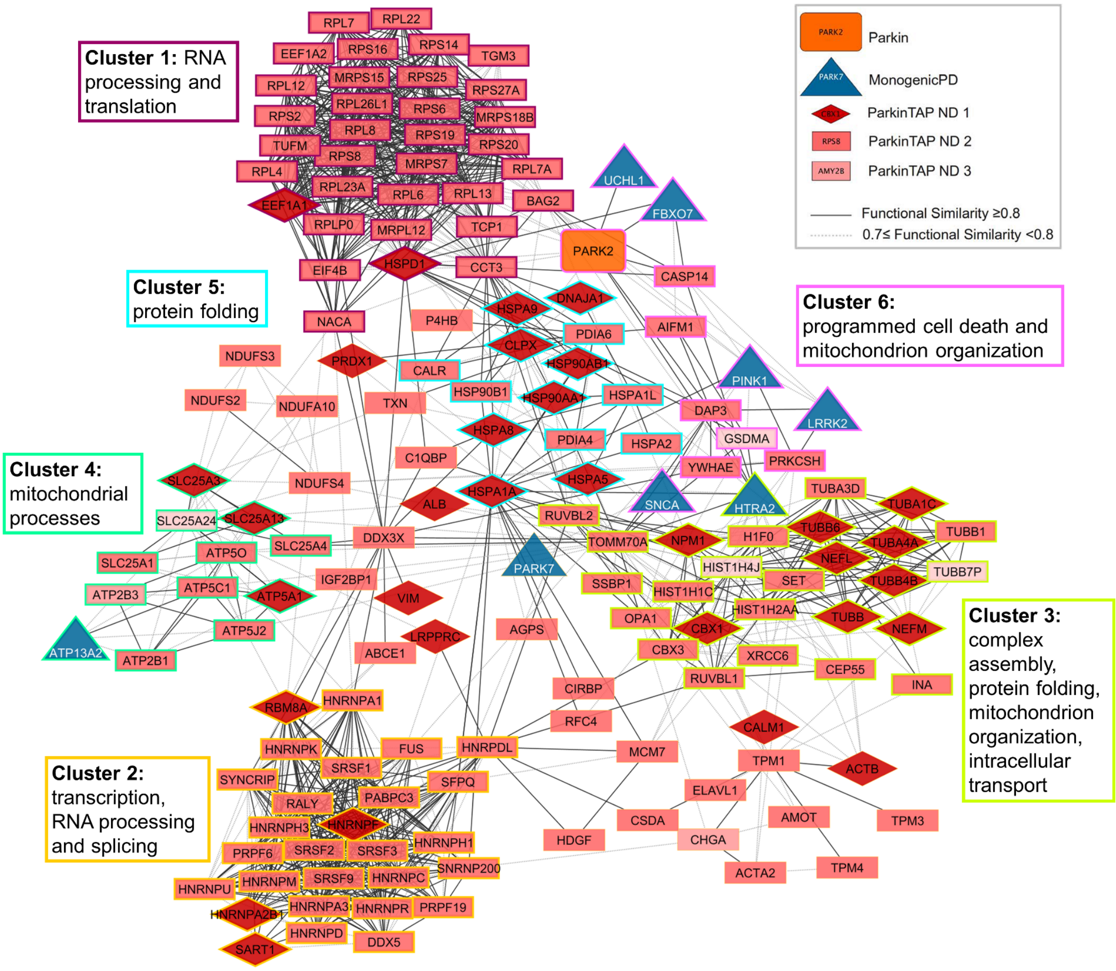


Figure 3.16: Functional similarity network of Parkin (orange), ParkinTAP candidates (red and pink nodes) and MonogenicPD proteins (blue nodes). The network contains 157 protein nodes that are connected by 1183 similarity edges. Edges indicate functional similarity between the connected proteins based on their Gene Ontology annotations above the cutoff of 0.7 (gray dotted lines) and 0.8 (solid black lines). Cluster membership is represented by node border colors as in this order: 1) violet, 2) orange, 3) yellow-green, 4) green-blue, 5) blue, 6) pink. Proteins that are not connected to any other node in the network are omitted. Figure adapted and extended from Zanon *et al.* (2013).

Furthermore, we constructed an FSN for the proteins from ParkinTAP, the MonogenicPD proteins and their interactors. Each edge represents strong functional similarity of two proteins based on annotated GO terms from the Biological Process Ontology as determined by the functional similarity measure rfunSim (Schlicker *et al.*, 2006). Significant rfunSim similarity values above the cutoff of 0.7 were retrieved using the FunSimMat web service (Schlicker and Albrecht, 2008, 2010). The network contains 157 proteins connected by 1183 similarity edges (886 with rfunSim ≥ 0.8) and is shown in Figure 3.16. The single connected component includes 149 ParkinTAP candidates and 8 MonogenicPD proteins.

We identified groups of functionally related proteins by using the Cytoscape plugin ClusterONE because it can find densely connected overlapping regions within a network and considers edges weighted by the functional similarity scores (Nepusz *et al.*, 2012). We identified six significant clusters in the FSN of ParkinTAP and MonogenicPD, whereas significance is defined by ClusterONE as $P \leq 0.05$. The clusters are highlighted in Figure 3.16. GO enrichment analysis with topGO (Alexa *et al.*, 2006) resulted in the following functional annotations for the proteins included in each cluster:

- Cluster 1 (34 proteins, P-value = 0.00): RNA processing and translation
- Cluster 2 (28 proteins, P-value = 0.00): transcription, RNA processing and splicing
- Cluster 3 (27 proteins, P-value = 5.62E-9): complex assembly, protein folding, mitochondrion organization, and cytoskeleton-dependent intracellular transport
- Cluster 4 (12 proteins, P-value = 9.34E-5): mitochondrial processes, like mitochondrial transport, mitochondrial ATP synthesis, and respiratory electron transport chain
- Cluster 5 (16 proteins, P-value = 0.012): protein folding
- Cluster 6 (15 proteins, P-value = 0.047): programmed cell death and mitochondrion organization

Furthermore, we observed that the ParkinTAP candidates are mostly similar to each other and that most of the MonogenicPD proteins (6 out of 8 in the FSN) are contained in only one cluster. Thus, we generated a subnetwork (Figure 3.17) consisting of only MonogenicPD proteins and the ParkinTAP candidates that are functionally similar to them. This network includes 38 proteins connected by 136 edges. ClusterONE identified four significant clusters ($P \leq 0.05$) in this network.

- Cluster 1 (18 proteins, P-value = 4.57E-7): programmed cell death and mitochondrion organization
- Cluster 2 (10 proteins, P-value = 1.34E-4): translation and protein folding
- Cluster 3 (15 proteins, P-value = 0.002): programmed cell death, mitochondrion organization, protein folding, and proteolysis

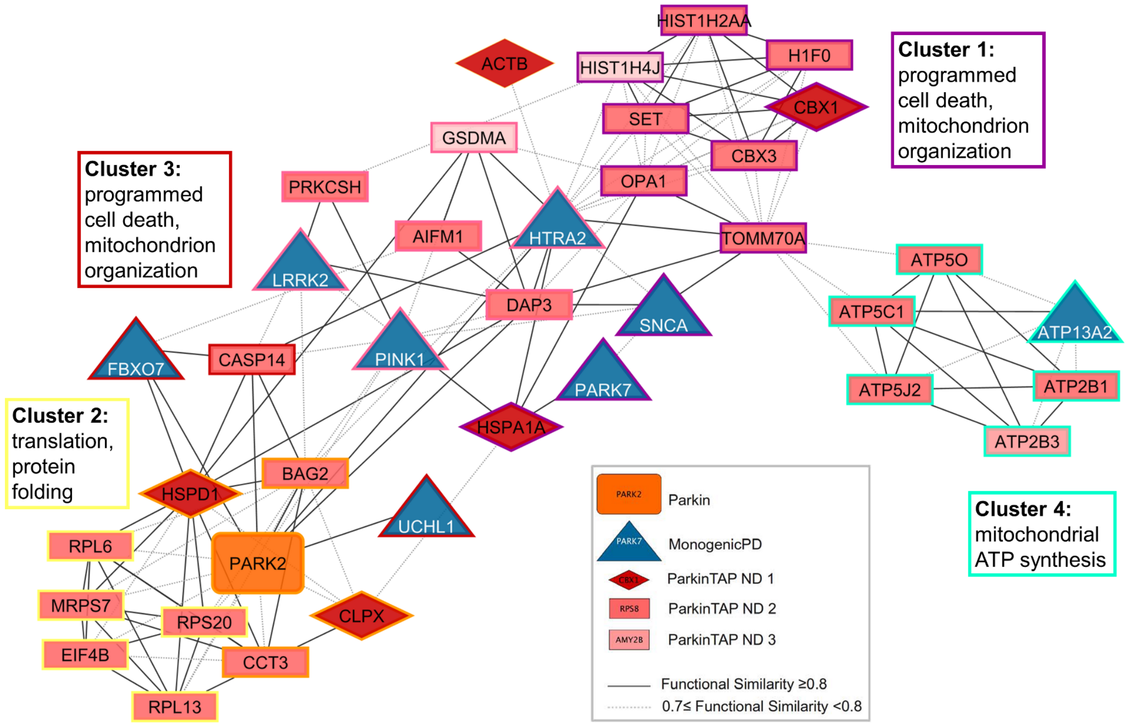


Figure 3.17: Functional similarity network of MonogenicPD proteins (blue) and their neighbors, ParkinTAP candidates (red and pink nodes) and Parkin (orange). The network contains 38 protein nodes connected by 136 similarity edges. Edges indicate functional similarity between the connected proteins based on their Gene Ontology annotations above the cutoff of 0.7 (gray dotted lines) and 0.8 (solid black lines). Cluster membership is represented by node border colors as in this order: 1) violet, 1&3) pink, 3) red, 2&3) orange, 2) yellow, 4) blue. Figure adapted and extended from Zanon *et al.* (2013).

- Cluster 4 (6 proteins, P-value = 0.002): mitochondrial ATP synthesis

Again, most MonogenicPD proteins cluster tightly together and are contained in clusters 1 and 3 (Figure 3.17). As can be seen both in the network and from the annotations, cluster 3 strongly overlaps with cluster 1 (7 nodes) and 2 (5 nodes).

Our collaborators from EURAC research made use of these network-based and other analyses to define a prioritization score for each ParkinTAP candidate (Zanon *et al.*, 2013). In particular, the scoring criteria included annotation of the candidates with GO processes enriched in the proteins causing monogenic forms of parkinsonism as well as the presence and number of direct protein interactions or strong functional similarity to PD-related proteins (MonogenicPD) or Parkin. In addition, the authors considered complementary experimental data from genetic interaction screens in *Drosophila melanogaster* and GWAS in humans. The promising candidates were involved in cell death processes, protein folding, the fission/fusion machinery, and the mitophagy pathway. Finally, Francisco S. Domingues and colleagues performed a co-immunoprecipitation experiment to verify two of the top ranking candidates, LRPPRC and TOMM70A. They were only able to confirm a physical interaction between Parkin and TOMM70A.

3.5.3 Summary and discussion

Since the identification of the Parkin gene several years ago (Kitada *et al.*, 1998) many research efforts have been devoted to the investigation of its gene product and its role within cellular pathways and processes. An important step in this direction is the identification of the complex network of interactions between Parkin and Parkin-binding proteins. Our collaboration partners at EURAC research performed a TAP/MS proteomic screen to determine 203 such candidate proteins (Parkin-TAP). We assisted them with the bioinformatics and, in particular, network-based analysis of the identified candidates.

First, we constructed a network of publicly available protein interactions for the ParkinTap proteins and the proteins known to be involved in monogenic parkinsonism (MonogenicPD). Most of the candidates interacted with Parkin through one intermediate protein, three of them were direct Parkin interactors and 40 were connected to MonogenicPD proteins. This finding suggested that these candidates might be part of a PD-specific pathways as proteins linked to the same disease tend to interact with each other and be involved in the same disease-related processes (Barabási *et al.*, 2011). Second, we analyzed the functional similarity network of ParkinTAP candidates and MonogenicPD proteins. Cluster analysis revealed that most of the candidate proteins are functionally similar amongst themselves and are members of 6 distinct functional groups associated with RNA processing, complex assembly, protein folding, intracellular transport, mitochondrial transport and ATP synthesis, and programmed cell death, respectively. One of the functional clusters was particularly interesting as it contained most of the MonogenicPD proteins. Candidates from this cluster were enriched with the same GO processes as the proteins already known to interact with Parkin, such as cell death, mitochondrion organization and protein folding. Our collaboration partners successfully verified one of the top candidates (TOMM70A) by a co-immunoprecipitation experiment.

Although this work has already given some new insights into the complex network of Parkin and its binding proteins, more high quality and comprehensive datasets will be needed to identify the shared disease pathways, their components and the perturbations that lead to the given phenotype. Again, network analysis has proven to be a valuable tool for prioritizing candidates and identifying groups of functionally related proteins that might represent disrupted pathways and processes. A network perspective on PD and other diseases might be the right means to providing new targets for the development of therapeutic interventions.

3.6 Conclusions

Many efforts are still devoted to the discovery of genes involved with specific phenotypes, in particular, diseases. High-throughput techniques are thus applied frequently to detect dozens or even hundreds of candidate genes. However, the experimental validation of many candidates is often an expensive and time-consuming task. Therefore, a great variety of computational approaches has been developed to support the identification of the most promising candidates for follow-up studies (Doncheva *et al.*, 2012b). The biomedical knowledge already available about the disease of interest and related genes is commonly exploited to find new gene-disease associations and to prioritize candidates for follow-up experimental studies.

In particular, many disease gene prioritization methods consider discriminative gene and protein properties and rank candidate genes according to their functional and phenotypic similarity or network proximity to known disease genes. In this context, functional information, manually curated or automatically derived annotations, often provides strong evidence for establishing links between diseases and relevant genes and proteins (Perez-Iratxeta *et al.*, 2002; Schlicker *et al.*, 2010). Many prioritization methods use protein interaction data as rich information source for finding relationships between gene products of candidate genes and disease genes (Kann, 2007; Ideker and Sharan, 2008; Baudot *et al.*, 2009; Vidal *et al.*, 2011; Wang *et al.*, 2011). In addition, the phenotypic similarity of diseases can help to increase the total number of known disease genes for less studied disease phenotypes (Lage *et al.*, 2007; Wu *et al.*, 2008; Li and Patra, 2010b; Vanunu *et al.*, 2010). Other sources of biological information frequently used by prioritization approaches are sequence properties, gene expression data, molecular pathways, functional orthology between organisms, and relevant biomedical literature (Kann, 2009; Tranchevent *et al.*, 2011; Piro and Di Cunto, 2012; Bromberg, 2013).

Building upon these methods, we designed an integrative network-based prioritization framework that combines different types of data and analysis techniques. So far, we focused on two data sources that have already proven to be very informative, protein interactions and functional annotations. First, we constructed phenotype-specific protein interaction (PINs) and functional similarity networks (FSNs) for a give list of candidate and known genes associated with a disease. The PINs contain known physical protein-protein interactions and protein complex memberships retrieved from the public database iRefIndex (Razick *et al.*, 2008), while the FSNs are based on pairwise similarity of genes based on their GO annotations as computed by the functional similarity measure rfunSim (Schlicker *et al.*, 2006) and retrieved from the FunSimMat web service (Schlicker and Albrecht, 2008). We analyzed these networks using clustering algorithms, centrality measures, subnetwork identification, and combinations of these techniques. We also proposed a method for assessing the overlap between similar phenotypes based on their joint FSN. We applied our framework to the prioritization of candidate genes and functional characterization of inflammatory bowel disease (Ellinghaus *et al.*, 2013b), primary sclerosing cholangitis (Liu *et al.*, 2013), and Parkinson's disease (Zanon *et al.*, 2013).

Inflammatory bowel disease is a chronic inflammatory disorder of the gastrointestinal tract that has been extensively studied in the last few years (Ellinghaus *et al.*, 2015). Although meta-analyses of genome-wide associations studies have established 163 IBD susceptibility loci (Barrett *et al.*, 2008; Franke *et al.*, 2010; Anderson *et al.*, 2011; Jostins *et al.*, 2012), there are still many unanswered questions about the etiology of IBD. In particular, the two IBD subtypes, Crohn's disease and ulcerative colitis, share many genetic associations although they are phenotypically very different. In order to assess their similarity, we constructed a joint FSN and performed several network-based analyses with the conclusion that there is a substantial functional overlap and no distinct functional groups between the two IBD subtypes. Furthermore, based on network connectivity, we selected a set of IBD candidate genes from different loci that are very likely to be involved in the same disrupted processes and pathways (such as cell proliferation, T cell activation, immune response) and thus, have a crucial role in the development of IBD.

Since the already known IBD risk factors explain only 13.6 % of the disease variance for CD and 7.5 % for UC (Jostins *et al.*, 2012), it is very likely that low-frequency and rare variants have a major contribution to heritability. Therefore, more and more studies make use of next-generation sequencing to study complex diseases such as IBD. Our collaboration partners performed exome sequencing of 42 CD patients and 5 healthy subjects and after several filtering, annotation, and follow-up genotyping stages, they identified two CD candidate genes, PRDM1 and NDP52 (Ellinghaus *et al.*, 2013b). Using network analysis, we were able to support the functional relationships of NDP52 with known IBD genes and autophagy-related genes. Furthermore, we performed a large-scale analysis of the candidates from the exome sequencing and suggested further IBD risk genes involved in programmed cell death, phosphorylation, and intracellular signaling.

In contrast to IBD, PSC is a severe liver disease with unknown etiology and no associated genes so far. Therefore, our collaboration partners performed a large genotyping study of PSC using the custom-made Immunochip array (Cortes and Brown, 2011) and identified 9 novel loci in addition to the 7 already known (Liu *et al.*, 2013). To prioritize the resulting list of PSC candidates, we put more effort into the development of a network-based method that does not need a priori knowledge about the involved genes. In an FSN of PSC candidates, we performed topology analysis to select one disease-relevant gene per locus and build a disease-specific network. Six out of seven candidate genes suggested by our method were also implied as causative by at least one additional prioritization method. Furthermore, we assessed the genetic and phenotypic overlap between PSC and IBD, which is diagnosed in 72 % of the PSC patients considered for this GWAS. By constructing and analyzing the joint FSN of PSC and IBD, we suggested a considerable functional overlap and interplay between the two diseases.

For the particular case of Parkinson's disease, a progressive neurodegenerative disorder of the central nervous system, the focus was on identifying interaction partners of the already known Parkin gene product (Kitada *et al.*, 1998; Hedrich *et al.*, 2004). Starting with 203 candidate proteins from a TAP/MS proteomic screen and a list of proteins involved in monogenic PD, we constructed a network based on known physical protein interactions and complex memberships and found evidence supporting the (almost) direct involvement of most candidates with Parkin (Zanon *et al.*, 2013). Furthermore, cluster analysis on the PD candidates FSN revealed that most candidates are functionally similar to each other with an exception of few proteins that are particularly similar to MonogenicPD proteins and are enriched with GO terms associated with Parkin (cell death, mitochondrion organization and protein folding). Further integrative analyses by our collaboration partners resulted in a ranking of the ParkinTAP candidates according to several functional and interaction properties and the experimental validation of one of the top candidates, TOMM70A.

Characterizing complex diseases by networks of interacting causal and candidate genes is a new trend in the field of disease gene prioritization and has been also considered in several recent studies. For instance, Menche *et al.* (2015) developed a computational framework to identify disease-specific modules within the human interactome and to explain some of the pathological relationships between diseases by the overlap of their disease modules. In a follow-up study, (Ghiassian *et al.*, 2015) proposed a new algorithm for disease module detection based on network properties derived after investigating the interaction network profile of 70 complex diseases. Furthermore, Tasan *et al.* (2015) presented a method for the identification of disease subnetworks consisting of functionally related genes spanning multiple GWA loci.

These approaches are very similar to our phenotypes-specific framework and thus confirm the significance to our approach. Different properties such as the number of already known disease genes or associated candidates as well as the availability of biological knowledge for the phenotype of interest still strongly influence the choice of approach. However, with the huge amounts and diversity of experimental and computationally derived data becoming available every day, we can expect a considerable improvement in the design and performance of candidate disease gene prioritization approaches in the near future. With regard to understanding the genetic and phenotypic similarities and differences between diseases, network-based methods already show great potential and should be considered for further development.

CHAPTER 4

Integrative visual analysis of protein sequence mutations

A natural and complementary step to identifying disease-related genes is understanding how these genes alter a given phenotype. One such mechanism are single nucleotide variants that have an effect on the structure and function of the respective gene product. So far, most computational methods focus on identifying whether a mutation will affect a protein or not, but there are not so many tools that can aid the interpretation of the resulting changes. Thus, in this chapter, we present a novel visual approach for analyzing residue mutations by showing various aspects of the biological information on different scales. We combine different biological visualizations and integrate them with molecular data derived from external resources. Thereby, we make use of the well-known software tools Cytoscape (Shannon *et al.*, 2003) and UCSF Chimera (Pettersen *et al.*, 2004), extend the Cytoscape plugin structureViz developed by our collaborator (Morris *et al.*, 2007), and combine them with our tools RINerator and RINalyzer (Doncheva *et al.*, 2011). In particular, we assess the impact of individual amino acid changes by the detailed analysis of the involved residue interactions. Our work won the overall favorite prize at the BioVis 2013 Data Analysis contest and was later published it in a peer-reviewed journal (Doncheva *et al.*, 2014). In particular, Section 4.1, 4.2, and 4.4 contain text that has been adapted and extended from Doncheva *et al.* (2014).

4.1 Introduction

Understanding and predicting the effect of amino acid mutations on the structure and function of a protein is still a challenging problem despite recent advances (Hecht *et al.*, 2013; Castellana and Mazza, 2013). In the case of multiple sequence changes, it is even more difficult to distinguish the mutations with a significant effect from the ones without. There are also several contradictory theories about the role of epistatic interaction between amino acid changes in molecular evolution (Breen *et al.*, 2012; McCandlish *et al.*, 2013; Weinreich *et al.*, 2013). A recent comparative genomics study by Jordan *et al.* (2015) identified several human disease mutations that occur in other species, where their benign effect is compensated by another amino acid change. Resistance-associated mutations in viral and bacterial proteins are also of particular interest for developing new therapies (Hughes and Andersson, 2015).

Many approaches that tackle this problem have been presented in the last couple of years as reviewed in (Thusberg *et al.*, 2011; Cooper and Shendure, 2011; Mah *et al.*, 2011; Capriotti *et al.*, 2012; Gnad *et al.*, 2013; Stefl *et al.*, 2013). Computational methods such as the well-known SIFT tool (Sim *et al.*, 2012) use evolutionary conservation derived from a multiple sequence alignment to predict that mutations of highly conserved residues have a considerable impact on function. Other methods such as the well-established PolyPhen2 tool (Adzhubei *et al.*, 2010) combine sequence features with structural and physico-chemical protein properties to assess the effect of a mutation. A notable disadvantage of most tools is that that they do not provide the user with a fine-grained control over the set of features used for the prediction, and the results are often difficult to interpret. In addition, those tools cannot easily cope with the speed at which new information on sequences, structures, and functions is made publicly available. Last but not least, mutations and their effects are considered as independent from each other, which is not the case in reality.

Thus, the BioVis community selected this area of research for the 2013 Data Analysis challenge (Ray *et al.*, 2014). The organizers posed the question how protein function depends on the underlying protein sequence and whether it is possible to predict the effect of sequence changes. They also encouraged the use of visualization and data integration as the key to solving the problem. In particular, given the sequence of a functionally defective triosephosphate isomerase (dTIM) with circa 100 mutations and its parent, the yeast triosephosphate isomerase (scTIM), the task was to identify a minimal set of mutations that abolish its function and to suggest rescue mutations that might restore its function. The dTIM sequence was the result of a lab experiment performed by Sullivan and Magliery, who also produced and validated some successful point rescue mutants (Sullivan *et al.*, 2011, 2012). An additional intriguing characteristic of the mutations was derived from the family consensus sequence of TIM proteins. This particular case demonstrates that amino-acid residues contribute to the function of the protein as a group and not as individual entities. Thus, the BioVis organizers aimed to encourage the development of tools that analyze mutational networks, e.g., the connections of multiple mutated residues (Ray *et al.*, 2014).

Recent studies have revealed that combining systems and structural biology could be very beneficial for both fields (Fraser *et al.*, 2013). Our previous efforts involved the development of a novel approach to investigating protein structure-function relationships based on interactive visual analysis of residue interaction networks (RINs) derived from the 3D protein structure (Doncheva *et al.*, 2011, 2012a). RINs provide additional insights into the structural and functional roles of interacting residues as evidenced by their successful application for investigating protein dynamics and engineering, structure-function relationships, protein and ligand binding (Csermely, 2008; Vishveshwara *et al.*, 2009; Greene, 2012; Di Paola *et al.*, 2013; Hu *et al.*, 2013; Yan *et al.*, 2014). An important aspect of studying the relationship between protein sequence, structure and function is the molecular characterization of the effect of protein mutations. To understand the functional impact of amino acid changes, the multiple biological properties of protein residues have to be considered together with their topological characteristics.

Therefore, for our entry to the BioVis 2013 data contest challenge, we focused on improving the integrative visualization of a wide variety of available information on sequences, structures and functions. Our objective was to provide the biological data for a manual visual analysis and interactive exploration by the user in an integrated fashion by making it accessible through a small number of carefully designed, linked views. In this way, the user is able to generate hypotheses based on a specific view (e.g. of the protein structure) in the context of the other linked views and the provided data. As there are many biological aspects of protein sequence mutations that might affect protein structure and function, we developed visualizations that provide different levels of detail and enriched them by mapping additional data onto the graphical representations. Our approach includes one-dimensional sequence views, three-dimensional protein structure views and two-dimensional views of residue interaction networks as well as aggregated views. The views are linked tightly and synchronized to reduce the cognitive load of the user when switching between them. In particular, the protein mutations are mapped onto the views together with further functional and structural information.

We aimed at a generic solution that is suitable for a wide range of proteins and will support a comprehensive analysis of the impact of mutations for a large class of sequence changes. This was accomplished by a visual analytics approach integrating several tools into a software suite freely available at the RINalyzer website (Doncheva *et al.*, 2015). We developed both the old and the new version of RINalyzer, except for the RINlayout and the abstract network generation, which were provided by Karsten Klein. We also implemented the new version of structureViz after extensive discussions with the previous developer John H. Morris. The integration with Pro-origami was provided by Michael Wybrow.

As detailed below, we applied our approach to the data provided for the BioVis 2013 Data Analysis contest. For this proof-of-concept study, we assessed the sequence changes between scTIM and dTIM by different visualizations of the protein structure together with further functional and structural information and by an exploratory analysis based on the complementary network views for both sequences. Although we did not use the term mutational networks and did not specifically study subnetworks of mutations, the use of our approach for this task is straightforward. After we performed the described analysis, we discussed and interpreted the final results with our collaborators Karsten Klein and Francisco S. Domingues.

Furthermore, we combined our visual analytics framework with some additional statistics on the structural and topological properties of residues to analyze a large set of known resistance-associated mutations in the HCV NS3 protease as provided by Christoph Welsch. The application of protease inhibitors for treating patients infected with HCV is often unsuccessful because of the appearance of such mutations. Therefore, it is crucial for future patient care to understand how and why resistance mutations arise. To this end, we analyzed 21 resistance-associated residues of the HCV NS3 protease and identified some interesting trends of their physico-chemical and topological properties with respect to important functional sites in the protease. We designed the project together with our collaboration partners Christoph Welsch and Francisco S. Domingues. Then we performed the implementation and analysis of the resulting findings.

4.2 Visual analytics approach

This section is an adapted and extended version of the content from Doncheva *et al.* (2014) and describes in detail our visual analytics approach presented at the BioVis data contest 2013.

4.2.1 General concept

Our visual analytics approach assists the user's reasoning about the biological impact of mutations by interactive visualizations of sequence and structure information enriched with additional biological knowledge such as evolutionary sequence conservation and functional annotations. To show the different aspects of the data, we combine the well-known 3D structure view and the one-dimensional sequence view with the 2D RIN view. In addition, we create simplified network representations to enable the user to focus on certain biological aspects, e.g. protein domains, secondary structure elements, and functional annotations.

Besides the sequence that is given as input, a variety of information is available that can be used to interpret the functional effects of sequence changes. This includes sequence conservation, which might point to highly conserved regions responsible for some function, protein domain information, functional annotations (e.g. on molecular binding), structural properties such as hydrophobicity and solvent accessible surface area, and already known mutations and their impact. We incorporated a number of sources for such information in our approach. The available data is then mapped as visual cues on top of the graphical representations of the protein structure and the RINs. In addition, we made use of the network representation provided by RINalyzer as well as the Cytoscape analysis capabilities to facilitate data exploration by filtering and combining the available information on individual residues.

Furthermore, to present sequence changes on the structure and residue interaction level simultaneously, we provide both a single cumulative network view and two

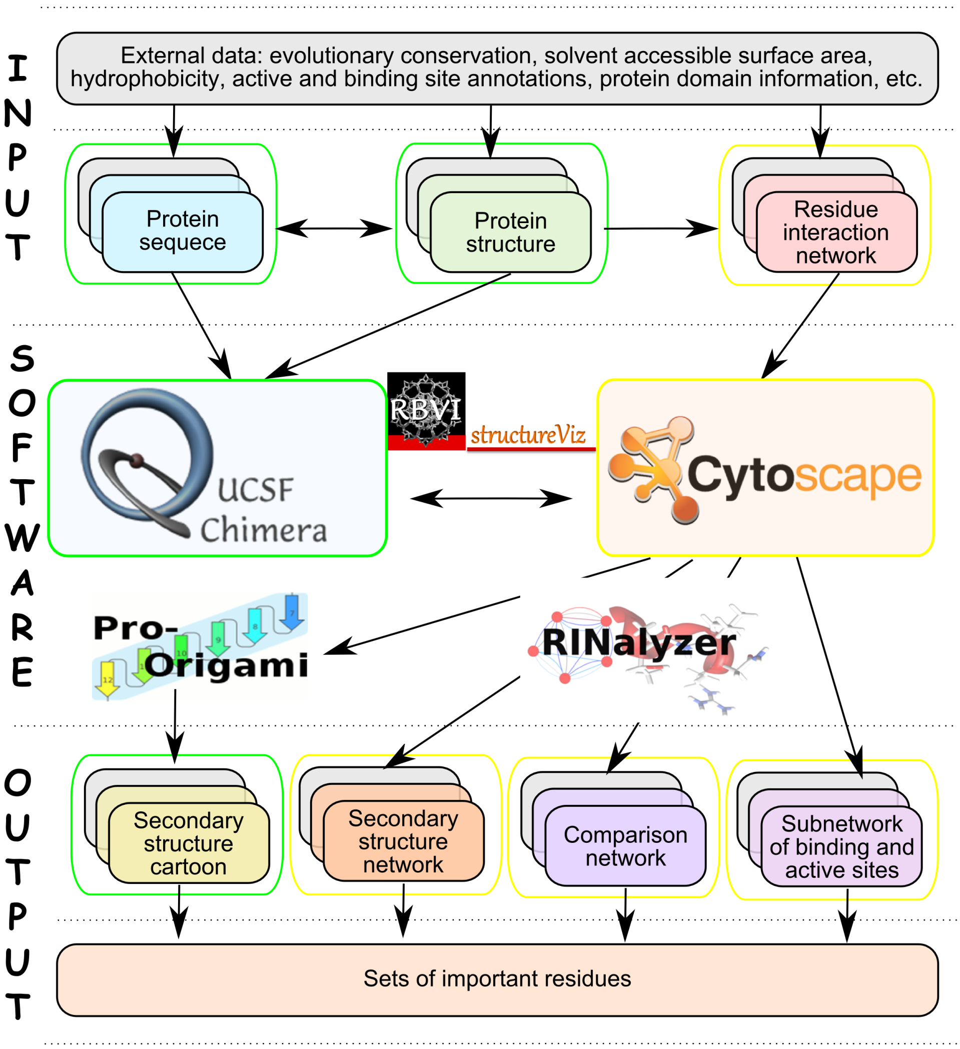

INPUT
External data: evolutionary conservation, solvent accessible surface area, hydrophobicity, active and binding site annotations, protein domain information, etc.
Protein sequece
Protein structure
Residue interaction network
SOFTWARE
UCSF Chimera
RBVI
structureViz
Cytoscape
Pro-Origami
RINalyzer
OUTPUT
Secondary structure cartoon
Secondary structure network
Comparison network
Subnetwork of binding and active sites
Sets of important residues

Figure 4.1: General analysis workflow. The workflow consists of three parts: input, software and output. The input consists of biological data, which might be protein sequences, structures, RINs as well as additional annotations and biological knowledge retrieved from external sources and databases (shown as gray background for each view). The middle part of the workflow shows the interactions between the different tools and which tool is responsible for the presentation of which data. The output consists of the different views with data mapped onto them and sets of important residues that can be identified through visual exploratory analysis of the available data. The yellow and green boundaries indicate the default selection color used by the different tools.

separate views of the parent and the defective mutant RINs side-by-side. While a single view facilitates the identification of changed sites, the dual view solution allows the user to identify the structural impact of the changes, for example, lost residue interactions might alter the protein structure.

A general analysis workflow is presented in Figure 4.1. Normally, the user starts with one or more experimentally determined protein structures and retrieves or generates RINs for them. In case only sequences are available, external tools for predicting the 3D structure could be used instead. External data such as evolutionary conservation and functional annotations need to be prepared in a format compatible with Cytoscape and the RIN specifications. Then the data is loaded by the user into Cytoscape and UCSF Chimera. Further views such as the secondary structure cartoon, the aggregated secondary structure network or the comparison network can be created from within Cytoscape. The sequences of the structures can be displayed and manipulated from within UCSF Chimera. Functional annotations and evolutionary conservation have to be imported manually into Cytoscape as node attributes of the RINs, while structural properties can be retrieved automatically from the protein structures currently opened in UCSF Chimera. These data can then be applied to create the visual cues and semi-automatically propagate them to the different views. Finally, by browsing and filtering the data in Cytoscape and UCSF Chimera, the user can identify relevant amino acids, in particular, mutated residues with a potentially strong effect on the protein function. Even if the visual analysis does not immediately reveal the functional consequences of mutations, our software will provide the user at least with very useful biological indications for the molecular analysis and further experiments.

4.2.2 Provided views

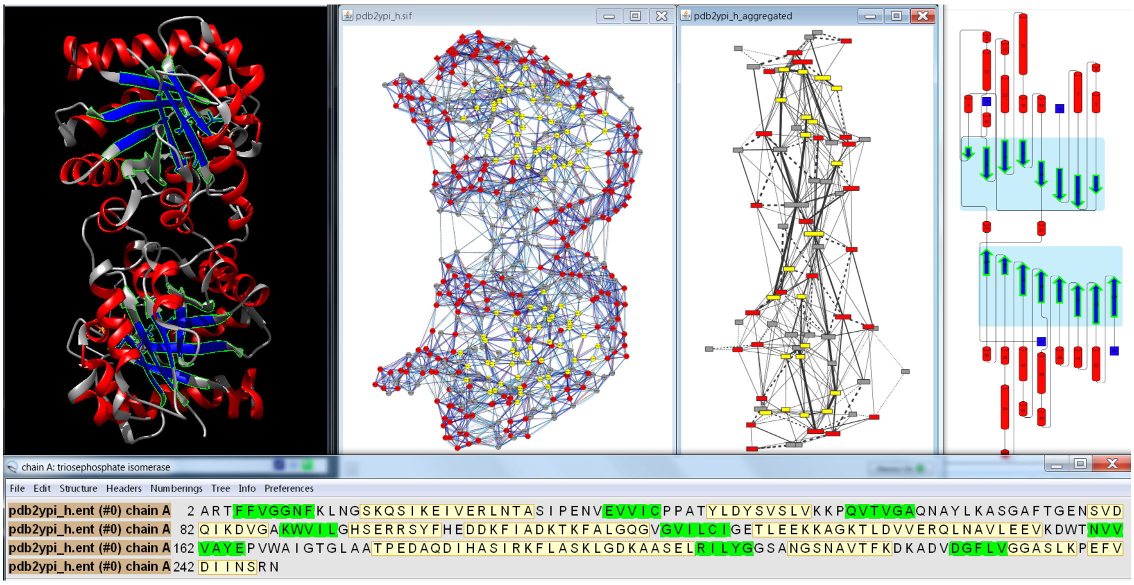
To offer the available information to the user on different levels of abstraction and to support interactive synchronized exploration (Figure 4.2), we selected the following suitable visualizations.

Structure and sequence

We used the standard representations of the three-dimensional (3D) structure and sequence of proteins as provided by UCSF Chimera (Pettersen *et al.*, 2004; Meng *et al.*, 2006) because sequence changes and their impact on the structure might give valuable insight. UCSF Chimera offers a variety of tools that support the interactive crosstalk between sequences and structures, affording advanced exploration of multiple sequence alignments, comparison of structures and incorporation of user-specific data. In particular, the user can study the amino acid changes between two sequences and their locations on the corresponding protein structures. It is also possible to construct a structure-based sequence alignment from the superposition of two structures. This deep integration of sequences and structures is further complemented by a multitude of molecular graphics features.

Residue interaction network

A two-dimensional (2D) residue interaction network (RIN) can be created for any given 3D protein structure by the RINerator package and then visualized with the help of RINalyzer (Doncheva *et al.*, 2011) within the Cytoscape platform (Shannon *et al.*, 2003). In the resulting visualization, network nodes represent amino-acid residues and edges depict non-covalent residue interactions. Such a network rep-


pdb2ypi_h.sif
pdb2ypi_h_aggregated
chain A: triosephosphate isomerase
File Edit Structure Headers Numberings Tree Info Preferences

```
pdb2ypi_h.ent (#0) chain A   2 ARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATYLDYSVSLVKKPQVTVGAQNAYLKASGAFTGENSVD
pdb2ypi_h.ent (#0) chain A  82 QIKDVGAKWVILGHSERRSYFHEDDKFIADKTKFALGQGVGVILCIGETLEEKKAGKTLDVVERQLNAVLEEVKDWTNVV
pdb2ypi_h.ent (#0) chain A 162 VAYEPVWAIGTGLAATPEDAQDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGFLVGGASLKPEFV
pdb2ypi_h.ent (#0) chain A 242 DIINSRN
```

Figure 4.2: Simultaneous visualization of biological information using different complementary views of scTIM. In particular, the three-dimensional structure and its sequence (top left and bottom, respectively) are shown with UCSF Chimera, the resulting two-dimensional view of the residue interaction network and the aggregated secondary structure network generated with RINalyzer are visualized in Cytoscape (top middle), and the cartoon image of the secondary structure elements is provided by Pro-origami (top right). Residue and network nodes are colored according to their secondary structure (strands in blue and helices in red). Strands that have been selected within UCSF Chimera are indicated by green boundary color in the structure view, by green background in the sequence view, by yellow node color in Cytoscape, and by green boundary color and blue background in Pro-origami. Figure first published in Doncheva *et al.* (2014).

resentation is very useful to demonstrate the impact of mutations at the detailed residue interaction level by highlighting the changes of local interactions as well as long-range interaction paths, e.g. indirect interactions between residues. To transfer the spatial localization information of the mutations from the structure view to the network view, we replaced the previous force-directed layout algorithm by a more appropriate stress minimization variant (Figure 4.2).

Aggregated views

We offer less complex, aggregated overviews that focus on functional or structural subunits like secondary structure elements and illustrate the location and distribution of the mutations on the protein structure. In particular, we utilized the cartoon view as provided by the Pro-origami web service (Stivala *et al.*, 2011). The main advantage of this view is that it gives a clear depiction of the chain and the secondary structure elements, while it leaves out the exact spatial location and the interrelations between those elements, which are provided by the other more detailed views. As the visual mapping from a RIN to the corresponding cartoon might be difficult for the user, a network representation that shows the RIN together with aggregated secondary structure elements can be created as an intermediate visualization.

The aggregated views are intended to give the user a quick overview on the mutation locations with respect to specific known structural or functional regions. While it would be possible to map additional information directly onto the network representation, the RIN might become quite complex for the user. Thus, we utilize views that aggregate regions based on secondary structures, protein domain information, or functional annotations. These views serve as an intermediate visualization when switching between the 3D structure view and the 2D RIN view.

The simple cartoon view provided by the Pro-origami web service reduces the complex 3D protein structure to the essential secondary and super-secondary structure information and presents it with an easily readable layout (Figure 4.2). Pro-origami provides SVG images, which are enriched with further information in the form of highlighted regions of interest such as the localization of mutated residues. As Pro-origami can decompose proteins into domains, we can also obtain a combined representation of secondary structure and protein domains within the cartoon view.

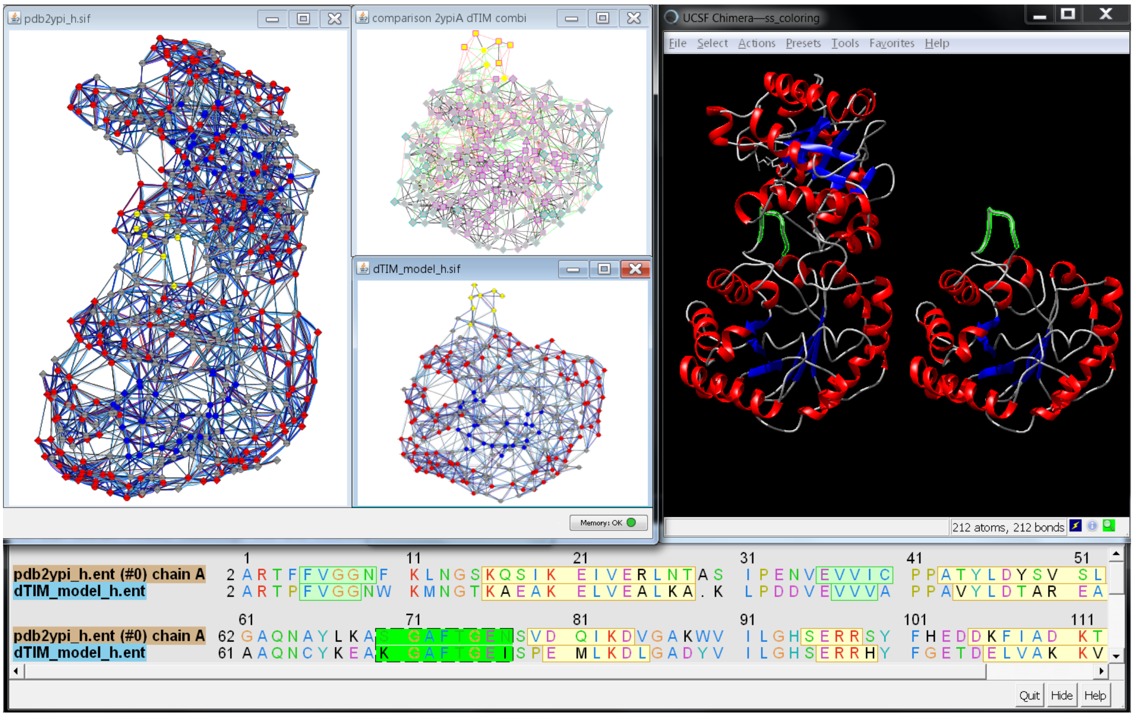
Comparison view

The representation of protein structures as RINs enables network comparison and alignment to explore the differences between parent and mutant structures further. Besides the comparison of two networks or structures side-by-side, we provide a comparison network view based on the alignment of the underlying sequences (Figure 4.3). In this view, each node represents a pair of aligned residues and two nodes are connected if the corresponding residues have a non-covalent interaction in either of the two compared RINs. In addition, visual cues are created to highlight interactions that were gained or lost upon amino acid change and the fraction of such interactions for each residue is estimated in order to quantify the mutational effect on protein structure and function.

Furthermore, to distinguish more or less likely mutations, we integrated the amino acid substitution scores from the Blosum62 matrix (Henikoff and Henikoff, 1992) in RINalyzer and assign a score to each mutated residue in the comparison network. Each score can be used to highlight sequence changes with a stronger impact on the protein.

4.2.3 Data enrichment and visual cues

Mapping of available knowledge onto the visualized sequences and structures is an important component of our visual analytics approach. The availability of this information in an easily accessible way for the user should facilitate the biological knowledge discovery considerably. To enrich the provided views, we extract additional structural and functional information from external databases and import the relevant data as node attributes in Cytoscape, which automatically associates them with the RIN and the protein structure. An additional benefit of this integra-



```
                              1                   11                  21                  31                  41                  51
pdb2ypi_h.ent (#0) chain A  2 A R T F F V G G N F   K L N G S K Q S I K   E I V E R L N T A S   I P E N V E V V I C   P P A T Y L D Y S V   S L
dTIM_model_h.ent            2 A R T P F V G G N W   K M N G T K A E A K   E L V E A L K A . K   L P D D V E V V V A   P P A V Y L D T A R   E A

                              61                  71                  81                  91                  101                 111
pdb2ypi_h.ent (#0) chain A 62 G A Q N A Y L K A S   G A F T G E N S V D   Q I K D V G A K W V   I L G H S E R R S Y   F H E D D K F I A D   K T
dTIM_model_h.ent           61 A A Q N C Y K E A K   G A F T G E I S P E   M L K D L G A D Y V   I L G H S E R R H Y   F G E T D E L V A K   K V
```

Figure 4.3: Side-by-side views versus comparison network view. The location of a set of residues is highlighted at the same time in all views, from left to right, the RIN of the 3D structure of scTIM, the comparison RIN, the RIN of the model of dTIM as generated by the SQWRL web server, the sequence alignment of the scTIM and dTIM sequences, and the corresponding 3D structures. The network nodes and residues are colored according to secondary structure (strands in blue and helices in red), except for the comparison RIN, where the node borders are colored according to conservation scores from ConSurf-DB (turquoise-to-pink coloring indicates variable-to-conserved sites). Selected nodes are shown in yellow color in the network views and with green boundary or green background in the structure and sequence view, respectively. Such a combination of views allows the user to study the structures and networks side-by-side or all at once in the comparison network.

tion is that it enables the use of the built-in Cytoscape functionality to create filters based on the imported data and to highlight the residue nodes with attribute values within a given range, e.g. with high or low conservation scores (see Figure 4.8).

The following information is regarded as potentially useful for analyzing the effect of mutations:

- Family conservation. ConSurf-DB (Goldenberg *et al.*, 2009) provides pre-computed profiles of evolutionary sequence conservation.
- Residue interactions. The RINerator package creates a network of non-covalent residue interactions such as contacts and hydrogen bonds for any 3D protein structure.
- Residue interaction counts and scores. RINerator also provides edge scores for

the strength of non-covalent interactions as well as a count of the interactions of a given type between two residues.

- Functional sites. Active and binding site information is retrieved manually from UniProtKB (The UniProt Consortium, 2014).

- Domain annotation. Protein domain information is obtained from the SCOP (Murzin *et al.*, 1995) online resource.

- Structural properties. Data for the solvent accessible surface area, secondary structure, hydrophobicity, and other residue properties is retrieved automatically from UCSF Chimera.

- Physico-chemical properties. A selected set of amino acid properties covering five major categories (polarity, secondary structure, molecular volume, codon diversity, electrostatic charge) as defined by Atchley *et al.* (2005) is automatically retrieved by RINerator from AAindex (Kawashima *et al.*, 1999; Kawashima and Kanehisa, 2000).

Functional residue annotations such as protein domain localization as well as binding and catalytic sites are important for identifying mutations that could have a direct impact on the function of the protein because they are in or near such sites. Structural properties of residues such as hydrophobicity, solvent accessible surface area, and polarity are used to characterize their potential effect on protein structure and function. Last but not least, evolutionary conservation information is crucial for distinguishing between residue changes in conserved (less tolerable of sequence changes) or variable regions.

The data used to enrich our visualizations is mapped as visual cues like color, shape, or line stroke in the network view and transferred to the other views where possible. We decided to control most visual properties via user-adjustable options with reasonable defaults. For example, different node shapes are used to distinguish the mutated residues in both the parent and the defective protein (Figure 4.7). Additionally, several visual styles are offered that map different functional and structural information on the views so that the user sees the distribution of corresponding values for the whole protein. Dark colors usually correspond to significant values such as strong hydrophobicity, large solvent accessible surface area or high number of changed residue interactions (Figure 4.3). For evolutionary conservation, the pink-to-turquoise coloring as applied by ConSurf-DB is used (Figure 4.8).

The visual cues are particularly useful for illustrating the changes in residue interactions due to the mutations in the comparison network view generated from the alignment of the respective sequences in UCSF Chimera. Residue interactions that are either lost or gained upon mutation are highlighted by differently colored and shaped lines (Figure 4.3). Residues that cannot be aligned are depicted by nodes with different node borders.

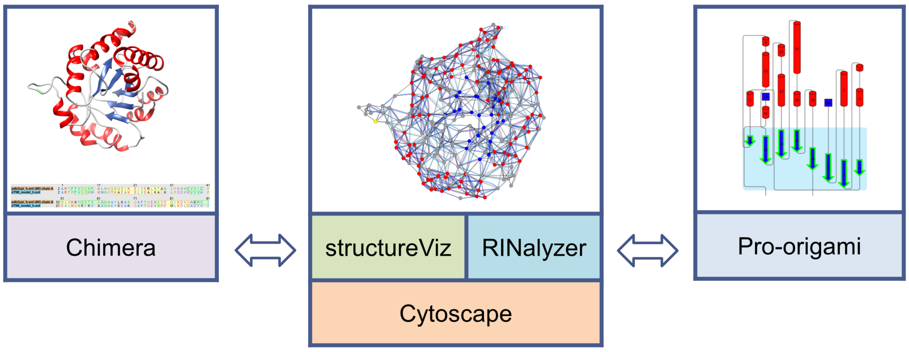


Figure 4.4: Overview of the involved tools and the corresponding visualizations. Figure first published in Doncheva *et al.* (2014).

4.2.4 Coordination of views

The linkage between the different views and the transfer of visual cues is maintained by several mechanisms. Regarding the interactive exploration, we propagate the selection of elements in one view to the others. We synchronize orientation and location between RINs and structures using a special layout algorithm that we developed for this purpose. In particular, we want to ensure a consistent use of information mapping and similar cues over all views.

To ease the user's cognitive load when switching between different views and tools, we link them in multiple important ways. For an interactive exploration, we implemented a global selection concept, that is, the selection of elements in one view leads to the immediate selection of their corresponding representatives in all other views. Our linkage concept also ensures the consistent use of information mapping and similar cues over all views, particularly, regarding the usage of colors.

Further coordination is achieved due to the synchronized orientation and location of the graphical representations in the different views. For instance, the user can freely explore the 3D structure within the UCSF Chimera window, e.g. by rotating the protein structure. The network view can then be adjusted according to the new orientation of the rotated structure by applying a 3D-structure based layout developed specifically for RINs (see Section 4.3.3 for more details).

4.3 Implementation details

All of the above is accomplished by a software suite (Figure 4.4) that integrates the freely available software tools Cytoscape (Shannon *et al.*, 2003), UCSF Chimera (Meng *et al.*, 2006), and Pro-origami (Stivala *et al.*, 2011) using our plugins RINalyzer (Doncheva *et al.*, 2011) and structureViz (Morris *et al.*, 2007). Cytoscape is an open-source software platform for data integration, analysis and visualization of complex (biological) networks (Shannon *et al.*, 2003). Its functionality can be further extended and tailored by users through the implementation of plugins or apps, from which over 200 are available nowadays. On the other hand, UCSF Chimera is a well-known program for interactive visualization and analysis of molecular structures (Meng *et al.*, 2006). Although linking Cytoscape and UCSF Chimera is programmatically not straightforward, the benefits of enriching biological networks with structural information as well as studying proteins from a network perspective are obvious. These two objectives have been accomplished by the plugins structureViz and RINalyzer, respectively.

Released in 2007, the structureViz plugin links the visualization of biological interaction networks, in particular, protein interaction networks, with the analysis and visualization of macromolecular structures provided by UCSF Chimera (Morris *et al.*, 2007). It supports the association between network nodes in Cytoscape and corresponding structures open in UCSF Chimera. For example, the user can explore the 3D structures of two physically interacting proteins in a network or, if resolved, the structure of their complex. The plugin structureViz also provides the Cytoscape Molecular Structure Navigator, a simplified tree-like interface for viewing the loaded structures and their residues and for accessing important UCSF Chimera functionality such as changing the display of models, chains, residues, selecting chemistry, and performing structure alignment.

In 2011, we released RINalyzer, a Cytoscape plugin that provides versatile and interactive structure analysis tools for RINs and enables dynamically linked 2D network and 3D structure views (Doncheva *et al.*, 2011). In particular, it allows for simultaneous, interactive 2D visualization and exploration of the RINs in Cytoscape and the corresponding molecular 3D structures in UCSF Chimera. Furthermore, RINalyzer offers the computation and illustration of a comprehensive set of weighted centrality measures for relating spatially distant residue nodes and discovering critical residues and their long-range interaction paths in protein structures. Another software feature is the network comparison of aligned protein structures by constructing a combined RIN, which enables the detailed comparative analysis of residue interactions in different proteins. In addition, RINalyzer facilitates the visual mapping of additional data, such as secondary structure, surface accessibility, evolutionary conservation, and structural reliability and flexibility onto RIN nodes and edges.

In order to implement the full linkage between Cytoscape and UCSF Chimera for our novel visual analytics approach, we made use of their new software versions. We also ported the plugins RINalyzer and structureViz to work with Cytoscape 3 and thereby linked them together. We kept the focus of the new structureViz2 on the interface between Cytoscape and UCSF Chimera, while the new RINalyzer2 version has extended visual analytics functionality for RINs. In particular, structureViz2 can communicate with UCSF Chimera not only by associating protein nodes with their structures, but on an additional level, that of residues. In this way, a RIN can be created from a loaded protein structure and the RIN nodes are linked with the protein residues. However, structureViz2 itself does not have any RIN-specific functionality besides retrieving the residue interactions from UCSF Chimera or enabling bidirectional selection and color transfer. Therefore, RINalyzer provides the structure-based layout, several visual styles, as well as the advanced network comparison and analysis functionality. Download links and further documentation can be found at the RINalyzer website (Doncheva *et al.*, 2015). More details on the implementation are given for each tool separately in the next sections.

4.3.1 Cytoscape 3.x series

In 2013, the Cytoscape consortium finally released the new 3.x series. This was a major step of redesign and reimplementation of the older 2.x series with focus on modular architecture and long-term maintainability. As before, the Cytoscape core distribution provides basic functionality for data integration, visualization and analysis, while additional features specific to particular biological questions are available as apps (called plugins for 2.x). Due to the major reorganization in the 3.x series, all 2.x plugins needed to be ported to apps by their developers.

The modular architecture of Cytoscape 3.x is based on a Open Service Gateway Initiative (OSGi) model, where each subset of functionality is represented by a separate API and implementation JAR files. In this way, the implementation may be switched without the need to change the interface to the service defined in the API. New apps can be implemented as bundle apps that can access the core Cytoscape functionality by utilizing the OSGi interfaces provided in the API.

Furthermore, Cytoscape 3.x provides a built-in and easy-to-use command line functionality. In this way, many features, both from the Cytoscape core and released apps, can be exposed to the user as commands or can be invoked by other apps. For each command, this is accomplished by creating a Cytoscape `TaskFactory` with two properties, the name of the command and a command namespace for a group of related commands such as `network` or `rinalyzer`. Each task factory has a `createTaskIterator()` method that is executed by one of the `TaskManager`s and invokes a `Task` to perform the needed operations. In addition, each task can have several `Tunables`, which are the command arguments that define the input needed by the user. An example for a command that creates a new network from a set of selected nodes in the current network is:

```
network create networkName="New network" source="current"
  nodeList="selected"
```

This command can also be executed by an app as demonstrated in Listing 4.1. For some arguments, such as the `source` network and the `nodeList`, there are special keywords, e.g. `current` and `selected`, respectively.

The new versions of structureViz and RINalyzer are implemented as bundle apps and utilize the Cytoscape 3.1.0 API. Most key features of the two apps are exposed as commands and can be called by other apps. This design also ensures the smooth interplay between the two apps as well as their communication with Cytoscape and UCSF Chimera. More details are given in the next sections.

```
// get the service registrar in the CyActivator class
CyServiceRegistrar registrar = getService(
    bundleContext, CyServiceRegistrar.class);

// get one of the task managers
SynchronousTaskManager tm = registrar.getService(
    SynchronousTaskManager.class);
// get the factory that creates a CommandExecutorTask
CommandExecutorTaskFactory cetf = registrar.getService(
    CommandExecutorTaskFactory.class);

// create a map of the arguments and their values
Map<String, Object> argMap = new HashMap<String, Object>();
argMap.put("networkName", "New network");
argMap.put("source","current");
argMap.put("nodeList","selected");

// assumes that this does not implements TaskObserver
if (cetf != null)
    tm.execute(cetf.createTaskIterator("network",
            "create", argMap, null), null);
```

Listing 4.1: Execute command task

4.3.2 The app structureViz

The app structureViz was initially released in 2007 as a Cytoscape 2.x plugin that links the visualization of biological networks in Cytoscape with visualization and analysis of protein structures in UCSF Chimera (Morris *et al.*, 2007). Together with the app developer John H. Morris, we released structureViz2 for Cytoscape 3.x in 2014. It was redesigned and extended with new functionality at the same time. Keeping the focus of structureViz on visualization, we improved the interactive interface with UCSF Chimera and additionally enabled the linking between residues in the 3D protein structure and nodes in a RIN. Among others, the synchronization of selection on all levels as well as the automatic association of networks, nodes or edges with structures in UCSF Chimera were newly implemented. Furthermore, we extended structureViz to the interactive generation and annotation of RINs from a residue selection in UCSF Chimera as described in more detail below.

Structure annotations

Annotating a network with structures entails the creation of new node attributes and the population of those attributes with structure identifiers. Once a structure is opened in UCSF Chimera, structureViz2 automatically associates it with all nodes that are annotated with that structure. The type of each attribute can be String (if multiple identifiers, a comma-separated list) or List (with each identifier given as a single string). Here, we will only mention the general type of identifiers.

A whole protein structure or a subset of it, such as one or more chains or residues, can be associated with a node as an attribute named `Structure, structure, pdb, pdbFileName, PDB ID,` or `biopax.xref.PDB`. The specification is of the form

```
modelName[.modelNumber]#[residueID][.chainID]
```

The `modelName` is either the 4-character PDB ID of the structure, or a path to a local file enclosed by quotation marks, or an URL enclosed by quotation marks. The `modelNumber` is the model number and only needs to be specified for PDB structures or files containing several different models, such as NMR structures. The `chainID` is a character used in the structure file to group residues by chain and should be included in the identifier for structures with more than one chain. The `residueID` may be a single-letter code and residue number such as `H263`, a three-letter code and number such as `His263`, or simply a residue number, such as `263`. The specifications can be given individually or as a comma-separated list. If no PDB identifier is given, the specification is assumed to apply to all currently open structures. Such annotations are usually created automatically for RINs, but can also be added manually. Here are some examples for residue nodes:

- `1hiv#25.A` ⇒ residue 25 in chain A of the structure with PDB identifier 1HIV.
- `"pdb1hiv_h.ent"#25.A` ⇒ residue 25 in chain A of the structure contained in file *pdb1hiv_h.ent*.
- `1abc.0#1.A` ⇒ residue 1 in chain A in model 0 of the structure with PDB identifier 1ABC.

A node can also be associated with a smiles structure if it has a node attribute called `Smiles`, `smiles`, or `SMILES` containing the corresponding smiles string. SMILES is a specification for representing chemical structures in a computer-readable form, e.g., only using ASCII characters (Anderson *et al.*, 1987; Weininger, 1988; Weininger *et al.*, 1989). This feature mainly provides a convenient way to load small molecule structures into UCSF Chimera, but can also be used to display the structure of individual amino-acid residues on their respective nodes.

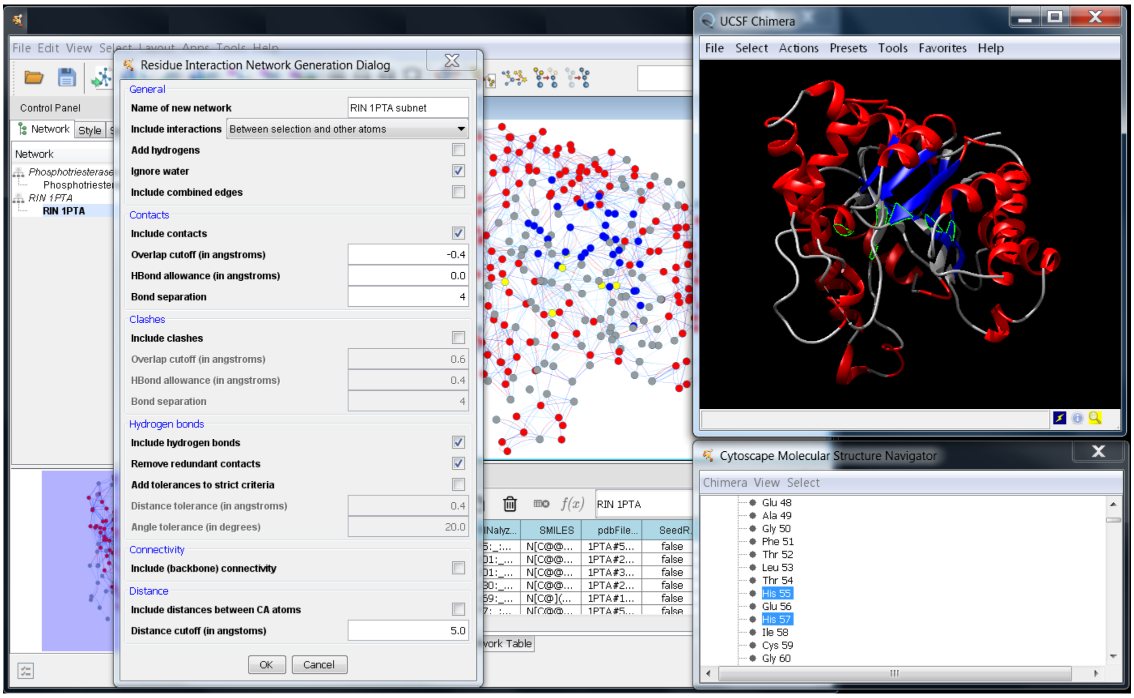
For the RINalyzer tasks, a node attribute in the following format is needed:

```
modelName:chainID:residueIndex:insertionCode:residueType
```

whereas missing values are substituted by an underscore according to the RIN specifications (`http://rinalyzer.de/docu/rins_spec.php`). This attribute should be named either `name` or `RINalyzerResidue`. These attributes are automatically created when generating a RIN from UCSF Chimera or importing a RIN using one of the available RINalyzer menus.

RIN generation

The generation of a RIN by structureViz can be initiated for any selection of residues in UCSF Chimera. The selection may include amino-acid residues, solvent



Residue Interaction Network Generation Dialog

General

Name of new network: RIN 1PTA subnet

Include interactions: Between selection and other atoms

Add hydrogens ☐

Ignore water ☑

Include combined edges ☐

Contacts

Include contacts ☑

Overlap cutoff (in angstroms): -0.4

HBond allowance (in angstroms): 0.0

Bond separation: 4

Clashes

Include clashes ☐

Overlap cutoff (in angstroms): 0.6

HBond allowance (in angstroms): 0.4

Bond separation: 4

Hydrogen bonds

Include hydrogen bonds ☑

Remove redundant contacts ☑

Add tolerances to strict criteria ☐

Distance tolerance (in angstroms): 0.4

Angle tolerance (in degrees): 20.0

Connectivity

Include (backbone) connectivity ☐

Distance

Include distances between CA atoms ☐

Distance cutoff (in angstoms): 5.0

OK Cancel

Figure 4.5: Generation of residue interaction networks with UCSF Chimera. In this image, the structure with PDB identifier 1PTA has been loaded into UCSF Chimera and used to create a RIN of all selected protein residues and user-specified interactions (`Apps` → `structureViz` → `Create Residue Network`). The residues are represented as nodes colored according to their secondary structure, and the colors were synchronized with the 3D structure (`Apps` → `structureViz` → `Synchronize Residue Colors`). The contacts between residues are shown as blue edges in the network view, while the hydrogen bonds are in red. A subset of nodes was selected in the structure (as well as in the other views automatically) and a new RIN is about to be generated based on the options set in the `Residue Interaction Network Generation Dialog`.

molecules, ligands, etc., and each entity is represented as a node in the resulting RIN in Cytoscape. The network edges correspond to non-covalent interactions between these entities such as van-der-Waals contacts or hydrogen bonds. Since in some cases the user may be interested in seeing the covalent peptide backbone interactions as edges in the RIN, we also retrieve these from the currently loaded structure in UCSF Chimera. As can be seen from the dialog in Figure 4.5, five types of edges are supported: contacts, clashes, hydrogen bonds, connectivity (backbone), and C_α distances. The parameters for each interaction type are initially set to the default values provided by UCSF Chimera (see the arguments of the `createRIN` command described in the next section). There are different interaction subtypes depending on whether an interaction occurs between the atoms in the main chain (mc), side chain (sc), water, etc., and the number of such interactions is stored in the edge attribute `NumberInteractions`. For each interaction, the interacting atoms identifiers as well as the distance/overlap are also included as attributes. For contact edges, the distance/overlap attribute equals to the minimum distance between the closest atoms; for hydrogen bonds, it is the distance between the H donor and ac-

```java
// get the registrar in the CyActivator class
CyServiceRegistrar registrar = getService(
    bundleContext, CyServiceRegistrar.class);

// get all needed managers
SynchronousTaskManager tm = registrar.getService(
    SynchronousTaskManager.class);
CyLayoutAlgorithmManager lm = registrar.getService(
    CyLayoutAlgorithmManager.class);

// get the RIN layout
CyLayoutAlgorithm rinlayout = lm.
    getLayout("rin-layout");
// execute the layout on all node views in
// the current network view netView
if (rinlayout != null)
  tm.execute(rinlayout.createTaskIterator(
    netView, rinlayout.getDefaultLayoutContext(),
    CyLayoutAlgorithm.ALL_NODE_VIEWS, null));
```

Listing 4.2: Apply RIN layout task

ceptor; for distance edges, it means the distance between the atoms; and for clashes, it corresponds to the maximum overlap between the atoms.

After a RIN is generated, all residue attributes available in UCSF Chimera are automatically transferred as node attributes in Cytoscape. Usually, they include secondary structure, hydrophobicity, residue coordinates, backbone and side chain angles, average B-factor, average occupancy and others. In addition, if the RINalyzer app is also installed, the RIN Layout considering the 3D coordinates of the residues as well as the default RINalyzer visual style are applied to the network view (see Figure 4.5). Listing 4.2 shows how the RIN layout can be invoked by structureViz.

The annotation of residue nodes with structural data from UCSF Chimera can also be performed on RINs that were generated by RINerator or another tool using the menu `Annotate Residue Network` as long as they have the correct node attributes for association with the structures in UCSF Chimera (see details above). Furthermore, the colors of the network nodes and structure residues can be synchronized using the menu `Synchronize Residue Colors`. The user decides on the direction of color transfer, whether from the current Cytoscape network to the associated models in UCSF Chimera or the other way around.

Commands

Finally, structureViz2 exports a number of commands. In general, each command includes several arguments, which are described in more detail on the structureViz website (Morris, 2015). The arguments and their values are specified as name-value pairs separated by an equals sign (=). For example, to send a command to UCSF

Chimera, the user might enter:

```
structureViz send command="select #0"
```

Note that the text arguments are placed within quotes. To annotate the current RIN with coordinates and secondary structure information from UCSF Chimera, the following command should be used:

```
structureViz annotateRIN network=current residueAttributes=
[SecondaryStructure, Coordinates]
```

Here, we will only list the commands related to RINs.

- `structureViz annotateRIN`: Annotate a residue interaction network (RIN) with the attributes of the corresponding residues in UCSF Chimera.
- `structureViz createRIN`: Create a residue interaction network from the current selection in UCSF Chimera. The different arguments are shown in the dialog for creating RINs (Figure 4.5).
- `structureViz syncColors`: Synchronize colors between residues and network nodes.
- `structureViz send`: Send command to UCSF Chimera.

4.3.3 The app RINalyzer

With the release of Cytoscape 3.x and the need to port the RINalyzer plugin to an app, we saw a great opportunity to extend it with new functionality. For this purpose, we surveyed the users of RINalyzer as well as our close collaborators in order to collect feedback on frequently used or still required features as well as suggestions for new functionality. In the process of porting the RINalyzer plugin to a Cytoscape 3.x app, we closely collaborated with John H. Morris, the developer of structureViz. Together, we redesigned both plugins to divide the functionality into meaningful units. We kept the focus of structureViz on the interface with UCSF Chimera and of RINalyzer on the visual analytics functionality for RINs. Besides considering well-known software design principles such as modularity, compatibility and exchangeability, we also cared much about the usability of our tools. One of our main goals was to provide as much as possible to the user in one single click, e.g. an appropriate and understandable visualization, but, at the same time, to make things customizable.

The new architecture of Cytoscape 3.x made it possible to easily use the functionality of structureViz from within RINalyzer. Thus, some of the RINalyzer menus, in particular, related to the interface with UCSF Chimera, actually call methods implemented by structureViz. Listing 4.3 shows two such examples.

```java
// get the registrar in the CyActivator class
CyServiceRegistrar registrar = getService(
  bundleContext, CyServiceRegistrar.class);

// get the task factory for the createRIN task
TaskFactory crtf = registrar.getService(
  TaskFactory.class,
  "(&(commandNamespace=structureViz)(command=createRIN))");
// execute the task for creating a RIN
// this call will invoke the create RIN dialog
if (crtf != null)
  insertTasksAfterCurrentTask(
    crtf.createTaskIterator());

// get the task factory for the annotateRIN task
NetworkTaskFactory atf = registrar.getService(
  NetworkTaskFactory.class,
  "(&(commandNamespace=structureViz)(command=annotateRIN))");
// execute the task for annotating a RIN
if (atf != null)
  insertTasksAfterCurrentTask(
    atf.createTaskIterator(network));
```

Listing 4.3: Invoke a structureViz task

New features and improvements

Interface to UCSF Chimera. One of the key features of RINalyzer is its ability to start UCSF Chimera and match the RIN shown in Cytoscape with its molecular structure viewed in UCSF Chimera. Thereupon, the selection of residues in UCSF Chimera leads to the selection of the corresponding nodes in the RIN view and vice versa. The menus in the new RINalyzer version are organized slightly differently than the previous versions. Furthermore, the new interface to UCSF Chimera is implemented in the structureViz2 app and basically runs in the background. structureViz keeps track of open structures and automatically associates them with the corresponding networks, nodes and edges in Cytoscape only if the nodes are correctly annotated (see details in Section 4.3.2). The required attributes are generated automatically for each RIN when importing a RIN from the RINdata web service or from a file as well as when creating a new RIN from UCSF Chimera.

Import and generation of RINs. There are several new ways for importing and generating RINs in Cytoscape. The `Import RIN from Web Service` feature allows the direct retrieval of RINs from our web service RINdata. It automatically imports the RIN with its associated attributes and opens the corresponding PDB structure in UCSF Chimera. The `Import RIN from File` option can be used to import any RIN, which is supported so far by RINalyzer or follows the RIN specifications. In this way, the attribute data required for associating the RIN with a protein structure in UCSF Chimera is generated automatically.

Last but not least, the completely new functionality to generate RINs from a selection of residues in UCSF Chimera is implemented in the structureViz app and can be invoked from RINalyzer as well. The selection can include amino-acid residues, solvent molecules, ligands, etc. Currently, five types of edges can be created: contacts, clashes, hydrogen bonds, connectivity (backbone), and C_α distances. As mentioned previously, the RINalyzer and structureViz apps can also transfer residue attributes from UCSF Chimera as node attributes to the corresponding RIN in Cytoscape. In particular, these attributes include secondary structure, residue coordinates, hydrophobicity, solvent accessible surface area (if already computed in UCSF Chimera), occupancy, etc.

RIN Layout. The RIN layout is specifically implemented for RINs and synchronizes orientation and location between a RIN and the corresponding structure. Our collaborator Karsten Klein developed a new stress-based layout method that minimizes the weighted mean square error between predefined distances for pairs of residues and the geometric distance in the layout. The layout is initialized using a projection of the 3D residue coordinates on a 2D plane. The stress is computed as a balanced combination of two factors, the flexible representation of the residue network and the user's spatial orientation using the fixed projection coordinates. The priority for the latter is increased over the course of the optimization. In order to emphasize the secondary structure, the distance error weights are larger for distances between residues within the same secondary structure element. Alternatively, the layout method can prioritize certain distances based on user-defined edge weights that represent additional structural or functional information.

Exploration of RINs. A few new features for exploring RINs have been implemented. In addition to creating a new network for a single or multiple chains in a RIN, RINalyzer can extract a new subnetwork consisting of the interface residues and their interactions between two or more chains. Interface residues are defined as residues with at least one non-covalent interaction to a residue in another chain. This functionality allows a more detailed analysis of protein binding and the effects of mutations on the function of the protein as shown in our case study on the BioVis 2013 Data Analysis contest. The Edge Distance Filter option is included in the RIN Visual Properties dialog and allows hiding edges between residues closer in sequence than a user-specified threshold.

Aggregated RINs. RINalyzer supports the generation of aggregated RINs based on a node attribute selected by the user. In the resulting network, each node represents a group of consecutive residues with the same characteristic (attribute value), for example, the same protein chain, domain or secondary structure element. The node width is proportional to the number of residues in this group and the node tooltip shows the group characteristic (attribute value) and the included residues. The non-covalent interactions between residues in one group and residues in another group are represented by one solid edge line. The interactions between residues in two consecutive groups are shown as a dashed edge line. The edge width is proportional to the number of interactions.

Comparison of RINs. The comparison functionality of RINalyzer has been greatly improved and linked to the structure alignment tool of UCSF Chimera. Thus, users can compare two RINs using the sequence-based structural alignment of the corresponding proteins in UCSF Chimera. The resulting comparison network highlights residue interactions present in either of the structures and allows backtracking these to the original structures. Such a network representation is instrumental for identifying changes in interaction patterns upon mutation as well as between different states of the same protein.

Alternatively to the structure alignment, the mapping of residue nodes can be provided as a FASTA alignment file or a simple node-to-node text mapping file. In the comparison RIN, the three different types of edges are shown as different visual cues: non-covalent residue interactions preserved in both structures are indicated by solid lines, and interactions present only in one of the structures are presented by dashed or dotted lines. In this combined network, there are also three types of nodes: nodes belonging to both networks and representing successfully aligned residues (black node border), as well as nodes that cannot be aligned, i.e., are contained only in one of the two networks/structures (green or red). The type of each node and edge is stored as an attribute called `BelongsTo`, and can have one of the three values: `net1`, `net2`, and `net1,net2` (previously known as `both`). `net1` always refers to the first (reference) network selected in the comparison, i.e., the first structure used in the alignment. Several additional node attributes are created for the comparison network. Among others, they include the fraction of adjacent edges belonging to either `net1` (`EdgeFracNet1`) or `net2` (`EdgeFracNet2`) and the fraction of adjacent edges belonging to `net1,net2` (`EdgeFracBoth`). If the amino acid type of the two aligned residues is not the same, a short identifier for the amino acid substitution (`Substitution`) as well as the amino acid substitution score from the BLOSUM62 matrix (`Blosum62SubstScore`) are also created.

Commands

Most of the RINalyzer functionality is not exposed as commands that can be used without the Cytoscape graphical user interface. This is mostly due to the design of RINalyzer as an app that interacts a lot with the user. However, RINalyzer uses most of the commands previously described for structureViz. Here, we list the two currently available commands that do not require a GUI mode and additional input from the user:

- `rinalyzer importRIN`: Import a RIN from a file, such as those generated by RINerator.
- `rinalyzer createAggregatedRIN`: Create an aggregated RIN.

setsApp

setsApp is a simple app for maintaining and manipulating sets of nodes and edges (Morris *et al.*, 2015b). It was inspired by the RINalyzer's node sets functionality. setsApp was implemented by Allan Wu under my joint supervision with John H. Morris during my stay at UCSF. The most recent version was greatly improved by Samad Lotia from the Gladstone Institutes.

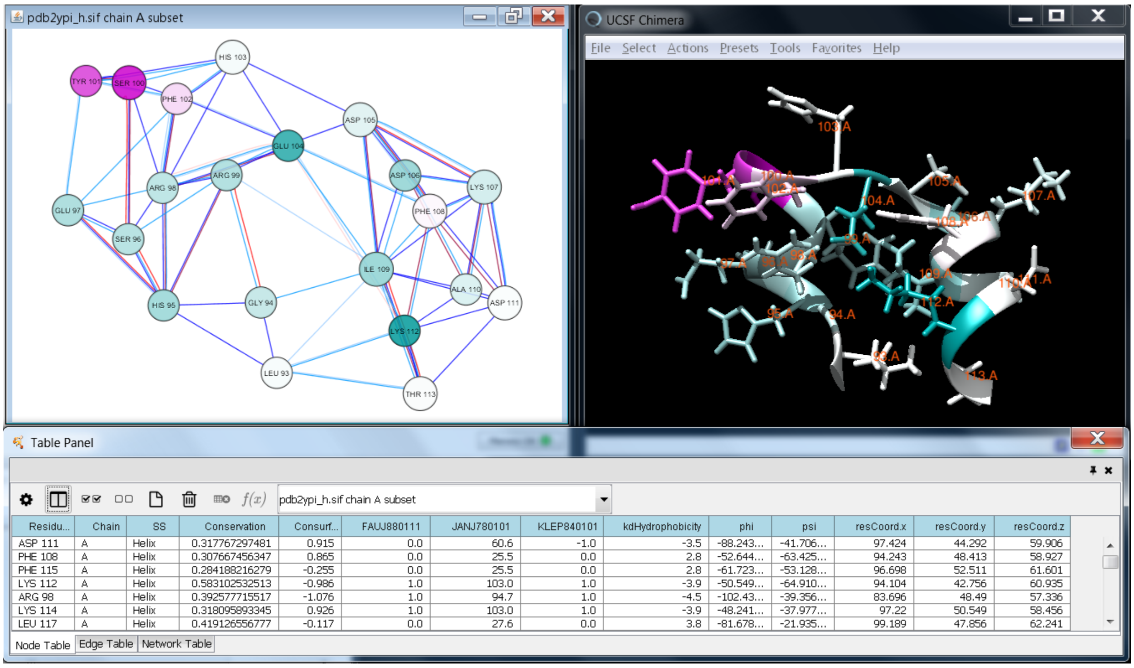
Basically, setsApp allows the user to create a set of nodes or edges from the current selections in the Cytoscape network view as well as to import a set from a file. In addition, nodes or edges with the same discrete attribute values can be grouped into separate sets, e.g., based on residue chain, secondary structure annotation, or interaction type. The user can perform standard set operations on the existing sets, such as union, intersection, and difference. In order to facilitate usage of the app's functionality, setsApp exports a number of commands with the namespace `setsApp` including `createSet`, `remove`, `import`, `export`, `addTo`, `removeFrom`, `rename`, `union`, `difference`, `intersect`. The commands and their arguments are described in more detail in Morris *et al.* (2015b).

4.3.4 UCSF Chimera

UCSF Chimera is a state-of-the art software tool for interactive visualization and analysis of molecular structures and related data. It is constantly updated and further developed by the RBVI team at UCSF. For the initial version of structureViz, the *ReadStdin* interface was implemented such that UCSF Chimera commands can be entered through standard input (`stdin`) and messages in response are sent to standard output (`stdout`). Beside the standard UCSF Chimera commands available to the users, a set of special commands was added. They include the commands `listen start models | selection`, which act as active listeners for changes in the current models or the residue selection, and several commands for listing models, chains, residues, atoms, or selected entities and their attributes. The first version of RINalyzer also made use of the *ReadStdin* interface to interact with UCSF Chimera.

For the new versions of structureViz and RINalyzer, a few additional commands were added upon our request. These are available since release 1.8 of UCSF Chimera. In particular, one of the new commands, `list distmat`, retrieves all distances between a set of atoms or residues and is used for the construction of a distance RIN, in which an edge exists between two residues if the distance between their C_α atoms is less than a user-specified cut-off. For adding backbone edges to RINs, the command `physicalchains` was included to report the starting and ending sequence numbers for physically connected chains. The command `list resattr` lists all available residue attributes for the current models and is employed for annotating RINs with structural data from UCSF Chimera.

Finally, the UCSF Chimera team also provided us with a RESTServer that allows the execution of UCSF Chimera commands through a REST (REpresentational State Transfer) interface. In this way, the communication between UCSF Chimera



Residu...	Chain	SS	Conservation	Consurf...	FAUJ880111	JANJ780101	KLEP840101	kdHydrophobicity	phi	psi	resCoord.x	resCoord.y	resCoord.z
ASP 111	A	Helix	0.317767297481	0.915	0.0	60.6	-1.0	-3.5	-88.243...	-41.706...	97.424	44.292	59.906
PHE 108	A	Helix	0.307667456347	0.865	0.0	25.5	0.0	2.8	-52.644...	-63.425...	94.243	48.413	58.927
PHE 115	A	Helix	0.284188216279	-0.255	0.0	25.5	0.0	2.8	-61.723...	-53.128...	96.698	52.511	61.601
LYS 112	A	Helix	0.583102532513	-0.986	1.0	103.0	1.0	-3.9	-50.549...	-64.910...	94.104	42.756	60.935
ARG 98	A	Helix	0.392577715517	-1.076	1.0	94.7	1.0	-4.5	-102.43...	-39.356...	83.696	48.49	57.336
LYS 114	A	Helix	0.318095893345	0.926	1.0	103.0	1.0	-3.9	-48.241...	-37.977...	97.22	50.549	58.456
LEU 117	A	Helix	0.419126556777	-0.117	0.0	27.6	0.0	3.8	-81.678...	-21.935...	99.189	47.856	62.241

Figure 4.6: Network view in Cytoscape of a RIN (left) generated for a selection of residues from the 3D structure of yeast TIM (PDB identifier 2YPI) in UCSF Chimera (right). The node table at the bottom displays the additional information retrieved by RINerator and structureViz. The network visualization is created using the RINalyzer app and synchronized with the structure using the structureViz app. The network nodes and the corresponding residues are colored according to the sequence conservation score of the TIM family (turquoise-to-pink coloring indicates variable-to-conserved sites). The network edges represent non-covalent residue interactions (blue for contacts and red for hydrogen bonds).

and other tools, such as Cytoscape can be drastically improved. The RESTServer is also designed as a replacement of the ReasStdin interface. The main issue with the latter is that it generates output to `stdout` together with the `listen` command. Thus, in order to parse the UCSF Chimera responses, structureViz needs to demultiplex the output into separate sources. In contrast, the RESTServer uses separate communication channels for notification and command execution. Since version 2 of structureViz still does not make use of the RESTServer, this will be the next main development step.

4.3.5 RINerator

Together with the first version of RINalyzer, we released RINerator, a package for the generation of user-defined RINs from a 3D protein structure. In contrast to previous simplistic interaction definition approaches based on spatial atomic distance between residues, RINerator enables a more realistic representation by considering different biochemical interaction types, such as hydrogen bonds and interatomic contacts, and even quantifying the strength of individual interactions. This is accomplished by performing the following steps:

1. All hydrogens are added to the protein structure by the *Reduce* program Word *et al.* (1999b).
2. The non-covalent interactions are identified using the *Probe* program Word *et al.* (1999a).
3. A residue interaction network is generated and saved in a Cytoscape-compatible file format.

More precisely, Probe identifies the interactions between amino-acid residues in a protein by evaluating their atomic packing. For this purpose, a small virtual probe (typically 0.25 Å) is rolled around the van-der-Waals surface of each atom and where the probe touches another non-covalently bonded atom, a contact dot is detected. The overlaps of van-der-Waals shells between non-polar atoms and hydrogen bonds are indicated by spikes with length l_{sp}. Dots without spikes have a length of 0. Then, the molecular goodness-of-fit of the interactions is measured using the dots and spikes. The hydrogen bonds and van-der-Waals overlaps are quantified by the volume of the overlap:

$$V(Overlap) = \sum_{Overlap\ dot} l_{sp}$$

$$V(HBond) = \sum_{Hbond\ dot} l_{sp}$$

The score of the contacts is measured for each atom pair by summing up over the contact dots. The non-overlapping van-der-Waals contacts are quantified by an error-function weighting:

$$w(gap) = \exp\left(-\left[\frac{gap}{err}\right]^2\right)$$

where the gap is the distance from the dot to the surface of the other atom and the error is equal to the probe radius (0.25 Å). This function assigns a higher score to close contacts than to distant or significantly overlapping ones, but small overlaps are still favorable.

The combined score is calculated using the formula:

$$combined\ score = \sum_{dots} [w(gap) + 4 \times V(Hbond) - 10 \times V(Overlap)]$$

Probe summarizes the scores for all atoms or residues in the structure in an output file. In the last step of the network generation method, an undirected weighted network with multiple edges that represent the non-covalent interactions identified by Probe is created.

Recently, we extended RINerator to retrieve biochemical amino acid properties from external resources, such as AAindex and ConSurfDB, and to calculate conservation scores from a user-specified multiple sequence alignment file. The latter was implemented by Olga Voitenko based on a reliable metric for quantifying the conservation of all amino acid positions in a multiple sequence alignment of a given protein as proposed by Valdar (2002). The conservation score is calculated considering the symbol diversity, the stereo-chemical diversity, and the fraction of gaps at each position in the multiple sequence alignment. The resulting conservation values are between 0 for position that is not conserved and 1 for a strictly conserved position. In case the user does not have a multiple sequence alignment at hand, the conservation scores calculated by ConSurfDB can be retrieved. A RIN colored according to conservation is shown in Figure 4.6.

AAindex is a database of numerical indices that represent various physico-chemical and biochemical amino acid properties derived from published literature (Kawashima *et al.*, 2008). The last version (9.1) contains 544 different amino acid indices. In order to summarize this huge amount of related data, Atchley *et al.* (2005) used multivariate statistical analysis to produce a small set of five patterns of amino acid variability that reflect polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. We selected representative indices to be retrieved automatically by RINerator for each of these five categories as follows:

- Factor I reflects several properties at the same time: the covariation in portion of exposed residues versus buried residues, non-bonded energy versus free energy, number of hydrogen bond donors, polarity versus non-polarity, and hydrophobicity versus hydrophilicity.
  - JANJ780101: Average accessible surface area (Janin and Wodak, 1978)
  - GRAR740102: Polarity (Grantham, 1974)
  - JURD980101: Modified Kyte-Doolittle hydrophobicity scale (Juretic *et al.*, 1998)
- Factor II relates to secondary structure.
  - ISOY800101: Normalized relative frequency of alpha-helix (Isogai *et al.*, 1980)
- Factor III represents molecular size or volume with focus on bulkiness, residue volume, average volume of a buried residue, side chain volume, and molecular weight.
  - GRAR740103: Volume (Grantham, 1974)
- Factor IV refers to relative amino acid composition in different proteins.
  - JOND920101: Relative frequency of occurrence (Jones *et al.*, 1992)
- Factor V reflects electrostatic charge, in particular, the isoelectric point and net charge.

  - FAUJ880111: Positive charge (Fauchere *et al.*, 1988)
  - FAUJ880112: Negative charge (Fauchere *et al.*, 1988)
  - KLEP840101: Net charge (Klein *et al.*, 1984)
  - ZIMJ680104: Isoelectric point (Zimmerman *et al.*, 1968)
- Other physico-chemical properties
  - JOND920102: Relative mutability (Jones *et al.*, 1992)

This list can be extended by further indices at any time. For each index, the values are assigned to the residues based on their amino acid type.

Finally, RINerator also computes the number and interaction strength for each interaction type individually. The retrieved and computed data is saved in a tab-delimited format with the RINalyzer node identifiers in the first column and each data entry in a consecutive column.

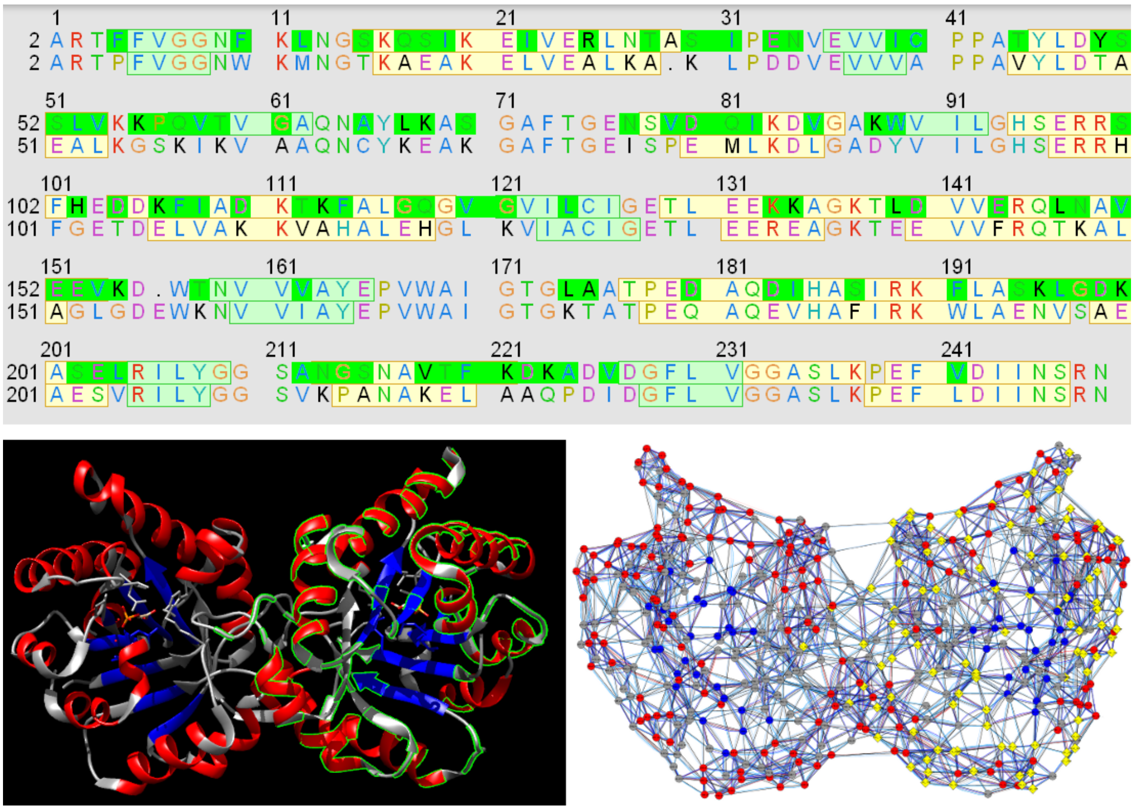
4.4 Contest use case

This section contains an adapted and extended version of the text from Doncheva *et al.* (2014). In particular, the effectiveness of our integrative visual analytics approach is illustrated with the help of a typical use case based on the data provided for the BioVis 2013 Data Analysis contest (Ray *et al.*, 2014). For the specific case in which a functionally defective protein sequence is given together with its parent sequence and structure, we perform a comprehensive assessment of the structural and functional impact of the sequence mutations and highlight the differences between the sequences in complementary views.

The BioVis contest data consisted of the 248-residue long sequence of the functionally defective triosephosphate isomerase mutant (dTIM), the 248-residue long sequence of its yeast parent (scTIM), the 3D structure of scTIM that is most similar to dTIM, full sequences of all known TIMs, and the hand-curated multiple sequence alignment used to create dTIM (Sullivan *et al.*, 2011). TIM is a moderately well conserved homodimer involved in glycolysis and efficient energy production in nearly every organism. Each of the subunits folds into an $\alpha\beta$ barrel, also called TIM barrel, a structural motif characterized by 8 outer α helices and 8 parallel inner β strands. TIM functions as en enzyme catalyzing the reversible interconversion of the dihydroxyacetone phosphate (DHAP) and D-glyceraldehyde 3-phosphate (DAP). The active site is located within the barrel and comprises of residues K12, H95, and E165 (Alber *et al.*, 1981).

4.4.1 Materials

For scTIM, we retrieved the 3D structure with PDB identifier 2YPI (Lolis and Petsko, 1990) from the RCSB Protein Data Bank (Rose *et al.*, 2013) and downloaded

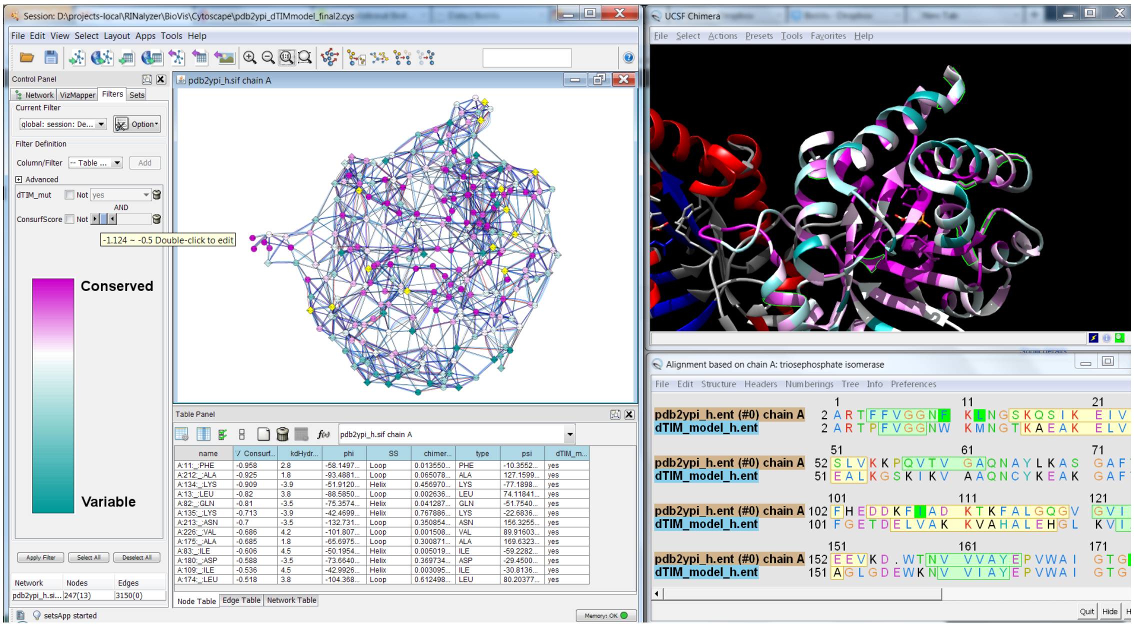


```
    1           11          21           31           41
  2 ARTFFVGGNF  KLNGSKQSIK  EIVERLNTAS    IPENVEVVIC  PPATYLDYS
  2 ARTPFVGGNW  KMNGTKAEAK  ELVEALKA.K    LPDDVEVVVA  PPAVYLDTA

   51           61          71           81           91
 52 SLVKKPQVTV  GAQNAYLKAS  GAFTGENSVD    QIKDVGAKWV  ILGHSERRS
 51 EALKGSKIKV  AAQNCYKEAK  GAFTGEISPE    MLKDLGADYV  ILGHSERRH

  101          111         121          131          141
102 FHEDDKFIAD  KTKFALGQGV  GVILCIGETL    EEKKAGKTLD  VVERQLNAV
101 FGETDELVAK  KVAHALEHGL  KVIACIGETL    EEREAGKTEE  VVFRQTKAL

  151          161         171          181          191
152 EEVKD.WTNV  VVAYEPVWAI  GTGLAATPED    AQDIHASIRK  FLASKLGDK
151 AGLGDEWKNV  VIAYEPVWAI  GTGKTATPEQ    AQEVHAFIRK  WLAENVSAE

  201          211         221          231          241
201 ASELRILYGG  SANGSNAVTF  KDKADVDGFL    VGGASLKPEF  VDIINSRN
201 AESVRILYGG  SVKPANAKEL  AAQPDIDGFL    VGGASLKPEF  LDIINSRN
```

Figure 4.7: Visualization of the sequence mutations in different views. The alignment of the scTIM and dTIM sequences (in this order) is shown in the UCSF Chimera sequence view tool (top) and is used to identify and highlight the differences, e.g. the mutations, by green boundaries in the protein structure of scTIM (bottom left) and by yellow diamonds in the corresponding RIN view (bottom right). Figure first published in Doncheva *et al.* (2014).

the precomputed RIN from the RINdata web service (Doncheva *et al.*, 2011). Since there is no experimentally resolved protein structure of dTIM, we used the SCWRL Server (Canutescu *et al.*, 2003) at BIC-JCSG with default settings and the parent structure as template to generate a three-dimensional model. A RIN for the defective mutant was created from the modeled structure by our RINerator package. In our analysis, we did not consider the sequence alignments. However, the recently released version of RINerator can be used to integrate the conservation information from the alignments into the network representation and the analysis workflow.

External data such as functional annotations, conservation information and structural properties was parsed and imported as attributes in Cytoscape to allow for mapping the data as visual cues on the network and structure views. The UCSF Chimera sequence tool was used to view, align and explore the parent and defective TIM sequences. Based on the sequence alignment, the nodes representing mutated residues were depicted as diamonds instead of circles (Figure 4.7). Especially mutations of residues buried in the structure or close to the functional sites might have a relatively strong impact on protein stability and function. Different node coloring schemes were prepared to map the different types of structural and functional in-



Session: D:\projects-local\RINalyzer\BioVis\Cytoscape\pdb2ypi_dTIMmodel_final2.cys
File Edit View Select Layout Apps Tools Help
Control Panel
Network VizMapper Filters Sets
Current Filter
global: session: De...
Option
Filter Definition
Column/Filter -- Table ...
Add
Advanced
dTIM_mut Not yes
AND
ConsurfScore Not
-1.124 ~ -0.5 Double-click to edit
Conserved
Variable
Apply Filter
Select All
Deselect All
Network Nodes Edges
pdb2ypi_h.si... 247(13) 3150(0)
setsApp started
pdb2ypi_h.sif chain A
Table Panel
pdb2ypi_h.sif chain A
Node Table
Edge Table
Network Table
Memory: OK
UCSF Chimera
File Select Actions Presets Tools Favorites Help
Alignment based on chain A: triosephosphate isomerase
File Edit Structure Headers Numberings Tree Info Preferences
Quit
Hide

name	▽ Consurf...	kdHydr...	phi	SS	chimer...	type	psi	dTIM_m...
A:11:_:PHE	-0.958	2.8	-58.1497...	Loop	0.013550...	PHE	-10.3552...	yes
A:212:_:ALA	-0.925	1.8	-93.4881...	Loop	0.065078...	ALA	127.1599...	yes
A:134:_:LYS	-0.909	-3.9	-51.9120...	Helix	0.456970...	LYS	-77.1898...	yes
A:13:_:LEU	-0.82	3.8	-88.5850...	Loop	0.002636...	LEU	74.11841...	yes
A:82:_:GLN	-0.81	-3.5	-75.3574...	Helix	0.041287...	GLN	-51.7540...	yes
A:135:_:LYS	-0.713	-3.9	-42.4699...	Helix	0.767886...	LYS	-22.6836...	yes
A:213:_:ASN	-0.7	-3.5	-132.731...	Loop	0.350854...	ASN	156.3255...	yes
A:226:_:VAL	-0.686	4.2	-101.807...	Loop	0.001508...	VAL	89.91603...	yes
A:175:_:ALA	-0.685	1.8	-65.6975...	Loop	0.300871...	ALA	169.6323...	yes
A:83:_:ILE	-0.606	4.5	-50.1954...	Helix	0.005019...	ILE	-59.2282...	yes
A:180:_:ASP	-0.588	-3.5	-73.6640...	Helix	0.369734...	ASP	-29.4500...	yes
A:109:_:ILE	-0.536	4.5	-42.9926...	Helix	0.003095...	ILE	-30.8136...	yes
A:174:_:LEU	-0.518	3.8	-104.368...	Loop	0.612498...	LEU	80.20377...	yes

```
                                  1                   11                  21
pdb2ypi_h.ent (#0) chain A    2   A R T F F V G G N F   K L N G S K Q S I K   E I V
dTIM_model_h.ent              2   A R T P F V G G N W   K M N G T K A E A K   E L V

                                  51                  61                  71
pdb2ypi_h.ent (#0) chain A   52   S L V K K P Q V T V   G A Q N A Y L K A S   G A F
dTIM_model_h.ent             51   E A L K G S K I K V   A A Q N C Y K E A K   G A F

                                  101                 111                 121
pdb2ypi_h.ent (#0) chain A  102   F H E D D K F I A D   K T K F A L G Q G V   G V I
dTIM_model_h.ent            101   F G E T D E L V A K   K V A H A L E H G L   K V I

                                  151                 161                 171
pdb2ypi_h.ent (#0) chain A  152   E E V K D . W T N V   V V A Y E P V W A I   G T G
dTIM_model_h.ent            151   A G L G D E W K N V   V I A Y E P V W A I   G T G
```

Figure 4.8: Mapping of conservation information onto the sequence, structure, and network representations. The nodes and residues in the RIN (top left) and chain A of scTIM (top right) are colored according to the conservation scores retrieved from ConSurf-DB (turquoise-to-pink coloring indicates variable-to-conserved sites). The network nodes that represent mutated residues with a high conservation score (F11, L13, Q82, I83, I109, K134, K135, L174, A175, D180, A212, N213, V226) are selected using two filters in Cytoscape (left) and highlighted in the network view by yellow color (top left) and in the other two views by green boundary around the structure (top right) or the amino acid letter (bottom right). Nodes that correspond to mutated residues are depicted as diamonds. Additional data annotated to the residue nodes is shown in the Cytoscape attribute browser as table (bottom left). Figure first published in Doncheva *et al.* (2014).

formation. This allowed us to identify relevant mutations with possible functional effects.

4.4.2 Results

Secondary structure and conservation. In the default secondary structure-colored view (Figure 4.7), we observe that most mutations are located on the surface of the protein, i.e., in helices (51 out of 100) and loops (45 out of 100), rather than in the interior consisting of strands (only 4). The conservation-colored view (Figure 4.8) indicates that residues in the protein exterior tend to be more variable in contrast to the ones in the interior, where the active site of the enzyme is located. Combining these two observations led us to the conclusion that most mutations are located in the variable regions on the surface of the protein. The mutated residues with highest conservation values (below −0.5) are F11, L13, Q82, I83, I109, K134, K135, L174, A175, D180, A212, N213, and V226, and they are good candidates for the functional deficit of the mutant structure.

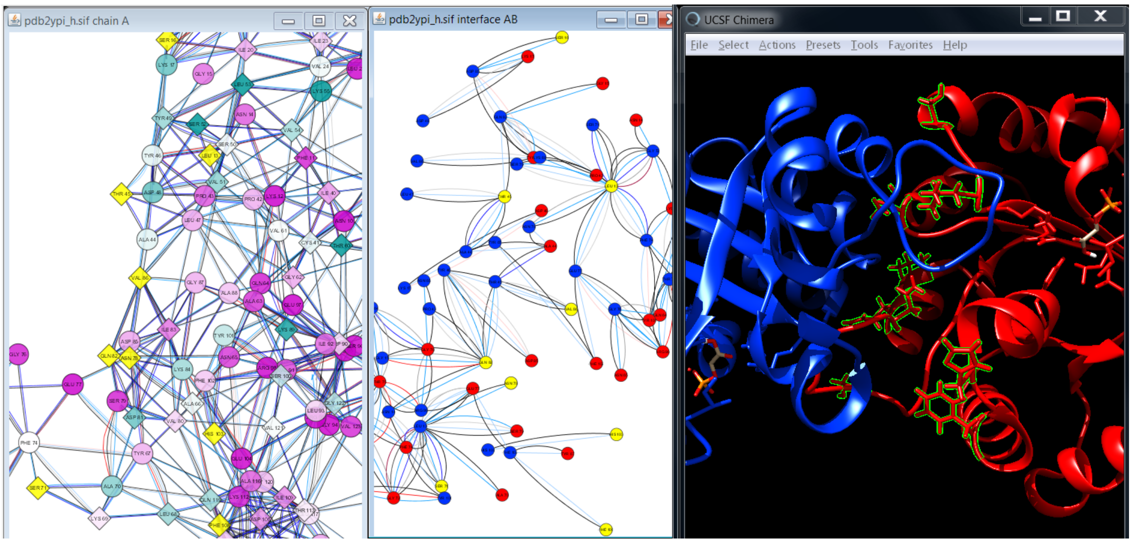

pdb2ypi_h.sif chain A
pdb2ypi_h.sif interface AB
UCSF Chimera
File Select Actions Presets Tools Favorites Help

Figure 4.9: Visualization of the dimer interface with focus on the mutated residues. The combined visualization of the conservation-colored RIN of chain A of scTIM (left), the residue nodes in the interface between chain A (red) and chain B (blue) of scTIM (middle), and the ribbon representation of scTIM are in the same colors as provided by UCSF Chimera (right). Mutations located in the dimer interface (V86, T45, S71, S16, Q82, N78, L13, H103, F108) are highlighted by yellow colored nodes in the network views and by green boundaries and ball-and-stick representations in the structure view. Nodes that correspond to mutated residues are depicted as diamonds. Figure first published in Doncheva *et al.* (2014).

Functional sites. Since scTIM functions as a dimer, another important aspect is the binding interface between the two monomers. We used RINalyzer to extract the residue interactions of the interface and visualize them in a separate network view. As can be seen in Figure 4.9, 9 out of the 69 residues are mutated (L13, S16, T45, S71, N78, Q82, V86, H103, F108). These changes might impair the dimer formation and thus affect the function of scTIM. Residues L13 and Q82 are particularly interesting as they are both conserved and in the dimer interface. A similar analysis can be performed with other functional sites. For instance, we found that none of the residues in the active or substrate binding site (N10, K12, H95, E165) are mutated. However, 24 residues possess direct non-covalent interactions with functionally important residues and thus could have a severe impact on their function if mutated. This is the case for the residues F11, L13, and C41, and this hypothesis is further strengthened by the fact that the first two of them are conserved.

Comparison network. The comparison network view provided further information about the location and nature of the mutations (Figure 4.10). From the overall distribution of red and green edges that indicate changes of non-covalent interactions, it is apparent that many mutations lead to a large number of differences primarily on the protein surface. Additionally, the active site residues form different interactions with their neighbors in the parent compared with the mutant structure.

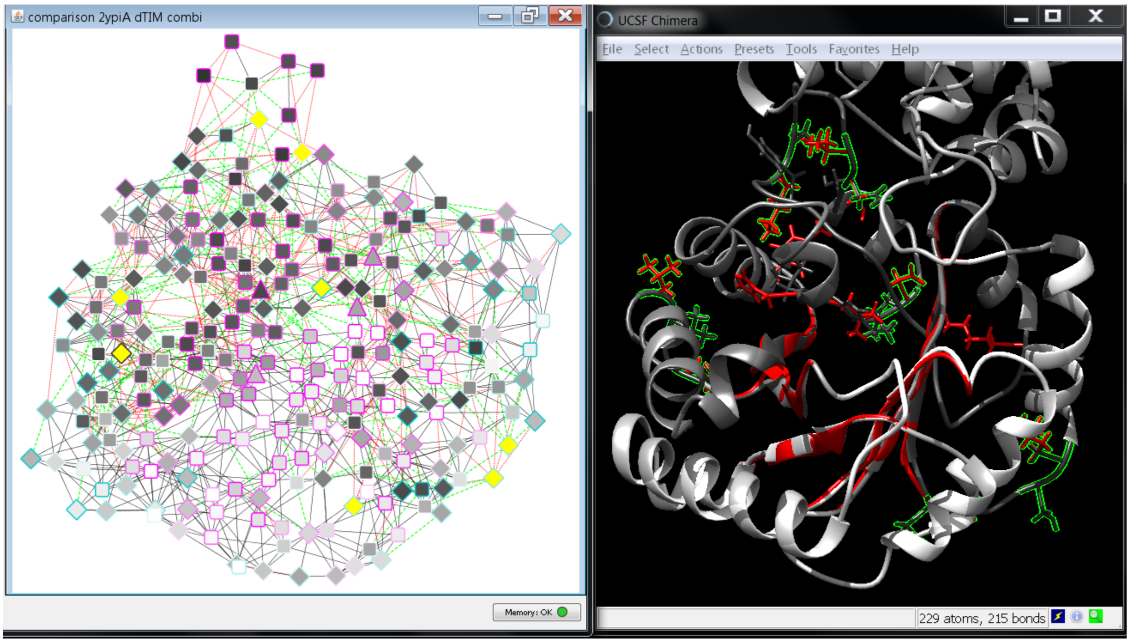

comparison 2ypiA dTIM combi
Memory: OK
UCSF Chimera
File Select Actions Presets Tools Favorites Help
229 atoms, 215 bonds

Figure 4.10: Highlighted mutations with important impact on residue interactions. A comparison network is shown in Cytoscape (left) and a visualization of the aligned structures (scTIM in gray, dTIM in red) in UCSF Chimera (right). In the network view, green dashed edges depict gained, and red dotted edges lost interactions. The network nodes are colored according to the fraction of adjacent interaction edges that do not change upon mutation (from white for all to gray for none), the node border colors represent the conservation score of the respective residue in the parent with turquoise-to-pink coloring for variable-to-conserved sites. Nodes with an amino acid mutation are shown as diamonds. The mutated residues with the largest impact on the residue interactions are highlighted by yellow colored nodes in the network views and by green boundaries and the ball-and-stick representations in the structure view. The mutations correspond to the following residue pairs based on the alignment of scTIM (chain A) and dTIM sequences: (A30, -), (S31, K30), (E34, D33), (N35, D34), (K56, GLY55), (G62, A61), (L68, K67), (S71, K70), (N78, I77), (K89, D88), (V154, L153), (-, E156). Figure first published in Doncheva *et al.* (2014).

Furthermore, there is an insertion (E156 in dTIM) and a deletion (A30 in scTIM) in the dTIM sequence in contrast to the parent sequence according to the sequence alignment in UCSF Chimera. However, they are not close to the active site or the dimer binding interface and thus the functional effect is difficult to judge. Finally, the residue nodes in Figure 4.10 are colored according to the fraction of interactions they gained or lost upon mutation. When combining this information with the conservation scores mapped to the node border colors, particularly interesting mutations can be found. Mutations with the largest change of local residue interactions are highlighted in Figure 4.10 (A30, S31, E34, N35, K56, G62, L68, S71, N78, K89, and V154 in scTIM and E156 in dTIM). Especially the mutated residues S71 and N78 are conspicuous because they are also located in the dimer interface.

Summary. By combining the different views and data in an interactive fashion, it was possible to pinpoint sets of residue mutations as candidates for having a pronounced effect on the enzymatic activity of dTIM. They were selected based on their conservation, distance to the active site, location on the dimer interface, or causing the largest change of local residue interactions. In particular, five residues (F11, F13, S71, N78, Q82) were contained in more than one of these sets. Further experimental validation will be needed to determine which mutations have to be replaced in the mutant by amino acids from the parent to rescue functionality. Other structural properties such as hydrophobicity, solvent accessible surface area or polarity can also be mapped onto the RIN view to characterize mutations with particular properties. Another strategy described in our previous work (Doncheva *et al.*, 2012a) would be the application of network topology analysis of the RIN for the detection of important residues.

4.4.3 Evaluation and related methods

All in all, six entries were submitted to the BioVis 2013 Data Analysis contest (Ray *et al.*, 2014). They were evaluated by six judges divided into two teams based on their domain of expertise. The Team Bio included experts in biology with knowledge of the specific TIM variant, while Team Vis were visual analytics experts. Each team selected one best entry and both teams together voted for the overall best. Our approach was selected as the overall favorite, while the contributions by Silveira *et al.* (2014) and Luciani *et al.* (2014) were the Team Bio and Team Vis favorites, respectively.

Silveira *et al.* developed the web-based tool VERMONT that visualizes the effect of mutations in a family of sequences using multiple-sequence-like views. Each panel focuses on different aspects such as the sequence conservation, the presence of residue contacts, the physicochemical properties of the individual amino-acid residues, or the topological properties of the residues in their contact network. Finally, the authors proposed an automatic method for identifying the damaging mutations using genetic algorithms and suggested one manually and one automatically inferred set of mutations as solution (Silveira *et al.*, 2014).

The best Vis contribution was an open-source tool called FixingTIM that integrates 3D structure and sequence data from distributed sources in one common interface of linked views. The main components are a side-by-side view of the scTIM/dTIM 3D structures and reference information from sources like PDB and a multiple-sequence-like view (trend image) of the TIM family with customized options for coloring and sorting that proof to be very helpful for exploratory analysis. After visual exploration, the authors, who included experienced structural biologists, suggested several residues around the active site as the most promising candidates (Luciani *et al.*, 2014).

The other three entries included two additional software tools and a more theoretical graph-based approach. The web-based tool Mu-8 focuses on visualizing the diversity of physico-chemical residue properties within a protein family and identifies the region between residues 150 and 156 as most notable for affecting the function of dTIM (Mercer *et al.*, 2014). ProfileGrid provides a simplified representation of sequence alignments as heat maps of residue frequencies and therefore, facilitates the visual comparison of a mutated sequence with the protein family (Roca, 2014). Knisley and Knisley (2014) developed an approach that represents 3D structures as hierarchical graphs with three layers (atoms, residues, and substructures) and assesses the effect of mutations based on the change in topological parameters computed for these graphs.

All these approaches have certain variations as well as common components compared to ours. In particular, we combined different views on the sequence and structure level such as FixingTIM and overlaid different physico-chemical properties on top of our visualizations in a qualitative and quantitative manner such as VERMONT, Mu-8, and ProfileGrid. We also provided a graph representation of the 3D structures on two of the three levels presented by Knisley and Knisley. However, in contrast to the other more simple tools, we integrated well-known and sophisticated visualization and analysis tools such as UCSF Chimera and Cytoscape and therefore, facilitated more complex visual exploratory analyses. Furthermore, our approach was closest to the idea of the organizers and data providers to consider multiple mutations together as mutational networks.

The contest organizers concluded in their overview paper (Ray *et al.*, 2014) that all entries presented valuable contributions to the presentation of differences and the assessment of mutations at protein family level. Several entries successfully identified some of the known rescue mutations and pointed to additional regions of interest for the biological domain experts. However, none of the approaches was able to identify the residue dependency networks violated by the dTIM sequence and the different ways to rescue the functionality of dTIM as suggested by Sullivan and colleagues after laborious manual bioinformatics and experimental work (Sullivan *et al.*, 2012). Thus, the submitted tools are just the beginning of future efforts in this field.

4.5 Drug resistance mutations

The hepatitis C virus (HCV) is blood-borne virus that can cause both acute and chronic asymptomatic hepatitis infection. About 130-150 million people worldwide are chronically infected with HCV and approximately 500,000 of them die each year from disease complications (World Health Organization (WHO), 2015; Lozano *et al.*, 2012). Until 2011, the standard combination therapy with peg-interferon plus ribavirin showed sustained virologic response in only 50 % of the genotype 1 infected patients (Manns *et al.*, 2001; Morgan *et al.*, 2010). Thus, several direct-acting antiviral agents (DAA) against HCV have been approved in the last few years and have significantly improved sustained virologic response rates of patients with chronic HCV infection (Ghany *et al.*, 2011; Schneider and Sarrazin, 2014). However, resistance-associated amino acid variants (RAV) to these DAAs have emerged and were shown to play an important role in treatment failure (Welsch, 2014). RAVs against protease inhibitors (PI) often impair the virus replication capacity and adversely affect infectious virus production. Since viral fitness of RAVs is essential for their selection from the virus population under drug pressure, it is very important to understand the molecular mechanisms that cause fitness deficits in NS3 protease RAVs.

Here, we investigated different structural and topological properties of RAVs, in particular, arising in the presence of DAAs against the HCV NS3/4A protease. For this purpose, we explored several topology measures on RINs together with other physico-chemical protein properties to find out what type of information they contain with respect to resistance development. Furthermore, we compared the profiles of resistance mutations to other relevant sets of residues, such as the catalytic site residues. In addition to our visual analytics approach, we also conducted a more quantitative evaluation of the dependencies between RAVs and different structural and topological properties.

4.5.1 Materials and methods

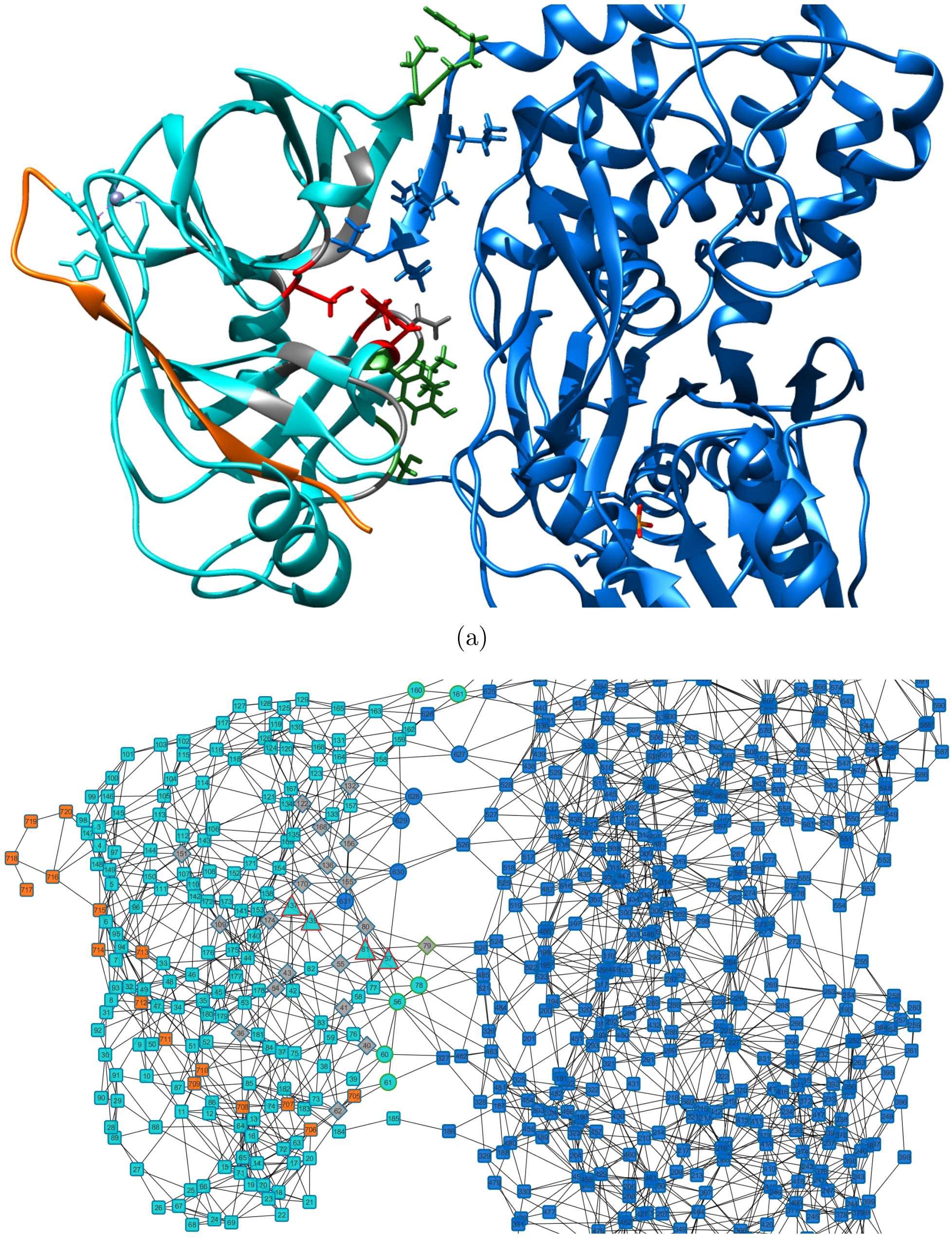
Structure and functional sites

The HCV NS3 protease cleaves the non-structural proteins from the viral polyprotein together with its cofactor NS4A. In addition, it comprises a helicase domain, which is involved in intracellular infectious virus particle assembly independently of the enzymatic activities. Thus, we selected two 3D structures for our analysis. The first structure consists of a protease and a helicase domain (PDB identifier 1CU1) (Yao *et al.*, 1999) and the second contains only the protease domain (PDB identifier 3KF2) (Cummings *et al.*, 2010). Then, we created RINs for protein chain A from the PDB structures 1CU1 and 3KF2 with the RINerator package.

Furthermore, we defined the following functional sites:

- Protease catalytic residues and oxyanion hole (*cat*): residues 57, 81, 137, 139 (Love *et al.*, 1996)
- Three putative domain-domain interaction sites in the protease domain (*ddip*): residues 56, 60, 61 (*ddip1*), 78, 79 (*ddip2*), 160, 161 (*ddip3*)

The *ddip* sites were previously identified by our cooperation partner Christoph Welsch based on the non-covalent interactions of residues in the protease domain with residues in the helicase domain of PDB structure 1CU1. Figure 4.11 shows the interaction interface between the protease and helicase domains with focus on the functional residues, which are highlighted in the structure and network representations. We integrated the functional sites into the computation of topological measures, e.g., we computed the shortest paths from mutated residues to these functional sites. Furthermore, we referred to them as a reference set of function-



(a)

(b)

Figure 4.11: Visualization of (a) the HCV NS3/4A protease structure (protease in cyan, helicase in blue, NS4A in orange) with PDB identifier 1CU1 as ribbon in UCSF Chimera and (b) the corresponding RIN in Cytoscape with focus on the functional and resistance-associated residues. The catalytic residues are represented as red sticks in the 3D structure and as red bordered triangles in the network, while the *ddip* residues are shown as green sticks and green bordered circles. The resistance-associated residues and nodes are colored in gray.

ally interesting residues and compared them to the residues exhibiting resistance mutations.

Resistance mutations

Table 4.1: Resistance mutations in HCV NS3/4A protease and the maximal fold change for *all*, only the *linear* or only the *cyclic* protease inhibitors. (RAV = resistance-associated amino acid variant, FC = EC_{50} fold change in replicon or infectious cell culture from wild type).

		All	Linear	Cyclic			All	Linear	Cyclic
Residue	RAV	Max FC	Max FC	Max FC	Residue	RAV	Max FC	Max FC	Max FC
V36	A	21.58	21.58	3	V151	A	0.9	0.9	-
	G	28.08	28.08	2.3	F154	Y	-	-	-
	L	3	3	2	R155	K	538	150	538
	M	7	7	2.1		T	460	5.2	460
	C	7.8	7.8	1.4		Q	267	4.1	267
	I	0.3	0.3	-		G	580	7.4	580
T40	A	1	-	1		M	30	5.6	30
Q41	R	6.2	1.5	6.2		S	418	4.1	418
	H	3.5	3.5	-		I	26	24	26
F43	S	44	18.78	44		N	39.8	-	39.8
	C	-	-	-	A156	S	18	9.6	18
	L	4	-	4		T	706	62	706
T54	A	12.28	12.28	1.1		V	2041	62	2041
	S	8.22	8.22	1		F	62	62	-
V55	A	1.6	1.6	-		N	93	93	-
	I	3	1.24	3	D168	A	900	1.1	900
R62	K	1	-	1		E	82	82	58
D79	E	1	-	1		H	160	-	160
Q80	R	9.3	1.09	9.3		I	-	-	-
	K	3	-	3		V	1700	1000	1700
	L	1	-	1		G	85	85	55.2
R109	K	3.86	3.86	0.9		N	20	-	20
S122	G	1	-	1		Y	622	-	622
	N	1	-	1		T	205	-	205
	R	3	-	3	I170	A	2.2	2.2	1.4
I132	V	2.4	2.4	-		T	5	4.61	5
K136	R	0.9	0.9	-		V	1	-	1
S138	T	-	-	-	N174	Y	1	-	1
# wt RAVs		12	8	13					
# intrm RAVs		2	2	3					
# notwt RAVs		38	26	28					
# undef RAVs		4	20	12					

By courtesy of Christoph Welsch (University Hospital Frankfurt), we used a comprehensive set of resistance variants associated with two classes of HCV protease inhibitors (PI): linear PIs (telaprevir, boceprevir, unk), and (macro)cyclic PIs (vaniprevir, ciluprevir, danoprevir, simeprevir, asunaprevir). For each variant listed in Table 4.1 and each PI, the 50 % effective concentration (EC_{50}) has been determined as the concentration of the PI required to cause a 50 % reduction in RNA replication of the virus in replicon or infectious cell culture. The fold change (FC) is determined as the ratio of EC_{50} values for the PI in the wild-type and mutant cell culture, whereas the latter always contains the wild-type HCV replicon RNA with a single point mutation corresponding to the resistance-associated variant. This extensive data collection originates from the publications by Welsch *et al.* (2008); Shimakami *et al.* (2011); Welsch *et al.* (2012a,b); Dvory-Sobol *et al.* (2012); Jiang *et al.* (2013); Lawitz *et al.* (2013); McPhee *et al.* (2012, 2013). Table B.1 contains a complete list of the resistance mutations with their respective FC values for each PI as well as the source of the data.

For example, Welsch *et al.* (2008) determined the EC_{50} value for the linear PI telaprevir and the resistance variant V36G in vitro using a wild-type HCV replicon assay and an assay with the respective point mutation. The point mutation was introduced into the wild-type HCV replicon RNA by a specialized site-directed mutagenesis kit. Then, Huh-7.5 cells were transfected with the wild-type and mutant HCV replicon RNA. After growing for 24 h, the cells were incubated with the PI for 48 h. Finally, the level of HCV RNA in the replicon cells was determined and the EC_{50} value was defined as the concentration of PI, at which the HCV RNA level was reduced by 50 % compared to the wild-type concentration.

For simplicity, we considered the largest FC (max FC) among all PIs (*All*) and within each class (*Linear* and *Cyclic*) for each resistant variant. Overall, there are 56 resistance variants with an FC above 0 for at least one of the drugs and they occur in 21 residues (see Table 4.1). From here on, we refer to these residues as resistance-associated residues.

Furthermore, we divided the 56 substitutions into three groups based on the maximal FC for each class of PIs. Residues with low FC (< 2.0) are considered as wild-type (*wt*), with FC between 2.0 and 3.0 as intermediate (*intrm*), and residues with FC > 3.0 as not wild-type (*notwt*). The resulting groups contain different number of residues and resistance variants, whereas the *notwt* group is the largest and the *intrm* group contains only very few RAVs. The exact numbers are given at the bottom of Table 4.1.

Physico-chemical properties

In order to evaluate a comprehensive set of physico-chemical amino acid properties, we considered the 11 previously described indices from AAindex, which are related to polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge (see Section 4.3.5). Evolutionary conservation scores were downloaded from ConSurfDB (Goldenberg *et al.*, 2009). We also added 4 properties using the information computed by UCSF Chimera for each structure, e.g., hydrophobicity, secondary structure, relative solvent-accessible surface area, and solvent-excluded surface area.

Furthermore, we retrieved data from several publicly accessible web servers. The PoPMuSiC web server predicts changes in protein stability as $\Delta\Delta G$ in kcal/mol (per mutation) and also provides information on solvent accessibility (Dehouck *et al.*, 2011, 2009). We included 3 properties based on the PoPMuSiC data: the minimal and maximal predicted change in ΔG for each residue as well as the fractional solvent accessibility. In addition, we retrieved data from the VADAR (Volume, Area, Dihedral Angle Reporter) web server, which provides multiple methods for quantitative assessment of protein structure quality (Willard *et al.*, 2003). This resulted in 3 additional properties: total residue accessible surface area, fractional residue accessible surface area, and excluded residue volume.

Topological parameters

We selected a set of well-known topological measures related to the importance of nodes for the local and global structure of the network. All measures were computed with the NetworkX package (Hagberg *et al.*, 2008). Here is a list with a short description.

- Degree centrality (fraction of nodes connected to a node)
- Average neighbor degree (average degree of all direct neighbors of a node)
- Clustering coefficient (fraction of possible triangles that exist between a node and its neighbors)
- Closeness centrality (inverse average distance to all other nodes)
- Betweenness centrality (sum of the fraction of all-pairs shortest paths
- Average shortest path distance to the 4 functional sites (*cat*, 3 *ddip* sites)
- Shortest path distance to 11 functional residues
- Number of direct interactions (4 interaction types (cnt, hbond, ovl, and combi))
- Strength of direct interactions (4 interaction types)

4.5.2 Results

We performed the analysis for both structures of the NS3/4A protease, e.g., with and without the helicase domain. Since the findings in both cases are very similar, here we will present only the results based on the structure without the helicase domain (see Figure B.4).

Dependencies among all properties

Altogether 56 properties were considered for the analysis of the 11 functional and 21 resistance-associated residues. First, we plotted the pair-wise correlation of all properties (Figure 4.12), in order to investigate their relationships and interdependencies as well as to check for expected and unexpected correlation between them. For example, the same properties from different resources, such as solvent accessibility and volume, should be very similar and we observe that they correlate very strongly. Furthermore, there are several groups of correlated properties that are not so obvious, but can be explained by their derivation. For example, the topological properties degree, betweenness and closeness centrality correlate with each other quite well. This means that nodes, which are more central according to closeness and betweenness and usually represent residues in the protein core, also have more direct neighbors than less central nodes, which correspond to surface residues. For this reason, we also find centrality measures to be strongly anti-correlated with solvent accessibility. We also observe high correlation between AA properties from the same category (factor) described by Atchley *et al.* (2005), e.g., net charge, positive charge, negative charge, and isoelectric point, which can be explained by the way these categories were constructed. There is a strong anti-correlation between the centrality measures and the shortest path distances since the latter are not normalized, but altogether they show the same trend, e.g. residues that are central to the RIN are also close to the catalytic site. Obviously, the shortest path distances to single residues correlate strongly with the average distance to the sites they belong to. Finally, the lack of strong correlation trends between the topological properties and AA properties indicates that these two groups complement each other well and should be considered together for characterizing residues and their effect on protein structure and function.

More interesting correlations are those between volume, degree and number of certain edges. The maximal predicted change in ΔG also correlates with most centrality measures and anti-correlates with the solvent accessible surface area properties. Finally, the lack of unexpectedly strong dependencies between topological and physico-chemical properties in the correlation plot suggests that combining these two different types of independent properties would be more beneficial than just focusing on one of the groups.

Residue properties

For each property, we plotted a histogram and a box-and-whisker plot of the values for all residues in the protein and highlighted the functional or resistance-associated residues, respectively. We aimed at visually identifying properties, for which the functional and/or resistance-associated residues group together and are also located in the histogram apart from the remaining residues, e.g., have extreme values. To quantify this, we also computed a z-score for each property and each functional or resistance-associated residue. The z-score of a raw value x is defined as $z = \frac{x-\mu}{\sigma}$, where μ is the mean and σ is the standard deviation. Overall, the histograms showed some emerging trends, but there were no clear groups of residues with property profiles that distinguish them from the remaining non-functional or non-resistant protein residues.

Both our visual inspection and the z-scores computation revealed that the four groups of functional sites have similar properties to the remaining protease residues. The distribution of values for two properties, conservation and closeness centrality, are shown in Figure 4.13. On average, the functional residues have values close to the

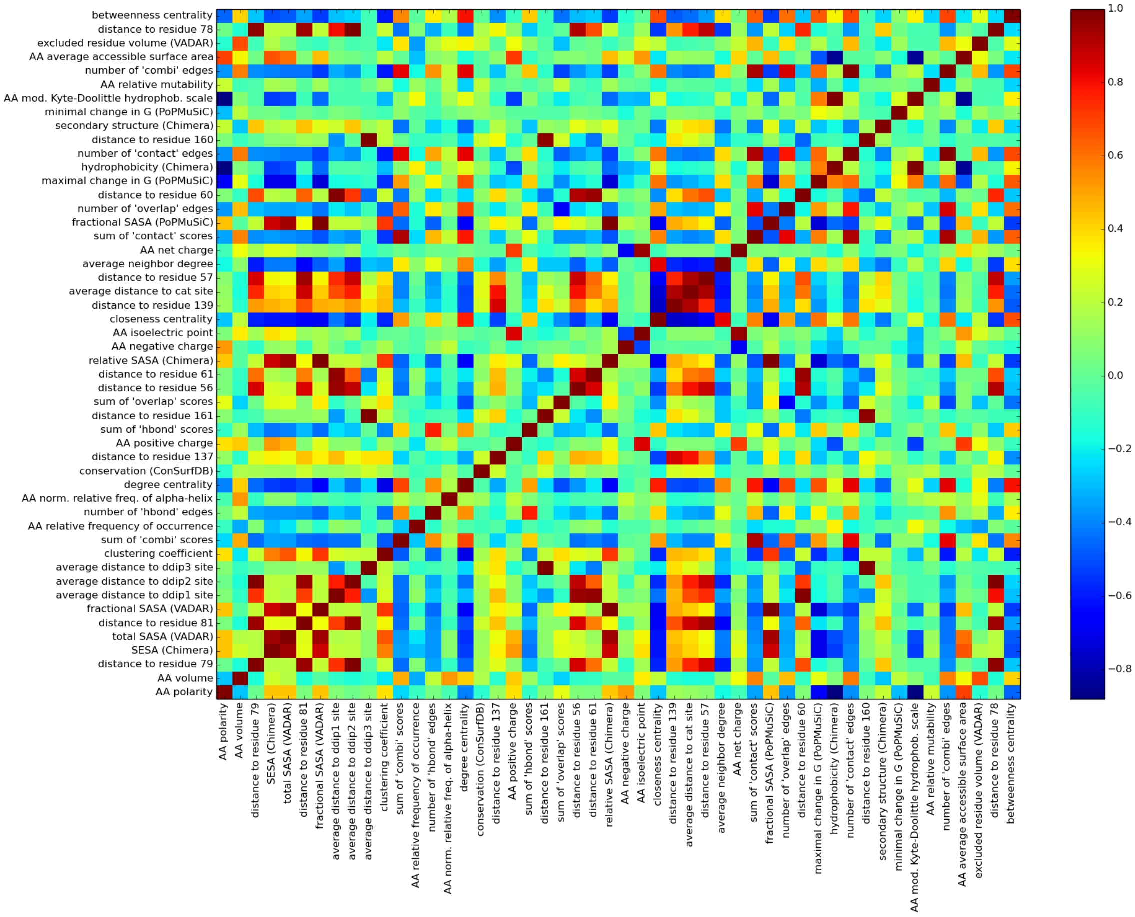
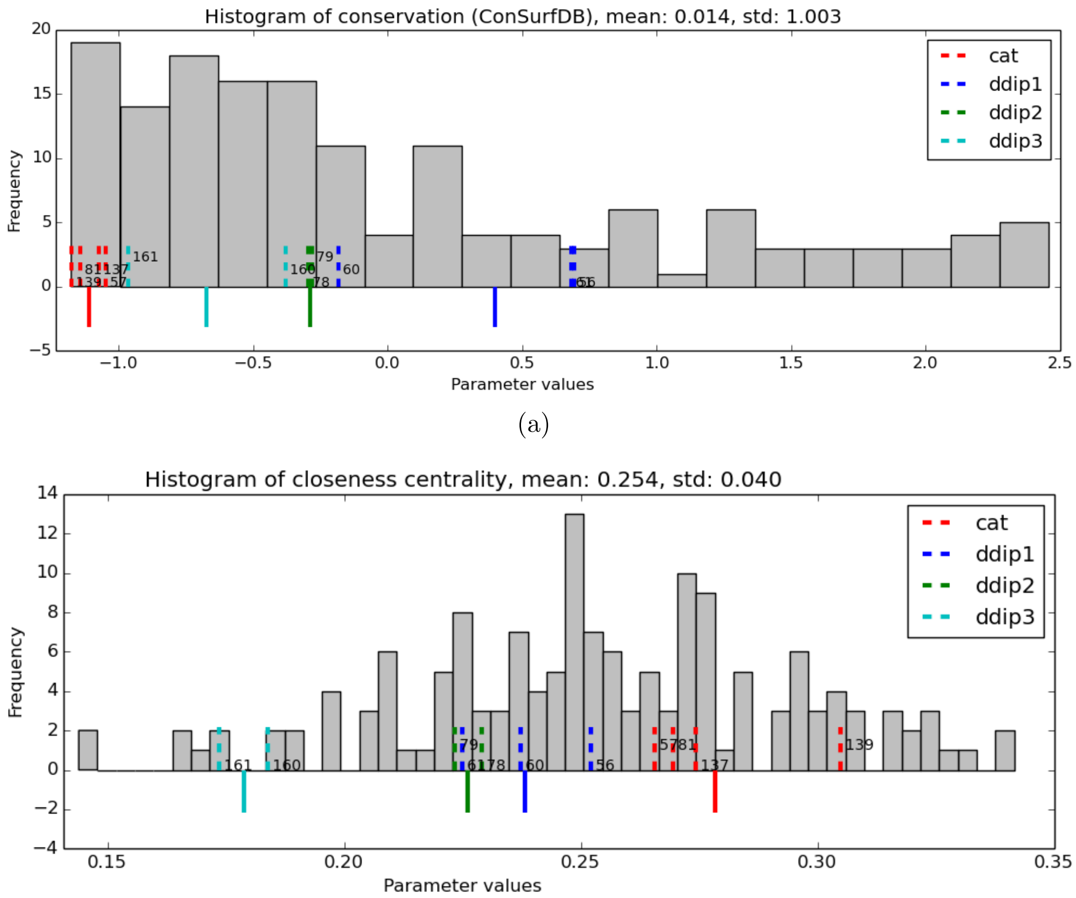


Figure 4.12: Pair-wise correlation of all properties for the structure that contains only the protease domain (PDB identifier 3KF2).

mean value. There are a few exceptions, such as the less conserved *ddip1* residues and the *ddip3* residues, which have low closeness values. However, what makes these two plots more unique in contrast to most other properties is that we can observe a grouping of the residues based on their site. In particular, the catalytic residues are highly conserved and have slightly larger closeness values than the average, while the *ddip* sites have different (lower than the average) closeness values and different levels of low conservation. Furthermore, most functional residues had low z-scores for most properties, e.g., their values were not significantly different from the overall distribution. This was not the case only for the properties related to distance to the functional sites since the catalytic, the *ddip1* and *ddip2* residues are located close to each other in the 3D structure and the RIN.

For the resistance-associated residues, we identified several properties that slightly distinguish them from the remaining protease residues (Figures 4.14, B.5 and B.6). In particular, the resistance-associated residues tend to be located close to the catalytic and *ddip1* and *ddip2* sites as well as to individual functional residues. They also have high betweenness centrality and usually at least 2 `hbond` edges. In addition, they are often uncharged and have either very high or very low hydrophobicity. However, we could not find a particular property that really distinguishes resistance-



(a)

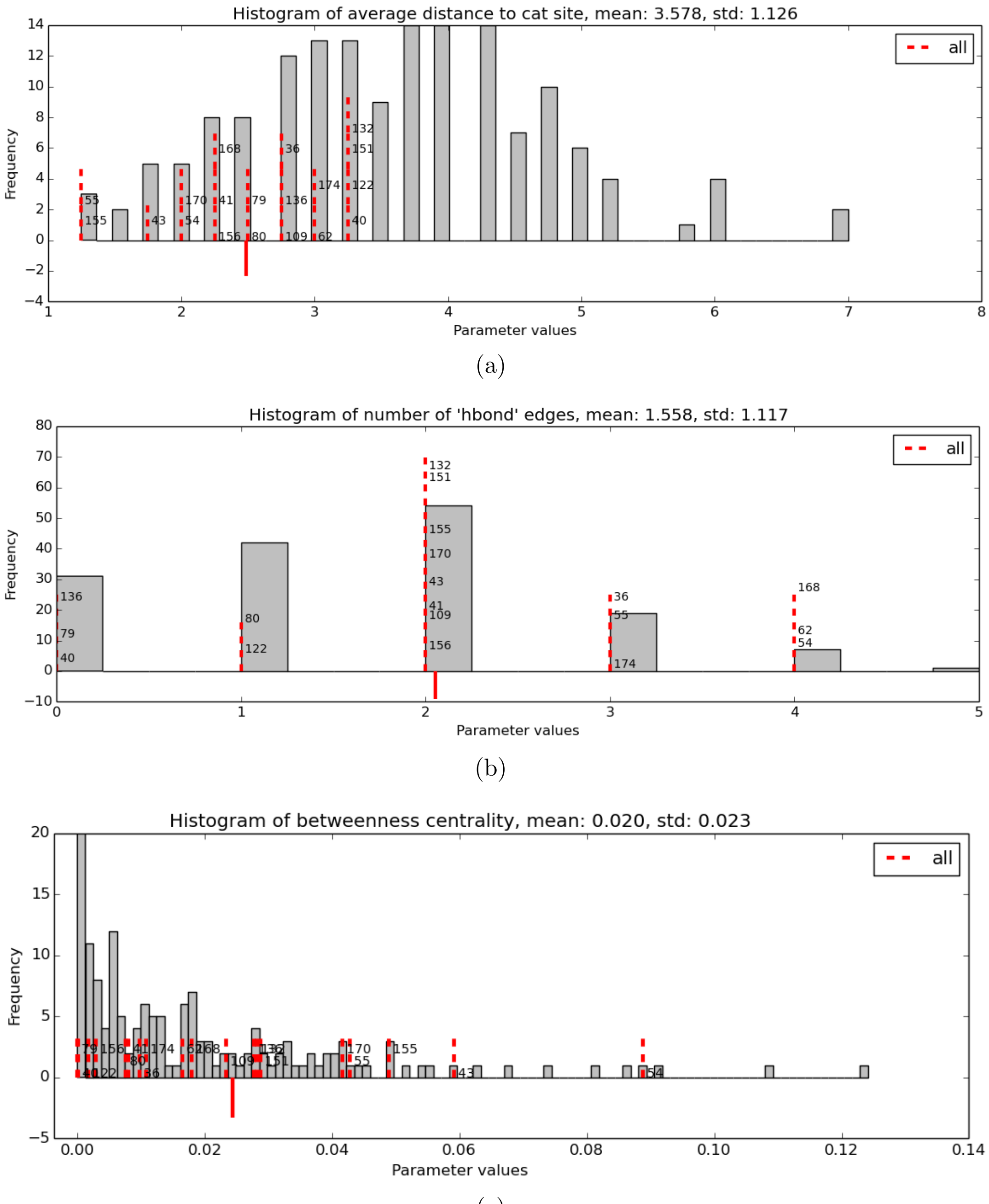
(b)

Figure 4.13: Distribution of residue values for selected properties: (a) conservation and (b) closeness centrality. The values of the functional residues are highlighted by vertical dotted lines in red for catalytic site, and blue, green and cyan for *ddip1*, *ddip2*, and *ddip3*, respectively. The vertical solid lines indicate the average value for each site. All vertical lines have the same size (proportional to the plot height).

associated residues from the remaining protease residues. We also did not observe distinct groups of resistance-associated residues in the histogram plots.

Overall, the computed z-scores for each property and each analyzed residue were low. Only three residues had z-scores above 3.0 (residue 54 for betweenness centrality, residues 79 and 168 for negative charge) and there were several residues with a z-score above 2.0, but at most six residues for one property. We selected all properties for which at least a few resistance-associated residues have a z-score > 2.0 and/or > 1.5. This resulted in 10 to 17 properties for the different structures and groups with considerable overlap between them. We also performed a multidimensional scaling (MDS) of these relevant properties (Borg and Groenen, 2005), but did not see a clear trend for group separation. The MDS was also not very stable as indicated by the high stress values, a measure of reliability.

In order to combine the topological and physico-chemical properties, we divided



(a)

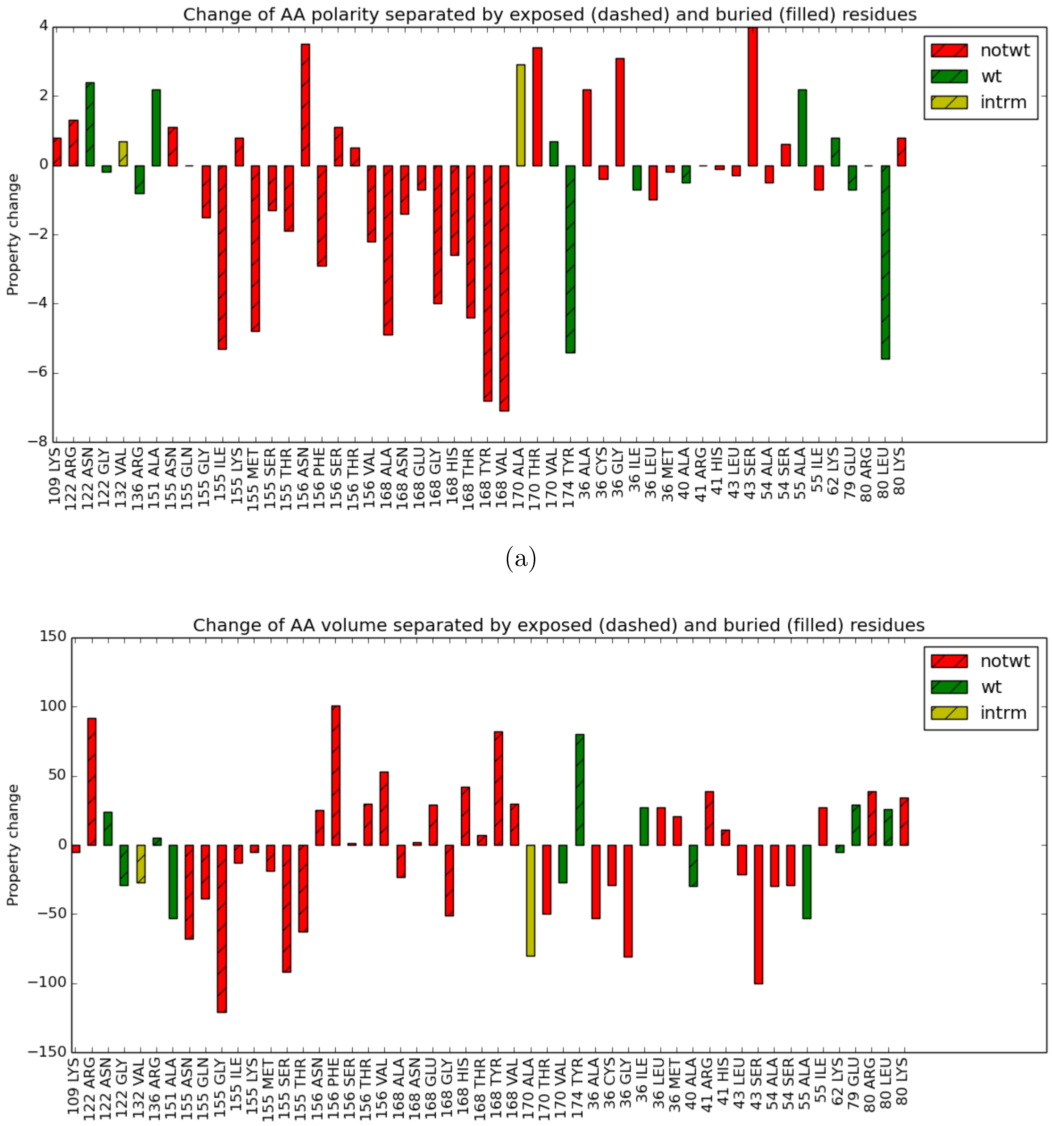
(b)

(c)

Figure 4.14: Distribution of residue values for selected properties: (a) distance to catalytic site, (b) number of hydrogen bonds, and (c) betweenness centrality. The values of the resistance-associated residues are highlighted by vertical dotted lines in red and the vertical solid line indicates the average value. All vertical lines have the same size (proportional to the plot height).

all residues into two groups based on their solvent accessible surface area (SASA): exposed with SASA > 0.1 and buried with SASA < 0.1. Thus, we generated two histograms for each property. We observed more or less the same trends for the resistance-associated residues as described above. We were also able to confirm the general assumptions about the dependencies between properties for exposed and buried residues. For example, buried residues have a higher degree centrality than exposed residues because they have more neighbors and, therefore, make more contacts with them (Figure B.7(a)) In addition, residues with smaller SASA are more hydrophobic as they are not exposed to the mostly hydrophilic solvent (Figure B.7(b)).

Resistance variants



(a)

(b)

Figure 4.15: Selected properties for resistance variants: (a) change of polarity and (b) change of volume. Wild-type (wt) variants are green, intermediate (intrm) variants yellow, and not wild-type (notwt) variants red. Exposed residues are indicated by dashed bars. All vertical lines have the same size (proportional to the plot height).

We performed a different type of analysis for the resistance variants. Since each residue might be associated with one or more wild-type or resistant variants, our goal was to find properties that would distinguish between these two groups of variants. Unfortunately, the topological properties can so far only be computed for residues and not for variants. The latter would require modeling of the residue mutation in the structure followed by optimization of side chains and the whole structure, and a new RIN generation.

Thus, we only considered the properties that are based on the amino acid type (all AA indices and the predicted stability change from PoPMuSiC) and for each property, we computed its change upon AA change. Then, we plotted these values for each resistance variant and highlighted the groups of *wt*, *intrm*, and *notwt* variants. Figure 4.15 provides two examples for the change of AA polarity and volume for each resistant variant considering the fold changes against all protease inhibitors. We observe that substitutions from all three groups (*wt*, *intrm*, and *notwt*) can cause a drastic change of the volume or polarity of the respective residue. If we only focus on individual residues, we can see more clear trends that might be related to the strength of their resistance in further analysis. For instance, all known resistant variants at residue 168 lead to a significant decrease in polarity, which might have an effect on the local interaction pattern of this residue, which is also close to the active site. However, at residue 80, we observe a decrease in polarity for the wild type variants and an increase for the resistant variant.

Summary and discussion

Using a combination of our visual analytics framework for the analysis of sequence mutations and some additional statistics, we performed a systematic investigation of a large set of known resistance mutations in the HCV NS3 protease. Such mutations arise after treatment of HCV infected patients with protease inhibitors and the characterization of these variants is crucial for future patient care (Welsch, 2014; Schneider and Sarrazin, 2014). Therefore, we analyzed the topological and physico-chemical properties of 56 resistant variants occurring in 21 residues of the NS3 protease and compared them to important functional sites as well as all other residues that have not been associated with resistance yet.

Overall, we identified some topological and physico-chemical properties that might be more specific to resistance-associated residues and variants, but at this stage they are not informative enough for an accurate prediction. There are different methodological limitations to our analysis. For instance, it would be very useful to be able to group the residues based on resistance and therefore look for properties that distinguish the group of resistance associated and wild-type residues. Unfortunately, this is not possible because most of these residue have both high and low resistance variants. On the other hand, if we only consider the RAVs, we have to deal with the fact that different resistance variants of the same residue have the same topological properties. Thus, it is difficult to perform the analysis on the residue and variant level at the same time and to combine all types of properties.

Another known issue is that there is not enough data on resistance variants for a comprehensive analysis and, in particular, on neutral variants that certainly do not affect PI treatment.

Our exploratory analysis revealed some interesting trends, but more data will be needed for a comprehensive characterization of RAVs in the HCV NS3 protease. Thus, the future development of this project strongly depends on the identification of further RAVs as well as neutral variants. Once, such data is available, a supervised statistical learning method can be trained on the data and used for prediction of resistance. In addition, it would be useful to perform feature selection on the set of properties as many of them correlate well with each other. A similar analysis can also be performed for other HCV proteins as well as for the proteins of the human immunodeficiency virus, where resistance is also a well-known challenge. We also recommend the exploratory analysis and functional characterization of single resistance mutations using RINs enriched with additional physico-chemical data as shown in our previous analysis of the dysfunctional TIM protein for the BioVis 2013 data contest.

4.6 Conclusions

In this chapter, we presented an integrative visual approach for analyzing the impact of sequence mutations on protein structure and function Doncheva *et al.* (2014). To understand the functional impact of amino acid changes, we combined biological visualizations providing different level of detail and enriched these graphical representations with molecular data derived from external resources. Our framework includes one-dimensional sequence views, three-dimensional protein structure views and two-dimensional views of residue interaction networks as well as aggregated secondary structure views, which are synchronized to reduce the cognitive load of the user when switching between them.

We accomplished this by improving our existing tools structureViz and RINalyzer and integrating them even better with the new versions of Cytoscape and UCSF Chimera. We combined the different visualizations in such a way that biological information can be exchanged between them and additional external data can be easily included. By enhancing molecular networks with structural information and further providing a network representation of residue interactions, our tools also facilitate an interactive multi-layered analysis of protein interactions and binding, allostery and drug resistance mechanisms, just to name a few. Since their release on June 30, 2014, structureViz and RINalyzer have been downloaded just over 650 times, and the two RINalyzer publications were cited 102 times (as of March 8, 2015). Overall, our computational framework is a big step towards bridging the gap between systems and structural biology (Fraser *et al.*, 2013).

We demonstrated the effectiveness of our approach on the data provided for the BioVis 2013 data contest, the sequence of a functionally defective triosephosphate isomerase mutant and its functional yeast parent. In particular, we mapped the protein mutations onto the views together with further functional and structural information and performed an exploratory analysis based on the complementary network views for the parent and mutant sequences. We also assessed the impact of individual amino acid changes by the detailed analysis and visualization of the involved residue interactions. Our contribution was voted the overall favorite by the contest committee, which included both biological and visual analytics experts (Ray *et al.*, 2014).

We also performed a systematic analysis of the physico-chemical, structural and topological properties of resistance mutations in the HCV NS3 protease. For instance, we explored the correlation of several different properties and the distribution of their values in the whole protein. Thereby, we compared the properties of residues known to have a functional role such as active site residues or to be associated with drug resistance with all other residues. In addition, we identified some physico-chemical properties that change significantly upon mutation and might be useful in distinguishing between resistant and neutral variants. Although our overall analysis revealed some interesting trends, we concluded that we need more data on resistance-associated variants as well as more sophisticated approaches to accurately predict them. Nevertheless, our visual analytics framework can be used to characterize the effect of already known or predicted resistance mutations on the function of viral proteins such as the HCV NS3 protease.

Recently, several new tools have been presented that implement a strategy similar to ours. Mosca *et al.* (2015) explored the role of disease mutations on binary human interactions using a structurally annotated interactome (Mosca *et al.*, 2013) and demonstrated the importance of combining different visualizations and levels of detail to gain further understanding of disease mechanisms. They also motivated the release of Structure-PPi (Vazquez *et al.*, 2015), which is a system for the annotation and interpretation of cancer-associated mutations at protein-protein interfaces. Furthermore, resources such as Aquaria (O'Donoghue *et al.*, 2015) further allow users to take full advantage of the wealth of structural information available today by annotating sequence queries with biologically relevant structural and functional data. Nevertheless, these tools still do not provide the same level of detail that is possible with our visual analytics approach centered on residue interaction networks.

Important future steps are the assessment of the usefulness and effectiveness of our approach and the improvement of the current implementation. For this purpose, we intend to collect more user feedback in a comprehensive evaluation. Important questions would be which visual cues are best suited for gaining insight into the impact of mutations, how they should be best mapped onto the sequence, structure, and network representations, and how they should be integrated into the visual layout. We can also make use of the new resources such as Aquaria and dSysMap to retrieve additional functional and structural annotations for our proteins of interest as well as to inquire about other important features. Another issue is the aggregation of network regions to reduce the visual complexity as only some of them might be of actual interest to assess the potential impact of mutations. In this way, patterns of mutations with specific functional consequences might become more apparent, in particular, when multiple proteins are analyzed. The software integration of the different tools can be further improved such that our approach can be realized in a more automated fashion. This includes better synchronization over linked views and automated retrieval of external data.

CHAPTER 5

Protein structure dynamics using RINs

Representing protein structures as networks of interacting residues can facilitate the study of structure-function relationships and give more insight into complex molecular mechanisms such as protein-protein and protein-ligand interactions (Greene, 2012; Di Paola *et al.*, 2013; Hu *et al.*, 2013; Yan *et al.*, 2014). In the previous chapter, we introduced a software suite that supports interactive, multi-layered visual analysis of protein structures and their interactions involved in protein binding, allostery, drug resistance and other molecular phenomena (Doncheva *et al.*, 2014). This approach is, however, still limited to single static snapshots of proteins and their interactions.

Therefore, this chapter presents a method to capture the dynamic nature of protein structures and their interactions by visualizing and analyzing ensembles of protein structures. The first section gives an introduction to protein dynamics and an overview of related work in the field. The next section describes the methodological work involving the definition of dynamic, weighted residue interaction networks (dRINs), their comparison and the ranking of interactions and residues. As a proof of concept, this approach was used for the visual exploratory analysis of data from molecular dynamics (MD) simulations to characterize the effect of sequence mutations. Additionally, an ensemble of docking structures (decoys) was analyzed using dRINs for the identification of the most frequent interface residues and interaction. We designed, implemented, and performed the described analysis. The docking project was inspired by discussions with John H. Morris and Dina Schneidman-Duhovny at UCSF.

5.1 Introduction

5.1.1 Biological Background

Proteins have a dynamic nature that can be described by the energy landscape of all possible conformations they populate (Frauenfelder *et al.*, 1991). Protein dynamics refers to the transition of a protein from one state to another by stochastic fluctuations of chemical bonds, collective residue motions, or global conformational changes (Henzler-Wildman and Kern, 2007; Orozco, 2014). These transition can happen at various spatial and temporal scales. Many molecular phenomena including (un)folding, allosteric signaling, enzyme catalysis, and protein complex assembly are strongly related to protein dynamics (Daggett, 2006; Gunasekaran *et al.*, 2004; Henzler-Wildman *et al.*, 2007; McGeagh *et al.*, 2011). Therefore, to further understand structure-function relationships, we also need to study the dynamic character of proteins.

The two common techniques for structure determination, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, already provide some insight into protein dynamics. X-ray diffraction data contains indirect information on flexibility and atomic displacement, while NMR methods directly deliver multiple transitions on different time scales at atomic resolution (Henzler-Wildman and Kern, 2007; Orozco, 2014). A recent Perspective paper by van den Bedem and Fraser (2015) argues that integrative structure biology, which combines evidence from multiple sources including NMR and crystallography, will bridge the gap between accurate high-resolution structures and dynamics. However, the richest source of information is still data generated from theoretical methods such as molecular dynamics.

Molecular dynamics is a computational technique for simulating the motions of particles such as atoms and molecules as a function of time (Adcock and McCammon, 2006). MD simulation of proteins were first performed in the seventies by McCammon *et al.* (1977) and have become a trusted and widely used tool for investigating protein dynamics. An MD simulation generates a trajectory of the system at atomic resolution under simulated experimental conditions. Thereby, the individual particles move as a result of their interactions with each other, and these movements are calculated by numerically solving Newton's equations of motion for this system. The forces between the particles as well as the potential energy of the system are described by a set of mathematical expressions and parameters, usually referred to as molecular mechanics force field. The choice of an appropriate energy function is crucial for the validity and stability of an MD simulation (Adcock and McCammon, 2006). Designing and running an MD simulation for a protein of interest is a complex task due to the large number of available force fields and software tools as well as the heavy demands on the underlying hardware. Recent theoretical and computational advances in the field are reviewed in more detail by Adcock and McCammon (2006); Klepeis *et al.* (2009); Zwier and Chong (2010); Dror *et al.* (2012); Benson and Daggett (2012); Orozco (2014).

Other computational techniques are applied to study particular molecular phenomena involving dynamics. For instance, the prediction of protein complexes for known protein interactions still represents one of the big challenges in the field of molecular docking (Ritchie, 2008; Vajda and Kozakov, 2009; Lensink and Wodak, 2010). Docking approaches predict the interactions between two molecules such as two proteins or a protein and a ligand. They usually operate in two main steps: (1) searching the conformational space and (2) scoring the selected models based on various criteria such as geometrical or physico-chemical compatibility. The output is a ranked list of predicted complexes, also referred to as decoys or models. Some of the widely used search algorithms and scoring functions as well as their combinations were reviewed by Halperin *et al.* (2002); Schneidman-Duhovny *et al.* (2004); Hildebrandt *et al.* (2008); Vajda *et al.* (2013).

Nowadays, there are numerous docking approaches and accompanying software available (Schneidman-Duhovny *et al.*, 2004; Moreira *et al.*, 2010; Rodrigues and Bonvin, 2014). A notable community-wide initiative for evaluating the performance of methods by blind prediction is the CAPRI contest (Critical Assessment of Prediction of Interactions), which originated in 2001 (Mendez *et al.*, 2003, 2005; Lensink *et al.*, 2007; Lensink and Wodak, 2010, 2013). Unfortunately, state-of-the-art methods include a near-native model, i.e., a good prediction, in the top 10 only in 30-40 % of the cases and even the recently introduced integrative docking approach by the Sali lab succeeds only in 42-82 % of the cases (Schneidman-Duhovny *et al.*, 2012). However, these approaches are more successful at predicting the correct interface even if the overall quality of the predicted complex is quite low. Lensink and Wodak (2010) evaluated the predictions of interface residues derived from the protein docking models submitted for 20 different CAPRI targets. They found that 70 % of the correctly predicted interface residues, i.e. with a precision and recall ≥ 50 %, are in models assessed as incorrect by the CAPRI guidelines and only 30 % in correct models. Furthermore, 24 % of the interfaces in incorrect models are actually correctly predicted.

As described above, both MD simulations and docking approaches generate ensembles of structures. We are particularly interested in finding the residue interaction similarities and differences between individual structures in such ensembles as well as between the ensembles. In this chapter, we present an approach for visualizing and analyzing ensembles of protein structures using dynamic residue interaction networks. Before we introduce the methodology and present the results, we give an overview on related approaches.

5.1.2 Related work

In the last few years, several studies and accompanying tools have been presented that combine the analysis of structural ensembles with graph theory towards understanding protein structure-function relationships. In particular, the group of Vishveshwara has extensively studied dynamic protein processes such as allostery in tRNA syntherases (Vishveshwara *et al.*, 2009), while the group of Xiaojun Yao has investigated inhibitor binding and drug resistance in Hepatitis C virus (HCV)

(Xue *et al.*, 2012, 2014a,b,c) and human immunodeficiency virus (HIV) (Xue *et al.*, 2013). In contrast to these two groups, which have considered different simulation time points individually in their analyses, the approaches by Sethi *et al.* (2009) and Tiberti *et al.* (2014) focused on the frequent interactions present in the whole ensemble to study binding and allostery. Furthermore, networks of residue interactions have helped in evaluating the prediction of protein structures (Chatterjee *et al.*, 2013) and protein complexes (Vangone *et al.*, 2012; Oliva *et al.*, 2013).

Intra and inter-molecular communications. Ghosh *et al.* (2007) first combined residue interaction networks and MD simulations by creating a network for each single snapshot. They used the definition of *protein structure networks* (PSNs) introduced by Vishveshwara and colleagues (Kannan and Vishveshwara, 1999; Brinda and Vishveshwara, 2005), where nodes represent residues and edges correspond to non-covalent interactions between them. In addition, the edges in PSNs are weighted by the strength of residue interaction, which depends on the number of interacting residue atoms and their type. Ghosh *et al.* (2007) computed different topological properties, such as degree distribution, shortest paths or connected component composition, for each time step and compared them to characterize the change of interactions between important residues during an equilibrium or unfolding simulation. Vishveshwara *et al.* (2009) described their approach and the various topological parameters of interest in a review article. So far, Vishveshwara and colleagues applied their methodology to study the process of unfolding of T4 lysozyme (Ghosh *et al.*, 2007) as well as the allosteric (shortest) pathways in methionyl tRNA synthetase (Ghosh and Vishveshwara, 2007), tryptophanyl tRNA synthetase (Hansia *et al.*, 2009; Bhattacharyya *et al.*, 2010), cysteinyl tRNA synthetase (Ghosh *et al.*, 2011), and pyrrolysyl tRNA synthetase (Bhattacharyya and Vishveshwara, 2011). This method was also applied by Blacklock and Verkhivker (2014a,b) to analyze allosteric regulation in the Hsp90 chaperones.

Recently, the standalone program PSN-Ensemble was released by the group of Vishveshwara (Bhattacharyya *et al.*, 2013). It can generate PSNs from structural ensembles and compute parameters related to the topological organization and long-range allosteric communication of the protein. In addition, the authors implemented a flexible weighing scheme derived from residue pair-wise cross-correlation and interaction energy to bring in dynamical and chemical knowledge into the network representation. The resulting networks are visualized on the protein structure opened in PyMOL (Schrödinger, LLC, 2010). The same group also previously developed the GraProStr web service (Vijayabaskar *et al.*, 2011) for the generation of PSNs from single structures and the computation of selected network parameters.

Wordom (Seeber *et al.*, 2011), a tool for molecular structure visualization and simulation, was extended to generate protein structure networks and to compute shortest communication paths as defined by Vishveshwara's group (Brinda and Vishveshwara, 2005; Vishveshwara *et al.*, 2009). Using Wordom, Mariani *et al.* (2013) investigated the communication pathways in transducin to understand the effect of missense mutation on its activity and, more globally, on the signaling network

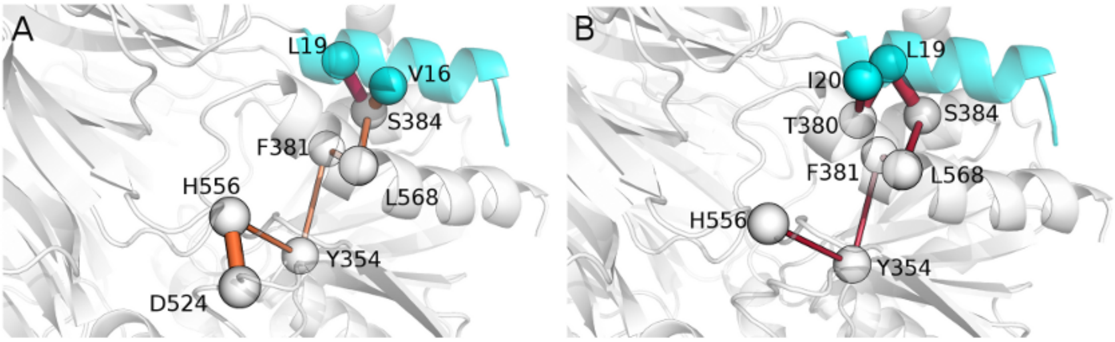


Figure 5.1: Shortest and highest frequency pathways between V16 (A), L19 (A), I20 (B) and the catalytic residues H556 and D524 are shown as sticks in the 3D structure visualization of the the hyperthermophilic acylaminoacyl peptidase. The stick width is proportional to the intensity of the correlation between the residues during the MD simulation. Reproduced with permission from Papaleo *et al.* (2012b).

involved in monogenic retinal diseases. Papaleo *et al.* (2012b) performed MD simulations and network analysis of the hyperthermophilic acylaminoacyl peptidase with a special focus on long-range communication paths (see Figure 5.1 for an example). Recently, the authors of Wordom published an improved method for investigating allosteric communication based on a mixed Protein Structure Network and Elastic Network Model-Normal Mode Analysis approach (Raimondi *et al.*, 2013) and implemented it in WebPSN, a user-friendly web server (Seeber *et al.*, 2014).

Two other tools were developed by Papaleo and colleagues for the purpose of bridging the gap between structural and network biology. xPyder is a plugin for PyMOL (Schrödinger, LLC, 2010) that analyzes interdependencies between residues, in particular, dynamical cross-correlation and non-covalent residue interactions (Pasi *et al.*, 2012). The plugin makes use of graph theory to analyze the data, usually represented as matrices, and visualizes the residue interactions on the protein structure in PyMOL. PyInteraph has been released recently as a stand-alone tool for creating PSNs for a specific type of residue interactions (only hydrophobic or side-chain, hydrogen or salt bridges) from MD simulations and structural ensembles and preforming network analysis on them (Tiberti *et al.*, 2014).

Combining dynamical cross-correlation and non-covalent residue interactions, Papaleo et al. identified a pattern of asymmetric flexibility in a cold-adapted homodimeric enzyme, the Vibrio alkaline phosphatase (Papaleo *et al.*, 2013) and established the crucial role for Loop 7 of E2 enzymes in E2-mediated steps of the ubiquitination cascade (Papaleo *et al.*, 2012a). Invernizzi *et al.* (2013) from the same group provided the first structural description of an intrinsically disordered domain of ataxin-3 and investigated its effect on ataxin-3 aggregation, while Lambrughi *et al.* (2012) used the xPyder plugin to identify intramolecular interactions that stabilize the compact conformations of the intrinsically disordered kinase-inhibitor domain of Sic1. Invernizzi *et al.* (2014) investigated the communication routes in ARID domains with the help of Wordom for creating PSNs from the MD simulation data and computing the shortest paths as well as xPyder for the visualization.

Allosteric regulation. Sethi *et al.* (2009) investigated the differences in binding modes of tRNA:protein complexes using MD simulations and *dynamical networks*. In the latter, an interaction between two residues is represented by an edge if it is present in the majority of the simulation and is weighted by the correlation between the monomers. Furthermore, they analyzed these networks by applying several graph theoretical algorithms such as shortest and suboptimal path identification and community search. This workflow was later implemented as an extension to the molecular visualization software VMD (Humphrey *et al.*, 1996) and called NetworkView (Eargle and Luthey-Schulten, 2012). Among other features, the JGROMACS Java library also aids the generation and analysis of dynamical networks from GROMACS data (Munz and Biggin, 2012).

The dynamical network approach was further applied to understand the allosteric immune escape pathways in the HIV-1 envelope glycoprotein by Sethi *et al.* (2013). Furthermore, Miao *et al.* (2013) analyzed the allosteric activation of the M2 muscarinic receptor by identifying significant differences in the network of residue interactions between the different forms. The dynamical networks of the enzyme imidazole glycerol phosphate synthase and the blood coagulation protease thrombin during allosteric regulation were also investigated by Manley *et al.* (2013) and Fuglestad *et al.* (2013), respectively. Vanwart *et al.* (2012) studied the effect of different network node definitions on the prediction of allosteric regulation in the imidazole glycerol phosphate synthase and concluded that the entire residue center of mass needs to be included in the detection of residue interactions. Scarabelli and Grant (2013, 2014) also compared the network of dynamic communication of several kinesin motor domains.

Resistance mechanisms in viral proteins. Recently, Xue *et al.* (2012, 2014a,c) assembled an analysis framework consisting of MD simulation, binding free energy calculation, free energy decomposition, substrate envelope analysis, and RIN analysis using RINalyzer (Doncheva *et al.*, 2011). They applied this framework to understand the structural and energetic basis of inhibitor and substrate binding with focus on drug resistance mechanisms. Thereby, a RIN was generated from the representative structure of each simulation (using the RING web service (Martin *et al.*, 2011)) and then visualized and analyzed in Cytoscape with the help of RINalyzer and NetworkAnalyzer (Doncheva *et al.*, 2012a). The main target of their studies was the HCV NS3/4A protein bound with the protease inhibitors ITMN-191 (Xue *et al.*, 2012), Vaniprevir and MK-5172 (Xue *et al.*, 2014a), as well as the allosteric inhibitor 4VA (Xue *et al.*, 2014c). An example for the RINs of the apo and inhibitor bound HCV NS3/4A protease is shown in Figure 5.2. In addition, Xue and colleagues analyzed the HCV NS5B polymerase complexed with the non-nucleoside inhibitors VX-222 and ANA598 (Xue *et al.*, 2014b) as well as the HIV-1 integrase and its integrase strand transfer inhibitors Raltegravir, Elvitegravir, and Dolutegravir (Xue *et al.*, 2013). Bhakat and colleagues performed the same type of analysis to understand the impact of resistance mutations on the HIV-1 reverse transcriptase (Bhakat *et al.*, 2014; Karubiu *et al.*, 2014), while Guariniello *et al.* (2014) studied to effect of mutations on the human selenoprotein M.

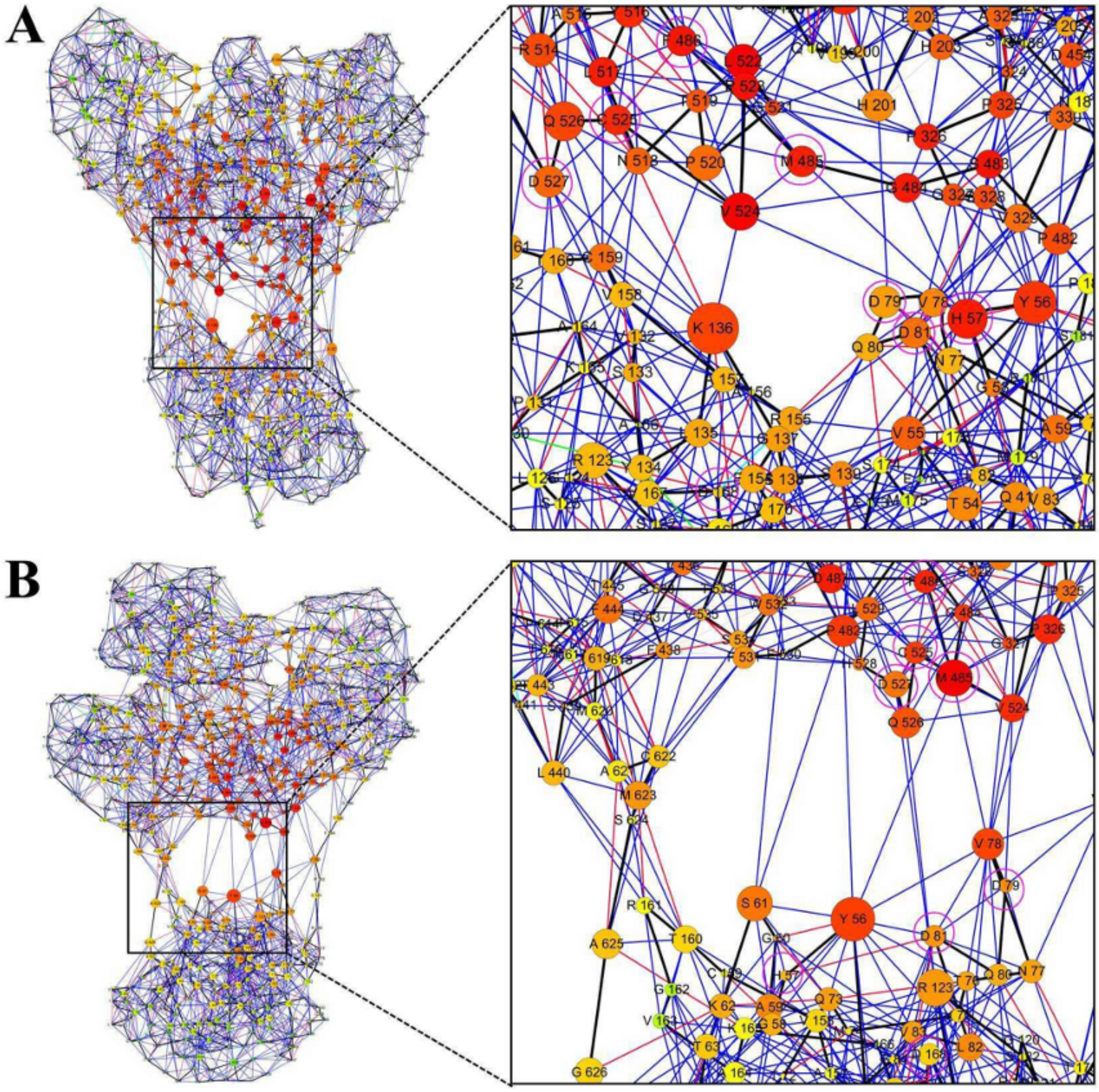


Figure 5.2: RIN of the representative protein structure derived from molecular dynamics simulation of the (A) apo and (B) inhibitor bound HCV NS3/4A protease. Residue nodes are colored according to their closeness values, while the node size corresponds to their betweenness values. Edges represent different types of non-covalent residue interactions. Reproduced with permission from Xue *et al.* (2014c).

Networks of disrupted residues. Furthermore, the group of V. Daggett developed a new tool called ContactWalker (Bromley *et al.*, 2013) to analyze differences in residue interactions between mutant and wild-type protein structures generated from MD simulations and to highlight networks of disrupted residues. For this purpose, residue contacts were identified for each snapshot and the change of contacts was computed to identify residues with significant 'residue occupancy difference'. Bromley *et al.* (2013) demonstrated the usefulness of their tool by analyzing the structural consequences of mutations to the α-tocopherol transfer protein. ContactWalker was released as part of the DIVE framework (Bromley *et al.*, 2014) for analysis of complex data on the example of the Dynameomics data warehouse (van der Kamp *et al.*, 2010).

Finally, a related method using slightly different data was recently presented by van den Bedem *et al.* (2013). CONTACT generates functional dynamic contact networks from raw high-resolution X-ray crystallography data. Thereby, conformationally heterogeneous residues and their contacts are summarized into separate networks and topological analysis of these networks reveals differences between room-temperature and cryogenic or wild-type and mutant data sets.

Protein structure prediction. The field of protein structure prediction has also benefited from the representation of protein structures as networks of residue interactions. Chatterjee et al. compared the network properties of native structures with predicted models (Chatterjee *et al.*, 2012) and used them to train an SVM for classifying (Chatterjee *et al.*, 2013) and ranking decoys (Ghosh and Vishveshwara, 2014). The method is called PSN-QA and is available as part of the GraProStr web service (Vijayabaskar *et al.*, 2011). Zhou and colleagues presented the integrated score function SVR_CAF, which consists of three scores combined by a support vector regression (Zhou *et al.*, 2014b). One of the scores is based on the average degree and shortest path length of the RINs generated from the contact energy calculations on the predicted models. In their next publication, Zhou *et al.* (2014a) provide a review on the application of amino acid network properties for the discrimination of native protein structures from decoys.

Protein complex prediction. At the same time, there have been some attempts at employing information derived from RINs to rank docking solutions. In 2008, Chang and colleagues performed a comparative study of different topological parameters (degree, clustering coefficient and characteristic path length) computed for RINs constructed from protein complexes (Chang *et al.*, 2008). Interestingly, for each complex, they created a hydrophobic and hydrophilic RIN that only consists of interactions between hydrophobic or hydrophilic residues, respectively. The authors concluded that protein complexes exhibit small-world network properties with large clustering coefficients and small characteristic path lengths. They demonstrated a successful combination of average degree and clustering coefficient into a simple scoring function as well as the further integration with energy-based terms for a more sophisticated scoring function. This work was further developed by Jiao and Chang (2011), who proposed the incorporation of contact energies as edge weights in the RINs and the computation of weighted degree and average nearest neighbors degree.

Recently, Khashan *et al.* (2012) developed the SPIDER scoring function by taking into account the geometry and frequency residue patterns identified by subgraph mining of native protein interfaces. Notably, to construct the residue interaction networks, they applied Almost-Delaunay Tessellation, a computational geometry technique previously used to discriminate native from non-native protein conformations by Krishnamoorthy and Tropsha (2003).

Furthermore, the group of Luigi Cavallo released the web-based tools CONSRANK and CONS-COCOMAPS for measuring the conservation of residue interactions on the interface of protein complexes (Vangone *et al.*, 2012) and ranking docking decoys based on their ability to match the most conserved interactions (Oliva *et al.*, 2013; Vangone *et al.*, 2013; Chermak *et al.*, 2014), respectively. In particular, the authors represented interface residue interactions as 2D contact maps and used these to visualize the consensus of multiple docking solutions. Their work is described in more detail in Section 5.4.

5.2 Methods and Implementation

Here, we introduce the term dynamic residue interaction network (dRIN), which describes a RIN generated not from a single protein structure but from a set of structures or a conformational ensemble. We are interested in the variation reflected by the individual structures and, at the same time, we greatly benefit from an overview representation that highlights the similarities and differences in non-covalent interactions between the amino-acid residues within the whole ensemble. Such an ensemble might be the result of an experimental technique like nuclear magnetic resonance spectroscopy (NMR) of a protein, a computational model of a protein structure or complex, or a molecular dynamics (MD) simulation. For NMR structures and MD simulations, an average representative structure can be computed, which, however, is only a static average view of the different conformations and does not show the individual variation. Thus, we propose the usage of dRINs as an alternative representation that is enriched with more information. In the following sections, we will formally define dRINs and give more details on how they can be generated from ensembles of protein structures.

5.2.1 Definition of dynamic RINs

Given a set of s structures of the same protein, we represent each structure by a RIN, in which the nodes correspond to the residues and the edges to the non-covalent residue interactions. Thus, each RIN consists of the same set of residue nodes R and an individual set of interaction edges $E_1, \ldots, E_s$ that is usually slightly different for each RIN. For simplicity, we will refer to the set of all edges contained in any of the s structures as $E_R = \bigcup_{i=1}^{s} E_i$.

For each edge $e \in E_R$, we denote the number of networks (structures) that contain this non-covalent interaction edge as s_e. Then we define the frequency of occurrence of the interaction edge e as

$$f(e) = \frac{s_e}{s}$$

with $f(e) \in [1/s, 1]$. $f(e) = 1$ when some non-covalent interaction occurs in all structures, and $f(e) = 1/s$ when the interaction is observed in only one of them. This score is also referred to as *interaction conservation* in the CONS-COCOMAPS

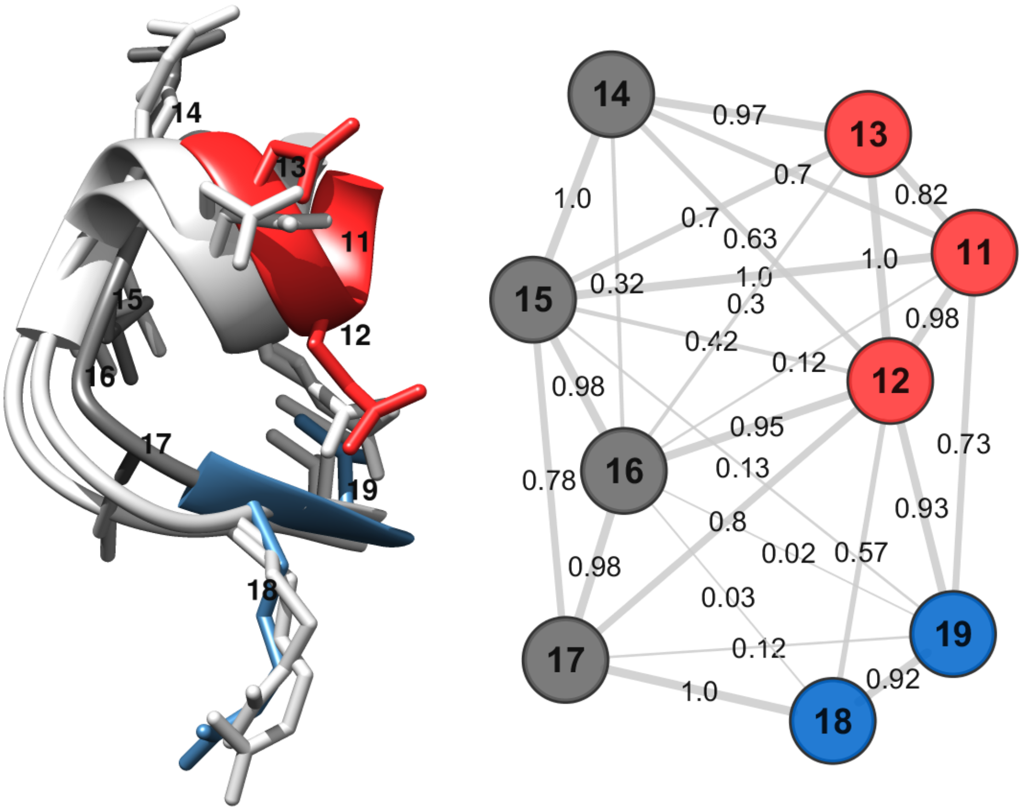


Figure 5.3: Example for a dynamic RIN. A region of the structure from three snapshots of an MD simulation is visualized in UCSF Chimera (left) and the dynamic RIN for the whole simulation is displayed in Cytoscape using RINalyzer (right). Residues are represented as nodes, and both are colored according to secondary structure (helices in red, sheets in blue, loops in gray). Non-covalent residue interactions are shown as edges labeled by their frequency of occurrence. The edge line thickness is also proportional to the edge weight. We observe that the non-covalent interactions between neighboring residues in sequence are very conserved, while the interactions between spatially close residues vary depending on the orientation of the side chains and the flexibility of the backbone. For instance, an interaction between residues 12 and 18 with a frequency of 0.57 is present only in half of the snapshots.

approach (Vangone *et al.*, 2012) and as *persistence* in the PyInteraph tool (Tiberti *et al.*, 2014).

Since the edges in a RIN can be weighted, e.g., by interaction strength, we define the weight of an edge $e \in E_i$ in the i-th RINs as $w_i(e)$ and can generalize the above definition for a weighted interaction frequency to

$$f_w(e) = \frac{\sum_{i=1}^{s} w_i(e)}{s \times \max_i w_i(e)}$$

with $f_w(e) \in [1/s, 1]$. The dRIN is built from the set of all residue nodes R and the union of the edge sets E_R, whereas each edge $e \in E_R$ is weighted by its frequency of occurrence $f(e)$. An example network is shown in Figure 5.3.

5.2.2 Ranking residues and interactions

Dynamic RINs provide a quantitative overview of the diversity of structures in the whole ensemble and, therefore, are a useful tool for studying the individual structures, their residues and interactions. First of all, we can rank the residue interactions according to their frequency of occurrence. On top of the list, we will have the most frequent interactions in the ensemble, while the least frequent ones will be at the bottom. We can then divide the interactions into two (or more) groups by setting appropriate thresholds. One group will represent the most similar interactions among the different conformations and the other group the most different ones. Depending on the chosen threshold and type of ensemble, these groups might also be interpreted as *conserved*, *stable*, or *reliable* and *variable*, *unstable*, or *unreliable*, respectively.

Furthermore, we can convert the edge frequencies into node scores and rank the residues according to these scores. In this way, we can identify residues with the most conserved interactions (top of the list) and at the same time, residues with the most variable interactions. For each node v, let $N(v)$ be the set of its neighboring nodes. We define the node score

$$g(v) = \frac{\sum_{w \in N(v)} f(e_{vw})}{|N(v)|}$$

where $f(e_{vw})$ is the interaction frequency of the edge between nodes v and w. As for the interactions, we can rank the nodes using this score and/or set a threshold and partition them into two (or more) groups. The group of nodes with high score will correspond to nodes that have on average the most frequent interactions, while the other group will contain residues with infrequent interactions.

Finally, we can rank the individual RINs generated from structures in the ensemble based on how many of the most frequent interactions they contain as defined by Oliva et al. (Oliva *et al.*, 2013). For each network $G_i = (R, E_i)$, we calculate the sum of interaction frequencies for its edges E_i and normalize it by the total number of edges:

$$h(G_i) = \frac{\sum_{e \in E_i} f(e)}{|E_i|}$$

This network score can be used as an alternative to finding a representative structure for the whole ensemble by choosing the one with the highest score.

5.2.3 Comparison of dynamic RINs

Apart from ranking residues and their interaction according to their abundance in the ensemble, we would also like to compare two different ensembles of structures/-conformations using their dRINs. For this purpose, we extend the comparison of RINs described in the previous chapter to weighted dynamic RINs. For two dRINs G_1 and G_2, we create a comparison network G_c, in which the nodes represent aligned residues and the set of edges is the union of edges $E_1 \cup E_2$. It is important to note

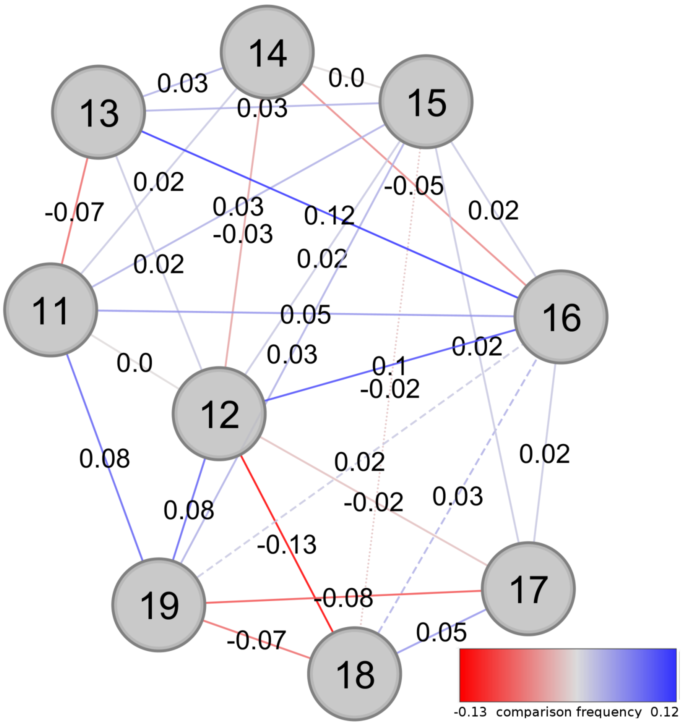


Figure 5.4: Example for a dynamic comparison RIN. The dynamic comparison RIN is displayed in Cytoscape using RINalyzer. Nodes represent aligned residues and are labeled by the respective residue number, while edges correspond to non-covalent residue interactions. The comparison frequency is the difference of interaction frequencies of the corresponding edges in the networks considered for the comparison. It is used to label the edges and color them in a red-white-blue gradient. In the cases, in which an edge was present in only one of the compared networks, the weight corresponds to the original frequency of occurrence and the edge is shown as a dashed (first network) or dotted (second network) line.

that, for each edge, we can have three different cases: (1) it belongs to both networks, (2) it belongs to the first network only, (3) it belongs to the second network only; and we keep track of this status. In the special case of weighted dynamic RINs, we need to transform the weights in such a way that we keep as much information about the frequencies as possible. Thus, we introduce the term comparison frequency, which can be computed using one of the suggested transformations of the original frequencies.

First, we denote the weight of a comparison network edge as $c(e)$. If this edge belongs to both networks, we have two weights for it $f_1(e)$ and $f_2(e)$. If it belongs to only one of the networks, we set the other weight to be equal to 0. An obvious transformation is to calculate the difference

$$c(e) = f_1(e) - f_2(e)$$

which results in weight values between -1 and 1. In this way, positive values are assigned to interaction edges that are more frequently seen in the first dRIN, while negative values point to interaction edges that are more frequent in the second dRIN. An example is shown in Figure 5.4. If do not consider the order of networks, but we are only interested in the absolute values, we can also generate the weights between 0 and 1 using the absolute difference:

$$c(e) = |f_1(e) - f_2(e)|$$

However, in both cases, the resulting weights are relative and do not provide information about the absolute frequency of the interaction. For example, if the weight of the comparison edge is 0.5, it might be the case that $f_1(e) = 0.5$ and $f_2(e) = 0$ or that $f_1(e) = 1$ and $f_2(e) = 0.5$.

To overcome this problem, we can compute the ratio of the weights defined as

$$c(e) = \frac{f_1(e)}{f_2(e)}$$

Now, we have values above 1 for interactions more frequent in G_1 and values below 1 for interactions more frequent in G_2. Since this transformation can only be used for $f_1(e) \neq 0$ and $f_2(e) \neq 0$, the weights for edges present only in one network remain unchanged and are not on the same scale as the transformed weights. In addition, we obtain weights in a very wide value range (from very small to very large numbers), depending on the difference in the interaction frequencies.

We can easily handle the second issue and render the values easier to compare using the logarithm, e.g.:

$$c(e) = \log \frac{f_1(e)}{f_2(e)}$$

Now, as in the first case, positive and negative edge weights are assigned to interaction edges more frequent in either the first or the second network, respectively, while edge weights of 0 indicate interactions with the same frequency in both networks. A possible solution to the first issue, e.g. distinguishing transformed from untransformed weights, is to increase or decrease these weights by 1 or -1, respectively.

5.2.4 Implementation details

To our knowledge, none of the previously published tools is capable of generating a comprehensive set of RINs for all structures in an ensemble as well as an overview network containing different types of non-covalent residue interactions. Thus, we have extended the RINerator package (Doncheva *et al.*, 2011) to create the dRINs presented in the previous section.

RINerator is a package that generates user-defined RINs from a 3D protein structure and enriches them with additional biochemical information. In contrast to previous simplistic interaction definition approaches based on spatial atomic distance between residues, *RINerator* enables a more realistic representation by considering different biochemical interaction types, such as hydrogen bonds and interactomic contacts as described in Section 4.3.5. Recently, we have extended *RINerator* to calculate conservation scores from a user-specified multiple sequence alignment file and to retrieve biochemical amino acid properties from external resources, such as AAindex and ConSurfDB. The resulting networks and accompanying data can be visualized in Cytoscape using the RINalyzer and structureViz apps.

For the purpose of generating dRINs, we have modified *RINerator* to be capable of handling multiple PDB structure files as input. This is accomplished in several steps:

1. The input is a directory of PDB structures and a list of chains/ligands to be included in the RIN.
2. For each structure the following steps are computed:
   - The program *Reduce* is used to add hydrogen atoms to the structure (Word *et al.*, 1999b);
   - The tool *Probe* is applied to identify contact dot surfaces (Word *et al.*, 1999a);
   - The contact dots are classified into pre-defined groups of non-covalent residue interactions (cnt, hbond, ovl, combi and mc, sc, all) [1] based on the atom type and the specific distance between the contact surfaces;
   - Residue interaction networks are saved in SIF format and respective attribute files in TSV format as defined by Cytoscape.
3. *RINerator* creates a network of all occurring interactions and calculates their frequencies, which are saved as edge attribute files.
4. The output consists of a dRIN and RINs for all structures with respective attribute files in TSV format.

Furthermore, we extended the *RINalyzer* app to support the comparison of weighted RINs, such as the dynamic RINs described above. The user can specify the numeric edge attribute to be used for the comparison as well as how the weight difference should be computed. For the letter, we provide the following four options:

- *difference*: the weight of the second network edge is subtracted from the weight of the first network edge
- *abs difference*: the absolute value of the *difference*
- *ratio*: the quotient of the weight of the first network edge and the second network edge
- *log ratio*: the logarithm of the *ratio*

[1] cnt = contact, hbond = hydrogen bond, ovl = overlap of van der Waals radii, combi = combined interaction; mc = main chain atoms, sc = side chain atoms, all = all (any) residue atoms

The resulting weight values are saved as a new edge attribute called *CompWeight* in the resulting network and can be highlighted in the network using an edge color gradient and different edge line types (see Figure 5.4).

To facilitate the analysis of interface residues and interactions, we defined a simple procedure for the generation of an interface RIN, which is implemented by both *RINerator* and *RINalyzer*. Given a list of protein chains, we retain each interaction between residues that belong to different chains. *RINerator* generates a text file containing the pair-wise interactions, while *RINalyzer* directly creates a new network with the appropriate visualization and takes into account the former network layout and visual style.

5.3 Analysis of MD simulation data

As described in Section 5.1.2, there have been a number of attempts to analyze MD simulation data with the help of residue interaction networks. However, most of them are very specific to the problem at hand and even if they provide analytical tools that might be applied to answer other biological questions, they usually still lack proper visualizations and the ability to integrate additional data. Here, we show how to address some of these issues by creating dRINs and visualizing and analyzing them within Cytoscape. Our approach provides both user-friendly visualizations that facilitate exploratory analysis as well as sophisticated analysis tools.

5.3.1 Data and experimental setup

The Dynameomics database is a comprehensive resource for molecular dynamics simulations of over 2000 protein systems (Beck *et al.*, 2008; Jonsson *et al.*, 2009; van der Kamp *et al.*, 2010). Its main goal is the characterization of the native state dynamics and the (un)folding pathways of representatives from all known protein folds. As of 2014, the database contains approximately 20 000 simulations of nearly 800 microseconds combined simulation time.

A further aim of the Dynameomics project is to facilitate the analysis of SNPs and their effect on the structure, stability, function, and dynamics of proteins. Especially interesting are mutations that cause structural disruptions at important, but distant locations in the protein, such as the active site. The database contains approximately 200 simulations of 31 proteins with single-point mutations and the corresponding wild-type simulations. In particular, Daggett and colleagues analyzed the catechol O-methyltransferase (Rutherford *et al.*, 2006, 2008a; Rutherford and Daggett, 2009), histamine N-methyltransferase (Rutherford *et al.*, 2008b), thiopurine S-methyltransferase (Rutherford and Daggett, 2008), DJ-1 (Anderson and Daggett, 2008), and the DNA glycosylase/β-lyase hOgg1 (Anderson and Daggett, 2009).

After acquiring access to the SQL database Dynameomics, we were able to retrieve simulation data for 6 different proteins. Each simulation was performed at 310K for approximately 31 ns in the micro canonical ensemble using the *in lucem* molecular mechanics (*il*mm) program (Beck *et al.*, 2010) and underwent an extensive set of standard analyses and quality control procedures. Multiple (on average three) simulations were performed for each wild-type and mutated structure.

For each simulation run, we created snapshots at every 0.5 ns. Simulation runs shorter than 31 ns were filtered out. We combined the data from the different runs into one dRIN per simulation using the RINerator package. The resulting dRINs were loaded into Cytoscape using the RINalyzer app. The comparison dRINs were generated by taking into account the difference in frequencies of occurrence between wild-type and mutant. All visualizations were generated with UCSF Chimera or Cytoscape.

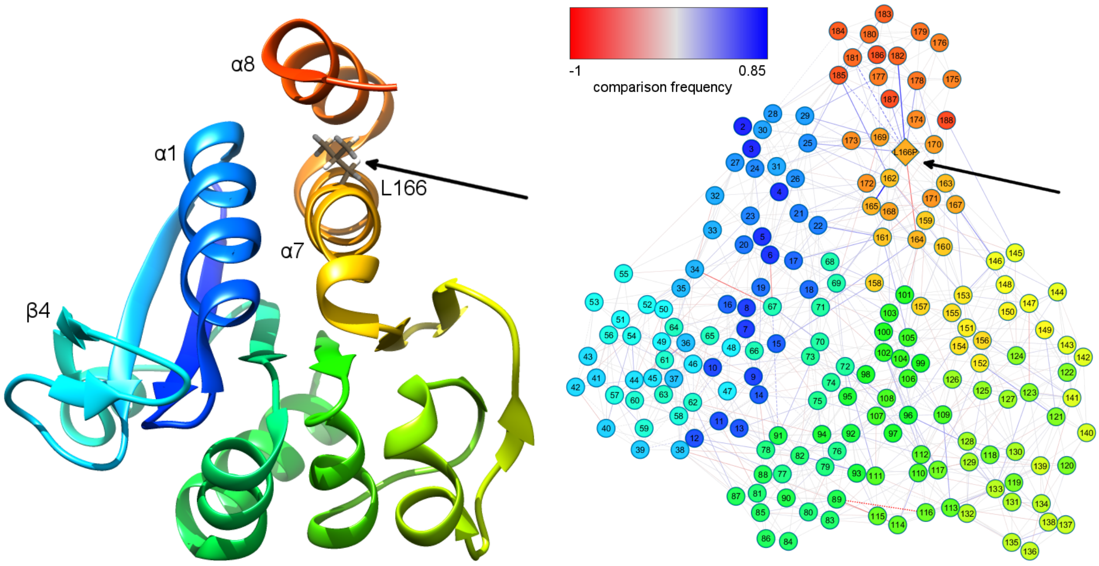
5.3.2 Results for L166P mutation in DJ-1

DJ-1, also known as PARK7, is an evolutionary conserved homodimer. The dimer interface is mainly formed by the $\alpha1$, $\alpha7$, $\alpha8$ helices and the $\beta4$ strand of both subunits (Figure 5.5(a)) (Tao and Tong, 2003). Among others, DJ-1 functions as a positive regulator of androgen receptor-dependent transcription (Takahashi *et al.*, 2001; Niki *et al.*, 2003; Xu *et al.*, 2005) and may act as a redox-sensitive chaperone or a sensor for oxidative stress (Shendelman *et al.*, 2004; Zhou *et al.*, 2006), thereby protecting neurons against oxidative stress and cell death (Xu *et al.*, 2005).

A mutation in the DJ-1 gene, which leads to a substitution of leucine to proline at position 166 in the protein structure, has been associated with early onset Parkinson's disease (Bonifati *et al.*, 2003). Different studies have concluded that this mutation leads to destabilization of the protein structure that prevents the formation of a dimer (Olzmann *et al.*, 2004; Görner *et al.*, 2004; Shendelman *et al.*, 2004; Malgieri and Eliezer, 2008) but may lead to the formation of higher-order protein complexes (Macedo *et al.*, 2003; Baulac *et al.*, 2004). Here, we studied the effect of mutation L166P on DJ-1 by comparing the dRINs created from MD simulations of the wild-type and mutant structures.

Anderson and Daggett (2008) performed 7 MD simulations of 31 ns length (3 wild-type runs and 4 runs with the L166P mutation) starting with the wild-type structure with PDB identifier 1PDV (Tao and Tong, 2003). They concluded that L166P has a destabilizing effect on several structural elements important for protein stability and dimerization, including the region around Cys-106, a residue important for the proposed chaperone function of DJ-1.

We combined the data from the 7 runs to generate one dRIN for the wild-type and one for the mutant simulations. Since Anderson and Daggett (2008) observed that the L166P simulations have a broader distribution of conformations and C_α-RMSD values than the wild-type (in the last 10 ns), we also generated dRINs for the last 10 ns and compared them regarding different aspects. Both networks contain the same number of nodes (187) and similar number of edges (985 vs. 964). The distributions of edge frequencies for the last 10 ns of the simulations are very similar



(a) (b)

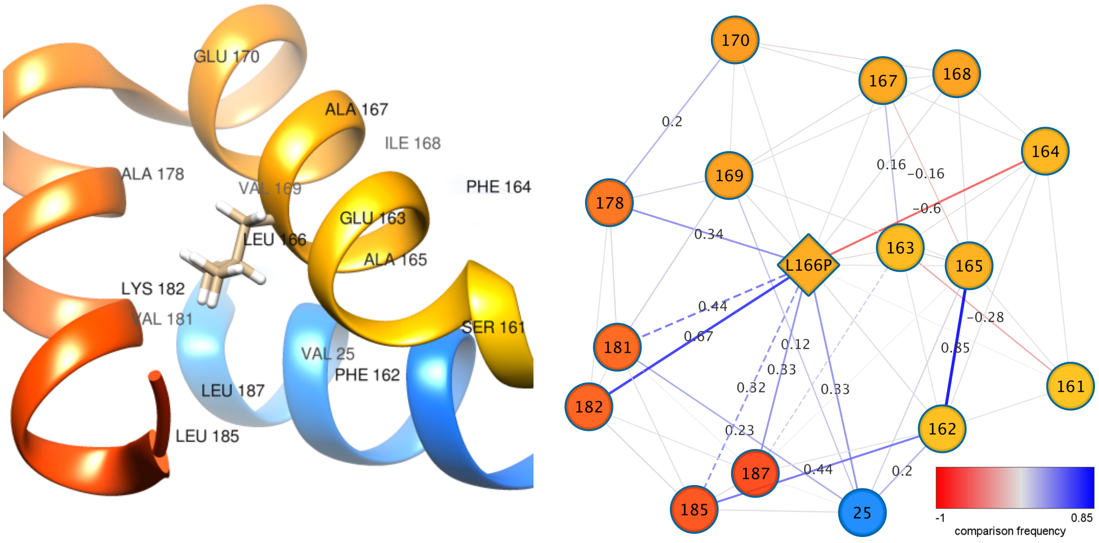
Figure 5.5: Ribbon representation of the DJ-1 structure (PDB identifier 1PDV) as shown in UCSF Chimera (a) and comparison dRIN visualized in Cytoscape (b). Rainbow coloring of the secondary structure elements was used for both ribbons and residue nodes. Residue 166 is highlighted by a stick representation in the structure and a diamond shape in the network view. Network edges correspond to non-covalent residue interactions that are more frequent in either the wild-type (blue line) or the mutant simulation (red line).

(Figures B.8 and B.9). So far these values do not support the expectations for a broader conformation distribution.

Furthermore, we performed a topological analysis on the networks with NetworkAnalyzer (Assenov *et al.*, 2008). We identified small differences between the wild-type and L166P dRINs for the following simple topological parameters: network radius (4 vs. 5), network diameter (7 vs. 8), characteristic path length (3.616 vs. 3.644), and average clustering coefficient (0.5 vs. 0.496). These results rather support the observations of Anderson and Daggett (2008) that the wild-type structures are more compact than the mutant ones.

In the next step of our analysis, we compared the wild-type and mutant dRINs for the whole simulations by computing the difference of edge frequencies. The resulting comparison network is shown in Figure 5.5(b). We color-coded the residues and network nodes based on secondary structure elements and highlighted L166P by a diamond shape to facilitate the visual exploratory analysis. At first sight, we can clearly notice a significant change in local residue interactions around the L166P mutation as indicated by the thick red and blue edges. Furthermore, there are a few edges throughout the whole structure that stand out and seem to be mostly located between nodes with different colors, e.g. belonging to different secondary structure elements. We focus on these two groups of edges for our further investigations.

A close-up on the region of residue 166 in the 3D structure and the comparison



(a) (b)

Figure 5.6: Close-up on the mutation L166P (a) in the DJ-1 wild-type structure (PDB identifier 1PDV) and (b) in the comparison dRIN. Rainbow coloring of the secondary structure elements was used for both ribbons and residue nodes. Residue 166 is shown as a stick representation in the structure and a diamond shape in the network view. Network edges correspond to non-covalent residue interactions that are more frequent in either the wild-type (blue line) or the mutant simulation (red line). Large difference in the frequency, e.g., above 0.1 or below -0.1, are shown by edge labels.

dRIN is shown in Figure 5.6. The high number of blue edges incident to residue 166 indicates that many of the interactions of this residue with its neighbors are primarily present in the wild-type simulations compared to the mutant simulations. In particular, these are interactions with residues in helices $\alpha 8$ or $\alpha 1$. At the same time, there is an increase of interactions of residue 166 with its neighbors in helix $\alpha 7$ upon mutation (red edges). Table 5.1 shows details for the nine edges with the largest difference in edge frequency, i.e., below -0.4 and above 0.4. Three of these edges involve residue 166 and two others are between its direct neighbors. The remaining four edges are located in two different regions of the network as can be seen in Figure 5.7 and are discussed in more detail below. Overall, these results support the hypothesis that the L166P mutation has a significant effect on the interactions between the secondary structure elements important for dimerization (helices $\alpha 1$, $\alpha 7$, $\alpha 8$).

Furthermore, we selected all edges with a comparison weight between -1.0 and -0.2 or between 0.2 and 1.0, i.e., corresponding to the interactions most frequent either in the mutant dRIN or in the wild-type dRIN, respectively, and highlighted them in the comparison network (Figure 5.7). In particular, residue 67, which is located in one of the central β-strands, switches its interaction partners from residues 70 and 91 in the wild-type simulation to residues 6 and 34 in the mutant simulation, respectively. These changes might lead to a disruption of the β-sheet structure and to a shift of helix $\alpha 1$. Such a shift is also likely to happen since the remaining

Table 5.1: Largest interaction differences between wild-type and mutant simulation

Interaction Edge		CompWeight	Edge Freq WT	Edge Freq Mut
A:162:_:PHE	A:165:_:ALA	0.85	0.98	0.13
A:166:_:LEU	A:182:_:LYS	0.67	0.86	0.19
A:166:_:LEU	A:181:_:VAL	0.44	0.44	0.00
A:162:_:PHE	A:185:_:LEU	0.44	0.44	0.00
A:89:_:LYS	A:114:_:ALA	-0.45	0.01	0.45
A:67:_:TYR	A:6:_:ALA	-0.47	0.35	0.82
A:34:_:THR	A:67:_:TYR	-0.59	0.32	0.91
A:166:_:LEU	A:164:_:PHE	-0.60	0.13	0.72
A:89:_:LYS	A:116:_:GLU	-1.00	0.00	1.00

residues from helix $\alpha1$ shown in this network (dark blue nodes) have interactions with other helices only in the wild-type simulations (blue edges to orange and green nodes). Another region with a large change in interaction frequency is located around residue 89. It makes more contacts with residues from the neighboring helix in the mutant simulation than in the wild-type. One possible explanation is that the whole structure widens slightly across an axis parallel to helices $\alpha1$ and $\alpha7$ as can be expected by a substitution to a proline in one of these helices (see Figure 5.5(a)).

Anderson and Daggett (2008) investigated the effects of the L166P substitution solely based on the data from the MD simulations. Thus, they focused on properties such as the distribution of Cα-RMSD deviations and fluctuations, stability of secondary structure elements (mostly helices), atomic contacts between selected residues as well as interactions on the dimer interface. They observed that the L166P mutation causes increased backbone mobility and a loss of helical content, which in turn most probably affects the formation of the dimer. Our topological analysis and comparison of the wild-type and mutant dRINs agrees with these findings, although we did not consider any physico-chemical properties of the residues and how they change after the mutation. Enriching the comparison dRIN with information of the change of residue properties would be a good direction for future investigations.

In addition, Anderson and Daggett (2008) discussed the role of residue 106 for the function of DJ-1 and the disruption of its interactions as having a significant effect. In this case, we could not draw any conclusions about the effect of the interaction changes on residue 106 from the dRIN comparison as its interactions do not change significantly between the two simulations. This might be a result of our data preprocessing step, where we combined the three simulation runs for the wild-type and mutant structure into one dRIN, respectively. It is likely that investigating the simulations runs one by one will reveal more detailed information about the changes in interactions.

Overall, we demonstrated that our RIN-based approach for comparison of MD simulations can be successfully applied to investigate the effect of mutations on protein structure. Not all of our conclusions overlap with those drawn from a typical MD

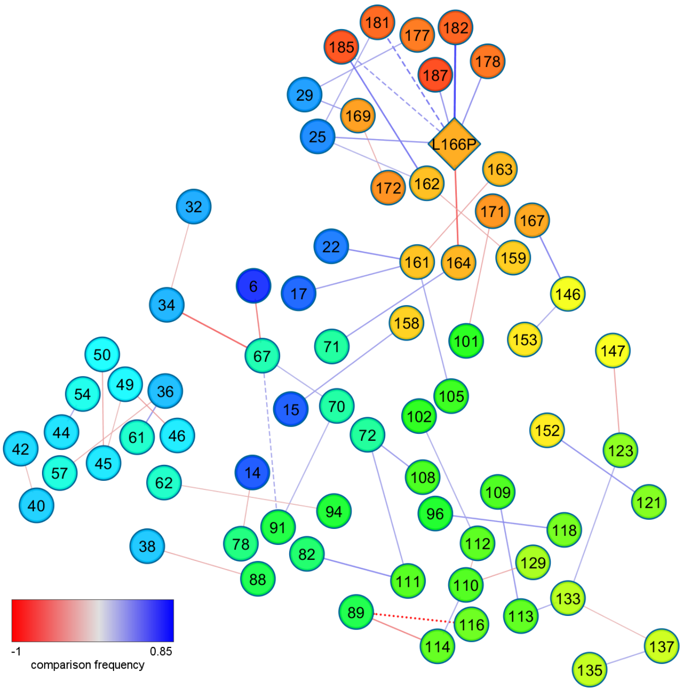


Figure 5.7: Interaction differences between wild-type and mutant simulation shown as edges with large difference in frequencies (above 0.2 and below -0.2). Network edges correspond to non-covalent residue interactions that are more frequent in either the wild-type (blue line) or the mutant simulation (red line). Network nodes are colored according to the secondary structure elements they belong to and residue 166 is highlighted by a diamond shape.

simulation analysis (Anderson and Daggett, 2008), but we are able to provide additional insights on the molecular mechanisms upon mutation in DJ-1 as well as easily explorable and interpretable visualizations.

5.4 Analysis of docking structures

Despite the large number of docking approaches and tools available nowadays, state-of-the-art methods include a near-native docking model, i.e., a good prediction, in the top 10 only in 30-40 % of the cases (Schneidman-Duhovny *et al.*, 2012). It has been observed that even if the overall predicted complex quality is not acceptable, many methods generate models with a native interface (up to 24 % based on an evaluation of the predictions for 20 CAPRI targets) (Lensink and Wodak, 2010). This suggests that predictions at the top ranks are more enriched in correct interfaces than in correct complexes.

Therefore, we propose to regard the set of docking solutions as an ensemble of different conformations of the same protein complex. In this way, we can apply the same dynamic RIN methodology as for MD simulations to find similarities and differences among the top-ranked docking models. In particular, we focused on the identification of the most frequent interactions and residues in the interface of the docked structures and the comparison with the true (native) interface. To evaluate our approach, we performed a large-scale analysis of three benchmark sets of different sizes and generated by different docking algorithms. We also reimplemented the recently released CONSRANK score for ranking docking solutions based on the contact (interaction) frequencies (Oliva *et al.*, 2013; Vangone *et al.*, 2013) and compared our results to it, where applicable.

5.4.1 Experimental setup

Essentially, each benchmark set comprises a number of decoys (docking solutions) for a set of targets (protein complexes). For each target, we have the structure of the native (true/correct) complex and among the decoys, there are a number of near-native models. The correctness of a solution is usually evaluated by computing the root-mean-square deviation (RMSD) of the target and decoy. Thereby, we distinguish between RMSD of the backbone atoms of the ligand (L_{RMSD}) and RMSD of the backbone atoms of the interface residues (I_{RMSD}). Many other properties, such as the fraction of native (f_{nat}) and non-native residue-residue contacts ($f_{non-nat}$) in the predicted and native complex, can also be considered in the evaluation (Mendez *et al.*, 2005).

In order to focus on the docking interfaces, for each decoy structure, we create an interface RIN (iRIN), which contains only non-covalent residue interactions between residues that belong to either of the two docked structures. For each target, we also create a dynamic interface RIN (diRIN), in which the residue interactions are weighted by their frequency of occurrence in all the decoy interfaces. Then we compare the diRIN to the correct (target) interface RIN. For the sake of simplicity, we only considered *combi* edges, which indicate a non-covalent interaction of any type between two residues. In order to directly compare our results to CONSRANK, we also created RINs with *dist* edges, where two nodes are connected by an edge if any of their atoms are closer than 5 Å.

Assuming we have enough near-native interfaces among our solutions, we expect that the most frequent residue interactions in the diRIN will correspond to the interactions in the interface of the target. To assess the prediction performance of our methods, we generate the receiver operating characteristic (ROC) curve with the frequency of occurrence as ranking criteria (labeled as *Ranked Int*). Thereby, at each rank, we plot the fraction of true predicted interactions out of the interactions above the current rank (TPR = true positive rate) vs. the fraction of false predicted interactions out of the interactions below the current rank (FPR = false positive rate). Here, we consider as true only the interactions that were successfully mapped between the target and decoy interfaces.

In addition, we perform the same analysis for the residues in the interface ranked by their involvement in the most frequent interactions (labeled as *Ranked Res*). This is less discriminative than looking at the individual residue interactions because the true interface residues might actually be in the top solutions, but their correct

Table 5.2: Summary of docking benchmark sets. (avg = average)

Benchmark set	Targets	Decoys (total)	Decoys (avg)	Near-native (avg)
DOCKGROUND	61	6605	108.3	9.7
CAPRI	6	1833	305.5	20.7
PatchDock	176	35200	200	2.0

interaction partners are not identified due to a small shift in the position.

Finally, we re-rank the docking solutions based on their ability to match the most frequent interactions (labeled as *Ranked Mod*) as suggested by CONSRANK (Oliva *et al.*, 2013) and as explained above. Here, we do not compare the decoy interface to the target interface based on residue interactions, but use the RMSD criteria to denote predicted near-native solutions as true positives.

Since docking solutions are usually ranked according to method-specific criteria, we make use of this ordering to analyze the *top 10, 25, 100* or *200/All* available solutions. Thereby, we aim at finding the smallest group of top solutions that contains enough information to reconstruct the near-native docking interface.

5.4.2 Benchmark datasets

We chose three different benchmark sets: DOCKGROUND (Liu *et al.*, 2008), CAPRI (Janin, 2007, 2010), and PatchDock (Schneidman-Duhovny *et al.*, 2005; Hwang *et al.*, 2010). Table 5.2 lists the basic characteristics of the three sets. The CAPRI dataset is very small with only 6 targets, but contains the highest number of near-native models. The DOCKGROUND dataset is more diverse and consists of 62 targets. We chose these two datasets because they were used for the evaluation of the CONSRANK (Oliva *et al.*, 2013) approach. In addition, we analyzed a larger dataset comprising docking predictions generated by the PatchDock server for 176 targets of varying difficulty. For this dataset, we focused on the identification of interface residues and interactions given different sets of top solutions.

DOCKGROUND benchmark

The DOCKGROUND benchmark set comprises 61 complexes from the Dockground unbound benchmark set (Liu *et al.*, 2008). For each complex, the top 100 non-native decoys with the highest surface complementarity scores as well as at least one near-native solution were included. As for the other datasets, we considered all models with $L_{RMSD} \leq 10$ Å as near-native. The docking decoys were built by the GRAMM-X docking web server (Tovchigrechko and Vakser, 2006). The underlying method first employs an FFT-based global search algorithm on a fine grid with a projection of a smoothed Lennard-Jones potential function to account for conformational changes. Then a rigid body minimization with the same potential is performed to select one representative prediction for each local minimum. Finally, the remaining models are re-scored based on a combination of empirical and standard force-field energy terms (Tovchigrechko and Vakser, 2005). The benchmark set is available for download from http://dockground.bioinformatics.ku.edu/.

CAPRI models

The CAPRI set consists of 6 targets (T24, T25, T26, T29, T32, and T36) with at least one prediction of acceptable quality (Janin, 2007, 2010). For each target, individual groups that develop docking procedures or automatic web servers can submit up to 10 predictions of the three-dimensional structure of the interacting proteins. Between 30 and 40 predictor groups usually participate in each round and, for the selected targets, there were 37, 37, 37, 39, 36, and 32 groups, respectively.

T24 and T25 were the complex between Arf1, a small G-protein, and the Arf1 binding domain (ArfBD) of ARHGap21 (Menetrey *et al.*, 2007). In T24, ArfBD was built by homology, while in T25, the bound version from the complex was used. As expected, the prediction results were quite poor for T24 and improved significantly improved for T25 (Lensink *et al.*, 2007). T26 was the TolB-Pal complex associated with the maintenance of the *E. coli* outer membrane (Bonsor *et al.*, 2007). Although the interface area is large and well buried, quite a few predictor groups submitted acceptable solutions, but, on average, the quality was not as good as for the previous target (Lensink *et al.*, 2007).

T29 was the Trm8/Trm82 complex that carries out guanine methylation in the tRNAs of the yeast *S. cerevisiae* (Leulliot *et al.*, 2008). T32 was a complex between the subtilisin Savinase, a microbial serine protease, and the α-amylase subtilisin inhibitor BASI (Micheelsen *et al.*, 2008). T36 was the association between the GH10 and CBM22 domains of the Xyn10B xylanase, an enzyme involved in plant cell wall degrading (Najmudin *et al.*, 2010). As can be seen from the number of near-native models in Table 5.2, the results for T29 and T32 were quite good and for T36 very poor, since the predictors failed to reproduce the association mode of the covalently linked domains (Lensink and Wodak, 2010).

The submitted predictions are classified by the CAPRI assessment (Mendez *et al.*, 2003; Lensink *et al.*, 2007) into the following quality categories: *high*, *medium*, *acceptable*, *incorrect*, and *clashes*, whereas the last group is excluded from the method evaluation. We considered solutions in one of the three categories *high*, *medium*, *acceptable* as near-native. This means that they have $f_{nat} \geq 0.1$ and $L_{RMSD} \leq 10$ Å or $I_{RMSD} \leq 4$ Å. The data is available for download from ftp://ftp.ebi.ac.uk/pub/databases/msd/capri/. In addition, we excluded models for which the residue identifiers cannot be mapped well to the target.

PatchDock models

This dataset contains the 176 targets from the Benchmark set 4 (Hwang *et al.*, 2010) and the top 200 decoys generated for each target by the PatchDock web server (Schneidman-Duhovny *et al.*, 2005). PatchDock employs a local shape feature matching technique for searching the conformational space (Schneidman-Duhovny *et al.*, 2003). Then the predictions are scored based on geometric fit and atomic desolvation energy. Finally, redundant solutions are excluded using RMSD-based (default threshold 4 Å) clustering. Since PatchDock is a geometry-based approach, it might rank wrong solutions higher than the correct one if they have significantly larger geometrical compatibility.

In contrast to the other benchmark sets considered in this work, this dataset contains on average only 2 near-native models within the best 200 predictions per target. This benchmark set was originally generated by Dina Schneidman-Duhovny who shared it with us in the course of our cooperation. For each target, we used up to several thousands of automatically generated decoys. However, because of the large number of targets, we considered only the top 200 decoys. We defined a solution as near-native if $L_{RMSD} \leq 10.0$ Å or $I_{RMSD} \leq 4.0$ Å.

5.4.3 Related method

The first approach to study the consensus of docking solutions, CONS-COCOMAPS, was introduced by Vangone *et al.* (2012). It relies on the idea of contact conservation defined as the count of residue contacts that are shared by more than one predicted complex. A contact is defined for a residue pair if any pair of their atoms is closer than a distance cut-off of 5 Å. The authors introduce the residue conservation rate CR as

$$CR_{kl} = \frac{nc_{kl}}{N}$$

where nc_{kl} is the number of models, where residues k and l are in contact, and N the total number of docking models. This score corresponds to our definition of interaction frequency. In the paper, the authors focused on the contact conservation score between two or more docking models as well as on the amount of inter-residue contacts that are conserved in a given fraction of models.

As a benchmark set, Vangone *et al.* used the CAPRI dataset introduced in the previous section. As expected, they observed that the pair-wise conservation score between individual models decreases with decreasing L_{RMSD} and that the inter-residue conservation rate for the models submitted by the same predictor is very variable (Vangone *et al.*, 2012). Furthermore, the success of the predictor might correlate with high conservation rate, but not necessarily. Most importantly, the authors noticed that several of the most conserved contacts correspond to native contacts. In contrast to our work, they did not comprehensively investigate whether the conservation rate can be used for the prediction of interface contacts and residues.

Building upon CONS-COCOMAPS, Vangone *et al.* (2013) and Oliva *et al.* (2013) conceived the CONSRANK algorithm for scoring docking solutions based on the conservation rate. The authors defined a score of average conservation of inter-residue contacts for each docking model i:

$$\overline{S}_i = \frac{\sum_1^{M_i} CR_{kl}}{M_i}$$

where M_i is the total number of inter-residue contacts in model i. Then the docking solutions are re-ranked based on this score. Vangone *et al.* evaluated their approach on three different datasets, two of which, CAPRI and DOCKGROUND, are also included in our evaluation. Oliva *et al.* demonstrated the performance of the approach on an extended set of CAPRI targets.

Both publications show that the method performs well on average according to the AUC (0.758, 0.654, 0.870, 0.819) and number of NL solutions in the top 10 rank positions (4.8, 1.7, 6.7, 5.0) for the RosettaDock, DOCKGROUND, CAPRI, and extended CAPRI dataset, respectively. Based on the results, Vangone *et al.* concluded that the approach performs better on diverse sets of docking solutions such as the CAPRI one (Vangone *et al.*, 2013). They also observed that the number of near-native models strongly influences the performance of the method, but does not clearly relate to the average conservation score $\overline{S}_i$ of the top solution. Furthermore, the authors showed that the cut-off distance used to define residue contacts did not have an effect on the performance. In addition, Oliva *et al.* demonstrated that, on the extended CAPRI dataset, CONSRANK significantly outperforms an RMSD-based consensus approach that ranks the models based on their average pair-wise similarity (measured as L_{RMSD}) to all other models (Oliva *et al.*, 2013).

5.4.4 Results for DOCKGROUND set

In the following, we will describe and discuss the results we obtain by analyzing the DOCKGROUND benchmark sets using dynamic RINs. We generated diRINs (dynamic interface RINs) from the *top 10, 25, 100* and *all* decoys of each target. It is important to note that the difference between *top 100* and *all* is that *all* contains near-native models that did not rank within the top 100 solutions according to the docking criteria. We considered two types of edges: *combi*, any non-covalent residue interaction, and *dist*, distance between the residues below 5 Å. Then we ranked the interactions and residues in the diRINs based on their frequency of occurrence and assessed how good the ranking performs in terms of identifying the true interface interactions and residues. Finally, we ranked the individual iRINs, e.g., the interface RIN for each decoy, based on the fraction of most frequent interactions they contain. We evaluated the performance of our dRIN-based method to rank near-native decoys on top of the list and compared the results to CONSRANK.

The results of our analyses on the DOCKGROUND set are summarized in Table 5.3 and Figure 5.8. We observe good performance of our method for ranking interactions, residues or models with best AUC values of 0.801 ± 0.162, 0.747 ± 0.125, and 0.658 ± 0.282, respectively. Notably, the standard deviation, in particular for lower N is very high. This might be explained by the fact that, for some targets, the docking approach performs well and ranks solutions with near-native interface on

Table 5.3: Average ranking performance on the DOCKGROUND set for the 61 targets and *top 10, 25, 100*, and *all* decoys available for each target as well as the number of all models and near-native models considered for each group. The performance for ranking models using *combi* and *dist* edges is given separately. The AUC values for ranking interactions, residues, or models for each target are plotted in Figures B.10, B.11, and B.12.

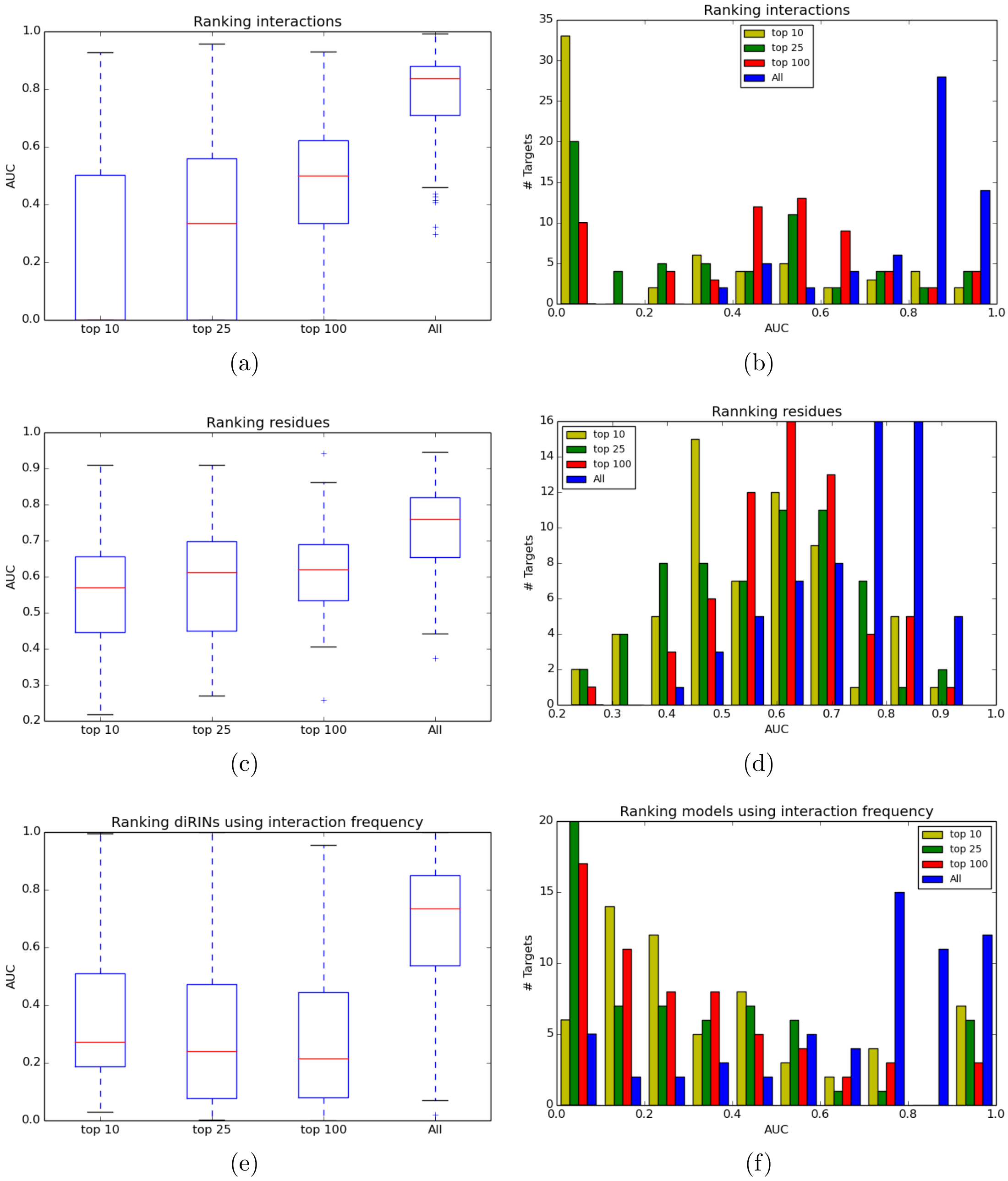
	average (stdev)			
	All models	top 100	top 25	top 10
# Models	108.279 (3.282)			
# Near-native	9.656 (7.305)	2.770 (7.263)	0.705 (2.929)	0.377 (1.380)
AUC Ranked Mod combi	0.658 (0.282)	0.288 (0.262)	0.325 (0.291)	0.379 (0.284)
AUC Ranked Mod dist	0.656 (0.281)	0.297 (0.259)	0.331 (0.302)	0.386 (0.300)
AUC CONSRANK	0.654 (0.281)			
AUC Ranked Int combi	0.779 (0.169)	0.472 (0.260)	0.342 (0.309)	0.262 (0.315)
AUC Ranked Int dist	0.801 (0.161)	0.491 (0.233)	0.367 (0.288)	0.321 (0.337)
AUC Ranked Res combi	0.732 (0.126)	0.622 (0.126)	0.576 (0.155)	0.560 (0.154)
AUC Ranked Res dist	0.747 (0.125)	0.639 (0.135)	0.604 (0.164)	0.576 (0.155)

high ranks, while for others it fails. Since our method is based on the frequency of interactions in the given set of docking decoys, it is likely that we find the wrong interface if it is more frequently contained in the incorrect solutions on top of the ranking list.

As can be expected, the number of near-native models increases with the threshold (top N), but it is noticeable that there are almost no near-native models in the top 10 / 25 solutions (Table 5.3). From an average of 9.7 near-native models in all decoys, there are only 2.8 in the top 100 decoys, and less than 1 (0.4 and 0.7) in the *top 10* and *25*, respectively. Such low numbers of near-native models are one of the reasons why the performance of our method for the *top 10, 25 and 100* models is significantly lower than for *all* solutions as also shown in the CONSRANK publication (Vangone *et al.*, 2013).

We observe that our method performs as well as CONSRANK (AUC 0.654) for both *dist* (AUC 0.656) and *combi* edges (AUC 0.658). Although the performance of our dRIN-based method for ranking decoys is slightly better with *combi* edges, it appears to be significantly better with *dist* edges for ranking interactions (AUC increases from 0.784 to 0.801) and it remains the same for residues (AUC 0.747). One possible interpretation is that distance edges are less specific than non-covalent interactions and thus better for matching the true interface interactions, but not for ranking the decoys. In contrast, the residue score, which is the average interaction frequency, is not affected by the type of edges. In the following, we focus on the results with *combi* edges as they are in general more informative and are an integral part of our method.

We compared the performance of our ranking method on the different sets of docking



(a) (b) (c) (d) (e) (f)

Figure 5.8: Box-and-whisker plot and distribution of AUC values for ranking interface interactions ((a) and (b)); interface residues ((c) and (d)); and models using interaction frequency ((e) and (f)) in the DOCKGROUND set. In each case, the AUC values are computed for all 61 targets and the *top 10, 25, 100* and *All* decoys available for each target.

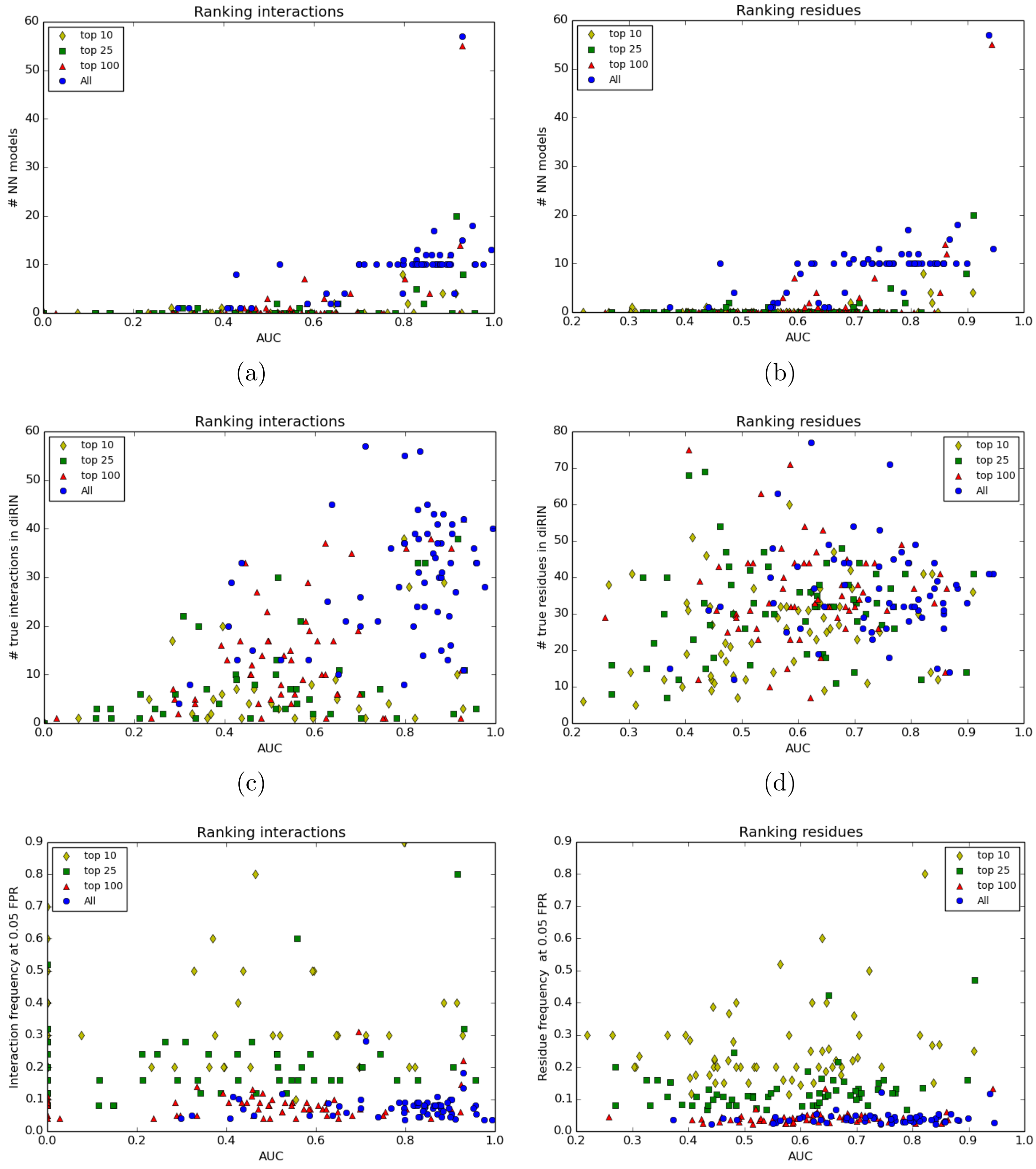
models (*top 10, 25, 100* and *all*). In contrast to our expectations, even the top 100 decoys do not seem to be enough to accurately identify the true interface interactions or rank the models based on the interaction frequency. This is not the case for ranking residues, but the performance is still not very convincing with AUC values of 0.560, 0.576 and 0.622 for *top 10, 25*, and *100* decoys, respectively. Here, we also observe high variance in the AUC values, which indicates that there are a few outliers for which the ranking succeeds exceptionally well. For ranking residues or interactions, the results improve with increasing number of decoys considered for the analysis (See Figures 5.8(a) and 5.8(c)). As expected, the more models are available, the easier it becomes to gather enough information about the true interface and identify it correctly.

Interestingly, this is not the case for ranking models, where the performance of our method slightly drops when more decoys are included in the analysis and drastically increases when all models are considered (Figure 5.8(e) and 5.8(f)). The main difference between the *top 100* and the *All* set is that the former contains less near-native models than the latter as can be easily recognized in Figure 5.8(b), Figure 5.8(d), and Figure 5.8(f). For both ranking interactions and residues, we see an increase in AUC values when the near-native models are included, but the change is more significant for the former. Thus, the ranking of models and interactions is effected by the ratio of near-native models, while the residues method is more robust in this regard. We should also keep in mind that this behavior is partially caused by the uneven number of near-native models compared to the incorrect models.

Furthermore, we studied the effect of near-native models, interface size and frequencies on the performance of our dRIN-based method (Figure 5.9). The plots of number of near-native models versus AUC values (see Figures 5.9(a) and 5.9(b)) indicate that the performance for identifying interface interactions and residues improves for targets with more near-native models, but there is no clear correlation between them, especially for residues. As expected, we also observe that even if the number of near-native models is very low, some models still contain enough information about the true interface, and AUC values can be as large as 0.9.

In addition, we plotted the number of true interface interactions and residues in the diRINs (see Figures 5.9(c) and 5.9(d)) versus AUC. For ranking interactions, we can see a clear trend of improvement when increasing the number of true interactions. However, this is not the case for ranking residues since the number of true residues does not seem to influence the performance of our method. We also investigated the dependency between the interaction and residue frequency scores at 0.05 FPR (top 5 % of the data) and the AUC values (see Figures 5.9(e) and 5.9(f)). Here, we can distinguish quite well between the different groups (*top 10, 25, 100, All*), since the scores decrease with the increase of the number of models considered for the analysis. However, no clear trend for the success of the approach emerges. Finally, we checked for a relationship between the number of residues / interactions in the diRIN and the performance of our ranking method, but did not observe any particular effect (data not shown).

More details for individual targets when considering *all* models are shown in Table 5.4. Here, we can clearly identify the targets, for which the performance of our method for ranking decoys using interaction frequency is worse as the ones that have the lowest number of near-native models. Of course, this is partly caused by the evaluation procedure and the unbalance between correct and incorrect cases, which are needed for the computation of TPR and FPR. If we exclude the entries



(a) (b)

(c) (d)

(e) (f)

Figure 5.9: Dependencies between performance for ranking interactions or residues and the number of near-native models per target (a) and (b), the number of true (target) interface interactions (c) and residues (d) in the diRIN, and the interaction (e) and residue (f) frequency scores at 0.05 FPR (top 5 %) in the DOCKGROUND set.

Table 5.4: Ranking performance on the DOCKGROUND set per structure including the number of models considered for the analysis as well the average number of residues (Res) and interactions (Int) in the target interface. Average values in parentheses were obtained for targets with > 2 near-native models (NM).

							AUC Ranked			AUC Ranked	
PDB	Chains	# Models	# NM	Res	Int	Int	Mod	Mod	CONSRANK	Int	Res
				combi	combi	dist	combi	dist	dist	combi	combi
1a2k	AB:C	102	2	33	38	48	0.350	0.295	0.295	0.585	0.556
1a2y	AB:C	110	10	32	31	44	0.557	0.541	0.536	0.701	0.462
1akj	AB:DE	110	10	38	35	60	0.898	0.893	0.894	0.739	0.842
1avw	A:B	110	10	40	49	71	0.681	0.660	0.656	0.848	0.846
1bth	LH:P	101	1	49	58	92	0.180	0.180	0.190	0.436	0.555
1bui1	B:C	110	10	26	28	57	0.865	0.860	0.854	0.879	0.714
1bui2	A:C	110	10	32	23	42	0.843	0.818	0.822	0.871	0.801
1bvn	P:T	110	12	44	43	73	0.748	0.766	0.770	0.879	0.682
1cho	E:I	110	15	14	11	18	0.731	0.673	0.688	0.930	0.868
1dfj	E:I	109	10	45	37	68	0.812	0.775	0.774	0.874	0.842
1e96	A:B	110	10	26	27	40	0.836	0.839	0.839	0.896	0.858
1ewy	A:C	110	10	29	22	45	0.603	0.638	0.624	0.700	0.716
1ezu	AB:C	110	10	56	68	102	0.190	0.190	0.190	0.711	0.745
1f51	AB:E	110	10	44	38	62	0.720	0.717	0.735	0.768	0.690
1f6m	A:C	110	10	39	39	56	0.324	0.367	0.360	0.525	0.628
1fm9	A:D	110	13	41	40	66	1.000	1.000	1.000	0.993	0.945
1g20	AB:EF	110	10	78	83	137	0.737	0.756	0.746	0.833	0.624
1g6v	A:K	108	8	26	27	45	0.361	0.313	0.286	0.428	0.603
1gpq	A:D	110	10	33	37	54	0.924	0.917	0.911	0.956	0.899
1gpw	A:B	110	18	37	37	68	0.990	1.000	1.000	0.952	0.883
1he1	A:C	110	13	37	39	56	0.967	0.981	0.979	0.829	0.743
1he8	A:B	101	1	17	16	30	0.000	0.000	0.000	0.323	0.373
1hxy	AB:D	102	2	20	17	35	0.235	0.145	0.145	0.652	0.637
1jps	LH:T	110	10	30	35	53	0.830	0.826	0.829	0.959	0.858
1ku6	A:B	110	10	45	50	74	0.738	0.708	0.706	0.797	0.769
1l9b	LMH:C	110	10	23	16	38	0.567	0.563	0.557	0.899	0.731
1ma9	A:B	110	10	71	65	112	0.850	0.857	0.849	0.798	0.763
1nbf	A:D	110	10	49	47	86	0.683	0.714	0.713	0.785	0.807
1ook	AB:G	104	4	25	22	31	0.538	0.570	0.552	0.667	0.580
1oph	A:B	110	10	35	43	62	0.911	0.896	0.888	0.904	0.857
1p7q	AB:D	104	4	38	34	54	0.418	0.440	0.440	0.628	0.683
1ppf	E:I	110	10	32	39	51	0.988	0.992	0.992	0.871	0.767
1r0r	E:I	110	12	34	44	70	0.706	0.720	0.722	0.903	0.813
1r4m	AB:I	101	1	55	63	95	0.070	0.070	0.070	0.414	0.654
1s6v	A:B	104	4	12	8	15	0.405	0.390	0.390	0.797	0.486
1t6g	A:C	110	57	41	43	69	0.955	0.951	0.951	0.929	0.938
1tmq	A:B	110	10	43	43	75	0.558	0.574	0.586	0.820	0.599
1tx6	A:I	110	10	45	50	76	0.506	0.483	0.482	0.871	0.663
1u7f	A:B	110	10	33	32	51	0.839	0.849	0.847	0.841	0.761
1uex	AB:C	101	1	32	27	40	0.090	0.210	0.210	0.461	0.646
1ugh	E:I	110	12	47	50	82	1.000	1.000	1.000	0.849	0.783
1w1i	A:F	104	4	32	34	56	0.775	0.743	0.742	0.911	0.787
1wej	LH:F	110	10	16	17	21	0.735	0.727	0.729	0.894	0.846
1wq1	R:G	110	12	44	44	91	0.716	0.693	0.697	0.861	0.796
1xd3	A:B	110	10	25	26	40	0.757	0.720	0.722	0.817	0.730
1xx9	A:CD	102	2	69	82	133	0.295	0.265	0.255	0.638	0.563
1yvb	A:I	110	10	29	30	50	0.650	0.650	0.650	0.878	0.819
1zy81	AB:K1	110	10	31	32	52	0.904	0.956	0.999	0.881	0.813
1zy82	AB:K2	110	10	31	32	47	0.988	0.996	0.955	0.976	0.858
2a5t	A:B	101	1	40	40	54	0.020	0.090	0.090	0.408	0.550
2bkr	A:B	110	11	54	62	90	0.782	0.787	0.788	0.828	0.697
2bnq	AB:DE	101	1	31	28	44	0.000	0.000	0.000	0.299	0.441
2btf	A:P	110	10	43	46	78	0.989	0.998	0.998	0.900	0.743
2ckh	A:B	110	10	26	25	37	0.726	0.719	0.722	0.827	0.756
2fi4	E:I	110	10	32	41	59	0.860	0.837	0.837	0.828	0.806
2goo	A:C	110	10	28	29	40	0.816	0.738	0.734	0.842	0.796
2kai	AB:I	110	11	34	42	64	0.713	0.708	0.700	0.798	0.724
2sni	E:I	110	10	35	46	71	0.729	0.700	0.692	0.863	0.834
3fap	A:B	110	10	18	15	26	0.738	0.737	0.734	0.839	0.761
3pro	A:C	110	17	44	44	83	0.926	0.966	0.967	0.866	0.795
3sic	E:I	110	10	40	50	74	0.816	0.820	0.817	0.884	0.880
average		108.279	9.656	36.525	38.016	60.869	0.658 (0.748)	0.656 (0.745)	0.654 (0.743)	0.779 (0.833)	0.732 (0.763)
stdev		3.282	7.305	36.192	15.447	24.790	0.282 (0.187)	0.281 (0.189)	0.281 (0.191)	0.169 (0.105)	0.126 (0.103)

with only 1 or 2 near-native models (*1a2k, 1bth, 1he8, 1hxy, 1r4m, 1uex, 1xx9, 2a5t, 2bnq*), we observe an increase of AUC from 0.658 to 0.748 for our approach and from 0.654 to 0.743 for CONSRANK. Still there are three outliers (*1ezu, 1f6m, 1g6v*), for which both approaches for ranking the docking models perform poor (AUC values below 0.5) even though the number of near-native models is as high as for most other targets (between 8 and 10). In contrast, for three targets (*1fm9, 1gpw, 1ugh*), the near-native models are ranked on top, resulting in an AUC of 1.0. We show the distribution of decoy I_{RMSD} for these six targets in Figure B.13. The main difference is that, for the unsuccessful targets, there are almost no other models with low I_{RMSD} apart from the near-native models, while there are many with high I_{RMSD}, e.g., the decoys are biased towards a wrong solution.

For ranking interactions and residues, AUC increases only by 0.05 and 0.03, respectively, when excluding models with low number of near-native models. This is expected given the conclusions from Figure 5.8 and 5.9 that the performance of our method for ranking interactions and residues increases with the number of models considered for the analysis and is not influenced by the fraction of near-native models to such an extent as for re-ranking the models. We also see that our approach performs better than a random predictor on most targets except those with only 1 near-native model.

Furthermore, Table 5.4 contains the number of interactions in the target interface defined by our approach using non-covalent interactions (*Int combi*) and by CONSRANK using only distance cutoff (*Int dist*), whereas the latter are significantly larger. In their paper, the authors of CONSRANK demonstrate that the distance cutoff does not have an effect on the performance of their method. Here, we observe the same trend when comparing our results to theirs. However, the clear difference in the number of interface interactions means that our approach focuses on closer, more reliable contacts between the residues.

5.4.5 Results for CAPRI models

As a proof of concept, we also analyzed the small CAPRI dataset. The results are shown in Table 5.5. The first striking difference to the DOCKGROUND set is the increased performance of our method for ranking models using the interaction frequency, with an AUC of 0.872, which also exceeds the CONSRANK AUC value of 0.870. The performance of our method is slightly better for ranking interactions (AUC 0.880) and significantly better for ranking residues (AUC 0.914). In addition, the average number of near-native models is 20.667 and of all models 305.5.

In Table 5.6, we have listed the AUC values for the single targets. As expected from the previous results, we observe a clear difference in the performance of our method for targets with low (T24, T36) and high number (T25, T26, T29, T32) of near-native models. However, these results are also in good agreement with the notes of the CAPRI team, who observed poor prediction performance of the participating groups for T24 and T36 and thus, declared them as difficult targets (Lensink *et al.*, 2007). In the case of T24, the ligand was a model built by homology. T36 was

Table 5.5: Average ranking performance on the CAPRI set for the 6 selected targets including average number of models considered for the analysis as well as the average number of near-native models.

	# Models	# Near-native	AUC Ranked Mod	AUC CONSRANK	AUC Ranked Int	AUC Ranked Res
average	305.500	20.667	0.872	0.870	0.880	0.914
stdev	35.365	15.565	0.188	0.210	0.126	0.090

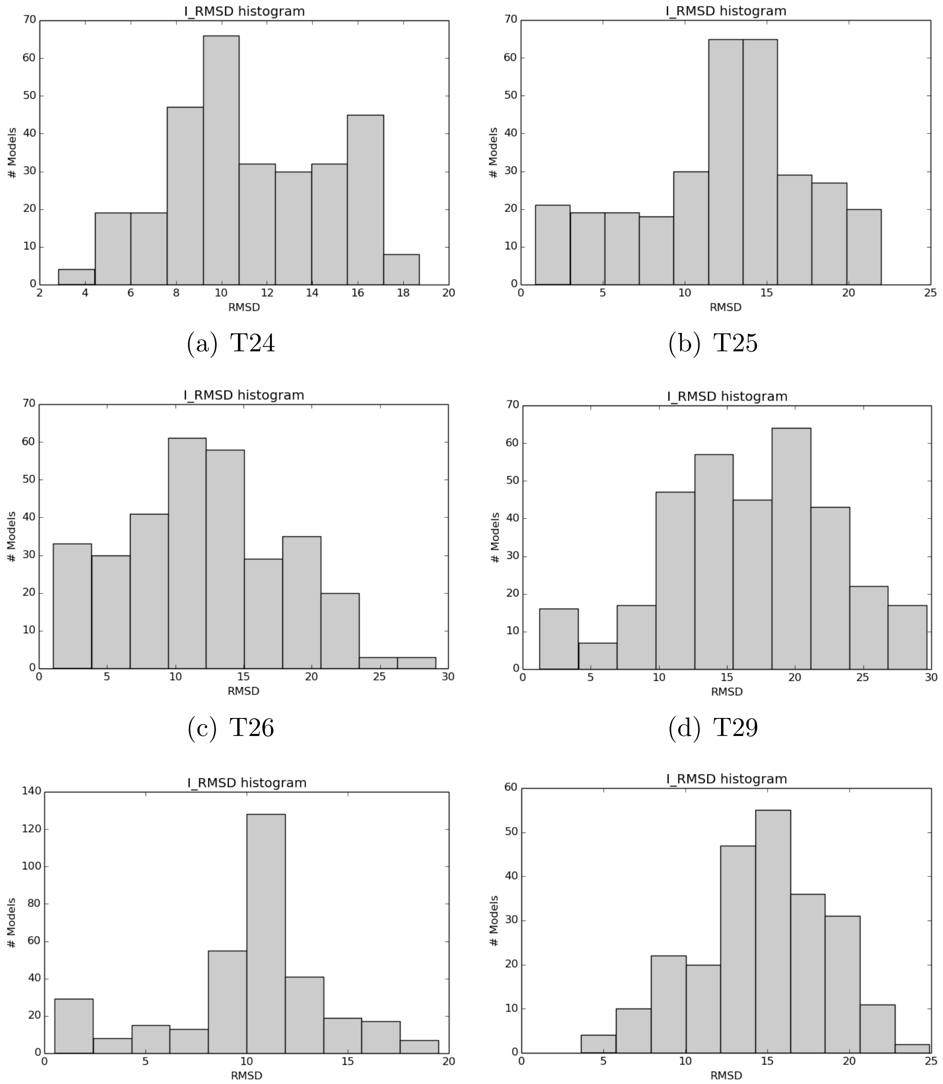
Table 5.6: Ranking performance on the CAPRI set per structure including number of models considered for the analysis as well as the average number of residues (Res) and interactions (Int) in the target interface. Values in parentheses were obtained after excluding near-native models from the analysis.

					AUC		AUC Ranked	
Target	# Models	# Near-native	Res	Int	Ranked Mod	CONSRANK	Int	Res
T24	302	4	33	28	0.753	0.811	0.801 (0.779)	0.874 (0.871)
T25	313	34	33	28	0.988	0.990	0.984 (0.825)	0.984 (0.961)
T26	313	34	53	55	0.984	0.986	0.942 (0.795)	0.973 (0.962)
T29	335	17	37	33	0.995	0.997	0.944 (0.682)	0.944 (0.847)
T32	332	34	46	50	0.974	0.969	0.951 (0.822)	0.963 (0.913)
T36	238	1	37	35	0.540	0.467	0.659 (0.596)	0.748 (0.737)
average	305.500	20.667	39.833	38.167	0.872	0.870	0.880 (0.750)	0.914 (0.882)
stdev	35.365	15.565	8.010	11.548	0.188	0.210	0.126 (0.092)	0.090 (0.085)

a complex of two covalently bound domains with a possibly unstable interface. Plotting the distribution of I_{RMSD} values for each target individually (Figure 5.10) also reveals that the fraction of submitted models with low interface RMSD is smaller for T24 and T36 than for the other targets.

Although the overall performance of our approach and CONSRANK is very similar, there are also some differences. The latter performs slightly better on three targets, significantly better on one, and worse on two. There are at least two factors that play a role. On the one hand, CONSRANK uses a distance cut-off of 5 Å to define edges, while we use non-covalent interactions, which is more stringent and results in fewer interactions as shown for the DOCKGROUND set (Table 5.4). On the other hand, due to the preprocessing criteria, the number of models considered for the analysis is slightly different between our dataset and the CONSRANK one.

Surprisingly, the results for the CAPRI set are strikingly better than for the DOCKGROUND set. This might be explained by the higher number of near-native models, although the near-native model versus all models rate is even slightly smaller for the CAPRI set than for the DOCKGROUND set (6.76 % vs. 8.92 %). In order to check this, we performed an additional analysis for each CAPRI target without including near-native models. Since this makes it impossible to compute an AUC for ranking decoys (no true positives), we only obtained values for ranking interactions and residues (see values in parentheses in Table 5.6). The results indicate that the overall quality of submitted models and not just the number of near-native models are important for our approach.



(a) T24

(b) T25

(c) T26

(d) T29

(e) T32

(f) T36

Figure 5.10: I_{RMSD} distribution for CAPRI targets.

The performance of our dRIN-based method for ranking interactions without near-native models decreases from an AUC of 0.88 to 0.75 on average for all targets. The results are most different for T29 (decrease of 0.26), which is not surprising since its RMSD distribution (Figure 5.10(d)) is very similar to those for T24 and T36 after removing the near-native model, e.g. low number of models with small I_{RMSD}. In contrast, the performance of our method for ranking residues decreases only by 0.03 on average. This might be explained by the robustness of the approach, i.e., residue frequency is computed as the sum of interaction frequencies with its direct neighbors.

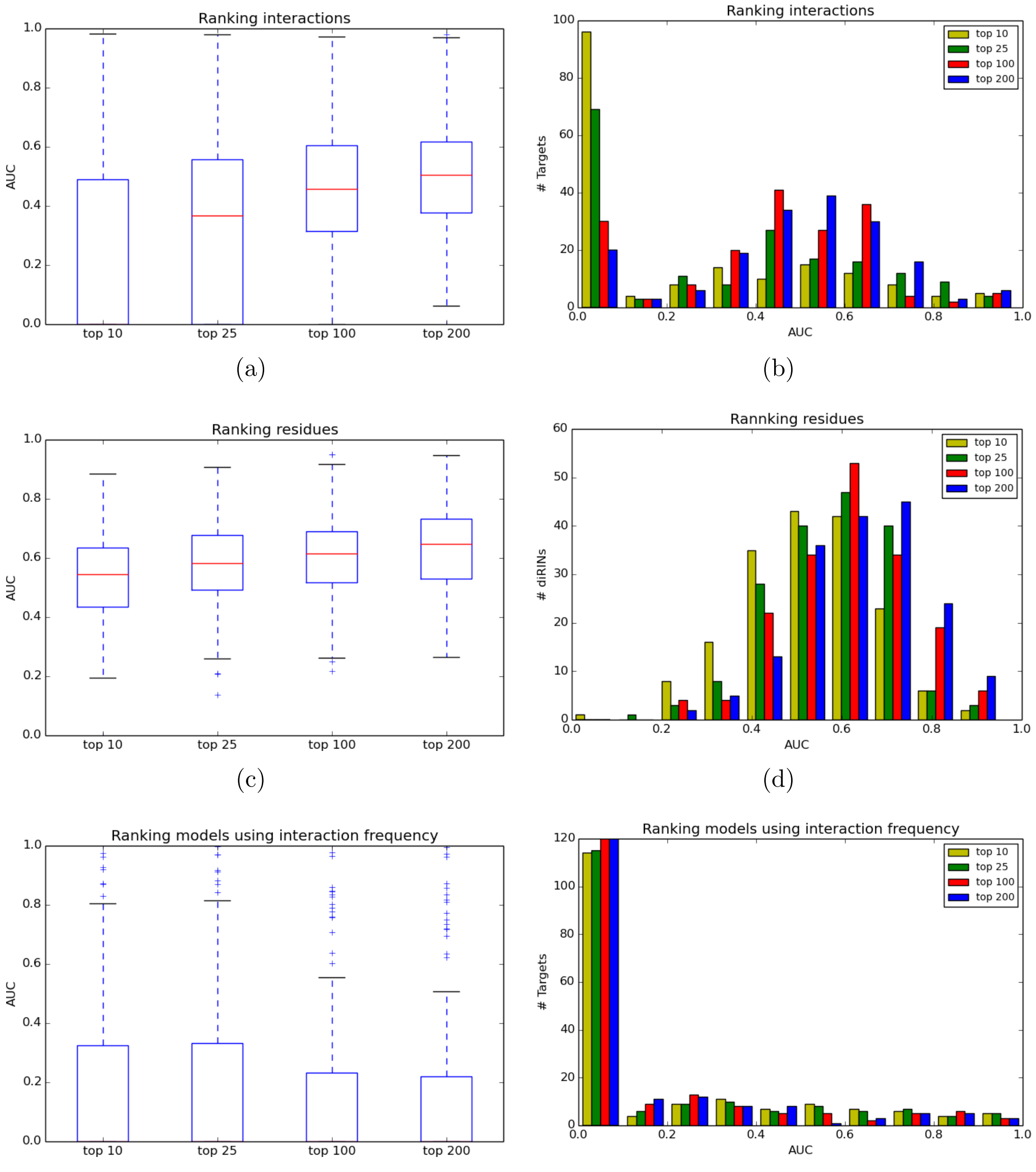
Another reason for the better performance of our method compared to the DOCKGROUND set might be the diversity of the CAPRI set. It contains the 10 best models selected and submitted by each predictor or server. Thus, the diRIN generated for each target can be seen as a consensus of the different docking approaches and not only as a consensus of the best solutions from one approach. This is in agreement with the conclusions in the CONSRANK paper that merging decoys from different docking programs improves the results (Vangone *et al.*, 2013). In addition, the excellent performance of our frequency-based method for identifying interface residues supports the observation of Lensink *et al.* that even incorrect models often contain the true interface residues (Lensink and Wodak, 2010).

5.4.6 Results for PatchDock models

Table 5.7: Average ranking performance on the PatchDock set for the *top 10, 25, 100*, and *200* decoys for each target as well as number of all models and near-native models considered for each group. The AUC values for ranking interactions, residues, or models for each target are plotted in Figures B.15, B.16, and B.17.

	average (stdev)			
	top 200	top 100	top 25	top 10
All 176 targets				
# Near-native	2.006 (5.379)	1.182 (3.823)	0.398 (1.607)	0.182 (0.786)
AUC Ranked Mod	0.142 (0.248)	0.145 (0.254)	0.178 (0.281)	0.182 (0.284)
AUC Ranked Int	0.477 (0.229)	0.422 (0.239)	0.326 (0.303)	0.237 (0.293)
AUC Ranked Res	0.637 (0.138)	0.607 (0.139)	0.576 (0.134)	0.532 (0.144)
64 targets with more than 0 near-native models				
# Near-native	5.516 (7.791)	3.250 (5.812)	1.094 (2.531)	0.500 (1.247)
AUC Ranked Mod	0.384 (0.276)	0.396 (0.281)	0.478 (0.267)	0.483 (0.257)
AUC Ranked Int	0.609 (0.177)	0.543 (0.202)	0.467 (0.280)	0.351 (0.307)
AUC Ranked Res	0.709 (0.130)	0.676 (0.140)	0.647 (0.130)	0.591 (0.127)
30 targets with more than 2 near-native models				
# Near-native	10.233 (9.394)	6.400 (7.328)	2.233 (3.360)	1.033 (1.671)
AUC Ranked Mod	0.542 (0.271)	0.480 (0.312)	0.529 (0.281)	0.535 (0.268)
AUC Ranked Int	0.710 (0.157)	0.636 (0.186)	0.581 (0.260)	0.442 (0.279)
AUC Ranked Res	0.750 (0.115)	0.713 (0.149)	0.670 (0.140)	0.617 (0.143)

Finally, we performed a large-scale analysis for models generated by the PatchDock web server for the Benchmark set 4 (Hwang *et al.*, 2010). For each of the 176 targets, we considered the *top 10, 25, 100*, and *200* decoys. In contrast to the other two datasets, this one is larger, has the lowest number of near-native models on average (from 0.182 in the *top 10* to 2.0 in the *top 200*), but is also more diverse with regard to the distribution of near-native models. In the *top 200*, there are 9 targets with 10 to 31 near-native models, 21 with more than 2, 34 with 1 or 2, and another 112 without any. In contrast, there are also 12 targets with 1 to 6 near-native models in the *top 10*.



(a) (b)

(c) (d)

(e) (f)

Figure 5.11: Box-and-whisker plot and distribution of AUC values for ranking interface interactions ((a) and (b)); interface residues ((c) and (d)); and models using interaction frequency ((e) and (f)) in the PatchDock set. In each case, the AUC values are computed for all 176 targets and the *top 10, 25, 100, 200* decoys available for each target.

The low number of near-native models is already an indicator for the overall performance of PatchDock. The web server performs well on some targets, but on average it fails to rank a near-native model in the top 10/25 and sometimes even 200 solutions. Nevertheless, the very low performance of our method (AUC value of 0.477, e.g., worse than a random method) for ranking interactions and the mediocre performance (AUC value of 0.637) for ranking residues are unexpected (see Table 5.7

and Figure 5.11). The results improve (average AUC values around 0.7) when considering only targets with at least 1 or 3 near-native models. As already seen for the other datasets, the performance of our ranking method increases with the increasing number of decoys considered for the analysis (Figure 5.11(a) and Figure 5.11(c)). In addition, the distribution of AUC values for the different groups (*top 10, 25, 100, 200*) indicates that there are more targets with an AUC above 0.8 using the *top 10* or *25* decoys than the *top 100* or *200* (Figure 5.11(b)). This is the other way around for residues (Figure 5.11(d)). The AUC values for each target and group of top models are shown in Figure B.15 and Figure B.16

In the particular case of ranking decoys, the AUC values can only be computed if there is at least one near-native model, which explains the extremely low average AUC values on the whole dataset (between 0.142 and 0.182). Therefore, we also included average performance of our method for the 64 targets with at least 1 near-native model and the 30 targets with at least 3 near-native models in Table 5.7. Although these results are closer to the values obtained for the other two datasets, they are still not better than a random method for ranking decoys (AUC values around 0.5). The box-and-whisker plot in Figure 5.11(e) is also in agreement with the previous results, i.e., the ranking performance of our method decreases with the inclusion of more decoys. Here, we also observe the same trend as for interactions, i.e., for a few targets the *top 10* or *25* decoys are already enough for good performance of our method (Figure 5.11(f) and Figure B.17).

We also plotted the AUC values for ranking interactions and residues versus the number of near-native models, interface size, and frequency score in the top 5 % (see Figure B.14). Again, we observe that the performance of our dRIN-based method tends to be better for targets with more near-native models, but there are also quite a few exceptions. Furthermore, neither the number of true interface interactions or residues nor the frequency score at 0.05 FPR have an effect on the ranking results.

5.4.7 Summary and discussion

Motivated by the observation that top-ranked non-native docking models often contain the native interface (Lensink and Wodak, 2010), we applied our RIN methodology to identify the native residues and interactions among a set of solutions. We constructed dynamic RINs for the *top N* ranked models and evaluated how well the most frequent interactions and their residues overlap with the true interface residues and interactions. If the native interface is well represented by the top-ranked models, we will detect this by the frequency of the corresponding interactions. Furthermore, given the most frequent interactions, we can re-rank the docking models based on the proportion of such interactions they contain. Altogether, we devised three different methods for analyzing an ensemble of ranked docking structures using dRINs. We evaluated the performance of our methods to identify interface residues and interactions as well as to rank docking solutions on three different benchmark sets.

The DOCKGROUND dataset comprised of 61 protein complexes and for each complex, the top 100 non-native docking models as well as 1 to 10 near-native solutions were available. We observed good performance of the dRIN methods for ranking interactions, residues and docking models (AUC values of 0.801 ± 0.162, 0.747 ± 0.125, and 0.658 ± 0.282, respectively), when *all* solutions were considered. Although the average AUC values dropped significantly (between 0.25 and 0.65) when smaller sets with less to no near-native models were included in the analysis, the general trends remained. On the one hand, the identification of interface residues performed best, followed by identifying interface interactions and ranking the models, and on the other hand, the more models were considered for the analysis, the better the AUC values were.

The second benchmark set contained 6 target complexes and, for each of them, around 300 predicted models submitted by different groups during the CAPRI initiative. Our three different dRIN-based methods performed very well in ranking the interface residues, interactions and docking solutions (AUC values of 0.872, 0.880, and 0.914, respectively). Even when all near-native models were excluded from the dataset, the average AUC values for identifying interface residues and interactions remained above 0.75. Most probably this is caused by the diversity and overall better quality of the CAPRI set compared to the DOCKGROUND set, since the former contains the top 10 models submitted by each group.

We also compared our results to the recently published CONSRANK algorithm (Vangone *et al.*, 2013; Oliva *et al.*, 2013), which ranks docking models based on the conservation of residue contacts (distance of at most 5 Å). For this purpose, also created dRINs based on the distance between the residues in addition to our dRINs, which contain non-covalent interaction edges. On both the DOCKGROUND and CAPRI dataset, we achieved similar to slightly better performance with our approach for ranking the models. We also confirmed that the type of RIN representation has only a marginal effect on the overall performance of our method for ranking models based on residue interaction frequency. Thus, it remains in the hand of the user to choose whether an atomic distance of 5 Å is a sufficient evidence for a residue interaction or to rely on a more sophisticated method that distinguishes between non-covalent interaction types such as ours.

Finally, we analyzed a larger dataset consisting of 176 docking targets with up to 200 models predicted by PatchDock. In contrast to the other two benchmark sets, this one contained on average the lowest number of near-native models, which also means that the docking method did not perform very well. Accordingly, the performance of our method for ranking residues, interactions and models was rather low, but significantly improved when only targets with more than 1 near-native models were included in the analysis (AUC values of 0.750, 0.710, and 0.542).

One of our goals was to find out whether the *top 10, 25, . . . , N* decoys are sufficient to predict the true interface. We observed that the performance of our ranking method based on the frequency of interactions in the set of models improves by increasing the number of models. Indeed, the *top 10* decoys might be enough if the docking performed very well, but usually more are needed for a successful prediction of the interface. Therefore, a possible solution is the integration of the *top 10/25* solutions from different docking algorithms as in the CAPRI dataset. In this way, the risk of bias towards a wrong solution from one of the methods might be avoided.

Unfortunately, we were not able to derive general criteria such as a threshold for interaction or residue frequency that would be indicative for the extent of the success of our dRIN-based method without any knowledge about the quality of the solutions. Our evaluation indicated that the performance of such ranking approaches strongly depends on the overall quality of the docking solutions and not so much on the exact number of near-native models as suggested by Oliva *et al.* (2013). However, the latter is usually a good indicator for the success of the docking algorithm.

Overall, we confirmed that docking models, which are assessed as incorrect by common criteria based on ligand or interface RMSD, actually contain enough information about the true interface residues and interactions. Therefore, an approach based on the identification of the most common interactions and residues in the set of decoys can be quite promising for predicting interface residues and interactions. We could also show that the most frequent interactions identified in a set of docking models can be used as a scoring function for the docking performance.

5.5 Conclusions

For a long time, proteins, their structures and interactions have been analyzed from a very static point of view, although they are actually very dynamic and usually populate several possible conformations. This is mostly caused by the difficulties arising from the attempts to study protein dynamics with known experimental structure determination techniques such as X-ray crystallography and NMR, but is slowly changing with the introduction and development of new integrative experimental approaches. Thus, theoretical and computational methods for simulating protein dynamics have been the richest source of information so far. MD simulations has become a trusted and widely used tool in the last 30 years. Recently, several groups combined the analysis of data from MD simulations with network theory in order to investigate protein dynamics from a new point of view. In particular, allosteric communication and drug resistance mechanisms have been studied extensively (see Section 5.1.2).

To capture the dynamic nature of protein structures and interactions, we developed a new method for visualizing and analyzing ensembles of protein structures by representing them as dynamic, weighted residue interaction networks (dRINs). Ensembles could result from an experimental technique such as NMR or a computational method like MD simulation or protein docking. Using residue interaction networks, we can analyze the variation reflected by the individual protein structures and, at the same time, identify non-covalent residue interactions shared by the different structures. Possible applications of our approach include the identification of structurally and functionally important residue interactions, the comparison of ligand-binding modes in protein interactions, as well as the characterization of protein mutations and their effect on structure and function.

We applied our approach in two different scenarios to address current challenges in structural bioinformatics. First, we analyzed MD simulation data to characterize the effect of residue mutations on protein structure and function. Thereby, we compared wild-type and mutant (L166P) simulations of DJ-1, a protein associated with Parkinson's disease. We pinpointed significant changes of the residue interaction pattern upon mutation of residue 166. Our topological analysis and visual exploration confirmed the findings of Anderson and Daggett (2008) that the L166P substitution most probably affects the dimerization of DJ-1 by disturbing the secondary structure of its neighboring helices.

Combining the dynamic RIN representation of the wild-type and mutant simulations with physico-chemical properties of the protein residues as described in Chapter 4 should provide additional hypotheses on the molecular mechanisms upon mutation. Our dRIN-based approach can also be applied to the remaining data deposited in the Dynameomics database as well as to MD simulation data of other proteins associated with diseases. The effects of protein mutations, in particular, located far from the active site, are also of special interest for studying drug resistance mutations, which occur often in viral proteins but are still not well understood. For this purpose, we have started a cooperation with Tomas Bastys and Olga V Kalinina (Max Planck Institute for Informatics) on the analysis of resistance mutations in HIV-1 protease using our methodology.

Furthermore, we analyzed sets of docking structures to determine the native interface residues and their interactions based on their frequency of occurrence. This analysis was based on the assumption that top-ranked docking solutions are more enriched in correct binding interfaces than in correct complexes. We also used the most frequent interactions as a scoring function for ranking the docking solutions as suggested by Vangone *et al.* (2013). We evaluated the performance of our dRIN-based methods for ranking residues, interactions and docking models on three different benchmark sets (DOCKGROUND, CAPRI and PatchDock). Overall, our analysis revealed that the top 100 solutions of a good docking method can contain enough information to correctly identify most of the native interactions and residues. However, the success of our method strongly depends on the quality of the docking models and thus, including the top 10 solutions from different methods, as it is the case for the CAPRI dataset, might be a good strategy to boost the performance of such methods.

In the docking field, there is a strong interest in new scoring methods and, more generally, new ways of analyzing and visualizing of docking models (Lensink and Wodak, 2014). Our approach of creating dynamic RINs from an ensemble of top-ranked solutions provides a completely new perspective on the available data. We already showed that, for docking methods with reliable predictions, we can apply it to successfully identify the native interface residues and interactions as well as to rescore the docking models. The network representation of such structure ensembles makes it possible to use further, more sophisticated graph analysis techniques such as clustering and community search. For example, we can estimate the heterogeneity of an ensemble of docking models solely using the dRIN interaction frequencies.

Here, we described a novel method for analysis and visualization of protein structures that explicitly takes into account their dynamic natures. Among others, the applications of our approach include the identification of structurally and functionally important residue interactions, the comparison of ligand-binding modes in protein interactions, as well as the characterization of protein mutations and their effect on structure and function.

CHAPTER 6

Conclusions

This chapter concludes the thesis by outlining the main methodological contributions and their application to biomedical data for understanding complex diseases. In addition, it provides future perspectives on the importance of network-based approaches in the field of disease gene prioritization and structural biology.

6.1 Summarizing remarks

Nowadays, we are still far from a complete understanding of human diseases, their origin, prevention, and cure. High-throughput techniques such as next-generation sequencing and mass spectrometry produce large and fast growing amounts of experimental data that cannot be handled without extensive computational power. Representing the complex relationships between biological entities such as genes, proteins, or residues, as networks has proven to be very useful for integrating, analyzing, prioritizing, and visualizing large-scale datasets, with the ultimate goal to gain more insight into complex cellular mechanisms (see Chapter 2).

The work performed in this thesis has two complementary viewpoints. On one hand, we focused on the development of novel methods and software tools on the interface between network biology, structural biology and medicine with focus on visual analytics. On the other hand, we collaborated with biological and medical experts to apply and adapt our methods to their data and gain new insights, in particular, for understanding complex diseases.

The investigation of less studied phenotypes with unknown causative genes is of great importance for current methodological development in the field of disease gene prioritization. Therefore, we designed a phenotype-specific framework for functional characterization and especially prioritization of candidate genes (see Chapter 3).

We generated networks of strong functional similarities between candidate genes based on their Gene Ontology annotations and integrated them with protein interaction data. We assessed the connectivity of the candidates using network topology analysis methods and selected promising disease genes to build phenotype-specific networks. We also characterized the functional overlap of similar phenotypes by analyzing and comparing their functional similarity networks. Using our framework, we identified several disease-relevant genes and processes for inflammatory bowel diseases, primary sclerosing cholangitis, and Parkinson's disease.

Since finding the causative disease genes does often not suffice, we also concentrated on the molecular characterization of sequence mutations and their effect on protein structure and function. To this end, we designed a software suite (including our tools RINalyzer and RINerator) that supports the interactive, multi-layered visual analysis of protein structures and their molecular function in protein binding, allostery, drug resistance and other interaction mechanisms (see Chapter 4). By representing protein structures as networks of interacting residues, integrating them with molecular data derived from external resources and applying network visualization and analysis techniques, we could facilitate the analysis of protein sequence mutations. In addition, we analyzed a large set of known resistance-associated mutations in the NS3 protease of the hepatitis C virus by combining our visual analytics framework with statistics on the structural and topological properties of the residues.

To capture the dynamic nature of protein structures and interactions, we also developed an approach to visualizing and analyzing ensembles of protein structures as generated by molecular dynamics (MD) simulations (see Chapter 5). We created dynamic, weighted residue interaction networks that account for the different protein conformations within the ensemble. We also facilitate the comparison of two ensembles within one network by highlighting the most similar and dissimilar residue interactions as well as the rate at which they are present in the ensembles. As a proof of concept, we applied our approach to characterize the effect of sequence mutations on protein structure and function using MD simulation data. Furthermore, we performed a thorough analysis of ensembles of docking structures (decoys) to evaluate their quality and aid in identifying the most probable docking interfaces.

In conclusion, we developed novel methods and software tools for the prioritization and functional characterization of genes and proteins associated with complex diseases. We also demonstrated how combining different types of data using biological networks and visual analytics can be instrumental in gaining more biological insight in the fields of medical and structural bioinformatics.

6.2 Perspectives

Large-scale interaction data are commonly represented as networks and analyzed by graph-theoretic methods that can characterize the topological network structure and its global and local interaction properties (Vidal *et al.*, 2011; Pavlopoulos *et al.*, 2011; Ideker and Krogan, 2012; Csermely *et al.*, 2013). In the last years, topological analysis of interaction networks has been an indispensable part of network-based approaches for disease gene prioritization (Vidal *et al.*, 2011; Barabási *et al.*, 2011; Jia and Zhao, 2014). Furthermore, structural biologists have recently used a network representation to study the interactions of residues in protein structures towards the understanding of complex protein structure-function relationships (Csermely, 2008; Vishveshwara *et al.*, 2009; Doncheva *et al.*, 2011). However, these are just examples for the power of biological networks as a tool for formulating new hypotheses and answering open questions in biomedical research.

With the increasing amounts of new data generated by high-throughput techniques (The ENCODE Project Consortium, 2012; Rolland *et al.*, 2014; Moignard *et al.*, 2015; Huttlin *et al.*, 2015; Sahni *et al.*, 2015), the field of network biology will need to advance further. Novel visualization, analysis and integration techniques will need to be developed in order to complement the existing ones. For example, the Cytoscape consortium has already made a step in this direction by releasing a new series of versions that support the analysis of large-scale datasets with novel features as well as by creating the Cytoscape App Store, which hosts a collection of 236 external apps developed to enhance Cytoscape with rich biological functionality (Saito *et al.*, 2012; Lotia *et al.*, 2013).

With respect to disease gene prioritization, network-based methods have already proven to be very useful and will continue to be exploited. The current trend as shown in recent work (Ellinghaus *et al.*, 2013b; Ghiassian *et al.*, 2015; Tasan *et al.*, 2015) is the identification of smaller disease-specific networks of interacting causal and candidate genes. However, these approaches can still be improved by the generation of more integrative networks using different data types as well as by the application of more appropriate graph algorithms for subnetwork detection. Our framework could be applied to the disease gene prioritization and functional phenotype comparison of other phenotypes as new data is generated (Parkes *et al.*, 2013; Andreassen *et al.*, 2015; Li *et al.*, 2015). Overall, we can expect a considerable improvement in candidate disease gene prioritization approaches in the near future due to the increasing amounts of data produced by high-throughput techniques every day.

Recent studies have revealed that combining network and structural biology could be very beneficial for both fields (Fraser *et al.*, 2013). In particular, residue interaction networks derived from the 3D protein structure have been successfully incorporated in the study of protein dynamics and engineering, protein and ligand binding, disease and drug resistance mutations (Csermely, 2008; Vishveshwara *et al.*, 2009; Di Paola *et al.*, 2013; Hu *et al.*, 2013; Yan *et al.*, 2014). As a recent study by Viswanathan *et al.* (2015) suggests, RINs can complement existing structure-based methods for therapeutic antibody discovery and anti-viral drug optimization. For this purpose, RIN-based approaches need to be adapted from assessing the general effect of residue mutations (Doncheva *et al.*, 2014) to characterizing more subtle resistance mechanisms (Hughes and Andersson, 2015). Furthermore, good gold-standard datasets need to be assembled through experimental and computational efforts to assure the proper validation of new (prediction) methods using RINs.

Another distinct advantage of RINs is that they could provide a novel way of investigating the epistatic interactions between residues (Harms and Thornton, 2013) and in particular, residue mutations (Ray *et al.*, 2014). By combining graph theory methods with physico-chemical properties of the amino-acid residues, a new approach can be developed to identify epistatic pairs, or even networks, of mutated residues and assess their putative effect on the protein. A good starting point would be to analyze the dataset recently presented by Jordan *et al.* (2015), who identified several hundreds of human disease mutations that are compensated by amino acid mutations in other species.

The further development, extension and adaptation of our software tools to the needs of structural biologists could enable novel biomedical applications. For example, we could gain more insight by supporting contact networks of conformationally heterogeneous residues as derived from high-resolution X-ray crystallography data by van den Bedem *et al.* (2013). Furthermore, the network representation of residue interactions utilizes a completely new way of comparing protein structures of different species and might reveal new insights into the evolutionary relationship between organisms. Such a comparison could also facilitate the analysis of different genotypes of the same viral species with respect to their differential resistance mutation patterns toward the same therapeutics (Sullivan *et al.*, 2013; Romano *et al.*, 2010; Hughes and Andersson, 2015). A promising study in this direction was recently presented by Flock *et al.* (2015), who combined residue contact networks with other computational techniques to reveal that different G protein-coupled receptors share a highly conserved mechanism of allosteric interaction.

Besides our work on dynamic RINs, several studies have provided evidence for the usefulness of RINs generated from a set of similar protein structures for representing the dynamic nature and diversity of proteins (Vishveshwara *et al.*, 2009; Sethi *et al.*, 2009; Xue *et al.*, 2012; Tiberti *et al.*, 2014; Seeber *et al.*, 2014). Future methodological improvements of this approach include a better and more realistic representation of the residue interactions in the MD simulation, for example, by taking into account the correlation or mutual information of residue interaction energies. Furthermore, more sophisticated graph analysis techniques such as clustering and community detection would need to be applied on the resulting RINs to investigate complex structure-function relationships. In the field of molecular docking, the application of dynamic RINs could answer the need to new scoring methods and, more generally, new ways of analyzing and visualizing the docking models (Lensink and Wodak, 2014).

Overall, integrative network-based approaches will continue to have a crucial role in the prioritization and functional characterization of genes, proteins and residues. The field of visual analytics has also gained more attention in the biological community addressing the importance of interactive visualization of big data sets. With the constantly evolving experimental techniques, new datasets and data types will appear that might just be best understood and characterized by combining the commonly known network representations with the novel techniques of visual analytics.

Bibliography

Abraham, A., Hassanien, A. E., and Snášel, V. (2010). *Computational social network analysis: trends, tools and research advances.* Springer-Verlag New York, Inc.

Adcock, S. A. and McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev*, 106(5):1589–1615.

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6:55.

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., *et al.* (2010). A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–249.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., *et al.* (2006). Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24(5):537–544.

Agapito, G., Guzzi, P. H., and Cannataro, M. (2013). Visualization of protein interaction networks: problems and solutions. *BMC Bioinformatics*, 14 Suppl 1:S1.

Alber, T., Banner, D., Bloomer, A., Petsko, G., Phillips, D., Rivers, P., and Wilson, I. (1981). On the three-dimensional structure and catalytic mechanism of Triosephosphate isomerase. *Philos Trans R Soc Lond B Biol Sci*, 293(1063):159–171.

Albert, R. (2005). Scale-free networks in cell biology. *J Cell Sci*, 118(Pt 21):4947.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev Mod Phys*, 74:47–97.

Albert, R., Jeong, H., and Barabasi, A. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382.

Alcaraz, N., Friedrich, T., Kötzing, T., Krohmer, A., Müller, J., Pauling, J., and Baumbach, J. (2012). Efficient key pathway mining: combining networks and OMICS data. *Integr Biol (Camb)*, 4(7):756–764.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607.

Almaas, E. (2007). Biological impacts and context of network theory. *J Exp Biol*, 210(Pt 9):1548.

Altelaar, A. F. M., Munoz, J., and Heck, A. J. R. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet*, 14(1):35–48.

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*, 37:D793D796.

Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., and Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *J Mol Biol*, 344(4):1135–1146.

Anderson, C. A., Boucher, G., Lees, C. W., Franke, A., D'Amato, M., Taylor, K. D., *et al.* (2011). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet*, 43(3):246–252.

Anderson, E., Veith, G., Weininger, D., and Environmental Research Laboratory (Duluth, Minn.) (1987). *SMILES, a line notation and computerized interpreter for chemical structures.* Environmental research brief. U.S. Environmental Protection Agency, Environmental Research Laboratory.

Anderson, P. C. and Daggett, V. (2008). Molecular basis for the structural instability of human DJ-1 induced by the L166P mutation associated with Parkinson's disease. *Biochemistry*, 47(36):9380–9393.

Anderson, P. C. and Daggett, V. (2009). The R46Q, R131Q and R154H polymorphs of human DNA glycosylase/beta-lyase hOgg1 severely distort the active site and DNA recognition site but do not cause unfolding. *J Am Chem Soc*, 131(27):9506–9515.

Andreassen, O. A., Desikan, R. S., Wang, Y., Thompson, W. K., Schork, A. J., Zuber, V., Doncheva, N. T., *et al.* (2015). Abundant genetic overlap between blood lipids and immune-mediated diseases indicates shared molecular genetic mechanisms. *PloS ONE*, 10(4):e0123057.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S. L., Ceol, A., Chautard, E., *et al.* (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods*, 8(7):528–529.

Arrell, D. and Terzic, A. (2010). Network systems biology for drug discovery. *Clin Pharmacol Ther*, 88(1):120–125.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284.

Astsaturov, I., Ratushny, V., Sukhanova, A., Einarson, M., Bagnyukova, T., Zhou, Y., *et al.* (2010). Synthetic lethal screen of an EGFR-centered network to improve targeted therapies. *Sci Signal*, 3(140):ra67.

Atchley, W. R., Zhao, J., Fernandes, A. D., and Drüke, T. (2005). Solving the protein sequence metric problem. *Proc Natl Acad Sci U S A*, 102(18):6395–6400.

Atias, N. and Sharan, R. (2012). Comparative analysis of protein networks: Hard problems, practical solutions. *Commun. ACM*, 55(5):88–97.

Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE*, 4(2):e4345.

Auerbach, D., Thaminy, S., Hottiger, M. O., and Stagljar, I. (2002). The post-genomic era of interactive proteomics: Facts and perspectives. *Proteomics*, 2(6):611–623.

Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y., and Borgwardt, K. M. (2013). Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29(13):i171–9.

Bader, G. D. and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2.

Bader, S., Kühner, S., and Gavin, A.-C. (2008). Interaction networks for systems biology. *FEBS Lett*, 582(8):1220–1224.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet*, 12(11):745–755.

Bandyopadhyay, S., Kelley, R., Krogan, N. J., and Ideker, T. (2008). Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol*, 4(4):e1000065.

Barabási, A.-L. (2007). Network medicine–from obesity to the "diseasome". *N Engl J Med*, 357(4):404–407.

Barabási, A. L. (2009). Scale-free networks: a decade and beyond. *Science*, 325(5939):412–3.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–12.

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68.

Barabási, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–13.

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., *et al.* (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet*, 40(8):955–962.

Barshir, R., Shwartz, O., Smoly, I. Y., and Yeger-Lotem, E. (2014). Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS Comput Biol*, 10(6):e1003632.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media.*

Batagelj, V. and Mrvar, A. (1998). Pajek – program for large network analysis. *Connections*, 21:4757.

Baudot, A., Gómez-López, G., and Valencia, A. (2009). Translational disease interpretation with molecular networks. *Genome Biol*, 10(6):221.

Baulac, S., LaVoie, M. J., Strahle, J., Schlossmacher, M. G., and Xia, W. (2004). Dimerization of Parkinson's disease-causing DJ-1 and formation of high molecular weight complexes in human brain. *Mol Cell Neurosci*, 27(3):236–246.

Beck, D., Alonso, D., and Daggett, V. (2000–2010). In lucem molecular mechanics (computer program). University of Washington, Seattle.

Beck, D. A. C., Jonsson, A. L., Schaeffer, R. D., Scott, K. A., Day, R., Toofanny, R. D., *et al.* (2008). Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations. *Protein Eng , Des Sel*, 21(6):353–368.

Bengtson, M.-B., Solberg, C., Aamodt, G., Jahnsen, J., Moum, B., Sauar, J., *et al.* (2009). Clustering in time of familial IBD separates ulcerative colitis from Crohn's disease. *Inflamm Bowel Dis*, 15(12):1867–1874.

Benson, N. C. and Daggett, V. (2012). A comparison of multiscale methods for the analysis of molecular dynamics simulations. *J Phys Chem B*, 116(29):8722–8731.

Bergmann, S., Ihmels, J., and Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, 2(1):E9.

Bergquist, A., Montgomery, S. M., Bahmanyar, S., Olsson, R., Danielsson, A., Lindgren, S., *et al.* (2008). Increased risk of primary sclerosing cholangitis and ulcerative colitis in first-degree relatives of patients with primary sclerosing cholangitis. *Clin Gastroenterol Hepatol*, 6(8):939–943.

Bhakat, S., Martin, A. J., and Soliman, M. E. (2014). An integrated molecular dynamics, principal component analysis and residue interaction network approach reveals the impact of M184V mutation on HIV reverse transcriptase resistance to lamivudine. *Mol Biosyst*, 10(8):2215–28.

Bhattacharyya, M., Bhat, C. R., and Vishveshwara, S. (2013). An automated approach to network features of protein structure ensembles. *Protein Sci*, 22(10):1399–416.

Bhattacharyya, M., Ghosh, A., Hansia, P., and Vishveshwara, S. (2010). Allostery and conformational free energy changes in human tryptophanyl-tRNA synthetase from essential dynamics and structure networks. *Proteins*, 78(3):506–17.

Bhattacharyya, M. and Vishveshwara, S. (2011). Probing the allosteric mechanism in pyrrolysyl-tRNA synthetase using energy-weighted network formalism. *Biochemistry*, 50(28):6225–6236.

Bindea, G., Galon, J., and Mlecnik, B. (2013). CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics*, 29(5):661–663.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., *et al.* (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093.

Blacklock, K. and Verkhivker, G. M. (2014a). Allosteric regulation of the Hsp90 dynamics and stability by client recruiter cochaperones: Protein structure network modeling. *PLoS ONE*, 9(1):e86547.

Blacklock, K. and Verkhivker, G. M. (2014b). Computational modeling of allosteric regulation in the Hsp90-chaperones: A statistical ensemble analysis of protein structure networks and allosteric communications. *PLoS Comput Biol*, 10(6):e1003679.

Bondy, J. A. (1976). *Graph Theory With Applications*. Elsevier Science Ltd., Oxford, UK, UK.

Bonifati, V., Rizzu, P., van Baren, M. J., Schaap, O., Breedveld, G. J., Krieger, E., *et al.* (2003). Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science*, 299(5604):256–259.

Bonsor, D. A., Grishkovskaya, I., Dodson, E. J., and Kleanthous, C. (2007). Molecular mimicry enables competitive recruitment by a natively disordered protein. *J Am Chem Soc*, 129(15):4800–4807.

Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications.* Springer.

Borgatti, S. P. (2005). Centrality and network flow. *Soc Networks*, 27(1):55–71.

Borgatti, S. P., Everett, M. G., and Johnson, J. C. (2013). *Analyzing social networks.* SAGE Publications Ltd.

Bottomly, D., Wilmot, B., Tyner, J. W., Eide, C. A., Loriaux, M. M., Druker, B. J., and McWeeney, S. K. (2013). HitWalker: variant prioritization for personalized functional cancer genomics. *Bioinformatics*, 29(4):509–510.

Boutros, M. and Ahringer, J. (2008). The art and design of genetic screens: RNA interference. *Nat Rev Genet*, 9(7):554–66.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J Math Sociol*, 25:163–177.

Brandes, U. and Erlebach, T. (2005). *Network Analysis: Methodological Foundations (Lecture Notes in Computer Science).* Springer-Verlag New York, Inc.

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., and Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538.

Breitkreutz, B.-J., Stark, C., and Tyers, M. (2003). Osprey: a network visualization system. *Genome Biol*, 4(3):R22.

Brinda, K. V. and Vishveshwara, S. (2005). A network representation of protein structures: implications for protein stability. *Biophys J*, 89(6):4159–70.

Bromberg, Y. (2013). Chapter 15: disease gene prioritization. *PLoS Comput Biol*, 9(4):e1002902.

Bromberg, Y. and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res*, 35(11):3823–3835.

Bromley, D., Anderson, P. C., and Daggett, V. (2013). Structural consequences of mutations to the alpha-tocopherol transfer protein associated with the neurodegenerative disease ataxia with vitamin E deficiency. *Biochemistry*, 52(24):4264–73.

Bromley, D., Rysavy, S. J., Su, R., Toofanny, R. D., Schmidlin, T., and Daggett, V. (2014). DIVE: a data intensive visualization engine. *Bioinformatics*, 30(4):593–595.

Broomé, U. and Bergquist, A. (2006). Primary sclerosing cholangitis, inflammatory bowel disease, and colon cancer. *Semin Liver Dis*, 26(1):31–41.

Brown, K. R., Otasek, D., Ali, M., McGuffin, M. J., Xie, W., Devani, B., *et al.* (2009). NAViGaTOR: Network analysis, visualization and graphing toronto. *Bioinformatics*, 25(24):3327–3329.

Butts, C. T. (2008). Social network analysis with sna. *J Stat Softw*, 24:6.

Callaway, D., Newman, M., Strogatz, S., and Watts, D. (2000). Network robustness and fragility: percolation on random graphs. *Phys Rev Lett*, 85(25):5468–5471.

Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):2001–2014.

Capriotti, E., Nehrt, N. L., Kann, M. G., and Bromberg, Y. (2012). Bioinformatics for personal genome interpretation. *Brief Bioinform*, 13(4):495–512.

Castellana, S. and Mazza, T. (2013). Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief Bioinform*, 14(4):448–459.

Chang, S., Jiao, X., Li, C. H., Gong, X. Q., Chen, W. Z., and Wang, C. X. (2008). Amino acid network and its scoring application in protein-protein docking. *Biophys Chem*, 134(3):111–8.

Chatterjee, S., Bhattacharyya, M., and Vishveshwara, S. (2012). Network properties of protein-decoy structures. *J Biomol Struct Dyn*, 29(6):606–622.

Chatterjee, S., Ghosh, S., and Vishveshwara, S. (2013). Network properties of decoys and CASP predicted models: a comparison with native protein structures. *Mol Biosyst*, 9(7):1774–88.

Chen, J., Aronow, B. J., and Jegga, A. G. (2009). Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, 10:73.

Chen, Y., Jiang, T., and Jiang, R. (2011a). Uncover disease genes by maximizing information flow in the phenome-interactome network. *Bioinformatics*, 27(13):i167–76.

Chen, Y., Wang, W., Zhou, Y., Shields, R., Chanda, S. K., Elston, R. C., and Li, J. (2011b). In silico gene prioritization by integrating multiple data sources. *PLoS ONE*, 6(6):e21137.

Chermak, E., Petta, A., Serra, L., Vangone, A., Scarano, V., Cavallo, L., and Oliva, R. (2014). CONSRANK: a server for the analysis, comparison and ranking of docking models based on inter-residue contacts. *Bioinformatics*.

Choura, M. and Rebaï, A. (2010). Application of computational approaches to study signalling networks of nuclear and Tyrosine kinase receptors. *Biol Direct*, 5:58.

Chuang, H. Y., Hofree, M., and Ideker, T. (2011). A decade of systems biology. *Annu Rev Cell Dev Biol*, 26:721–744.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140.

Ciriello, G. and Guerra, C. (2008). A review on models and algorithms for motif discovery in protein-protein interaction networks. *Brief Funct Genomic Proteomic*, 7(2):147–156.

Clark, C. and Kalita, J. (2014). A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics*, 30(16):2351–2359.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., *et al.* (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*, 2(10):2366–82.

Collins, A. (2015). The genomic and functional characteristics of disease genes. *Brief Bioinform*, 16(1):16–23.

Conrad, K., Roggenbuck, D., and Laass, M. W. (2014). Diagnosis and classification of ulcerative colitis. *Autoimmun Rev*, 13(4-5):463–466.

Cooper, G. M. and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet*, 12(9):628–640.

Cortes, A. and Brown, M. A. (2011). Promise and pitfalls of the Immunochip. *Arthritis research & therapy*, 13(1):101.

Costa, P. R., Acencio, M. L., and Lemke, N. (2010). A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC Genomics*, 11(Suppl 5):S9.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., *et al.* (2010). The genetic landscape of a cell. *Science*, 327(5964):425–431.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., *et al.* (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*, 39(suppl 1):D691–D697.

Crotty, S., Johnston, R. J., and Schoenberger, S. P. (2010). Effectors and memories: Bcl-6 and Blimp-1 in T and B lymphocyte differentiation. *Nat Immunol*, 11(2):114–120.

Csárdi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, page 1695.

Csermely, P. (2008). Creative elements: network-based predictions of active centres in proteins and cellular and social networks. *Trends Biochem Sci*, 33(12):569–576.

Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther*, 138(3):333–408.

Cummings, M. D., Lindberg, J., Lin, T.-I., de Kock, H., Lenz, O., Lilja, E., *et al.* (2010). Induced-fit binding of the macrocyclic noncovalent inhibitor TMC435 to its HCV NS3/NS4A protease target. *Angewandte Chemie (International ed. in English)*, 49(9):1652–1655.

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A. R., Simonis, N., *et al.* (2009). Literature-curated protein interaction datasets. *Nat Methods*, 6(1):39–46.

Daggett, V. (2006). Protein folding-simulation. *Chem Rev*, 106(5):1898–1916. PMID: 16683760.

Dand, N., Sprengel, F., Ahlers, V., and Schlitt, T. (2013). BioGranat-IG: a network analysis tool to suggest mechanisms of genetic heterogeneity from exome-sequencing data. *Bioinformatics*, 29(6):733–741.

De Bie, T., Tranchevent, L. C., Van Oeffelen, L. M., and Moreau, Y. (2007). Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–32.

De Las Rivas, J. and Fontanillo, C. (2012). Protein-protein interaction networks: unraveling the wiring of molecular machines within the cell. *Brief Funct Genomics*, 11(6):489–496.

De Lau, L. M. L. and Breteler, M. M. B. (2006). Epidemiology of Parkinson's disease. *Lancet Neurol*, 5(6):525–535.

Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., and Rooman, M. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, 25(19):2537–2543.

Dehouck, Y., Kwasigroch, J. M., Gilis, D., and Rooman, M. (2011). PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, 12:151.

del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol Syst Biol*, 2:2006.0019.

Deng, M., Zhang, K., Mehta, S., Chen, T., and Sun, F. (2003). Prediction of protein function using protein-protein interaction data. *J Comput Biol*, 10(6):947–960.

Dezso, Z., Nikolsky, Y., Nikolskaya, T., Miller, J., Cherba, D., Webb, C., and Bugrim, A. (2009). Identifying disease-specific genes based on their topological significance in protein networks. *BMC Syst Biol*, 3:36.

Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., and Giuliani, A. (2013). Protein contact networks: an emerging paradigm in chemistry. *Chem Rev*, 113(3):1598–1613.

Diestel, R. (2012). *Graph Theory, 4th Edition*, volume 173 of *Graduate texts in mathematics*. Springer.

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–231.

Dobrin, R., Zhu, J., Molony, C., Argman, C., Parrish, M. L., Carlson, S., *et al.* (2009). Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol*, 10(5):R55.

Doncheva, N., Klein, K., Domingues, F., and Albrecht, M. (2015). RINalyzer website. http://www.rinalyzer.de/.

Doncheva, N. T., Assenov, Y., Domingues, F. S., and Albrecht, M. (2012a). Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc*, 7(4):670–685.

Doncheva, N. T., Kacprowski, T., and Albrecht, M. (2012b). Recent approaches to the prioritization of candidate disease genes. *WIREs Systems Biology and Medicine*, 4(5):429–442.

Doncheva, N. T., Klein, K., Domingues, F. S., and Albrecht, M. (2011). Analyzing and visualizing residue networks of protein structures. *Trends Biochem Sci*, 36(4):179–182.

Doncheva, N. T., Klein, K., Morris, J. H., Wybrow, M., Domingues, F. S., and Albrecht, M. (2014). Integrative visual analysis of protein sequence mutations. *BMC Proceedings*, 8(Suppl 2 Proceedings of the 3rd Annual Symposium on Biol):S2.

Dong, J. and Horvath, S. (2007). Understanding network concepts in modules. *BMC Syst Biol*, 1:24.

Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys*, 41:429–52.

Du Plessis, L., Skunca, N., and Dessimoz, C. (2011). The what, where, how and why of gene ontology – a primer for bioinformaticians. *Brief Bioinform*, 12(6):723–735.

Dunham, W. H., Mullin, M., and Gingras, A.-C. (2012). Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics*, 12(10):1576–1590.

Dvory-Sobol, H., Wong, K. A., Ku, K. S., Bae, A., Lawitz, E. J., Pang, P. S., Harris, J., Miller, M. D., and Mo, H. (2012). Characterization of resistance to the protease inhibitor gs-9451 in hepatitis c virus-infected patients. *Antimicrobial agents and chemotherapy*, 56(10):52895295.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 613–622, New York, NY, USA. ACM.

Eargle, J. and Luthey-Schulten, Z. (2012). NetworkView: 3D display and analysis of protein.RNA interaction networks. *Bioinformatics*, 28(22):3000–1.

Ellinghaus, D., Bethune, J., Petersen, B.-S., and Franke, A. (2015). The genetics of Crohn's disease and ulcerative colitis – status quo and beyond. *Scand J Gastroenterol*, 50(1):13–23.

Ellinghaus, D., Folseraas, T., Holm, K., Ellinghaus, E., Melum, E., Balschun, T., *et al.* (2013a). Genome-wide association analysis in primary sclerosing cholangitis and ulcerative colitis identifies risk loci at GPR35 and TCF4. *Hepatology*, 58(3):1074–1083.

Ellinghaus, D., Zhang, H., Zeissig, S., Lipinski, S., Till, A., Jiang, T., *et al.* (2013b). Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. *Gastroenterology*, 145(2):339–347.

Emig, D., Salomonis, N., Baumbach, J., Lengauer, T., Conklin, B. R., and Albrecht, M. (2011). AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res*, 38(Web Server issue):W755–W762.

Erdös, P. and Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ Math Inst Hung Acad Sci*, 5:17–61.

Fauchere, J., Charton, M., Kier, L., Verloop, A., and Pliska, V. (1988). Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int J Pept Protein Res*, 32(4):269–278.

Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., *et al.* (2013). SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol*, 7:7.

Feldman, I., Rzhetsky, A., and Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci U S A.*, 105(11):4323–8.

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am Nat*, 125(1):pp.1–15.

Fermi, G., Perutz, M. F., Shaanan, B., and Fourme, R. (1984). The crystal structure of human deoxyhaemoglobin at 1.74 A resolution. *J Mol Biol*, 175(2):159–174.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., *et al.* (2014). Pfam: the protein families database. *Nucleic Acids Res*, 42(Database issue):D222–30.

Flannick, J., Novak, A., Srinivasan, B. S., McAdams, H. H., and Batzoglou, S. (2006). Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res*, 16(9):1169–1181.

Flock, T., Ravarani, C. N. J., Sun, D., Venkatakrishnan, A., Kayikci, M., Tate, C. G., *et al.* (2015). Universal allosteric mechanism for Ga activation by GPCRs. *Nature*, 524(7564):173–179.

Folseraas, T., Liaskou, E., Anderson, C. A., and Karlsen, T. H. (2014). Genetics in PSC: What Do the "Risk Genes" Teach Us? *Clinical reviews in allergy & immunology.*

Folseraas, T., Melum, E., Rausch, P., Juran, B. D., Ellinghaus, E., Shiryaev, A., *et al.* (2012). Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. *J Hepatol*, 57(2):366–375.

Forno, L. (1996). Neuropathology of Parkinson's disease. *J Neuropathol Exp Neurol*, 55(3):259–272.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., *et al.* (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*, 41(Database issue):D808–15.

Franke, A., Balschun, T., Karlsen, T. H., Sventoraityte, J., Nikolaus, S., Mayr, G., *et al.* (2008). Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat Genet*, 40(11):1319–1323.

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., *et al.* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*, 42(12):1118–1125.

Franke, L., Van Bakel, H., Fokkens, L., De Jong, E. D., Egmont-Petersen, M., and Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78(6):1011–25.

Fraser, J. S., Gross, J. D., and Krogan, N. J. (2013). From systems to structure: bridging networks and mechanism. *Mol Cell*, 49(2):222–231.

Frauenfelder, H., Sligar, S., and Wolynes, P. (1991). The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603.

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nat Rev Genet*, 10(4):241–51.

Freeman, L. (1979). Centrality in social networks: conceptual clarification. *Soc Networks*, 1(215-239).

Freeman, L. (2006). *The Development of Social Network Analysis*. Empirical Press, Vancouver.

Freudenberg, J. and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18(suppl 2):S110–5.

Frishman, D., Valencia, A., Frishman, D., Albrecht, M., Blankenburg, H., Bork, P., *et al.* (2008). Protein-protein interactions: analysis and prediction. In *Modern genome annotation*, pages 353–410. Springer-Verlag, Vienna, Austria.

Fuglestad, B., Gasper, P. M., McCammon, J. A., Markwick, P. R. L., and Komives, E. A. (2013). Correlated motions and residual frustration in thrombin. *J Phys Chem B*, 117(42):12857–12863.

Fung, D. C. Y., Li, S. S., Goel, A., Hong, S.-H., and Wilkins, M. R. (2012). Visualization of the interactome: what are we looking at? *Proteomics*, 12(10):1669–1686.

Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software - Practice and Experience*, 30(11):1203–1233.

Garcia-Garcia, J., Guney, E., Aragues, R., Planas-Iglesias, J., and Oliva, B. (2010). Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics*, 11:56.

Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., *et al.* (2010). Visualization of omics data for systems biology. *Nat Methods*, 7(3 Suppl):S56–68.

Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Res*, 43(Database issue):D1049–56.

Gerasch, A., Faber, D., Küntzer, J., Niermann, P., Kohlbacher, O., Lenhof, H.-P., and Kaufmann, M. (2014). BiNA: a visual analytics tool for biological network data. *PLoS ONE*, 9(2):e87397.

Ghany, M. G., Nelson, D. R., Strader, D. B., Thomas, D. L., Seeff, L. B., and American Association for Study of Liver Diseases (2011). An update on treatment of genotype 1 chronic hepatitis C virus infection: 2011 practice guideline by the American Association for the Study of Liver Diseases. *Hepatology*, 54(4):1433–1444.

Ghiassian, S. D., Menche, J., and Barabási, A.-L. (2015). A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol*, 11(4):e1004120.

Ghosh, A., Brinda, K., and Vishveshwara, S. (2007). Dynamics of Lysozyme structure network: Probing the process of unfolding. *Biophysical Journal*, 92(7):2523–2535.

Ghosh, A., Sakaguchi, R., Liu, C., Vishveshwara, S., and Hou, Y. M. (2011). Allosteric communication in cysteinyl-tRNA synthetase: a network of direct and indirect readout. *J Biol Chem*, 286(43):37721–31.

Ghosh, A. and Vishveshwara, S. (2007). A study of communication pathways in methionyl-tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proc Natl Acad Sci U S A*, 104(40):15711–6.

Ghosh, S. and Vishveshwara, S. (2014). Ranking the quality of protein structure models using sidechain based network properties [v1; ref status: indexed, http://f1000r.es/2eu]. *F1000Research*, 3(17).

Gilad, Y. and Pritchard, J. K. (2010). eQTL resources from the Gilad/Pritchard group.

Girvan, M. and Newman, M. (2002). Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–7826.

Glossop, N., Lyons, L., and Hardin, P. (1999). Interlocked feedback loops within the Drosophila circadian oscillator. *Science*, 286(5440):766–768.

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., and Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*, 14(Suppl 3):S7.

Goh, K., Cusick, M., Valle, D., Childs, B., Vidal, M., and Barabási, A. (2007). The human disease network. *Proc Natl Acad Sci U S A.*, 104(21):8685–90.

Goldenberg, O., Erez, E., Nimrod, G., and Ben-Tal, N. (2009). The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res*, 37(Suppl 1):D323–D327.

Görner, K., Holtorf, E., Odoy, S., Nuscher, B., Yamamoto, A., Regula, J. T., *et al.* (2004). Differential effects of Parkinson's disease-associated mutations on stability and folding of DJ-1. *J Biol Chem*, 279(8):6943–6951.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864.

Greenamyre, J. T. and Hastings, T. G. (2004). Biomedicine. Parkinson's - divergent causes, convergent mechanisms. *Science*, 304(5674):1120–1122.

Greene, L. H. (2012). Protein structure networks. *Brief Funct Genomics*, 11(6):469–478.

Gu, H., Zhu, P., Jiao, Y., Meng, Y., and Chen, M. (2011). PRIN: a predicted rice interactome network. *BMC Bioinformatics*, 12:161.

Guala, D., Sjölund, E., and Sonnhammer, E. L. L. (2014). MaxLink: network-based prioritization of genes tightly linked to a disease seed set. *Bioinformatics*, 30(18):2689–2690.

Guariniello, S., Colonna, G., Raucci, R., Costantini, M., Di Bernardo, G., Bergantino, F., *et al.* (2014). Structure-function relationship and evolutionary history of the human selenoprotein M (SelM) found over-expressed in hepatocellular carcinoma. *Biochim Biophys Acta*, 1844(2):447–56.

Gulbahce, N. and Lehmann, S. (2008). The art of community detection. *Bioessays*, 30(10):934–938.

Gunasekaran, K., Ma, B., and Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins: Struct, Funct, Bioinf*, 57(3):433–443.

Guney, E., Garcia-Garcia, J., and Oliva, B. (2014). GUILDify: a web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms. *Bioinformatics*, 30(12):1789–1790.

Guo, X., Gao, L., Wei, C., Yang, X., Zhao, Y., and Dong, A. (2011). A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. *PLoS ONE*, 6(9):e24171.

Gustafsson, M., Nestor, C. E., Zhang, H., Barabási, A.-L., Baranzini, S., Brunak, S., *et al.* (2014). Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome Med*, 6(10):82.

Guzzi, P. H., Mina, M., Guerra, C., and Cannataro, M. (2012). Semantic similarity analysis of protein data: assessment with biological features and issues. *Brief Bioinform*, 13(5):569–585.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA.

Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). statnet: software tools for the representation, visualization, analysis and simulation of network data. *J Stat Softw*, 24:1.

Hansia, P., Ghosh, A., and Vishveshwara, S. (2009). Ligand dependent intra and inter subunit communication in human tryptophanyl-tRNA synthetase as deduced from the dynamics of structure networks. *Mol Biosyst*, 5(12):1860–72.

Harms, M. J. and Thornton, J. W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet*, 14(8):559–571.

Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52.

Hawkins, R. D., Hon, G. C., and Ren, B. (2010a). Next-generation genomics: an integrative approach. *Nat Rev Genet*, 11(7):476–86.

Hawkins, T., Chitale, M., and Kihara, D. (2010b). Functional enrichment analyses and construction of functional similarity networks with high confidence function prediction by PFP. *BMC Bioinformatics*, 11:265.

Hecht, M., Bromberg, Y., and Rost, B. (2013). News from the protein mutability landscape. *J Mol Biol*, 425(21):3937–3948.

Hedrich, K., Eskelson, C., Wilmot, B., Marder, K., Harris, J., Garrels, J., *et al.* (2004). Distribution, type, and origin of Parkin mutations: review and case studies. *Mov Disord*, 19(10):1146–1157.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450(7172):964–72.

Henzler-Wildman, K. A., Lei, M., Thai, V., Kerns, S. J., Karplus, M., and Kern, D. (2007). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916.

Hildebrandt, A., Kohlbacher, O., and Lenhof, H.-P. (2008). *Modeling Protein-Protein and Protein-DNA Docking*, pages 601–650. Wiley-VCH Verlag GmbH.

Hirschfield, G. M., Karlsen, T. H., Lindor, K. D., and Adams, D. H. (2013). Primary sclerosing cholangitis. *Lancet*, 382(9904):1587–1599.

Hirschhorn, J. N. and Gajdos, Z. K. (2011). Genome-wide association studies: results from the first few years and potential implications for clinical medicine. *Annu Rev Med*, 62:11–24.

Hoehn, M. and Yahr, M. (1967). Parkinsonism: onset, progression and mortality. *Neurology*, 17(5):427–442.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res*, 39(18):e119.

Holm, L. and Sander, C. (1996). Mapping the protein universe. *Science*, 273(5275):595–603.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*, 4(11):682–690.

Hou, L., Chen, M., Zhang, C. K., Cho, J., and Zhao, H. (2014). Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. *Hum Mol Genet*, 23(10):2780–2790.

Hu, G., Zhou, J., Yan, W., Chen, J., and Shen, B. (2013). The topology and dynamics of protein complexes: insights from intra- molecular network theory. *Curr Protein Pept Sci*, 14(2):121–132.

Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D., and Delisi, C. (2005). VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res*, pages W352–7.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57.

Hughes, D. and Andersson, D. I. (2015). Evolutionary consequences of drug resistance: shared principles across diverse targets and organisms. *Nat Rev Genet*, 16(8):459–471.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD – Visual Molecular Dynamics. *J Mol Graph*, 14:33–38.

Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., and DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief Bioinform*, 13(3):281–291.

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., and O'Donovan, C. (2015). The goa database: Gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063.

Huttenhower, C., Haley, E. M., Hibbs, M. A., Dumeaux, V., Barrett, D. R., Coller, H. A., and Troyanskaya, O. G. (2009). Exploring the human genome with functional maps. *Genome Res*, 19(6):1093–106.

Huttlin, E., Ting, L., Bruckner, R., Gebreab, F., Gygi, M., Szpyt, J., *et al.* (2015). The BioPlex Network: A systematic exploration of the human interactome. *Cell*, 162(2):425–440.

Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). Protein-protein docking benchmark version 4.0. *Proteins*, 78(15):3111–4.

Hwang, T., Atluri, G., Xie, M., Dey, S., Hong, C., Kumar, V., and Kuang, R. (2012). Co-clustering phenome-genome for phenotype classification and disease gene discovery. *Nucleic Acids Res*, 40(19):e146.

Hwang, T., Zhang, W., Xie, M., Liu, J., and Kuang, R. (2011). Inferring disease and gene set associations with rank coherence in networks. *Bioinformatics*, 27(19):2692–2699.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Mol Syst Biol*, 8:565.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–40.

Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Res*, 18(4):644–52.

Ideker, T., Thorsson, V., Ranish, J., Christmas, R., Buhler, J., Eng, J., *et al.* (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934.

Ikeda, F. and Dikic, I. (2008). Atypical ubiquitin chains: new molecular signals. 'protein modifications: Beyond the usual suspects' review series. *EMBO Rep*, 9(6):536–542.

Imai, Y., Soda, M., Hatakeyama, S., Akagi, T., Hashikawa, T., Nakayama, K. I., and Takahashi, R. (2002). CHIP is associated with Parkin, a gene responsible for familial Parkinson's disease, and enhances its ubiquitin ligase activity. *Mol Cell*, 10(1):55–67.

International Human Genome Sequencing Consortium, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Invernizzi, G., Lambrughi, M., Regonesi, M. E., Tortora, P., and Papaleo, E. (2013). The conformational ensemble of the disordered and aggregation-protective 182-291 region of ataxin-3. *Biochim Biophys Acta*, 1830(11):5236–47.

Invernizzi, G., Tiberti, M., Lambrughi, M., Lindorff-Larsen, K., and Papaleo, E. (2014). Communication routes in ARID domains between distal residues in helix 5 and the DNA-binding loops. *PLoS Comput Biol*, 10(9):e1003744.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., *et al.* (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci U S A*, 107(33):14621–14626.

Isogai, Y., Némethy, G., Rackovsky, S., Leach, S., and Scheraga, H. (1980). Characterization of multiple bends in proteins. *Biopolymers*, 19(6):1183–1210.

Ivanov, S. and Roy, C. R. (2009). NDP52: the missing link between ubiquitinated bacteria and autophagy. *Nat Immunol*, 10(11):1137–1139.

Janin, J. (2007). The targets of CAPRI rounds 6-12. *Proteins*, 69(4):699–703.

Janin, J. (2010). Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst*, 6(12):2351–62.

Janin, J. and Wodak, S. (1978). Conformation of amino acid side-chains in proteins. *J Mol Biol*, 125(3):357–386.

Javed, A., Agrawal, S., and Ng, P. C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat Methods*, 11(9):935–937.

Jegga, A. G., Schneider, L., Ouyang, X., and Zhang, J. (2011). Systems biology of the autophagy-lysosomal pathway. *Autophagy*, 7(5):477–489.

Jeong, H., Mason, S., Barabási, A., and Oltvai, Z. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z., and Barabási, A. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.

Jia, P. and Zhao, Z. (2014). Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Hum Genet*, 133(2):125–138.

Jia, P., Zheng, S., Long, J., Zheng, W., and Zhao, Z. (2011). dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27(1):95–102.

Jiang, M., Mani, N., Lin, C., Ardzinski, A., Nelson, M., Reagan, D., *et al.* (2013). In vitro phenotypic characterization of hepatitis C virus NS3 protease variants observed in clinical studies of telaprevir. *Antimicrob Agents Chemother*, 57(12):6236–6245.

Jiang, R., Gan, M., and He, P. (2011). Constructing a gene semantic similarity network for the inference of disease genes. *BMC Syst Biol*, 5 Suppl 2:S2.

Jiao, X. and Chang, S. (2011). Scoring function based on weighted residue network. *Int J Mol Sci*, 12(12):8773–86.

Jimenez-Sanchez, G., Childs, B., and Valle, D. (2001). Human disease genes. *Nature*, 409(6822):853–5.

Jones, D., Taylor, W., and Thornton, J. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS*, 8(3):275–282.

Jonsson, A. L., Scott, K. A., and Daggett, V. (2009). Dynameomics: a consensus view of the protein unfolding/folding transition state ensemble across a diverse set of protein folds. *Biophys J*, 97(11):2958–2966.

Jonsson, P. F. and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–7.

Jordan, D. M., Frangakis, S. G., Golzio, C., Cassa, C. A., Kurtzberg, J., *et al.* (2015). Identification of cis-suppression of human disease mutations by comparative genomics. *Nature*, 524(7564):225–229.

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., *et al.* (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124.

Junker, B. and Schreiber, F. (2008). *Analysis of biological networks.* John Wiley & Sons, Inc., Hoboken, NJ, USA.

Juretic, D., Lucic, B., Zucic, D., and Trinajstic, N. (1998). Protein transmembrane structure: recognition and prediction by using hydrophobicity scales through preference functions. In Parkanyi, C., editor, *Theoretical Organic Chemistry*, volume 5 of *Theoretical and Computational Chemistry*, pages 405–445. Elsevier.

Kacprowski, T. (2011). Disease gene prioritization by combining network information and functional knowledge. Masters thesis, Universität des Saarlandes, Saarbrücken.

Kacprowski, T., Doncheva, N. T., and Albrecht, M. (2013). NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, 29(11):1471–1473.

Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). ConsensusPathDB – a database for integrating human functional interaction networks. *Nucleic Acids Res*, 37(Database issue):D623–8.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27–30.

Kann, M. G. (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform*, 8(5):333–46.

Kann, M. G. (2009). Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief Bioinform*, 11(1):96–110.

Kannan, N. and Vishveshwara, S. (1999). Identification of side-chain clusters in protein structures by a graph spectral method. *J Mol Biol*, 292(2):441–464.

Karlsen, T. H. and Boberg, K. M. (2013). Update on primary sclerosing cholangitis. *J Hepatol*, 59(3):571–582.

Karlsen, T. H., Franke, A., Melum, E., Kaser, A., Hov, J. R., Balschun, T., *et al.* (2010a). Genome-wide association analysis in primary sclerosing cholangitis. *Gastroenterology*, 138(3):1102–1111.

Karlsen, T. H. and Kaser, A. (2011). Deciphering the genetic predisposition to primary sclerosing cholangitis. *Semin Liver Dis*, 31(2):188–207.

Karlsen, T. H., Schrumpf, E., and Boberg, K. M. (2010b). Update on primary sclerosing cholangitis. *Digestive and Liver Diseasei*, 42(6):390–400.

Karni, S., Soreq, H., and Sharan, R. (2009). A network-based method for predicting disease-causing genes. *J Comput Biol*, 16(2):181–9.

Karubiu, W., Bhakat, S., and Soliman, M. E. S. (2014). Compensatory role of double mutation N348I/M184V on nevirapine binding landscape: insight from molecular dynamics simulation. *Protein J*, 33(5):432–446.

Kawashima, S. and Kanehisa, M. (2000). AAindex: Amino Acid index database. *Nucleic Acids Res*, 28(1):374.

Kawashima, S., Ogata, H., and Kanehisa, M. (1999). AAindex: Amino Acid index database. *Nucleic Acids Res*, 27(1):368–369.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: Amino Acid index database, progress report 2008. *Nucleic Acids Res*, 36(Database issue):D202–5.

Keeling, M. J. and Eames, K. T. D. (2005). Networks and epidemic models. *J R Soc Interface*, 2(4):295–307.

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., *et al.* (2009). Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181.

Kelder, T., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2010). Finding the right questions: exploratory pathway analysis to enhance biological discovery in large datasets. *PLoS Biol*, 8(8):4414–4426.

Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., *et al.* (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–72.

Khashan, R., Zheng, W., and Tropsha, A. (2012). Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. *Proteins*, 80(9):2207–17.

Khor, B., Gardet, A., and Xavier, R. J. (2011). Genetics and pathogenesis of inflammatory bowel disease. *Nature*, 474(7351):307–317.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., *et al.* (2014). A draft map of the human proteome. *Nature*, 509(7502):575–581.

Kitada, T., Asakawa, S., Hattori, N., Matsumine, H., Yamamura, Y., Minoshima, S., *et al.* (1998). Mutations in the Parkin gene cause autosomal recessive juvenile parkinsonism. *Nature*, 392(6676):605–608.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.

Kitano, H. (2004). Biological robustness. *Nat Rev Genet*, 5(11):826–837.

Klein, P., Kanehisa, M., and DeLisi, C. (1984). Prediction of protein function from sequence properties. discriminant analysis of a database. *Biochim Biophys Acta*, 787(3):221–226.

Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol*, 19(2):120–7.

Knisley, D. and Knisley, J. (2014). Seeing the results of a mutation with a vertex weighted hierarchical graph. *BMC Proceedings*, 8(Suppl 2 Proceedings of the 3rd Annual Symposium on Biol):S7.

Köhler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P., and Philippi, S. (2006). Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–90.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*, 82(4):949–58.

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., *et al.* (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*, 42(Database issue):D966–74.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., *et al.* (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*, 85(4):457–464.

Korioth, F., Gieffers, C., Maul, G., and Frey, J. (1995). Molecular characterization of NDP52, a novel protein of the nuclear domain 10, which is redistributed upon virus infection and interferon treatment. *J Cell Biol*, 130(1):1–13.

Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. (2004). Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc Natl Acad Sci U S A.*, 101(42):15148–53.

Krishnamoorthy, B. and Tropsha, A. (2003). Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 19(12):1540–1548.

Krishnan, A., Zbilut, J. P., Tomita, M., and Giuliani, A. (2008). Proteins as networks: usefulness of graph theory in protein science. *Curr Protein Pept Sci*, 9(1):28–38.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., *et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637643.

Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W., and Przulj, N. (2010). Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society, Interface / the Royal Society*, 7(50):1341–1354.

Kuchaiev, O. and Przulj, N. (2011). Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396.

Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., *et al.* (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–16.

Lambrughi, M., Papaleo, E., Testa, L., Brocca, S., De Gioia, L., and Grandori, R. (2012). Intramolecular interactions stabilizing compact conformations of the intrinsically disordered kinase-inhibitor domain of Sic1: a molecular dynamics investigation. *Frontiers in Physiology*, 3(435).

Langfelder, P. and Horvath, S. (2008). WGCNA: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

Lawitz, E., Sulkowski, M., Jacobson, I., Kraft, W. K., Maliakkal, B., Al-Ibrahim, M., Gordon, S. C., Kwo, P., Rockstroh, J. K., Panorchan, P., Miller, M., Caro, L., Barnard, R., Hwang, P. M., Gress, J., Quirk, E., and Mobashery, N. (2013). Characterization of vaniprevir, a hepatitis c virus ns3/4a protease inhibitor, in patients with hcv genotype 1 infection: safety, antiviral activity, resistance, and pharmacokinetics. *Antiviral research*, 99(3):214220.

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res*, 21(7):1109–21.

Lee, J. M. and Sonnhammer, E. L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res*, 13(5):875–82.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., *et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804.

Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nat Rev Genet*, 14(3):168–178.

Lemetre, C., Zhang, Q., and Zhang, Z. D. (2013). SubNet: a Java application for subnetwork extraction. *Bioinformatics*, 29(22):2958–2958.

Lensink, M. F., Mendez, R., and Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins*, 69(4):704–18.

Lensink, M. F. and Wodak, S. J. (2010). Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins*, 78(15):3085–3095.

Lensink, M. F. and Wodak, S. J. (2013). Docking, scoring, and affinity prediction in CAPRI. *Proteins*, 81(12):2082–2095.

Lensink, M. F. and Wodak, S. J. (2014). Score_set: A CAPRI benchmark for scoring protein complexes. *Proteins: Struct, Funct, Bioinf*, 82(11):3163–3169.

Letovsky, S. and Kasif, S. (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19 Suppl 1:i197–204.

Letunic, I., Doerks, T., and Bork, P. (2015). SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res*, 43(Database issue):D257–60.

Leulliot, N., Chaillet, M., Durand, D., Ulryck, N., Blondeau, K., and van Tilbeurgh, H. (2008). Structure of the yeast tRNA m7G methylation complex. *Structure*, 16(1):52–61.

Li, Y., Huang, T., Xiao, Y., Ning, S., Wang, P., Wang, Q., *et al.* (2013). Prioritising risk pathways of complex human diseases based on functional profiling. *Eur J Hum Genet*, 21(6):666–672.

Li, Y. and Patra, J. C. (2010a). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–24.

Li, Y. and Patra, J. C. (2010b). Integration of multiple data sources to prioritize candidate genes using discounted rating system. *BMC Bioinformatics*, 11 Suppl 1:S20.

Li, Y. R., Li, J., Zhao, S. D., Bradfield, J. P., Mentch, F. D., Maggadottir, S. M., *et al.* (2015). Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat Med*, 21(9):1018–1027.

Liekens, A. M., De Knijf, J., Daelemans, W., Goethals, B., De Rijk, P., and Del-Favero, J. (2011). BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol*, 12(6):R57.

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabó, G., Rual, J. F., *et al.* (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, 125(4):801–14.

Lima-Mendez, G. and van Helden, J. (2009). The powerful law of the power law and other myths in network biology. *Mol Biosyst*, 5(12):1482–93.

Limviphuvadh, V., Tanaka, S., Goto, S., Ueda, K., and Kanehisa, M. (2007). The commonality of protein interaction networks determined in neurodegenerative disorders (NDDs). *Bioinformatics*, 23(16):2129–2138.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Linghu, B., Snitkin, E. S., Hu, Z., Xia, Y., and Delisi, C. (2009). Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol*, 10(9):R91.

Liu, J. Z. and Anderson, C. A. (2014). Genetic studies of Crohn's disease: past, present and future. *Best Pract Res Clin Gastroenterol*, 28(3):373–386.

Liu, J. Z., Hov, J. R., Folseraas, T., Ellinghaus, E., Rushbrook, S. M., Doncheva, N. T., *et al.* (2013). Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat Genet*, 45(6):670–677.

Liu, S., Gao, Y., and Vakser, I. A. (2008). DOCKGROUND protein-protein docking decoy set. *Bioinformatics*, 24(22):2634–5.

Lolis, E. and Petsko, G. (1990). Crystallographic analysis of the complex between triosephosphate isomerase and 2-phosphoglycolate at 2.5-A resolution: implications for catalysis. *Biochemistry*, 29(28):6619–6625.

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., and Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26(18):2347–2348.

López-Bigas, N. and Ouzounis, C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*, 32(10):3108–14.

Lorenz, W., Alba, R., Yu, Y., Bordeaux, J., Simões, M., and Dean, J. (2011). Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (P. taeda L.). *BMC Genomics*, 12:264.

Lotia, S., Montojo, J., Dong, Y., Bader, G. D., and Pico, A. R. (2013). Cytoscape app store. *Bioinformatics*, 29(10):1350–1351.

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., *et al.* (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403):361–367.

Love, R., Parge, H., Wickersham, J., Hostomsky, Z., Habuka, N., Moomaw, E., *et al.* (1996). The crystal structure of hepatitis C virus NS3 proteinase reveals a trypsin-like fold and a structural zinc binding site. *Cell*, 87(2):331–342.

Lowe, H. J. and Barnett, G. O. (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA*, 271(14):1103–8.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., *et al.* (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, 380(9859):2095–2128.

Luciani, T., Wenskovitch, J., Chen, K., Koes, D., Travers, T., and Marai, G. E. (2014). Fixing-TIM: interactive exploration of sequence and structural data to identify functional mutations in protein families. *BMC Proceedings*, 8(Suppl 2 Proceedings of the 3rd Annual Symposium on Biol):S3.

Macedo, M. G., Anar, B., Bronner, I. F., Cannella, M., Squitieri, F., Bonifati, V., *et al.* (2003). The DJ-1 L166P mutant protein associated with early onset Parkinson's disease is unstable and forms higher-order protein complexes. *Hum Mol Genet*, 12(21):2807–2816.

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*, 10(8):565–77.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449.

Magger, O., Waldman, Y. Y., Ruppin, E., and Sharan, R. (2012). Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput Biol*, 8(9):e1002690.

Maggiora, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *J Med Chem*, 57(8):3186–3204. PMID: 24151987.

Mah, J. T., Low, E. S., and Lee, E. (2011). In silico SNP analysis and bioinformatics tools: a review of the state of the art to aid drug discovery. *Drug Discovery Today*, 16(17 - 18):800–809.

Malgieri, G. and Eliezer, D. (2008). Structural effects of Parkinson's disease linked DJ-1 mutations. *Protein Sci*, 17(5):855–868.

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A*, 100(21):11980–11985.

Manley, G., Rivalta, I., and Loria, J. P. (2013). Solution NMR and computational methods for understanding protein allostery. *J Phys Chem B*, 117(11):3063–3073.

Manns, M. P., McHutchison, J. G., Gordon, S. C., Rustgi, V. K., Shiffman, M., Reindollar, R., *et al.* (2001). Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. *Lancet*, 358(9286):958–965.

Manolio, T. A. (2010). Genome-wide association studies and assessment of the risk of disease. *N Engl J Med*, 363(2):166–76.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Mariani, S., DellOrco, D., Felline, A., Raimondi, F., and Fanelli, F. (2013). Network and atomistic simulations unveil the structural determinants of mutations linked to retinal diseases. *PLoS Comput Biol*, 9(8):e1003207.

Marras, C., Lohmann, K., Lang, A., and Klein, C. (2012). Fixing the broken system of genetic locus symbols: Parkinson disease and dystonia as examples. *Neurology*, 78(13):1016–1024.

Martin, A. J., Vidotto, M., Boscariol, F., Di Domenico, T., Walsh, I., and Tosatto, S. C. (2011). RING: networking interacting residues, evolutionary information and energetics in protein structures. *Bioinformatics*, 27(14):2003–5.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, 296(5569):910.

Masoudi-Nejad, A., Schreiber, F., and Kashani, Z. R. M. (2012). Building blocks of biological networks: a review on major network motif discovery algorithms. *IET systems biology*, 6(5):164–174.

McCammon, J., Gelin, B., and Karplus, M. (1977). Dynamics of folded proteins. *Nature*, 267(5612):585–590.

McCandlish, D. M., Rajon, E., Shah, P., Ding, Y., and Plotkin, J. B. (2013). The role of epistasis in protein evolution. *Nature*, 497(7451):E12; discussion E23.

McGeagh, J. D., Ranaghan, K. E., and Mulholland, A. J. (2011). Protein dynamics and enzyme catalysis: insights from simulations. *Biochim Biophys Acta*, 1814(8):1077–92.

McPhee, F., Friborg, J., Levine, S., Chen, C., Falk, P., Yu, F., Hernandez, D., Lee, M. S., Chaniewski, S., Sheaffer, A. K., and Pasquinelli, C. (2012). Resistance analysis of the hepatitis c virus ns3 protease inhibitor asunaprevir. *Antimicrobial agents and chemotherapy*, 56(7):36703681.

McPhee, F., Hernandez, D., Yu, F., Ueland, J., Monikowski, A., Carifa, A., Falk, P., Wang, C., Fridell, R., Eley, T., Zhou, N., and Gardiner, D. (2013). Resistance analysis of hepatitis c virus genotype 1 prior treatment null responders receiving daclatasvir and asunaprevir. *Hepatology (Baltimore, Md.)*, 58(3):902911.

Mehlhorn, K. and Näher, S. (1999). *LEDA: a platform for combinatorial and geometric computing.* Cambridge University Press.

Melum, E., Franke, A., Schramm, C., Weismüller, T. J., Gotthardt, D. N., Offner, F. A., *et al.* (2011). Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci. *Nat Genet*, 43(1):17–19.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224).

Mendez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*, 52(1):51–67.

Mendez, R., Leplae, R., Lensink, M. F., and Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60(2):150–69.

Menetrey, J., Perderiset, M., Cicolari, J., Dubois, T., Elkhatib, N., El Khadali, F., *et al.* (2007). Structural basis for ARF1-mediated recruitment of ARHGAP21 to Golgi membranes. *EMBO J*, 26(7):1953–1962.

Meng, E., Pettersen, E., Couch, G., Huang, C., and Ferrin, T. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, 7(1):339.

Mercer, J., Pandian, B., Lex, A., Bonneel, N., and Pfister, H. (2014). Mu-8: visualizing differences between proteins and their families. *BMC Proceedings*, 8(Suppl 2 Proceedings of the 3rd Annual Symposium on Biol):S5.

Merico, D., Gfeller, D., and Bader, G. D. (2009). How to visually interpret biological data using networks. *Nat Biotechnol*, 27(10):921–924.

Meyniel-Schicklin, L., de Chassey, B., André, P., and Lotteau, V. (2012). Viruses and interactomes in translation. *Mol Cell Proteomics*, 11(7):M111.014738.

Miao, Y., Nichols, S. E., Gasper, P. M., Metzger, V. T., and McCammon, J. A. (2013). Activation and dynamic network of the m2 muscarinic receptor. *Proc Natl Acad Sci U S A*, 110(27):10982–10987.

Micheelsen, P. O., Vévodová, J., De Maria, L., Ostergaard, P. R., Friis, E. P., Wilson, K., and Skjøt, M. (2008). Structural and mutational analyses of the interaction between the barley alpha-amylase/subtilisin inhibitor and the subtilisin savinase reveal a novel mode of inhibition. *J Mol Biol*, 380(4):681–690.

Millán, P. P. (2013). Visualization and analysis of biological networks. *Methods in molecular biology (Clifton, N.J.)*, 1021:63–88.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., *et al.* (2004). Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet*, 14(10):719–732.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., *et al.* (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol*, 33(3):269–276.

Molodecky, N. A., Soon, I. S., Rabi, D. M., Ghali, W. A., Ferris, M., Chernoff, G., *et al.* (2012). Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*, 142(1):46–54.e42; quiz e30.

Mordelet, F. and Vert, J. P. (2011). ProDiGe: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, 12:389.

Moreau, Y. and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet*, 13(8):523–536.

Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2010). Protein-protein docking dealing with the unknown. *J Comput Chem*, 31(2):317–342.

Morgan, T. R., Ghany, M. G., Kim, H.-Y., Snow, K. K., Shiffman, M. L., De Santo, J. L., *et al.* (2010). Outcome of sustained virological responders with histologically advanced chronic hepatitis C. *Hepatology*, 52(3):833–844.

Morris, J. (2015). structureViz website. http://www.cgl.ucsf.edu/cytoscape/structureViz2/index.shtml.

Morris, J. H., Huang, C. C., Babbitt, P. C., and Ferrin, T. E. (2007). structureViz: linking Cytoscape and UCSF Chimera. *Bioinformatics*, 23(17):2345–2347.

Morris, J. H., Knudsen, G. M., Verschueren, E., Johnson, J. R., Cimermancic, P., Greninger, A. L., and Pico, A. R. (2014). Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. *Nat Protoc*, 9(11):2539–2554.

Morris, J. H., Kuchinsky, A., and Pico, A. (2015a). Analysis and visualization of biological networks with Cytoscape.

Morris, J. H., Lotia, S., Wu, A., Doncheva, N. T., Albrecht, M., and Ferrin, T. E. (2015b). setsApp for Cytoscape: Set operations for Cytoscape nodes and edges [version 2; referees: 3 approved]. *F1000Research*, 3(149).

Morris, J. H., Meng, E. C., and Ferrin, T. E. (2010). Computational tools for the interactive exploration of proteomic and structural data. *Mol Cell Proteomics*, 9(8):1703–1715.

Morriswood, B., Ryzhakov, G., Puri, C., Arden, S. D., Roberts, R., Dendrou, C., *et al.* (2007). T6BP and NDP52 are myosin VI binding partners with potential roles in cytokine signalling and cell adhesion. *J Cell Sci*, 120(Pt 15):2574–2585.

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. *Nat Methods*, 10(1):47–53.

Mosca, R., Tenorio-Laranga, J., Olivella, R., Alcalde, V., Céol, A., Soler-López, M., and Aloy, P. (2015). dSysMap: exploring the edgetic role of disease mutations. *Nat Methods*, 12(3):167168.

Mueller, L. A. J., Kugler, K. G., Dander, A., Graber, A., and Dehmer, M. (2011). QuACN: an R package for analyzing complex biological networks quantitatively. *Bioinformatics*, 27(1):140–141.

Munz, M. and Biggin, P. C. (2012). JGromacs: a Java package for analyzing protein simulations. *J Chem Inf Model*, 52(1):255–9.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540.

Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1:i302–310.

Najmudin, S., Pinheiro, B. A., Prates, J. A. M., Gilbert, H. J., Romão, M. J., and Fontes, C. M. G. A. (2010). Putting an N-terminal end to the Clostridium thermocellum xylanase Xyn10B story: crystal structure of the CBM22-1-GH10 modules complexed with xylohexaose. *J Struct Biol*, 172(3):353–362.

Navlakha, S. and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–63.

Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2015). CyToStruct: Augmenting the network visualization of cytoscape with the power of molecular viewers. *Structure*, 23(5):941–948.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods*, 9(5):471–472.

Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Soc Networks*, 27:39–54.

Niki, T., Takahashi-Niki, K., Taira, T., Iguchi-Ariga, S. M. M., and Ariga, H. (2003). DJBP: a novel DJ-1-binding protein, negatively regulates the androgen receptor by recruiting histone deacetylase complex, and DJ-1 antagonizes this inhibition by abrogation of this complex. *Mol Cancer Res*, 1(4):247–261.

Nitsch, D., Gonçalves, J. P., Ojeda, F., De Moor, B., and Moreau, Y. (2010). Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, 11:460.

Nitsch, D., Tranchevent, L., Gonçalves, J., Vogt, J., Madeira, S., and Moreau, Y. (2011). PINTA: a web server for network-based gene prioritization from expression data. *Nucleic Acids Res*, 39(Web Server issue):W334–8.

Nitsch, D., Tranchevent, L. C., Thienpont, B., Thorrez, L., Van Esch, H., Devriendt, K., and Moreau, Y. (2009). Network analysis of differential expression for the identification of disease-causing genes. *PLoS ONE*, 4(5):e5526.

Nussbaum, R. L. and Ellis, C. E. (2003). Alzheimer's disease and Parkinson's disease. *N Engl J Med*, 348(14):1356–1364.

O'Donoghue, S. I., Sabir, K. S., Kalemanov, M., Stolte, C., Wellmann, B., Ho, V., *et al.* (2015). Aquaria: simplifying discovery and insight from protein structures. *Nat Methods*, 12(2):98–99.

Oliva, R., Vangone, A., and Cavallo, L. (2013). Ranking multiple docking solutions based on the conservation of inter-residue contacts. *Proteins*, 81(9):1571–84.

Olzmann, J. A., Brown, K., Wilkinson, K. D., Rees, H. D., Huai, Q., Ke, H., *et al.* (2004). Familial Parkinson's disease-associated L166P mutation disrupts DJ-1 protein folding and function. *J Biol Chem*, 279(9):8506–8515.

O'Madadhain, J., Fisher, D., White, S., and Boey, Y. (2003). The JUNG (Java Universal Network/Graph) Framework. *Technical Report UCI-ICS*, 03-17.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: generalizing degree and shortest paths. *Soc Networks*, 32(3):245–251.

Orozco, M. (2014). A theoretical view of protein dynamics. *Chem Soc Rev*, 43(14):5051–66.

Oti, M. and Brunner, H. G. (2007). The modular nature of genetic diseases. *Clin Genet*, 71(1):1–11.

Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *J Med Genet*, 43(8):691–8.

Panni, S. and Rombo, S. E. (2015). Searching for repetitions in biological networks: methods, resources and tools. *Brief Bioinform*, 16(1):118–136.

Papaleo, E., Casiraghi, N., Arrigoni, A., Vanoni, M., Coccetti, P., and De Gioia, L. (2012a). Loop 7 of E2 enzymes: an ancestral conserved functional motif involved in the E2-mediated steps of the ubiquitination cascade. *PLoS ONE*, 7(7):e40786.

Papaleo, E., Renzetti, G., Invernizzi, G., and Asgeirsson, B. (2013). Dynamics fingerprint and inherent asymmetric flexibility of a cold-adapted homodimeric enzyme. A case study of the Vibrio alkaline phosphatase. *Biochim Biophys Acta*, 1830(4):2970–2980.

Papaleo, E., Renzetti, G., and Tiberti, M. (2012b). Mechanisms of intramolecular communication in a hyperthermophilic acylaminoacyl peptidase: A molecular dynamics investigation. *PLoS ONE*, 7(4):e35686.

Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T., and Iliopoulos, I. (2015). Protein-protein interaction predictions using text mining methods. *Methods (San Diego, Calif.)*, 74:47–53.

Parkes, M., Cortes, A., van Heel, D. A., and Brown, M. A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet*, 14(9):661–673.

Pasi, M., Tiberti, M., Arrigoni, A., and Papaleo, E. (2012). xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. *J Chem Inf Model*, 52(7):1865–74.

Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Phys Rev Lett*, 86(14):3200–3203.

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., and Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Min*, 4:10.

Pavlopoulos, G. A., Wegener, A.-L., and Schneider, R. (2008). A survey of visualization tools for biological network analysis. *BioData Min*, 1:12.

Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nat Genet*, 31(3):316–9.

Pers, T. H., Hansen, N. T., Lage, K., Koefoed, P., Dworzynski, P., Miller, M. L., *et al.* (2011). Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. *Genet Epidemiol*, 35(5):318–32.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 5(7):e1000443.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–1612.

Pico, A. R., Bader, G. D., Demchak, B., Guitart Pla, O., Hull, T., Longabaugh, W., *et al.* (2014). The Cytoscape app article collection. *F1000Research*, 3:138.

Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008). WikiPathways: pathway editing for the people. *PLoS Biol*, 6(7):e184.

Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E., and Ziv-Ukelson, M. (2005). Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–3408.

Piro, R. M. and Di Cunto, F. (2012). Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J*, 279(5):678–96.

Poirel, C. L., Rahman, A., Rodrigues, R. R., Krishnan, A., Addesa, J. R., and Murali, T. (2013). Reconciling differential gene expression data with molecular interaction networks. *Bioinformatics*, 29(5):622–629.

Przulj, N. (2011). Protein-protein interactions: making sense of networks via graph-theoretic modeling. *Bioessays*, 33(2):115–123.

Przytycka, T. M., Singh, M., and Slonim, D. K. (2010). Toward the dynamic interactome: it's about time. *Brief Bioinform*, 11(1):15–29.

Radivojac, P., Peng, K., Clark, W. T., Peters, B. J., Mohan, A., Boyle, S. M., and Mooney, S. D. (2008). An integrated approach to inferring gene-disease associations in humans. *Proteins*, 72(3):1030–7.

Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G., and Schwartz, J. (2010). Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst Biol*, 4:114.

Ragusa, M., Avola, G., Angelica, R., Barbagallo, D., Guglielmino, M., Duro, L., *et al.* (2010). Expression profile and specific network features of the apoptotic machinery explain relapse of acute myeloid leukemia after chemotherapy. *BMC Cancer*, 10:377.

Raimondi, F., Felline, A., Seeber, M., Mariani, S., and Fanelli, F. (2013). A mixed protein structure network and elastic network model approach to predict the structural communication in biomolecular systems: The PDZ2 domain from tyrosine phosphatase 1E as a case study. *J Chem Theory Comput*, 9(5):2504–2518.

Ramírez, F., Lawyer, G., and Albrecht, M. (2012). Novel search method for the discovery of functional relationships. *Bioinformatics*, 28(2):269–276.

Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T., and Albrecht, M. (2007). Computational analysis of human protein interaction networks. *Proteomics*, 7(15):2541–2552.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551.

Ray, W. C., Rumpf, R. W., Sullivan, B., Callahan, N., Magliery, T., Machiraju, R., *et al.* (2014). Understanding the sequence requirements of protein families: insights from the BioVis 2013 contests. *BMC Proceedings*, 8(Suppl 2 Proceedings of the 3rd Annual Symposium on Biol):S1.

Raychaudhuri, S. (2011). Mapping rare and common causal alleles for complex human diseases. *Cell*, 147(1):57–69.

Raychaudhuri, S., Plenge, R. M., Rossin, E. J., Ng, A. C. Y., International Schizophrenia Consortium, Purcell, S. M., *et al.* (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic snp associations and rare deletions. *PLoS Genet*, 5(6):e1000534.

Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9:405.

Reiss, S., Rebhan, I., Backes, P., Romero-Brey, I., Erfle, H., Matula, P., *et al.* (2011). Recruitment and activation of a lipid kinase by hepatitis C virus NS5A is essential for integrity of the membranous replication compartment. *Cell host & microbe*, 9(1):32–45.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14^{th} International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Resnik, P. (1998). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

Ritchie, D. W. (2008). Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci*, 9(1):1–15.

Rives, A. W. and Galitski, T. (2003). Modular organization of cellular networks. *Proc Natl Acad Sci U S A*, 100(3):11281133.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 83(5):610–615.

Robinson, P. N., Köhler, S., Oellrich, A., Sanger Mouse Genetics Project, Wang, K., Mungall, C. J., *et al.* (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*, 24(2):340–348.

Roca, A. (2014). ProfileGrids: a sequence alignment visualization paradigm that avoids the limitations of sequence logos. *BMC Proceedings*, 8(Suppl 2 Proceedings of the 3rd Annual Symposium on Biol):S6.

Rodrigues, J. P. G. L. M. and Bonvin, A. M. J. J. (2014). Integrative computational modeling of protein interactions. *FEBS J*, 281(8):1988–2003.

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., *et al.* (2014). A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226.

Romano, K. P., Ali, A., Royer, W. E., and Schiffer, C. A. (2010). Drug resistance against HCV NS3/4A inhibitors is defined by the balance of substrate recognition versus inhibitor binding. *Proc Natl Acad Sci U S A*, 107(49):20986–20991.

Roque, A. C. A. and Lowe, C. R. (2008). Affinity chromatography: history, perspectives, limitations and prospects. *Methods in molecular biology (Clifton, N.J.)*, 421:1–21.

Rose, P. W., Bi, C., Bluhm, W. F., Christie, C. H., Dimitropoulos, D., Dutta, S., *et al.* (2013). The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res*, 41(D1):D475–D482.

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., *et al.* (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet*, 7(1):e1001273.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., *et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178.

Rutherford, K., Alphandéry, E., McMillan, A., Daggett, V., and Parson, W. W. (2008a). The V108M mutation decreases the structural stability of catechol O-methyltransferase. *Biochim Biophys Acta*, 1784(7-8):1098–1105.

Rutherford, K., Bennion, B. J., Parson, W. W., and Daggett, V. (2006). The 108M polymorph of human catechol O-methyltransferase is prone to deformation at physiological temperatures. *Biochemistry*, 45(7):2178–2188.

Rutherford, K. and Daggett, V. (2008). Four human thiopurine s-methyltransferase alleles severely affect protein structure and dynamics. *J Mol Biol*, 379(4):803–814.

Rutherford, K. and Daggett, V. (2009). A hotspot of inactivation: The A22S and V108M polymorphisms individually destabilize the active site structure of catechol O-methyltransferase. *Biochemistry*, 48(27):6450–6460.

Rutherford, K., Parson, W. W., and Daggett, V. (2008b). The histamine N-methyltransferase T105I polymorphism affects active site structure and dynamics. *Biochemistry*, 47(3):893–901.

Saari, D. G. (1999). Explaining all three-alternative voting outcomes. *Journal of Economic Theory*, 87:313–355.

Saarinen, S., Olerup, O., and Broomé, U. (2000). Increased frequency of autoimmune diseases in patients with primary sclerosing cholangitis. *Am J Gastroenterol*, 95(11):3195–3199.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J., Coulombe-Huntington, J., Yang, F., *et al.* (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3):647–660.

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., *et al.* (2012). A travel guide to Cytoscape plugins. *Nat Methods*, 9(11):1069–1076.

Scarabelli, G. and Grant, B. (2014). Kinesin-5 allosteric inhibitors uncouple the dynamics of nucleotide, microtubule, and neck-linker binding sites. *Biophysical Journal*, 107(9):2204–2213.

Scarabelli, G. and Grant, B. J. (2013). Mapping the structural and dynamical features of kinesin motor domains. *PLoS Comput Biol*, 9(11):e1003329.

Scardoni, G., Petterlini, M., and Laudanna, C. (2009). Analyzing biological network parameters with CentiScaPe. *Bioinformatics*, 25(21):2857–2859.

Schlicker, A. and Albrecht, M. (2008). FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res*, 36(Database issue):D434–9.

Schlicker, A. and Albrecht, M. (2010). FunSimMat update: new features for exploring functional similarity. *Nucleic Acids Res*, 38(Database issue):D244–8.

Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7:302.

Schlicker, A., Huthmacher, C., Ramírez, F., Lengauer, T., and Albrecht, M. (2007a). Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23(7):859–865.

Schlicker, A., Lengauer, T., and Albrecht, M. (2010). Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*, 26(18):i561–7.

Schlicker, A., Rahnenführer, J., Albrecht, M., Lengauer, T., and Domingues, F. S. (2007b). GO-Tax: investigating biological processes and biochemical activities along the taxonomic tree. *Genome Biol*, 8(3):R33.

Schneider, M. D. and Sarrazin, C. (2014). Antiviral therapy of hepatitis C in 2014: do we need resistance testing? *Antiviral Res*, 105:64–71.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res*, 33(Web Server issue):W363–7.

Schneidman-Duhovny, D., Inbar, Y., Polak, V., Shatsky, M., Halperin, I., Benyamini, H., *et al.* (2003). Taking geometry to its edge: Fast unbound rigid (and hinge-bent) docking. *Proteins: Struct, Funct, Bioinf*, 52(1):107–112.

Schneidman-Duhovny, D., Nussinov, R., and Wolfson, H. J. (2004). Predicting molecular interactions in silico: II. protein-protein and protein-drug docking. *Curr Med Chem*, 11(1):91–107.

Schneidman-Duhovny, D., Rossi, A., Avila-Sakar, A., Kim, S. J., Velázquez-Muriel, J., Strop, P., *et al.* (2012). A method for integrative structure determination of protein-protein complexes. *Bioinformatics*, 28(24):3282–3289.

Schreiber, S., Rosenstiel, P., Albrecht, M., Hampe, J., and Krawczak, M. (2005). Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat Rev Genet*, 6(5):376–88.

Schrödinger, LLC (2010). The PyMOL molecular graphics system, version 1.3r1.

Schuierer, S., Tranchevent, L., Dengler, U., and Moreau, Y. (2010). Large-scale benchmark of Endeavour using MetaCore maps. *Bioinformatics*, 26(15):1922–3.

Seeber, M., Felline, A., Raimondi, F., Mariani, S., and Fanelli, F. (2014). WebPSN: a web server for high-throughput investigation of structural communication in biomacromolecules. *Bioinformatics*.

Seeber, M., Felline, A., Raimondi, F., Muff, S., Friedman, R., Rao, F., *et al.* (2011). Wordom: a user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *J Comput Chem*, 32(6):1183–94.

Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. (2009). Dynamical networks in tRNA:protein complexes. *Proc Natl Acad Sci U S A*, 106(16):6620–5.

Sethi, A., Tian, J., Derdeyn, C. A., Korber, B., and Gnanakaran, S. (2013). A mechanistic understanding of allosteric immune escape pathways in the HIV-1 envelope glycoprotein. *PLoS Comput Biol*, 9(5):e1003046.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., *et al.* (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504.

Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24(4):427–433.

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., *et al.* (2005). Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A*, 102(6):1974–1979.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol Syst Biol*, 3:88.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, 31(1):64–68.

Shendelman, S., Jonason, A., Martinat, C., Leete, T., and Abeliovich, A. (2004). DJ-1 is a redox-dependent molecular chaperone that inhibits alpha-synuclein aggregate formation. *PLoS Biol*, 2(11):e362.

Shimakami, T., Welsch, C., Yamane, D., McGivern, D. R., Yi, M., Zeuzem, S., and Lemon, S. M. (2011). Protease inhibitor-resistant hepatitis C virus mutants with reduced fitness from impaired production of infectious virus. *Gastroenterology*, 140(2):667–675.

Shimbel, A. (1953). Structural parameters of communication networks. *Bull Math Biophys*, 15:501–507.

Shimura, H., Hattori, N., Kubo, S., Mizuno, Y., Asakawa, S., Minoshima, S., *et al.* (2000). Familial Parkinson disease gene product, Parkin, is a ubiquitin-protein ligase. *Nat Genet*, 25(3):302–305.

Shlomi, T., Segal, D., Ruppin, E., and Sharan, R. (2006). QPath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, 7:199.

Silveira, S. A., Fassio, A. V., Gonçalves-Almeida, V. M., de Lima, E. B., Barcelos, Y. T., Aburjaile, F. F., *et al.* (2014). VERMONT: Visualizing mutations and their effects on protein physico-chemical and topological property conservation. *BMC Proceedings*, 8(Suppl 2 Proceedings of the 3rd Annual Symposium on Biol):S4.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*, 40(W1):W452–W457.

Singleton, M., Guthery, S., Voelkerding, K., Chen, K., Kennedy, B., Margraf, R., *et al.* (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am J Hum Genet*, 94(4):599–610.

Skrabanek, L., Saini, H. K., Bader, G. D., and Enright, A. J. (2008). Computational prediction of protein-protein interactions. *Mol Biotechnol*, 38(1):1–17.

Smedley, D., Köhler, S., Czeschik, J. C., Amberger, J., Bocchini, C., Hamosh, A., *et al.* (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, 30(22):3215–3222.

Speicher, N. K. (2010). Network analysis of viral host factors. Bachelors thesis, Universität des Saarlandes, Saarbrücken.

Spillantini, M., Crowther, R., Jakes, R., Hasegawa, M., and Goedert, M. (1998). Alpha-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with lewy bodies. *Proc Natl Acad Sci U S A*, 95(11):6469–6473.

Spirin, V. and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, 100(21):1212312128.

Srivas, R., Hannum, G., Rushcheinski, J., Ono, K., Wang, P., Smoot, M., and Ideker, T. (2011). Assembling global maps of cellular function through integrative analysis of physical and genetic networks. *Nat Protoc*, 6(9):1308–23.

Srivastava, B., Mells, G. F., Cordell, H. J., Muriithi, A., Brown, M., Ellinghaus, E., *et al.* (2012). Fine mapping and replication of genetic risk loci in primary sclerosing cholangitis. *Scand J Gastroenterol*, 47(7):820–826.

Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R., and Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. *J Mol Biol*, 425(21):3919–3936.

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., and Doyle, J. (2004). Robustness of cellular functions. *Cell*, 118(6):675–685.

Stenson, P., Ball, E., Howells, K., Phillips, A., Mort, M., and Cooper, D. (2008). Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet*, 45(2):124–126.

Stivala, A., Wybrow, M., Wirth, A., Whisstock, J. C., and Stuckey, P. J. (2011). Automatic generation of protein structure cartoons with Pro-origami. *Bioinformatics*, 27(23):3315–3316.

Stobbe, M., Pico, A., Hanspers, K., van Iersel, M., Kelder, T., Digles, D., *et al.* (2013). TCA Cycle (Homo sapiens), Pathway:WP78, Revision:70014.

Su, G., Morris, J. H., Demchak, B., and Bader, G. D. (2014). Biological network exploration with Cytoscape 3. *Current Protocols in Bioinformatics*, 47:8.13.1–8.13.24.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550.

Suderman, M. and Hallett, M. (2007). Tools for visually exploring biological networks. *Bioinformatics*, 23(20):2651–2659.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.

Sullivan, B. J., Durani, V., and Magliery, T. J. (2011). Triosephosphate isomerase by consensus design: dramatic differences in physical properties and activity of related variants. *J Mol Biol*, 413(1):195–208.

Sullivan, B. J., Nguyen, T., Durani, V., Mathur, D., Rojas, S., Thomas, M., *et al.* (2012). Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J Mol Biol*, 420(4-5):384–399.

Sullivan, J. C., De Meyer, S., Bartels, D. J., Dierynck, I., Zhang, E. Z., Spanks, J., *et al.* (2013). Evolution of treatment-emergent resistant variants in telaprevir phase 3 clinical trials. *Clin Infect Dis*, 57(2):221–229.

Susser, S., Welsch, C., Wang, Y., Zettler, M., Domingues, F. S., Karey, U., *et al.* (2009). Characterization of resistance to the protease inhibitor boceprevir in hepatitis C virus-infected patients. *Hepatology*, 50(6):1709–1718.

Suthram, S., Beyer, A., Karp, R. M., Eldar, Y., and Ideker, T. (2008). eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol Syst Biol*, 4(162):162.

Suthram, S., Dudley, J. T., Chiang, A. P., Chen, R., Hastie, T. J., and Butte, A. J. (2010). Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput Biol*, 6(2):e1000662.

Suthram, S., Shlomi, T., Ruppin, E., Sharan, R., and Ideker, T. (2006). A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, 7:360.

Swint-Kruse, L. (2004). Using networks to identify fine structural differences between functionally distinct protein states. *Biochemistry (Mosc)*, 43(34):10886–10895.

Takahashi, K., Taira, T., Niki, T., Seino, C., Iguchi-Ariga, S., and Ariga, H. (2001). DJ-1 positively regulates the androgen receptor by impairing the binding of PIASx alpha to the receptor. *J Biol Chem*, 276(40):37556–37563.

Tao, X. and Tong, L. (2003). Crystal structure of human DJ-1, a protein associated with early onset Parkinson's disease. *J Biol Chem*, 278(33):31372–31379.

Tasan, M., Musso, G., Hao, T., Vidal, M., MacRae, C. A., and Roth, F. P. (2015). Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nat Methods*, 12(2):154–159.

Tetali, P. (1991). Random walks and effective resistance of networks. *J Theoret Probab*, 4:101–109.

Thanki, N., Rao, J. K., Foundling, S. I., Howe, W. J., Moon, J. B., Hui, J. O., *et al.* (1992). Crystal structure of a complex of HIV-1 protease with a dihydroxyethylene-containing inhibitor: comparisons with molecular modeling. *Protein Sci*, 1(8):1061–1072.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

The UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, 42(D1):D191–D198.

Theocharidis, A., van Dongen, S., Enright, A. J., and Freeman, T. C. (2009). Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nat Protoc*, 4(10):1535–1550.

Thurston, T. L. M., Ryzhakov, G., Bloor, S., von Muhlinen, N., and Randow, F. (2009). The TBK1 adaptor and autophagy receptor NDP52 restricts the proliferation of ubiquitin-coated bacteria. *Nat Immunol*, 10(11):1215–1221.

Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat*, 32(4):358–368.

Tiberti, M., Invernizzi, G., Lambrughi, M., Inbar, Y., Schreiber, G., and Papaleo, E. (2014). PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. *J Chem Inf Model*, 54(5):1537–51.

Till, A., Lipinski, S., Ellinghaus, D., Mayr, G., Subramani, S., Rosenstiel, P., and Franke, A. (2013). Autophagy receptor CALCOCO2/NDP52 takes center stage in Crohn disease. *Autophagy*, 9(8):1256–1257.

Tovchigrechko, A. and Vakser, I. A. (2005). Development and testing of an automated approach to protein docking. *Proteins*, 60(2):296–301.

Tovchigrechko, A. and Vakser, I. A. (2006). GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res*, 34(Web Server issue):W310–4.

Tranchevent, L. C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., *et al.* (2008). ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res*, 36(Web Server issue):W377–84.

Tranchevent, L. C., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. *Brief Bioinform*, 12(1):22–32.

Turner, F. S., Clutterbuck, D. R., and Semple, C. A. M. (2003). POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*, 4(11):R75.

Ulitsky, I. and Shamir, R. (2007). Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol*, 1:8.

Vajda, S., Hall, D. R., and Kozakov, D. (2013). Sampling and scoring: A marriage made in heaven. *Proteins: Struct, Funct, Bioinf*, 81(11):1874–1884.

Vajda, S. and Kozakov, D. (2009). Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol*, 19(2):164–170.

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins*, 48(2):227–241.

van den Bedem, H., Bhabha, G., Yang, K., Wright, P. E., and Fraser, J. S. (2013). Automated identification of functional dynamic contact networks from X-ray crystallography. *Nat Methods*, 10(9):896–902.

van den Bedem, H. and Fraser, J. S. (2015). Integrative, dynamic structural biology at atomic resolution – it's about time. *Nat Methods*, 12(4):307–318.

van der Kamp, M. W., Schaeffer, R. D., Jonsson, A. L., Scouras, A. D., Simms, A. M., Toofanny, R. D., *et al.* (2010). Dynameomics: a comprehensive database of protein dynamics. *Structure*, 18(4):423–435.

Van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., and Leunissen, J. A. M. (2006). A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14(5):535–42.

Vangone, A., Cavallo, L., and Oliva, R. (2013). Using a consensus approach based on the conservation of inter-residue contacts to rank CAPRI models. *Proteins*, 81(12):2210–20.

Vangone, A., Oliva, R., and Cavallo, L. (2012). CONS-COCOMAPS: a novel tool to measure and visualize the conservation of inter-residue contacts in multiple docking solutions. *BMC Bioinformatics*, 13 Suppl 4:S19.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*, 6(1):e1000641.

Vanwart, A. T., Eargle, J., Luthey-Schulten, Z., and Amaro, R. E. (2012). Exploring residue component contributions to dynamical network models of allostery. *J Chem Theory Comput*, 8(8):2949–2961.

Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*, 21(6):697–700.

Vazquez, M., Valencia, A., and Pons, T. (2015). Structure-PPi: a module for the annotation of cancer-related single-nucleotide variants at protein–protein interfaces. *Bioinformatics*, 31(14):2397–2399.

Vendruscolo, M., Paci, E., Dobson, C., and Karplus, M. (2001). Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409(6820):641–645.

Vidal, M., Cusick, M. E., and Barabási, A. L. (2011). Interactome networks and human disease. *Cell*, 144(6):986–98.

Vijayabaskar, M. S., Niranjan, V., and Vishveshwara, S. (2011). GraProStr – Graphs of Protein Structures: A tool for constructing the graphs and generating graph parameters for protein structures. *The Open Bioinformatics Journal*, 5:53–58.

Villaveces, J. M., Koti, P., and Habermann, B. H. (2015). Tools for visualization and analysis of molecular networks, pathways, and -omics data. *Advances and Applications in Bioinformatics and Chemistry*, 8:11–22.

Vishveshwara, S., Ghosh, A., and Hansia, P. (2009). Intra and inter-molecular communications through protein structure network. *Curr Protein Pept Sci*, 10(2):146–60.

Viswanathan, K., Shriver, Z., and Babcock, G. J. (2015). Amino acid interaction networks provide a new lens for therapeutic antibody discovery and anti-viral drug optimization. *Curr Opin Virol*, 11:122–129. Viral pathogenesis Preventive and therapeutic vaccines.

von Muhlinen, N., Thurston, T., Ryzhakov, G., Bloor, S., and Randow, F. (2010). NDP52, a novel autophagy receptor for ubiquitin-decorated cytosolic bacteria. *Autophagy*, 6(2):288–289.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18(7):1283–1292.

Wang, L., Matsushita, T., Madireddy, L., Mousavi, P., and Baranzini, S. E. (2015). PINBPA: Cytoscape app for network analysis of GWAS data. *Bioinformatics*, 31(2):262–264.

Wang, Q., Zhang, S., Pang, S., Zhang, M., Wang, B., Liu, Q., and Li, J. (2014). GroupRank: rank candidate genes in ppi network by differentially expressed gene groups. *PLoS ONE*, 9(10):e110406.

Wang, X., Gulbahce, N., and Yu, H. (2011). Network-based methods for human disease gene prediction. *Brief Funct Genomics*, 10(5):280–93.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2.

Weckwerth, W., Loureiro, M. E., Wenzel, K., and Fiehn, O. (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc Natl Acad Sci U S A*, 101(20):7809–7814.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf Comput Sci*, 28(1):31–36.

Weininger, D., Weininger, A., and Weininger, J. L. (1989). SMILES. 2. algorithm for generation of unique smiles notation. *J Chem Inf Comput Sci*, 29(2):97–101.

Weinreich, D. M., Lan, Y., Wylie, C. S., and Heckendorn, R. B. (2013). Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Genet Dev*, 23(6):700–707.

Welsch, C. (2014). Genetic barrier and variant fitness in hepatitis C as critical parameters for drug resistance development. *Drug Discovery Today*, 11:19–25.

Welsch, C., Domingues, F. S., Susser, S., Antes, I., Hartmann, C., Mayr, G., *et al.* (2008). Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of the hepatitis C virus. *Genome Biol*, 9(1):R16.

Welsch, C., Schweizer, S., Shimakami, T., Domingues, F. S., Kim, S., Lemon, S. M., and Antes, I. (2012a). Ketoamide resistance and hepatitis C virus fitness in Val55 variants of the NS3 serine protease. *Antimicrob Agents Chemother*, 56(4):1907–1915.

Welsch, C., Shimakami, T., Hartmann, C., Yang, Y., Domingues, F. S., Lengauer, T., *et al.* (2012b). Peptidomimetic escape mechanisms arise via genetic diversity in the ligand-binding site of the hepatitis C virus NS3/4A serine protease. *Gastroenterology*, 142(3):654–663.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M. M., *et al.* (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587.

Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R. F., Sykes, B. D., and Wishart, D. S. (2003). VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res*, 31(13):3316–3319.

Williamson, K. D. and Chapman, R. W. (2015). Primary sclerosing cholangitis: a clinical update. *Br Med Bull*, 114(1):53–64.

Wittkop, T., Emig, D., Truss, A., Albrecht, M., Bocker, S., and Baumbach, J. (2011). Comprehensive cluster analysis with Transitivity Clustering. *Nat Protoc*, 6(3):285–95.

Wong, A. S. L., Cheung, Z. H., and Ip, N. Y. (2011). Molecular machinery of macroautophagy and its deregulation in diseases. *Biochim Biophys Acta*, 1812(11):1490–1497.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., *et al.* (1999a). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol*, 285(4):1711–1733.

Word, J. M., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (1999b). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol*, 285(4):1735–1747.

World Health Organization (WHO) (2015). HCV Fact Sheet. http://www.who.int/mediacentre/factsheets/fs164/en/. Accessed on July 24, 2015.

Wu, C., Zhu, J., and Zhang, X. (2012). Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinformatics*, 13:182.

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Mol Syst Biol*, 4(189):189.

Xie, Z. and Klionsky, D. J. (2007). Autophagosome formation: core machinery and adaptations. *Nat Cell Biol*, 9(10):1102–1109.

Xu, J. and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 22(22):2800–5.

Xu, J., Zhong, N., Wang, H., Elias, J. E., Kim, C. Y., Woldman, I., *et al.* (2005). The Parkinson's disease-associated DJ-1 protein is a transcriptional co-activator that protects against neuronal apoptosis. *Hum Mol Genet*, 14(9):1231–1241.

Xue, W., Ban, Y., Liu, H., and Yao, X. (2014a). Computational study on the drug resistance mechanism against HCV NS3/4A protease inhibitors vaniprevir and MK-5172 by the combination use of molecular dynamics simulation, residue interaction network, and substrate envelope analysis. *J Chem Inf Model*, 54(2):621–33.

Xue, W., Jiao, P., Liu, H., and Yao, X. (2014b). Molecular modeling and residue interaction network studies on the mechanism of binding and resistance of the HCV NS5B polymerase mutants to VX-222 and ANA598. *Antiviral Res*, 104:40–51.

Xue, W., Jin, X., Ning, L., Wang, M., Liu, H., and Yao, X. (2013). Exploring the molecular mechanism of cross-resistance to HIV-1 integrase strand transfer inhibitors by molecular dynamics simulation and residue interaction network analysis. *J Chem Inf Model*, 53(1):210–22.

Xue, W., Wang, M., Jin, X., Liu, H., and Yao, X. (2012). Understanding the structural and energetic basis of inhibitor and substrate bound to the full-length NS3/4A: insights from molecular dynamics simulation, binding free energy calculation and network analysis. *Mol Biosyst*, 8(10):2753–65.

Xue, W., Yang, Y., Wang, X., Liu, H., and Yao, X. (2014c). Computational study on the inhibitor binding mode and allosteric regulation mechanism in hepatitis C virus NS3/4A protein. *PLoS ONE*, 9(2):e87077.

Yamada, T. and Bork, P. (2009). Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nat Rev Mol Cell Biol*, 10(11):791.

Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G., and Shen, B. (2014). The construction of an amino acid network for understanding protein structure and function. *Amino Acids*, 46(6):1419–1439.

Yang, H., Robinson, P. N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods*, 12(9):841–843.

Yang, P., Li, X., Wu, M., Kwoh, C.-K., and Ng, S.-K. (2011). Inferring gene-phenotype associations via global protein complex network propagation. *PLoS ONE*, 6(7):e21502.

Yang, Z. and Klionsky, D. J. (2010). Mammalian autophagy: core molecular machinery and signaling regulation. *Curr Opin Cell Biol*, 22(2):124–131.

Yao, N., Reichert, P., Taremi, S., Prosise, W., and Weber, P. (1999). Molecular views of viral polyprotein processing revealed by the crystal structure of the hepatitis C virus bifunctional protease-helicase. *Structure*, 7(11):1353–1363.

Yao, X., Hao, H., Li, Y., and Li, S. (2011). Modularity-based credible prediction of disease genes and detection of disease subtypes on the phenotype-gene heterogeneous network. *BMC Syst Biol*, 5(1):79.

Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L., and Vidal, M. (2007). Drug-target network. *Nat Biotechnol*, 25(10):1119–1126.

Yook, S.-H., Oltvai, Z. N., and Barabási, A.-L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942.

Yoon, J., Blumer, A., and Lee, K. (2006). An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics*, 22(24):3106.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., *et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4):e59.

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., *et al.* (2011a). Next-generation sequencing to generate interactome datasets. *Nat Methods*, 8(6):478–480.

Yu, S., Tranchevent, L. C., De Moor, B., and Moreau, Y. (2011b). *Kernel-based data fusion for machine learning methods and applications in bioinformatics and text mining*, volume 345. Springer.

Zanon, A., Rakovic, A., Blankenburg, H., Doncheva, N. T., Schwienbacher, C., Serafin, A., *et al.* (2013). Profiling of parkin-binding partners using tandem affinity purification. *PLoS ONE*, 8(11):1–17.

Zhang, S., Jin, G., Zhang, X.-S., and Chen, L. (2007). Discovering functions and revealing mechanisms at molecular level from biological networks. *Proteomics*, 7(16):2856–2869.

Zhang, W., Sun, F., and Jiang, R. (2011). Integrating multiple protein-protein interaction networks to prioritize disease genes: a bayesian regression approach. *BMC Bioinformatics*, 12(1):1–10.

Zhao, J., Yang, T.-H., Huang, Y., and Holme, P. (2011). Ranking candidate disease genes from gene expression and protein interaction: a Katz-centrality based approach. *PLoS ONE*, 6(9):e24306.

Zhernakova, A., van Diemen, C. C., and Wijmenga, C. (2009). Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat Rev Genet*, 10(1):43–55.

Zhou, J., Yan, W., Hu, G., and Shen, B. (2014a). Amino acid network for the discrimination of native protein structures from decoys. *Curr Protein Pept Sci*, 15(6):522–528.

Zhou, J., Yan, W., Hu, G., and Shen, B. (2014b). SVR_CAF: an integrated score function for detecting native protein structures among decoys. *Proteins*, 82(4):556–64.

Zhou, W., Zhu, M., Wilson, M. A., Petsko, G. A., and Fink, A. L. (2006). The oxidation state of DJ-1 regulates its chaperone activity toward alpha-synuclein. *J Mol Biol*, 356(4):1036–1048.

Zimmerman, J., Eliezer, N., and Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol*, 21(2):170–201.

Zwier, M. C. and Chong, L. T. (2010). Reaching biological time-scales with all-atom molecular dynamics simulations. *Curr Opin Pharmacol*, 10(6):745–52.

APPENDIX A

Protocol for network analysis and visualization

This chapter is an adapted version of text contained in Doncheva *et al.* (2012a) and describes three workflows based on the *NetworkAnalyzer* and *RINalyzer* plugins for Cytoscape, a popular software platform for networks. *NetworkAnalyzer* was initially developed by Yassen Assenov and has become a standard Cytoscape tool for comprehensive network topology analysis. In addition, *RINalyzer* provides methods for exploring residue interaction networks derived from protein structures. We developed RINalyzer and extended NetworkAnalyzer as well as designed and implemented the presented workflows. As outlined in Figure A.1, the first workflow uses *NetworkAnalyzer* to perform a topological analysis of biological networks. The second workflow applies *RINalyzer* to study protein structure and function and to compute network centrality measures. The third workflow combines *NetworkAnalyzer* and *RINalyzer* to compare residue networks. The full protocol can be completed in ~ 2 h.

Experimental design

Topological network analysis

In particular, NetworkAnalyzer supports the characterization of molecular networks in terms of scale-free and small-world properties, modularity and hierarchical structure (Yamada and Bork, 2009; Albert, 2005; Almaas, 2007; Barabási and Oltvai, 2004), the identification of important network nodes and edges based on topological parameters (Welsch *et al.*, 2008; Astsaturov *et al.*, 2010; Ragusa *et al.*, 2010; Lorenz *et al.*, 2011), and the comparison of networks with regard to their topology (Radrich *et al.*, 2010; Choura and Rebaï, 2010; Gu *et al.*, 2011; Yu *et al.*, 2011a). Since its ini-

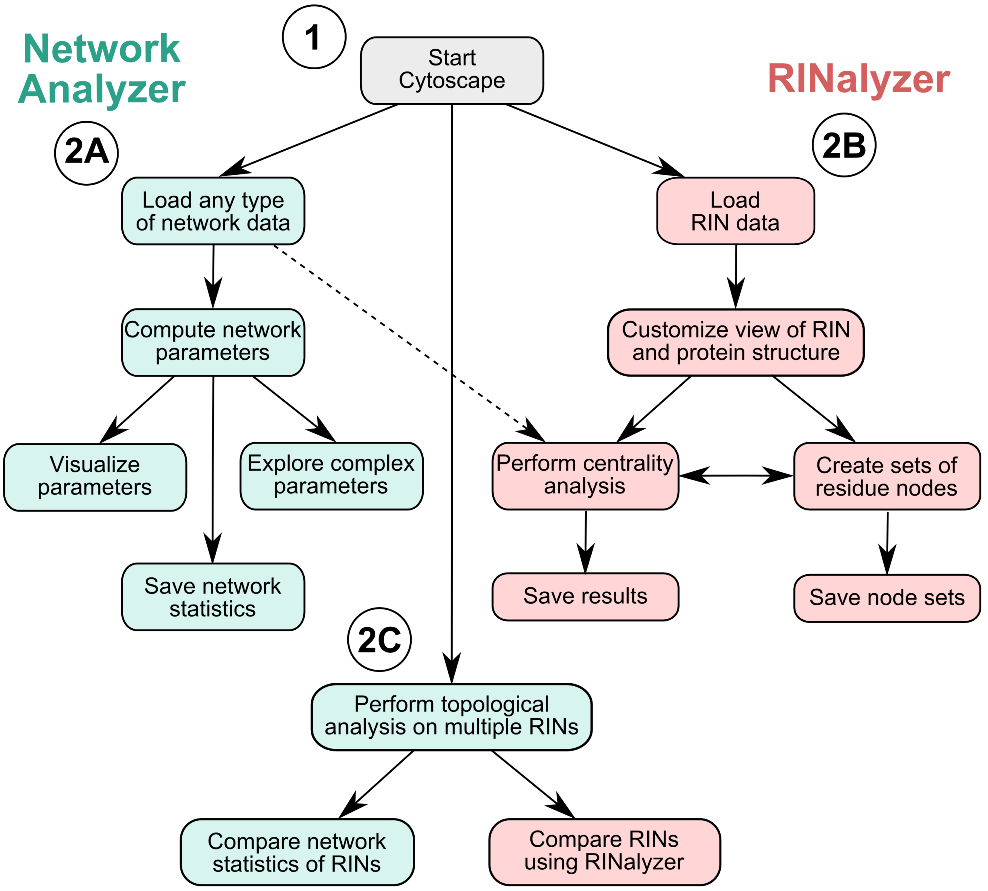

Network Analyzer
1
Start Cytoscape
RINalyzer
2A
2B
Load any type of network data
Load RIN data
Compute network parameters
Customize view of RIN and protein structure
Visualize parameters
Explore complex parameters
Perform centrality analysis
Create sets of residue nodes
Save network statistics
Save results
Save node sets
2C
Perform topological analysis on multiple RINs
Compare network statistics of RINs
Compare RINs using RINalyzer

Figure A.1: Outline of the protocol for network analysis and visualization. This protocol starts with launching Cytoscape (Step 1) and consists of three major workflows: (Step 2A) topological analysis of biological networks; (Step 2B) interactive visual analysis of residue networks; (Step 2C) comparison of residue networks. Steps colored in blue are performed with NetworkAnalyzer and those in pink with RINalyzer. The dotted line represents an optional step that connects the two workflows, which is not described in detail in this protocol. Figure first published in Doncheva *et al.* (2012a).

tial release in 2007, NetworkAnalyzer has been extended by additional features and topological parameters and is widely used in academia and industry as indicated by thousands of software downloads. Recently, this plugin became an integral part of each standard installation of Cytoscape, and its source code was published under the GNU Lesser General Public License. The workflow in Appendix A Step 2A describes how to use the NetworkAnalyzer plugin to perform a topological analysis on an unweighted network loaded into Cytoscape, as well as how to process and visualize the results.

Basically, NetworkAnalyzer calculates many simple topological parameters, such as clustering coefficient, number of connected components, diameter and radius, centralization, number of shortest paths, average shortest path length, average number of neighbors, density, heterogeneity (only for undirected networks), number of isolated nodes, number of self-loops and number of multi-edge node pairs. In addition, the following complex topological parameters are computed by NetworkAnalyzer: average clustering coefficient distribution, shortest path length distribution, betweenness centrality versus number of neighbors, closeness centrality versus number of neighbors and stress centrality distribution. The degree distribution, topological coefficients, shared neighbors distribution and neighborhood connectivity distribution are computed for undirected networks only, whereas the in-degree, out-degree and three different types of neighborhood connectivity are used for directed networks. The complete set of simple and complex parameters is referred to as network statistics in NetworkAnalyzer. As described in the next section, RINalyzer also computes several centrality measures for weighted networks and provides further options for the visual exploration of the results.

The computed topological parameters are represented as single values, histograms or scatter plots and can be visualized in the Cytoscape network view by corresponding node and edge size as well as color choice. For example, the degree might correspond to the node size and the clustering coefficient might determine the node color (Step 2A(x)). Complex topological parameters are depicted as histograms or scatter plots. The user can easily customize various visual settings as well as switch between histograms or scatter plots of the computed distributions and between linear or logarithmic scales of the x and y axes. In addition, a power law can be fitted to the degree distribution to illustrate whether the analyzed network has scale-free properties (Step 2A(vi)). Finally, both displayed charts and network statistics can be saved to files (Steps 2A(ix) and (xi)).

Interactive visual analysis of residue networks

A residue interaction network (RIN) consists of nodes that represent protein residues and edges that correspond to non-covalent interactions between residues. In particular, RINalyzer is currently the only tool that supports the simultaneous view of a RIN in 2D and the corresponding protein structure in 3D by connecting Cytoscape to the UCSF Chimera molecular structure viewer (Pettersen *et al.*, 2004). RINalyzer also provides versatile user options, such as the computation of weighted network centrality measures to highlight biologically important residues and the network comparison of superimposed protein structures to study differing residue interactions. This new structure analysis approach can be very useful in a number of biological and medical application scenarios.

The workflow in Appendix A Step 2B explains the use of the Cytoscape plugin RINalyzer and its features for analyzing and visualizing RINs. It is divided into the following major steps (Figure A.1): retrieving and loading RIN data into Cytoscape (Step 2B(i-v)); customizing RIN and 3D structure views (Step 2B(vi-xi)); creating, managing and saving sets of residue nodes (Step 2B(xii-xviii)); performing centrality analysis, exploring and saving the results (Step 2B(xix-xxix)).

The workflow starts with the retrieval of residue interaction data for a protein of interest from the web interface to our RINdata database (Step 2B(i)). It contains RINs generated by means of the RINerator software (Doncheva *et al.*, 2011) for over

50,000 protein structures from the Protein Data Bank (PDB) (Rose *et al.*, 2013). In contrast to previous approaches that define residue interactions on the basis of spatial atomic distance between residues, RINerator distinguishes different residue interaction types and quantifies the strength of individual interactions, which results in an undirected weighted network with multiple interaction edges. To this end, RINerator first adds hydrogens to the 3D protein structure by using the Reduce tool (Word *et al.*, 1999b) and then samples contacts on the van der Waals surface of each atom by using Probe (Word *et al.*, 1999a).

In a RIN, the nodes represent the protein residues and the edges between them represent the non-covalent interactions identified by Probe. The edges are labeled with an interaction type and subtype. Possible types are interatomic contact (cnt), hydrogen bond (hbond), overlapping van der Waals radii (ovl) and generic residue interaction (combi), whereas the subtypes indicate interactions between main chains (mc) and side chains (sc) of the amino-acid residues. Each edge is weighted with the respective score for the interacting residues as computed by Probe and the weight is proportional to the strength of the interaction. The resulting RIN and additional information (such as edge weights) are stored in the Cytoscape default formats, the simple interaction format (SIF) for the network, and the edge attribute (EA) files for the edge weights. Thus, each RIN is accompanied by the original PDB file with hydrogens added, and two edge attribute files.

Once both the RIN and the corresponding protein structure are imported (after Step 2B(iv)), RINalyzer establishes a bidirectional connection between Cytoscape and the 3D structure viewer UCSF Chimera. In particular, when the user selects nodes of a RIN in the Cytoscape network view, the corresponding residues in the protein structure are automatically highlighted in UCSF Chimera, and vice versa. RIN nodes can be colored according to secondary structure based on the data retrieved from UCSF Chimera, and the node colors can be synchronized with the residue colors in UCSF Chimera. In addition, the user is able to show or hide different types of interaction edges such as backbone and hydrogen bonds. The visual RIN settings that can be customized by the user are listed in Box 2. Notably, a RIN-specific 2D layout can be applied to the network view that takes the current 3D structure coordinates into account.

The subsequent visual exploration of RINs often includes the study of the molecular interactions of active site residues and binding residues. For this purpose, RINalyzer offers a user interface for creating and modifying sets of residue nodes. In particular, the user can apply it to identify the interacting residues in the binding interface of two distinct protein domains (Step 2B(xv)) or to highlight different sets of residues such as active site residues (Step 2B(xvii)) in both the network and the 3D structure view.

We also show how to use RINalyzer for the computation of weighted centrality measures and the identification of central nodes in a RIN (Step 2B(xxi-xxvii)). To this end, RINalyzer calculates the following centrality measures: weighted degree; shortest path closeness and betweenness; current flow closeness and betweenness; random walk closeness and betweenness. Here, a crucial point is the choice of the appropriate user settings for the centrality analysis. As the edge weights in a RIN are proportional to the strength of the represented residue interaction, the weights need to be converted to distance scores such that smaller values are assigned to edges that represent stronger interactions for the shortest path computation. For each computed centrality measure, RINalyzer offers three different ways to examine the results: (i) inspecting the raw values in a sortable table, (ii) highlighting selected nodes in the network view or (iii) saving the values in a tab-delimited format for further processing. The presented workflow particularly focuses on the second option (ii), which involves a filter to select nodes with centrality values in a given numerical range (Step 2B(xxvi)). This functionality allows the user to create sets of best-scoring residue nodes for further investigations of their functional and structural characteristics in both the network view and the 3D protein structure.

Comparison of residue networks

The workflow in Appendix A Step 2C introduces one possible application scenario that combines NetworkAnalyzer and RINalyzer. We compiled a small data set consisting of four RINs that are generated from the four subunits of the deoxyhemoglobin structure (Fermi *et al.*, 1984). First, the batch analysis option of NetworkAnalyzer is used to compute the network statistics of these RINs and to compare their topologies (Step 2C(ii-v)). Second, two RINs that represent the two different subunits of deoxyhemoglobin are compared with each other using RINalyzer (Step 2C(vi-xiv)).

This comparison requires an additional structure alignment of the two 3D protein structures from the user and eventually results in a combined RIN. The comparison network contains different types of edges and nodes according to the preserved residue interactions and the aligned residues. The type of each node and edge is stored as an attribute, which can be used to visually adjust the network view. Thus, the user can easily highlight and investigate the identified similarities and differences between the two RINs and the corresponding protein structures.

Materials

Hardware requirements

- Personal computer with Internet access and web browser (e.g., Mozilla Firefox, Microsoft Internet Explorer or Google Chrome); we also recommend a screen with resolution of at least 1024 x 768 pixels and a three-button mouse.

Software requirements

- Java Standard Edition, version 6 (download from `http://www.java.com/`)

- Cytoscape, version 2.8 (Cytoscape can be installed following the steps provided in the Cytoscape protocol (Cline *et al.*, 2007)
- *NetworkAnalyzer* (included in the Cytoscape 2.8 installation as a core plug-in)
- *RINalyzer* (download and installation instructions for *RINalyzer* are available at `http://rinalyzer.de/docu/install.php`)
- UCSF Chimera, version 1.5 (instructions for its installation are available at `http://www.cgl.ucsf.edu/chimera/download.html`)

Data

- Sample data sets required for this protocol are provided as supplementary files at this url. The human protein interaction network (Supplementary Data 1) was published in the recent interactome screening study by Yu *et al.* (Yu *et al.*, 2011a). The set of four RINs (Supplementary Data 2) was generated using the RINerator package (`http://rinalyzer.de/rinerator.php`) and represents the four subunits of the deoxyhemoglobin structure with the PDB identifier 4HHB (Fermi *et al.*, 1984).

Step 2A: Topological analysis of biological networks

(i) *Download data.* Here we perform the topological analysis of the protein-protein interaction network from Yu *et al.* (2011a) (Supplementary Data 1). First, download the file Supplementary Data 1 to a local directory.

(ii) *Import network data* (for details, see the Cytoscape protocol (Cline *et al.*, 2007)). In the Cytoscape main window, go to the menu option `File` → `Import Network (multiple file types)`. Select the option `Local` for `Data Source Type` and click the `Select` button. Navigate to the directory that contains *Supplementary Data 1* and select the file. Confirm the selection by clicking the `Open` button. Then click the `Import` button to import this network into the current Cytoscape session. When the network is successfully loaded, a summary window will appear. Click the `Close` button of this window and return to the Cytoscape main window.

(iii) *Apply network layout.* To apply a specific layout to the network, go to the menu option `Layouts` → `yFiles` → `Organic`. The network view can be enlarged by clicking the `Maximize` button in the upper right corner of the network view window.

(iv) *Run NetworkAnalyzer.* To initiate the analysis, go to the menu option `Plugins` → `Network Analysis` → `Analyze Network`. *NetworkAnalyzer* can perform topological analysis on directed networks as well as on undirected networks. Therefore, the user can choose how the edges should be interpreted. As this network is undirected, select the option `Treat the network as undirected` and click the `OK` button to start the analysis. A `Progress` dialog will appear. The analysis time depends on the size of the network and the amount of memory assigned to the Cytoscape application. The `Cancel` button can be used at any time to stop the analysis. ?TROUBLESHOOTING

(v) *View results.* The results window appears after the analysis is completed. The first tab shows the computed simple parameters, e.g., the clustering coefficient and the average shortest path length. The remaining tabs display complex network parameters such as degree and shortest

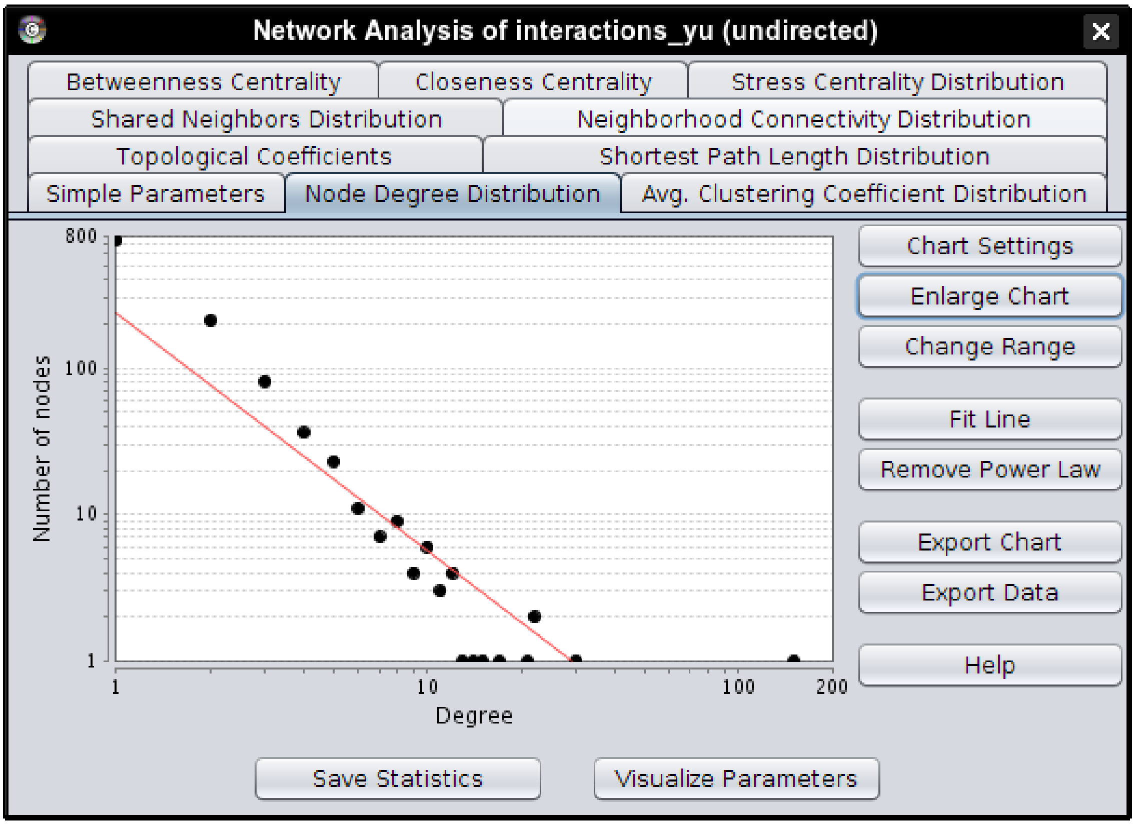


Figure A.2: Screenshot of network statistics computed by *NetworkAnalyzer*. The depicted node degree distribution is derived from the undirected protein-protein interaction network from Yu et al (Yu *et al.*, 2011a). The red line represents a fitted power law, which indicates that the analyzed network is scale-free. The tabs below the dialog title lead to the display of histograms or scatter plots of the complex topological parameters computed by *NetworkAnalyzer*. The buttons on the right side provide the user with a variety of options for customizing the view as well as for exporting the displayed charts and the underlying data. Figure first published in Doncheva *et al.* (2012a).

path distributions. All topological parameters are described in more detail in Box 1. Select the tab `Node Degree Distribution`. The node degree distribution is depicted in a log-log plot. The x axis enumerates the degrees of nodes in the network and the y axis shows the frequency of nodes with a given degree.

(vi) *Fit a power law.* The degree distribution of many biological networks is known to approximate a power law. Click on the button `Fit Power Law` to fit a power law to the distribution. A warning message will inform you that only points with positive coordinate values are considered for the fit. Confirm this message by clicking the `OK` button. After a short delay, the dialog `NetworkAnalyzer - Fitted Function` appears. It reports the fitted power law constants, the correlation between the given data points and the corresponding points on the fitted curve, and the R-squared value as a measure of fit quality between 0 and 1 (the higher the value, the better the fit). Click the `OK` button to close the dialog and see the fitted power law in the chart (A.2).

(vii) *Explore charts.* Click the button `Enlarge Chart` to open the distribution plot in a separate, enlarged window. Almost all nodes in the network have a degree of < 30. The dot near the lower right corner of the plot indicates that there is only one node with degree 151, which hereafter we call hub node because of this exceptional number of protein interactions. Close the window.

(viii) *Customize charts.* Click on the button `Chart Settings` to rename the axes in the tab `Axes`, show or hide the grid lines in the tab `Gridlines`, change the shape and color of chart points in the tab `Histogram`. Click the `OK` button to apply changes of the settings or the `Cancel` button to close the dialog without saving the changes.

(ix) *Export charts.* Every chart in the results window can be saved to a file. To save the current chart as an image, click the `Export Chart` button. Adjust the image size by entering your preferred values in the two displayed text fields and confirm it by the `Save` button. Navigate to the directory where you want to save the image and select the file type from the drop-down menu. Finally, click the `Save` button. In addition, it is possible to export the visualized data for further processing in a different application. For example, select the tab `Betweenness centrality`. This scatter plot displays the correlation between node degree and betweenness centrality in the studied network. Every node in the network is represented by a point. The x axis gives the node degree and the y axis the betweenness. Click the button `Export data` and enter a file name (including extension) to store the values of these topological parameters. After clicking the `Save` button, the newly created tab-separated text file will contain a table of the degree and betweenness centrality values for every node in the network. This file can be easily imported in external software applications such as a spreadsheet tool for further analysis or processed by other programs.

(x) *Visualize topological parameters.* In Step 2A(vii), we identified a hub node in the network; now we are interested in locating it in the network view. Thus, we will visually map the node degree to node size in the network view. Click the button `Visualize Parameters` in the results dialog of *NetworkAnalyzer*. In the `Map node size to` drop-down menu, select `Degree`. Nodes with a low degree should be displayed as small circles in contrast to nodes with a high degree. To this end, select the option `Low values to small sizes`. In addition, it is possible to map the degree or any other computed topological parameter to the node color. Choose `ClusteringCoefficient` in the drop-down menu on the right side and select the option `Low values to bright colors`. Nodes with low clustering coefficient will now be green and nodes with high clustering coefficient will be red. Finally, confirm the mapping choice by clicking the `Apply` button. This results in changed network visualization (A.3). If necessary, move the network statistics dialog to the right corner of your screen or close it in order to see the updated network view. The hub node is now clearly visible as the largest circle in the network view. The large number of green-colored nodes indicates that most nodes have low clustering coefficient, i.e., the neighbors of most nodes do not tend to interact with each other. To obtain an even better view of the nodes, zoom into it by applying the button `Zoom` in on the toolbar or using the mouse scroll wheel.

(xi) *Save network statistics.* Close the analysis results window. A warning message appears that the computed network statistics have not been saved. Click the `Yes` button to close the statistics window without saving the results. To recompute the network statistics at a later time point, just run *NetworkAnalyzer* again. Alternatively, the results can be saved to and reloaded from a text file to avoid re-computation. For this purpose, click on the button `Save Statistics`. Enter a file name to store the network statistics in a file with the extension *.netstats* and click the `Save` button to confirm it.

(xii) *(Optional) Perform centrality analysis.* In addition to *NetworkAnalyzer*, *RINalyzer* can be applied to perform centrality analysis on the loaded network. *RINalyzer* supports weighted networks and computes several weighted centrality measures additionally (Box 3). To use *RINalyzer* now, continue with Step 2B(xx).

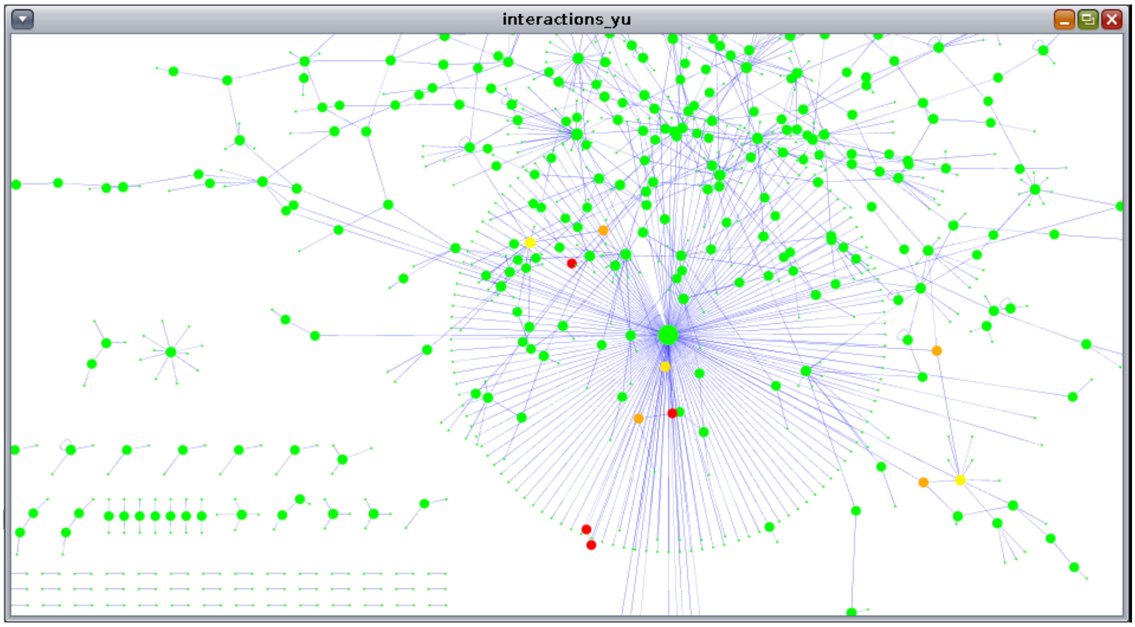


Figure A.3: Mapping topological parameters to the network view. In the PPI network from Yu et al. (Yu *et al.*, 2011a), the node degree and clustering coefficient are mapped to node size and node color, respectively. Nodes of low degree appear as small circles, whereas nodes of high degree are enlarged. Additionally, nodes with low clustering coefficient are depicted in green, and nodes with high clustering coefficient in red. Figure first published in Doncheva *et al.* (2012a).

Step 2B: Interactive visual analysis of residue networks

(i) *Retrieve RIN data.* Identify a protein of interest with an experimentally determined 3D structure deposited in the PDB. For example, we have chosen the HIV-1 protease (Thanki *et al.*, 1992) with the PDB identifier 1HIV. Start a web browser and go to the RINdata website (`http://rinalyzer.de/rindata.php`) to download the corresponding RIN data. Enter the PDB identifier 1HIV in the search form and click the button `Retrieve RIN data`. If RIN data are available for this PDB identifier, a download link is provided. Click on this link and download the file to a local directory. The downloaded RIN data are a zipped archive that contains multiple files: a PDB file with the 3D protein structure of the original PDB file (as retrieved from the PDB) with added hydrogens (*pdb1hiv_h.ent*); a SIF file containing the RIN for all chains in the PDB file (*pdb1hiv_h.sif*); an edge attribute file with edge weights reflecting the strength of the interactions between residues (*pdb1hiv_h_intsc.ea*); and an edge attribute file with edge weights representing the number of interactions between residues (*pdb1hiv_h_nrint.ea*). Unzip all files from the archive.

(ii) *Import network into Cytoscape.* In the Cytoscape main window, go to the menu option `File` → `Import` → `Network (multiple file types)`. Select the option `Local` for `Data Source Type` and click the `Select` button. Navigate to the directory that contains the extracted RIN files and select the network SIF file, e.g., *pdb1hiv_h.sif*. Confirm the selection by clicking the `Open` button and then click the `Import` button. When the network is successfully loaded, a summary window will appear. Click the `Close` button of this window and return to the Cytoscape main window. The network view can be enlarged by clicking the `Maximize` button in the upper right corner of the network view window.

(iii) *Import edge attributes into Cytoscape.* Import the edge weights representing the number of interactions between residues, as they are needed in Step 2B(xxii) for the network centrality analysis. Go to the menu option `File` → `Import` → `Edge Attributes`. Navigate to the directory that contains the RIN files. Select the edge attribute file *pdb1hiv_h_nrint.ea* and click the `Open` button. When the attributes are successfully loaded, a summary window will appear. Click the `Close` button of this window and return to the Cytoscape main window.

(iv) *Open protein structure in UCSF Chimera.* Go to the menu option `Plugins` → `RINalyzer` → `Protein Structure` → `Open structure` from file in the Cytoscape main window and navigate to the directory that contains the RIN files. Select the PDB file (*pdb1hiv_h.ent*) and click the `Open` button. It may take a while until UCSF Chimera is launched and the 3D structure is loaded. Afterwards, a summary window about the internally performed mapping between network nodes and structure entities will appear. Click the `Close` button of this window. ?TROUBLESHOOTING

(v) *Explore protein structure.* Use the mouse to move and scale the protein structure in the main UCSF Chimera window. By default, the left mouse button controls rotation, the middle mouse button controls XY translation and the right mouse button controls scaling. While holding down the `Ctrl` key, use the left mouse button to select residues of interest by clicking on them or to drag out a selection area (sweep out an area before releasing the left mouse button).

(vi) *Show protein backbone.* To see the protein backbone in the 3D structure and to add backbone edges to the RIN, select `Plugins` → `RINalyzer` → `Protein Structure` → `Show backbone` in the Cytoscape main window. This option automatically adds protein backbone edges to the RIN in Cytoscape and also invokes the display of the ribbon representation for the corresponding 3D protein structure in UCSF Chimera.

(vii) *Apply RIN layout.* Layout the RIN in Cytoscape according to the 3D structure view in UCSF Chimera by selecting the menu option `Plugins` → `RINalyzer` → `Layout` → `RIN Layout`. Click on the icon `1:1` in the Cytoscape toolbar to see the whole network. As the graphics details of the network view are normally not displayed when the network is zoomed-out, select the menu option `View` → `Show Graphics Details`. ?TROUBLESHOOTING

(viii) *Customize RIN view.* Go to the menu `Plugins` → `RINalyzer` → `Visual Properties` to choose in the tab `General & Nodes` how the node label should be displayed. For example, if only residue index and type are selected, the node labels are updated accordingly. In the tab `Edges`, the visible edge types can be selected. The network view is updated automatically each time an edge type box is checked or unchecked. In the same tab, the option `Straighten edge lines` controls whether multiple edges are drawn as straight parallel lines or not. When satisfied with the customized settings, confirm them by clicking the `Apply` button and click the `Close` button of the dialog `RIN Visual Properties`. In the resulting network view, the nodes are colored according to secondary structure and the edges according to interaction type. More details about the different visual properties can be found in Box 2.

(ix) *Synchronize colors between views.* After customizing the visual properties of a RIN, nodes are usually colored according to secondary structure. To transfer the node colors to the corresponding residues in UCSF Chimera, go to the menu option `Plugins` → `RINalyzer` → `Protein Structure` → `Sync 3D view colors`. The resulting network and 3D structure views should be the same as in A.4.

(x) *(Optional) Show only protein backbone.* Now we want to look at only the protein backbone in both the network view and the 3D structure view. If the protein backbone is not yet visible in the 3D structure and the RIN, select `Plugins` → `RINalyzer` → `Protein Structure` → `Show backbone` in the Cytoscape main window. Then go to the menu option `Actions` → `Atoms/Bonds` → `hide` in the UCSF Chimera window to hide all atoms in the 3D structure view. In the Cytoscape main window, go to the menu option `Plugins` → `RINalyzer` → `Visual Properties` → `Edges`. Uncheck the boxes next to all edge types except of the backbone edges. Show the edges again by checking the boxes next to the edge types in the dialog `RIN Visual Properties`. When the edges are added to the network, they are visualized as curved lines. Click the `Apply` button to straighten them. The atoms in the

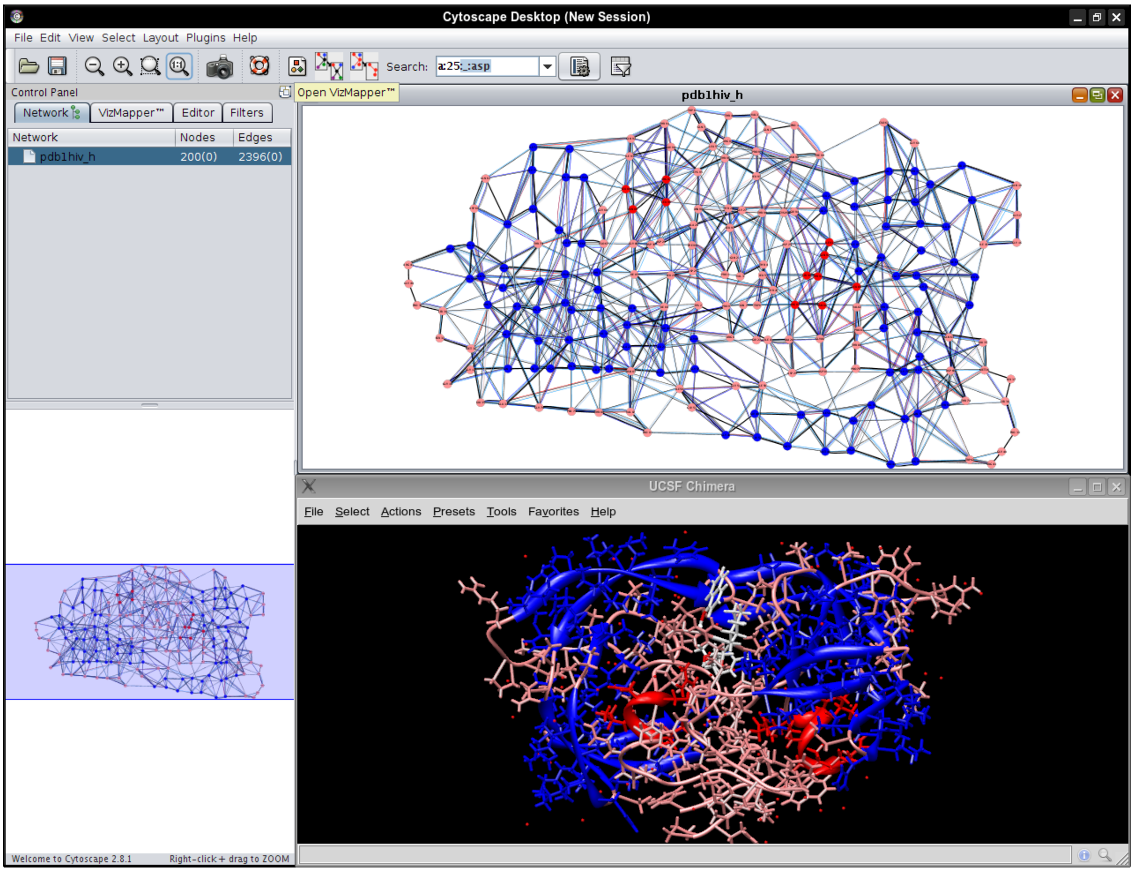

Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Search: a:25:_:asp
Control Panel
Network
VizMapper™
Editor
Filters
Network
Nodes
Edges
pdb1hiv_h
200(0)
2396(0)
Open VizMapper™
pdb1hiv_h
UCSF Chimera
File Select Actions Presets Tools Favorites Help
Welcome to Cytoscape 2.8.1
Right-click + drag to ZOOM

Figure A.4: Simultaneous view of RIN and 3D protein structure by *RINalyzer*. The RIN of the HIV-1 protease (PDB identifier 1HIV (Thanki *et al.*, 1992)) is displayed in Cytoscape (top), whereas the molecular graphics visualization of the 3D protease structure is shown in UCSF Chimera (bottom). All RIN nodes and the corresponding residues are colored according to secondary structure: blue for helices and red for strands. The various types of non-covalent residue interactions correspond to different edge colors: interatomic contacts are in blue; hydrogen bonds in red; overlapping van der Waals radii in gray; and the backbone in black. Figure first published in Doncheva *et al.* (2012a).

3D structure can be depicted again by executing `Actions` → `Atoms/Bonds` → `show` in the UCSF Chimera window.

(xi) *(Optional) Hide protein backbone.* The backbone can be hidden in both views by clicking the menu item `Plugins` → `RINalyzer` → `Protein Structure` → `Hide backbone` in the Cytoscape main window.

(xii) *Create sets of residue nodes.* *RINalyzer* provides an interface to manage node sets. To open it, click on the menu option `Plugins` → `RINalyzer` → `Manage Node Sets`. The `RINalyzer Node Sets` panel appears as the last tab in the `Cytoscape Control Panel`. New node sets can be created in different ways. For instance, to create a set that contains the currently selected residues in UCSF Chimera, switch to the UCSF Chimera window and click `Select` → `Chain` → `A` to select all residues in chain A. Selected residues are colored in green, and the corresponding nodes in Cytoscape are also selected automatically (yellow). In the panel `RINalyzer Node Sets`, go to the menu option `File` → `New` → `Set from selected nodes` to create a set that contains the nodes corresponding to currently selected residues in UCSF Chimera. Insert a name for the set to be created, e.g., `Chain A`, and click `OK` to confirm it. The same actions can be repeated to create a second set named `Chain B` that contains all nodes corresponding to residues in chain B.

(xiii) *Select set nodes in the network view.* To see all set nodes selected in the network view, use the option `Select nodes` in the context menu of the set (right-click the set name). To clear the current node selection, click on the background in the network view window.

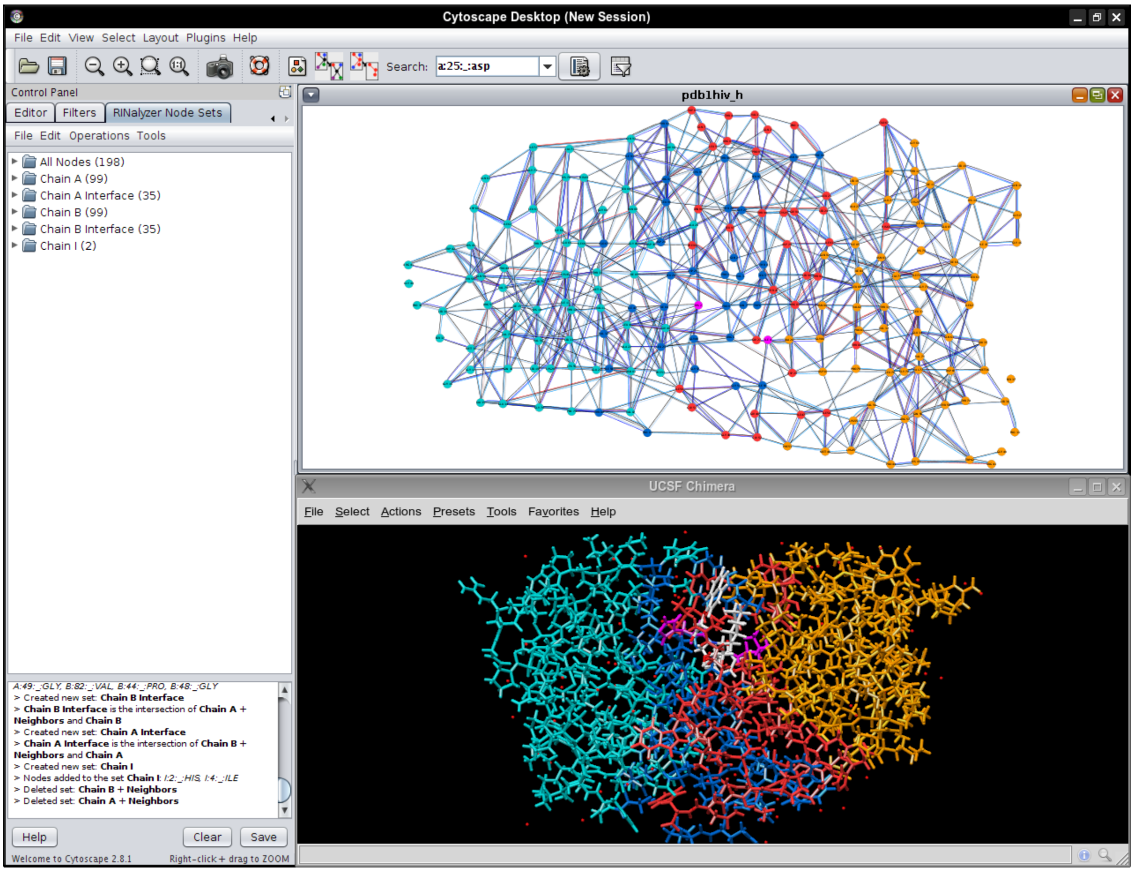
(xiv) *Add active site nodes to a set.* It is known that the active site residues of the HIV-1 protease are ASP 25, THR 26 and GLY 27 in chains A and B (Thanki *et al.*, 1992). To create a set with the active site residues of the HIV-1 protease for use in the centrality analysis in Step 2B(xxvii), go to the menu option `File` → `New` → `Empty set` in the Cytoscape panel `RINalyzer Node Sets`. Enter the name `Active site` and click the `OK` button to confirm. Go to the `Search` field in the Cytoscape toolbar and start inserting the node identifier `a:25:_:asp`. As a result of this insertion, a single hit should appear in the drop-down menu of the search field. Press `Enter` to select the node. In the panel `RINalyzer Node Sets`, go to the menu option `Edit` → `Add nodes`, and the selected node will be added to the currently selected node set, which should be `Active site`. Repeat the same actions for the remaining five active site residues: `a:26:_:thr; a:27:_:gly; b:25:_:asp; b:26:_:thr; and b:27:_:gly`. The set `Active site` should eventually contain six nodes. It is possible to color the active site nodes and the corresponding residues in the 3D structure as will be shown in Step 2B(xvii).

(xv) *Identify residue nodes on the interface of chain A.* We can use the interface `RINalyzer Node Sets` to identify which residues from chain B interact with chain A. Right-click the node set `Chain A` and execute the menu option `Select Nodes`. Afterwards, in the Cytoscape menu, go to the menu option `Select` → `Nodes` → `First Neighbors of Selected Nodes`. This operation may take several seconds and it is finished when the neighboring residues are highlighted in yellow. Back in the panel `RINalyzer Node Sets`, go to the menu option `File` → `New` → `Set from selected nodes` to create a set that contains all nodes corresponding to chain A and their neighbors. Enter the set name, e.g., `Chain A and neighbors`, and click the `OK` button to confirm it. Now, all nodes in this new set that do not belong to chain A are the nodes from chain B that interact with nodes from chain A. To extract these nodes, we need to build the intersection of the sets `Chain B` and `Chain A and neighbors`. The interface `RINalyzer Node Sets` supports typical set operations such as the union and intersection of sets. To create the intersection of two sets, select both by left-clicking while pressing the `Ctrl` key (or the `Command` key for Mac users) and go to the menu option `Operations` → `Intersection`. This action will create a new set that is the intersection of the two selected sets. Enter a name for the new set, e.g., `Chain B Interface`, and click the `OK` button to confirm it.

(xvi) *Identify residue nodes on the interface of chain B.* To create a node set `Chain A Interface`, select the nodes in the set `Chain B`; then select their first neighbors using the Cytoscape option `Select` → `Nodes` → `First Neighbors of Selected Nodes` and create a node set `Chain B and neighbors`, and finally, build the intersection of the set `Chain A` and the set `Chain B and neighbors` to create the node set `Chain A Interface` as described in Step 2B(xv).

(xvii) *Color set nodes and corresponding residues.* We can highlight different sets in the network view by changing the visual properties of the corresponding set nodes, e.g., by coloring them in a different color. Right-click the node set `Chain A` to access its context menu and select the menu option `Visual Mapping Bypass` → `Node Color`. Choose a color and click `OK` to color all set nodes in the network view. In addition, select the option `Sync 3D view colors` from the context menu to color the corresponding residues in UCSF Chimera with the same color. It is possible to repeat the same actions for the node sets `Chain A Interface`, `Chain B` and `Chain B Interface`. In the end, the network and 3D structure could look like the image shown in A.5.

(xviii) *Save node sets for further analysis.* Select all sets by left-clicking them while pressing the `Ctrl` key or by clicking the first set and then clicking the last set while holding the `Shift` key pressed. In the panel `RINalyzer Node Sets`, go to the menu option `File` → `Save`


Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Search: a:25:_:asp
Control Panel
Editor Filters RINalyzer Node Sets
File Edit Operations Tools
pdb1hiv_h
UCSF Chimera
File Select Actions Presets Tools Favorites Help
Help
Clear
Save
Welcome to Cytoscape 2.8.1
Right-click + drag to ZOOM

All Nodes (198)
Chain A (99)
Chain A Interface (35)
Chain B (99)
Chain B Interface (35)
Chain I (2)

```
A:49:_:GLY, B:82:_:VAL, B:44:_:PRO, B:48:_:GLY
> Created new set: Chain B Interface
> Chain B Interface is the intersection of Chain A +
Neighbors and Chain B
> Created new set: Chain A Interface
> Chain A Interface is the intersection of Chain B +
Neighbors and Chain A
> Created new set: Chain I
> Nodes added to the set Chain I: I:2:_:HIS, I:4:_:ILE
> Deleted set: Chain B + Neighbors
> Deleted set: Chain A + Neighbors
```

Figure A.5: Node Sets interface of RINalyzer. This user interface offers four menus to create, load, save, and modify sets of residue nodes (left). It also supports typical set operations such as union and intersection. RINalyzer keeps track of all operations performed with each node set (left bottom). The nodes in the network view of the RIN for the HIV-1 protease (top right) and the residues in the corresponding 3D protein structure (bottom right) are colored according to the node set they belong to: light blue for `Chain A`; dark blue for `Chain A Interface`; orange for `Chain B`; red for `Chain B Interface`; and pink for `Chain I`. The node set `Chain A Interface` is a subset of `Chain A`. This screenshot is taken after finishing Step 2B(ix) of the protocol. Figure first published in Doncheva *et al.* (2012a).

`selected set(s)`. Enter a file name and click `Save`. Close the resulting dialog that informs you about the successfully performed action.

(xix) *Prepare network for centrality analysis.* Make sure that the backbone edges in the network are hidden, as they are only meant to aid with the visual analysis of the RIN. To hide them, go to the menu option `Plugins` → `RINalyzer` → `Protein Structure` → `Hide backbone`. Hiding the backbone edges in the RIN will concomitantly hide the ribbons in the 3D structure view. Therefore, if you do not see the 3D structure any more, switch to UCSF Chimera and go to the menu option `Actions` → `Atoms/Bonds` → `show` to display the atoms. In addition, the 1HIV structure contains a third chain I that represents an inhibitor bound to the protease. One might want to remove or hide the corresponding RIN nodes before performing the centrality analysis. In order to select this chain I, go to the menu option `Select` → `Chain` → `I` in UCSF Chimera. Then switch to the Cytoscape main window and go to the menu option `Edit` → `Delete Selected Nodes and Edges` to delete the selected nodes.

(xx) *Handle disconnected network nodes.* Make sure the network is connected. The HIV-1 protease RIN contains two nodes, `A:40:_:GLY` and `B:37:_:SER`, which are not connected to any other node in the network. Thus, when the centrality analysis is started, a warning message will appear that the network has more than one connected components. In such cases, shortest path centrality measures are computed for each connected component independently, but current flow and random walk centralities are not computed at all. There are two possible solutions to deal with this issue: proceed with the analysis by clicking the `Yes` button, keeping in mind that these nodes are disconnected from all other nodes in the network; alternatively, cancel the analysis by clicking the `No` button, select the two disconnected nodes in the network view and delete them by clicking `Edit` → `Delete Selected Nodes and Edges`.

(xxi) *Select root nodes for analysis.* The centrality analysis can be started only if a set of nodes (root set) is selected in the network view. *RINalyzer* computes each centrality measure with respect to the root set (Box 3). For example, the weighted degree of a node is computed by counting its neighbors that are contained in the root set and that are within a given distance cutoff from the node of interest. The first-time user might just select all nodes by clicking `Select` → `Nodes` → `Select all nodes` in the Cytoscape main window. This action can take a few seconds because both the nodes in the network and the residues in the 3D structure are selected.

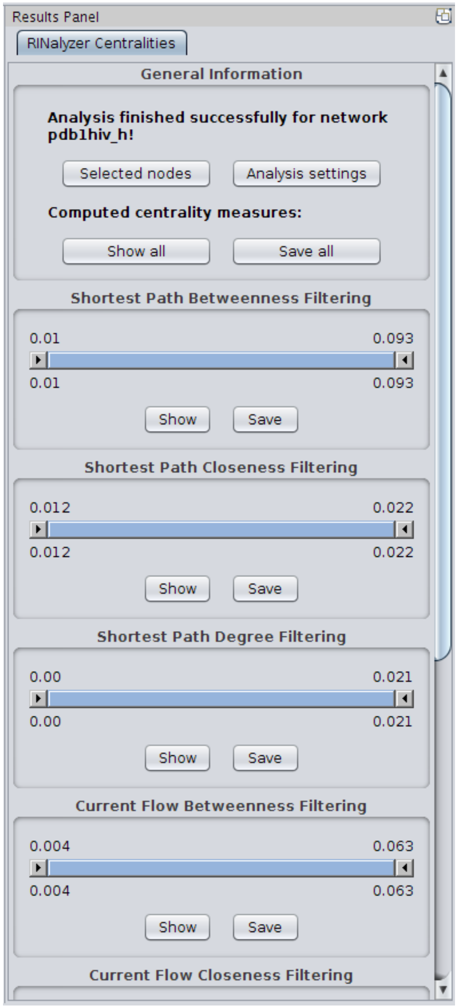
(xxii) *Perform centrality analysis.* Start the analysis from the menu option `Plugins` → `RINalyzer` → `Analyze Network`. A dialog that contains different analysis settings will appear. The settings are described in detail in Box 3. In this dialog, choose which centrality measures will be computed by checking the corresponding boxes. If the network is not connected, only shortest path centralities will be computed. Here we choose the edge-weight attribute `NrInteractions` (a value representing the number of interactions between two residues). Furthermore, we select the option `Average weight` to consider the average weight of multiple edges between node pairs as well as the max-value method for converting the averaged weight scores into distances. If we set the cutoff to 10, the nodes that are connected by high-weighted paths with other nodes will get high weighted degrees. After customizing the settings, click the `Analyze` button. The dialog `Progress` will appear and show the progress of the analysis. The time for computing the centrality measures depends on the size of the network and the number of selected nodes. The analysis can be canceled at any time by clicking the `Cancel` button in the dialog. ?TROUBLESHOOTING

(xxiii) *View RINalyzer results panel.* After successful computation, the results are shown in the `RINalyzer Centralities` tab of the `Cytoscape Results Panel` (A.6). It consists of at least three different sections; the first one entitled `General Information` contains general analysis information, and each consequent section provides access to the values of a centrality measure. Go to the menu option `View` → `Hide Data Panel` to free more display space for the centrality analysis results.

(xxiv) *Explore analysis settings.* The panel `General Information` keeps track of the analyzed network, the set of selected nodes and the analysis settings chosen for the current analysis run. Click the button `Selected Nodes` to select the root set nodes in the network view. Click the button `Analysis Settings` to see the settings for this run.

(xxv) *View centrality values.* Navigate to the panel `Shortest Path Closeness Filtering` to become familiar with the three possible ways of examining the results. To view the raw centrality values, click the `Show` button. A table with two columns will be displayed. The first column contains the names of all nodes in the network and the second one the corresponding closeness centrality values. Clicking on the name of the second column sorts the rows according to the values in the cells. An up-arrow indicates descending order and a down-arrow ascending order.

(xxvi) *Select best-scoring residue nodes.* Use the selection filter to see those nodes in the network view that have centrality values in a specified range. The numbers above the slider show the lowest and the highest bounds for the selection filter, 0.012 and 0.022, respectively. The numbers below the slider are the current bounds. Move the left end of the slider to the



Analysis finished successfully for network pdb1hiv_h!

Figure A.6: Centrality analysis results by RINalyzer. The panel `RINalyzer Centralities` consists of different sections. The first section contains general analysis information and allows the user to show all computed centrality values in a table or to save them to a file. The other sections provide access to the values of each centrality measure. In particular, the user can apply a selection filter to select nodes with centrality values in a given range, view the values in a table, or save them to a file. This figure shows the results of the centrality analysis performed in Step 2B(xii) and explored in Steps 2B(xiii) and (xiv). Figure first published in Doncheva *et al.* (2012a).

right. The left number should change and move to the right as you move the slider. At the same time, you should see that the node selection in the network changes. Stop when you reach 0.02 to see the nodes in the network view that have a closeness centrality above 0.02. Double-click the slider and a window will appear in which the lower and upper bound for the selection range can be inserted manually. Set the lower bound to 0.021 and click the `OK` button to set the new selection bounds.

(xxvii) *(Optional) Create set of best-scoring active site nodes.* If the panel `RINalyzer Node Sets` is open, go to the menu item `File` → `New` → `Set from selected nodes` to create a new set from the selected nodes. Insert a name for the new set, e.g., `Best SPC`, and click the `OK` button to confirm it. Now, we can determine which active site residues have best centrality scores, and vice versa. Select both the set `Active site` created in Step 2B(xiv) and the set `Best SPC`. Then go to `Operations` → `Intersection`. Insert a name for the new set, e.g., `Best-scoring active site residues`, and click the `OK` button to confirm it. The resulting set should contain four out of the six active site residues.

(xxviii) *Save centrality values.* Save the centrality values by clicking on the Save button in the panel `Shortest Path Closeness Filtering`. Navigate to the directory where you want to save the file. Insert the name of the file, e.g., *pdb1hiv_spc*, and click the Save button to confirm it. *RINalyzer* will automatically add the extension *.centstats* to the file name.

(xxix) *Show and save all results.* The values of all computed centrality measures can be visualized in a table or saved to a file by using the button `Show All` or `Save All`, respectively.

(xxx) *Save RIN data and view.* The current network, the loaded attributes, the current network view and the node sets can be stored as a Cytoscape session file, which can be opened again later on. Go to the menu option `File` → `Save` and select a directory. Enter a session file name and confirm it by clicking the `Save` button.

Step 2C: Comparison of residue networks

(i) *Download data.* Here we analyze four RINs that represent the four subunits of human deoxyhemoglobin (chains A, B, C and D in the PDB structure with identifier 4HHB). For this purpose, download Supplementary Data 2 and unzip the files from the archive into a new directory, e.g., *rins*.

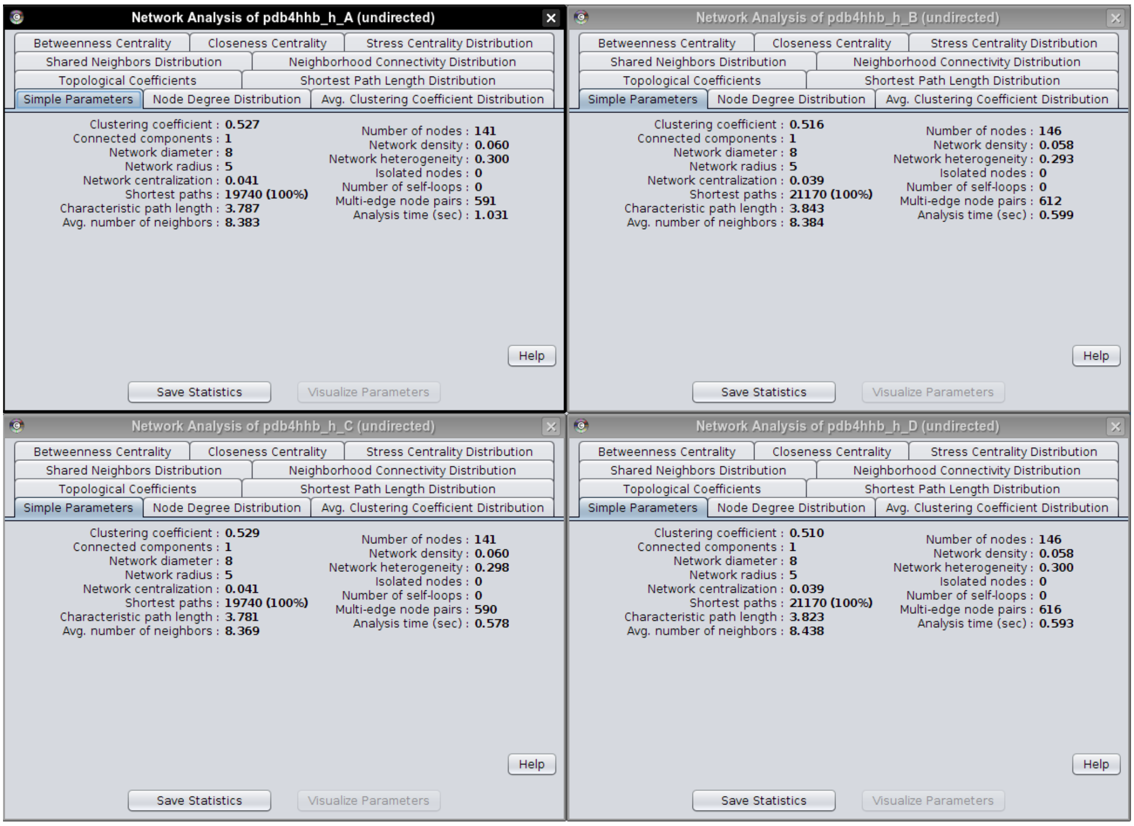
(ii) *Specify batch analysis settings.* To start the batch analysis dialog, go to the menu option `Plugins` → `Network Analysis` → `Batch Analysis`. Select the input directory by clicking on the first `Select Directory` button in the dialog `Batch Analysis`. Navigate to the input directory, *rins* in our case and click the `Open` button to confirm it. The input directory should contain network files that can be loaded into Cytoscape. The output directory will contain all analysis results after the batch analysis. In order to avoid file overwriting, *NetworkAnalyzer* requires that the output directory is empty before the batch analysis starts. Therefore, create a new directory for the output, e.g., *output*, using an external file browser. In the dialog `Batch Analysis`, click the second `Select Directory` button. Navigate to the output directory, and click the `Open` button to confirm it. As RINs are undirected, we do not need to consider all network interpretations. Select the option `Consider networks as undirected`.

(iii) *Perform batch analysis.* Click the button `Start Analysis`. A dialog appears that displays the progress of the batch analysis. Depending on the number of networks and their size, this might be a very time-consuming step. The batch analysis can be canceled at any time by clicking the `Cancel` button in the progress dialog.

(iv) *View batch analysis results.* After the analysis is complete, the button `Show Results` is enabled. Click on it to see the dialog `Batch Analysis - Results`. The dialog contains a table of all topological analyses performed. Every row in the results table lists the loaded network, its interpretation and the resulting network statistics file, which was saved to the output directory.

(v) *Load network statistics in Cytoscape.* Clicking on a network name and a statistics file name will load the network and the topological analysis results in Cytoscape, respectively. Load all four statistics files and compare the simple parameters computed for each network (A.7). We can notice that the two α subunits, networks A and C, are very similar to each other. This is also the case for the two β subunits, networks B and D. However, there are apparent differences between the network parameters for the RINs of the α and β subunits. Close the network statistics dialogs to finish the results inspection.

(vi) *Load networks into Cytoscape.* Click on the networks `pdb4hhb_h_A.sif` and `pdb4hhb_h_B.sif` to load them into Cytoscape for the next steps. You can now close the dialog `Batch Analysis - Results`.


Network Analysis of pdb4hhb_h_A (undirected)
Network Analysis of pdb4hhb_h_B (undirected)
Network Analysis of pdb4hhb_h_C (undirected)
Network Analysis of pdb4hhb_h_D (undirected)
Betweenness Centrality
Closeness Centrality
Stress Centrality Distribution
Shared Neighbors Distribution
Neighborhood Connectivity Distribution
Topological Coefficients
Shortest Path Length Distribution
Simple Parameters
Node Degree Distribution
Avg. Clustering Coefficient Distribution
Help
Save Statistics
Visualize Parameters

Network Analysis of pdb4hhb_h_A (undirected)

```
        Clustering coefficient : 0.527
       Connected components : 1
          Network diameter : 8                     Number of nodes : 141
            Network radius : 5                    Network density : 0.060
     Network centralization : 0.041          Network heterogeneity : 0.300
            Shortest paths : 19740 (100%)           Isolated nodes : 0
 Characteristic path length : 3.787          Number of self-loops : 0
  Avg. number of neighbors : 8.383         Multi-edge node pairs : 591
                                            Analysis time (sec) : 1.031
```

Network Analysis of pdb4hhb_h_B (undirected)

```
        Clustering coefficient : 0.516
       Connected components : 1
          Network diameter : 8                     Number of nodes : 146
            Network radius : 5                    Network density : 0.058
     Network centralization : 0.039          Network heterogeneity : 0.293
            Shortest paths : 21170 (100%)           Isolated nodes : 0
 Characteristic path length : 3.843          Number of self-loops : 0
  Avg. number of neighbors : 8.384         Multi-edge node pairs : 612
                                            Analysis time (sec) : 0.599
```

Network Analysis of pdb4hhb_h_C (undirected)

```
        Clustering coefficient : 0.529
       Connected components : 1
          Network diameter : 8                     Number of nodes : 141
            Network radius : 5                    Network density : 0.060
     Network centralization : 0.041          Network heterogeneity : 0.298
            Shortest paths : 19740 (100%)           Isolated nodes : 0
 Characteristic path length : 3.781          Number of self-loops : 0
  Avg. number of neighbors : 8.369         Multi-edge node pairs : 590
                                            Analysis time (sec) : 0.578
```

Network Analysis of pdb4hhb_h_D (undirected)

```
        Clustering coefficient : 0.510
       Connected components : 1
          Network diameter : 8                     Number of nodes : 146
            Network radius : 5                    Network density : 0.058
     Network centralization : 0.039          Network heterogeneity : 0.300
            Shortest paths : 21170 (100%)           Isolated nodes : 0
 Characteristic path length : 3.823          Number of self-loops : 0
  Avg. number of neighbors : 8.438         Multi-edge node pairs : 616
                                            Analysis time (sec) : 0.593
```

Figure A.7: Network statistics of the four subunits of human deoxyhemoglobin. The batch analysis results dialog allows users to open the network statistics of the analyzed networks. The computed simple topological parameters for the networks representing all subunits of the human deoxyhemoglobin are displayed as follows: chain A (top left), chain B (top right), chain C (bottom left), and chain D (bottom right). Figure first published in Doncheva *et al.* (2012a).

(vii) *Retrieve structure alignment file.* *RINalyzer* offers the functionality to compare two RINs based on a superposition alignment of the corresponding 3D protein structures. Here we compare two of the networks loaded in the previous step, i.e., one α subunit and one β subunit of the human protein deoxyhemoglobin (PDB identifier 4HHB). Start a web browser and navigate to the `RCSB PDB Protein Comparison Tool` website (`http://www.rcsb.org/pdb/workbench/workbench.do`). Insert the PDB identifier `4HHB` in the text field for `ID 1` and choose chain A by selecting `4HHB.A` in the drop-down menu. Insert the same identifier in the text field for `ID 2` and select `4HHB.B` in the drop-down menu. Then, in the drop-down menu `Select Comparison Method`, choose the `jCE algorithm` and click the `Compare` button. In the `Structure Alignment View` page, scroll down to the panel `Download Alignment`. Right-click the link `Download XML` and select the option `Save Link As`. Navigate to the directory where the file should be saved, enter a name for it (e.g., *4hhba_vs_4hhbb.xml*), and click the `Save` button to confirm it. Close the *Protein Comparison Tool*.

(viii) *Perform RIN comparison.* To compare RINs using *RINalyzer*, go to `Plugins` → `RINalyzer` → `Compare RINs`. Select `pdb4hhb_h_A` as the first network and `pdb4hhb_h_B` as the second network. Then, enter a name for the resulting comparison network (e.g., comparison). Click the `...` button and navigate to the alignment file downloaded in Step 2C(vii). Confirm its selection by clicking the `Open` button. Next, click the `Compare` button to perform the actual comparison. A new network with 148 nodes and 2405 edges is created. This combined RIN consists of three types of nodes: nodes that represent aligned residues according to the structure superposition, and two types of nodes that correspond to residues that were not

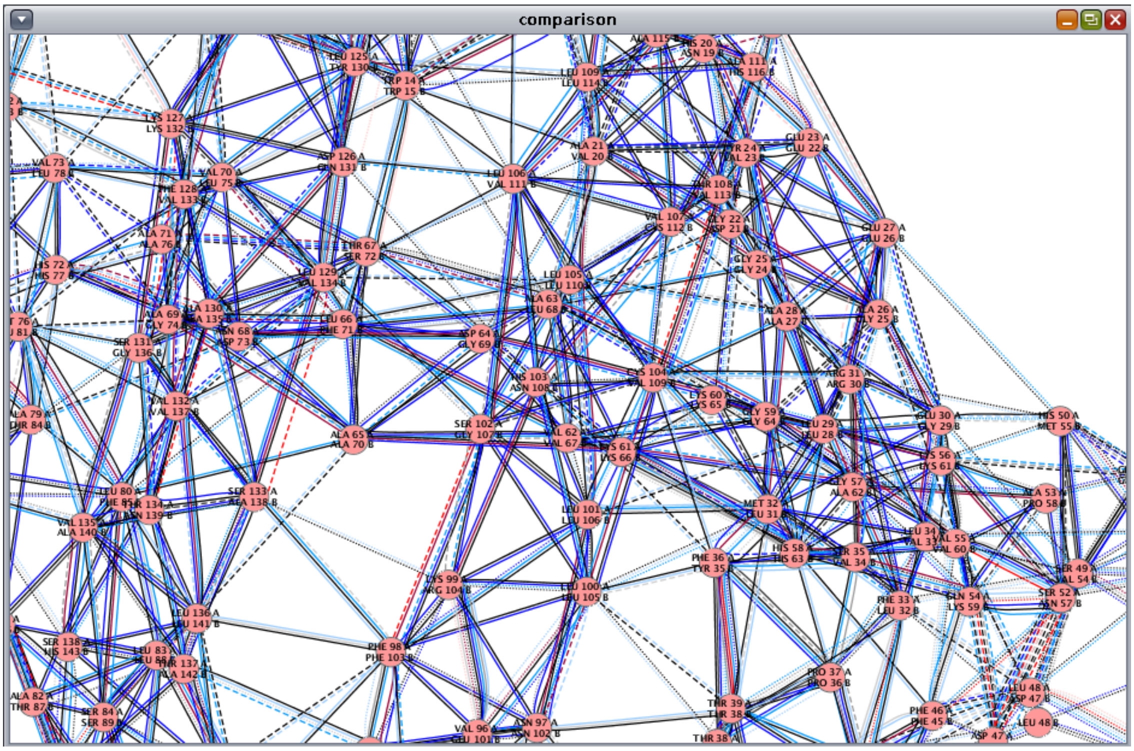


Figure A.8: Comparison network generated by RINalyzer. The combined network resulted from the comparison of the two RINs that represent one of the α and one of the β subunits of human deoxyhemoglobin (PDB identifier 4HHB (Fermi *et al.*, 1984), chains A and B). Edge colors refer to the interaction type, i.e., interatomic contacts in blue; hydrogen bonds in red; and overlaps in gray. Edge line styles correspond to noncovalent residue interactions that are preserved in both subunits (solid lines), present only in the α subunit (dashed lines) or only in the β subunit (dotted lines). Figure first published in Doncheva *et al.* (2012a).

aligned by the superposition and belong to the first or second network. The network also contains three different edge line styles: solid lines for interactions present in both networks, dashed lines for interactions from the first network and dotted lines for interactions from the second network. The type of each node and edge is stored as an attribute named `BelongsTo` and represented by one of the following three values: `net1`, `net2` or `both`. The value `net1` refers to the first RIN selected in the comparison, and the value `net2` to the second RIN.

(ix) *Adjust network view.* Maximize the network view window and show the graphics details from the Cytoscape menu (`View` → `Show Graphics Details`).

(x) *Apply network layout.* Apply the yFiles organic layout (`Layout` → `yFiles` → `Organic`) and the RIN visual properties(`Plugins` → `RINalyzer` → `Visual Properties`). The resulting network should look as in A.8.

(xi) *Hide interaction edges.* First, we want to reduce the visual complexity by showing fewer edges. Go to the menu option `Plugins` → `RINalyzer` → `Visual Properties` and select the `Edges` tab. Hide all edges except `combi:all_all` by unchecking the boxes next to each edge type and close the dialog `RIN Visual Properties` by clicking the `Close` button.

(xii) *Color nodes and edges.* Now, we color the nodes and edges according to the network they belong to. In the `Cytoscape Control Panel`, go to the tab `VizMapper` and double-click the field `Edge Color`. Select the edge attribute `BelongsTo` from the drop-down menu for edge color values and the mapping type `Discrete Mapping` from the mapping type drop-down menu. A list that contains the three `BelongsTo` attribute values `net1`, `net2` and `both` will appear. For each attribute value, do the following: click the field next to the attribute value and the button `...` will appear; click this button, select a color and click the `OK` button to confirm it. Repeat the same actions for mapping the node color using the `BelongsTo` node attribute.

(xiii) *Customize node labels.* It is also possible to change the node labels by clicking the field next to the visual property `Node Label`. Then select the attribute `CombinedLabel` and the mapping type `Passthrough Mapping`. The node attribute `CombinedLabel` contains node labels composed of the labels of the aligned nodes from the compared networks.

(xiv) *Explore comparison network.* After the mapping is applied to the network view, it should look as in Supplementary Figure 10. Zoom in using the `+` button in the Cytoscape toolbar to observe the residue interaction differences between the superimposed α and β subunits of deoxyhemoglobin.

Troubleshooting

Troubleshooting advice for basic problems that may occur during the procedure is given in Table A.1.

Further information about using Cytoscape can be found in the documentation at `http://www.cytoscape.org/documentation_users.html` and via the helpdesk mailing list. Tutorials and documentation about UCSF Chimera are available at `http://plato.cgl.ucsf.edu/chimera/docindex.html` and questions can be addressed to the users' mailing list (chimera-users@cgl.ucsf.edu). *RINalyzer* and *NetworkAnalyzer* documentations can be found at `http://www.rinalyzer.de/documentation.php` and at `http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/index.html`, respectively.

Table A.1: Troubleshooting table.

Step	Problem	Possible reason	Solution
1	Cytoscape does not start.	Java is not installed properly.	Make sure that Java version 6 is installed. Java can be downloaded from http://www.java.com/
2A(iv), 2B(xxii)	The analysis takes very long or seems to be frozen.	Cytoscape has run out of memory.	Increase the memory for the Cytoscape program. One way to do this is to start Cytoscape from the command line and use the `-Xmx` option to set the memory size. To this end, open a command line window, navigate to the Cytoscape directory and type `java -Xms10m Xmx1500M -jar cytoscape.jar -p plugins` to start Cytoscape with 1,500 MB of memory. For alternative ways to increase the memory, see the Cytoscape Wiki.
2B(i)	There is no RIN data for a protein.	The RINdata database does not contain precomputed RINs for all PDB identifiers.	Download and apply the package RINerator to generate the RIN data. Alternatively, the RING web server (Martin *et al.*, 2011) can be used to create different types of RINs.

Continued on next page

Step	Problem	Possible reason	Solution
2B(iv)	UCSF Chimera does not start.	The path to UCSF Chimera is not configured properly.	Open the dialog `Cytoscape Preferences Editor` (`Edit` → `Preferences` → `Properties`). Click the `Add` button and enter `Chimera.chimeraPath` as the name of the property. Click `OK` and enter the path to the UCSF Chimera application. Save the new preferences by clicking the option `Make Current Cytoscape Properties Default` at the bottom of the dialog.
2B(vii)	RINLayout is not applied to the network.	The 3D structure corresponding to the current network is not loaded in UCSF Chimera.	Load the protein structure corresponding to the current network using the menu option `Plugins` → `RINalyzer` → `Protein Structure` → `Open structure from file`.
		More than one protein structure is loaded in UCSF Chimera.	Close all protein structures opened in UCSF Chimera except for the structure that corresponds to the current network, using the menu option `Plugins` → `RINalyzer` → `Protein Structure` → `Close`.

Anticipated results

Here we discuss the results obtained by following each of the three workflows described in this protocol.

Step 2A: Topological analysis of biological networks

The application of *NetworkAnalyzer* on the protein-protein interaction network from Yu *et al.* (Yu *et al.*, 2011a) (Supplementary Data 1) produces a comprehensive set of topological network parameters. The network exhibits scale-free behavior because a power law $k^{-\alpha}$ with $\alpha = 1.62$ can be fitted to the node degree distribution. Furthermore, such an α value is indicative for a hub-and-spoke network with one hub being connected to a large fraction of nodes. Indeed, the network contains one hub protein with an exceptionally high node degree (151 interactions). The visual exploration of the network view after mapping the clustering coefficient to node color suggests that only a few nodes have clustering coefficients larger than 0. This means that the proteins in the network do not tend to form clusters with their interaction partners.

Step 2B: Interactive visual analysis of residue networks

The RIN generated from the protein structure of the HIV-1 protease (PDB identifier 1HIV) contains 200 nodes and 2199 edges. The nodes can be divided into three groups according to the protein chain: 99 nodes for residues in chain A; 99 nodes for residues in chain B; and two nodes for chain I. By using the interface *RINalyzer Node Sets*, we could identify the residues in the interface between chains A and B of the protein structure. In all, 35 residues from chain A interact with 35 residues from chain B (dark blue and red nodes in Supplementary Fig. 5, respectively).

Furthermore, we performed a centrality analysis of the RIN of the HIV-1 protease to highlight central nodes. The best-scoring nodes according to weighted shortest path closeness (i.e., centrality values > 0.21) were saved in a node set. The overlap between this node set (seven nodes) and the node set representing the protease active site (six nodes) is four nodes. When studying the single centrality values in a table sorted from highest to lowest closeness, we observed that the four active site residues have the best ranks.

Step 2C: Comparison of residue networks

The RINs (Supplementary Data 2) generated from the four subunits of human deoxyhemoglobin (chains A, B, C and D in the PDB structure with identifier 4HHB) are of similar size: 141 nodes and 1,885 edges for chain A; 146 nodes and 1,935 edges for chain B; 141 nodes and 1887 edges for chain C; and 146 nodes and 1971 edges for chain D. As one might expect, the analysis performed with *NetworkAnalyzer* (Supplementary Fig. 9) indicates that the RINs of chains A and C, the two α subunits, have almost identical simple network parameters such as clustering coefficient, network centralization, number of shortest paths, characteristic path length and network density. The same holds for the RINs of chains B and D, the two β subunits. However, the difference between the simple parameter values for chains A and B is, for most parameters, larger than the difference between the same subunits. The complete set of both simple and complex network parameters can be compared further using the network statistics files generated by *NetworkAnalyzer*.

To compare the individual residue interactions in the two RINs of chains A and B, we used *RINalyzer*, which generates a combined comparison network based on the superposition alignment of the corresponding 3D structures. The comparison network contains 148 nodes and 2405 edges. Of the 148 nodes, two represent residues in chain A and 7 nodes residues in chain B; the remaining 139 nodes correspond to the aligned residues. The number of edges that correspond to non-covalent interactions that are identical in both subunits (1415 edges) is considerably higher than the number of nonidentical edges (470 and 520 for chains A and B, respectively). These numbers reflect structural similarities and differences of the two subunits. When visually exploring the simplified comparison network (Supplementary Fig. 10), we can recognize the large number of edges that represent non-covalent interactions identical in both subunits (523 solid edge lines) and the rather small number of interactions present either in the α subunit (68 dashed edge lines) or in the β subunit (89 dotted edge lines) of deoxyhemoglobin. The nonidentical edges can be seen mainly in the network part that contains nodes of unaligned residues. Dashed or dotted edges between aligned residue nodes indicate that the corresponding residues form functionally distinct interactions in the two homologous, structurally very similar subunits.

APPENDIX B

Supplementary figures and tables

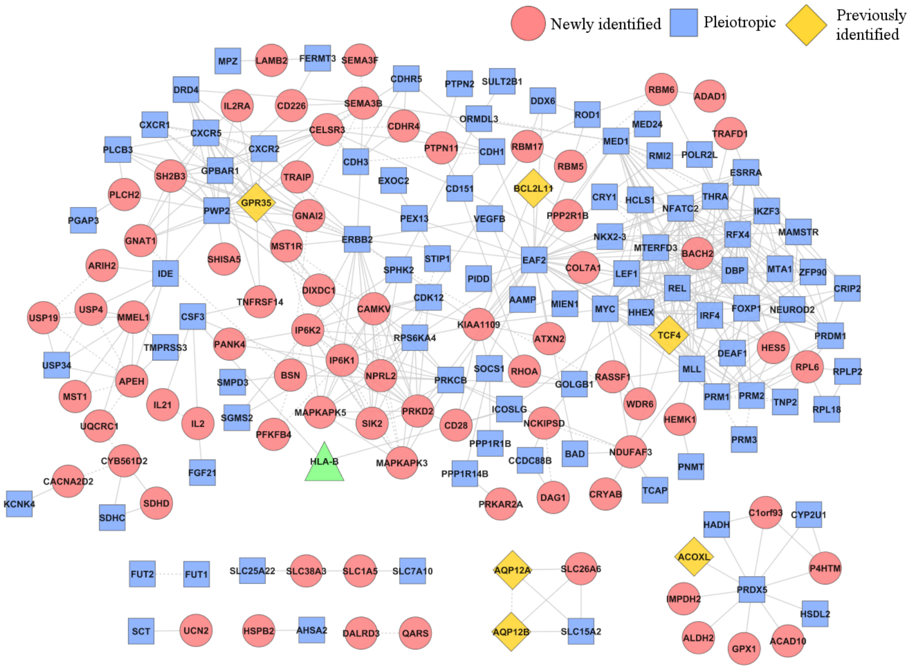


Figure B.1: Functional similarity network of genes within 0.1cM of identified PSC susceptibility regions (red circles), pleiotropic PSC loci (blue rectangles), previously identified loci (orange diamonds) and a representative of the HLA locus (green triangle). The network contains 177 out of the 341 considered genes and these are connected by 511 edges (51 within the same locus and 460 between different loci). Grey edges indicate strong functional similarity based on Gene Ontology annotations and connect genes either from different loci (solid edge lines) or within the same locus (dashed edge lines). Figure first published in (Liu *et al.*, 2013).

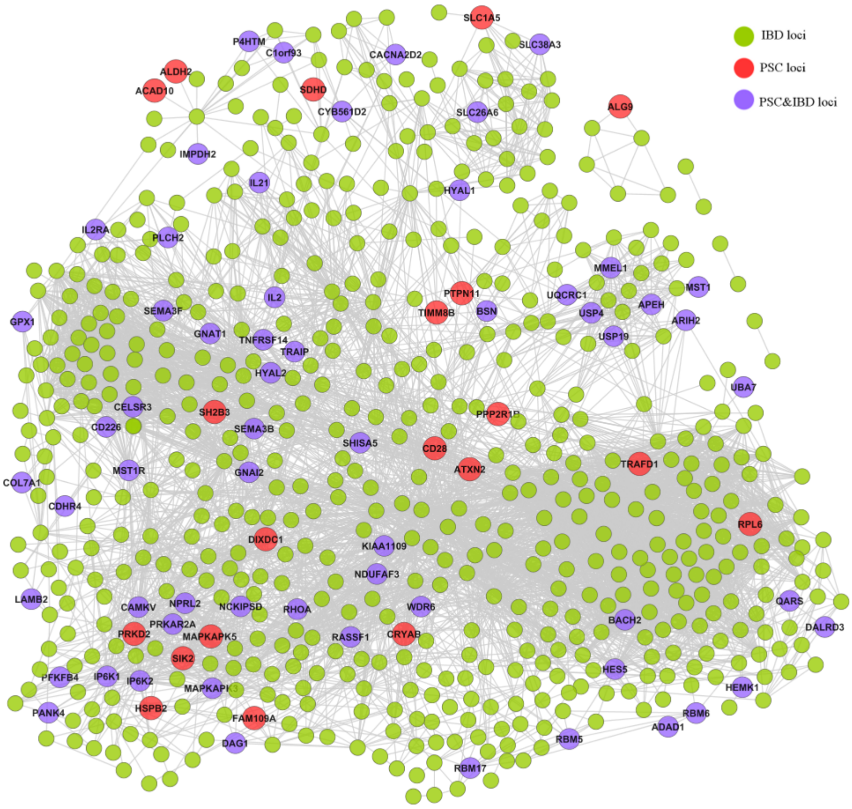
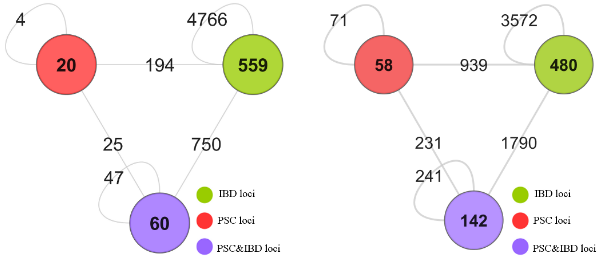


Figure B.2: Functional similarity network of the genes within 0.1cM of identified PSC susceptibility regions and the 163 confirmed IBD loci. The network contains 20 PSC only (red circles), 559 IBD only (green circles), and 60 PSC&IBD loci genes (violet circles) and these are connected by 5786 edges. Grey edges indicate strong functional similarity based on Gene Ontology annotations.

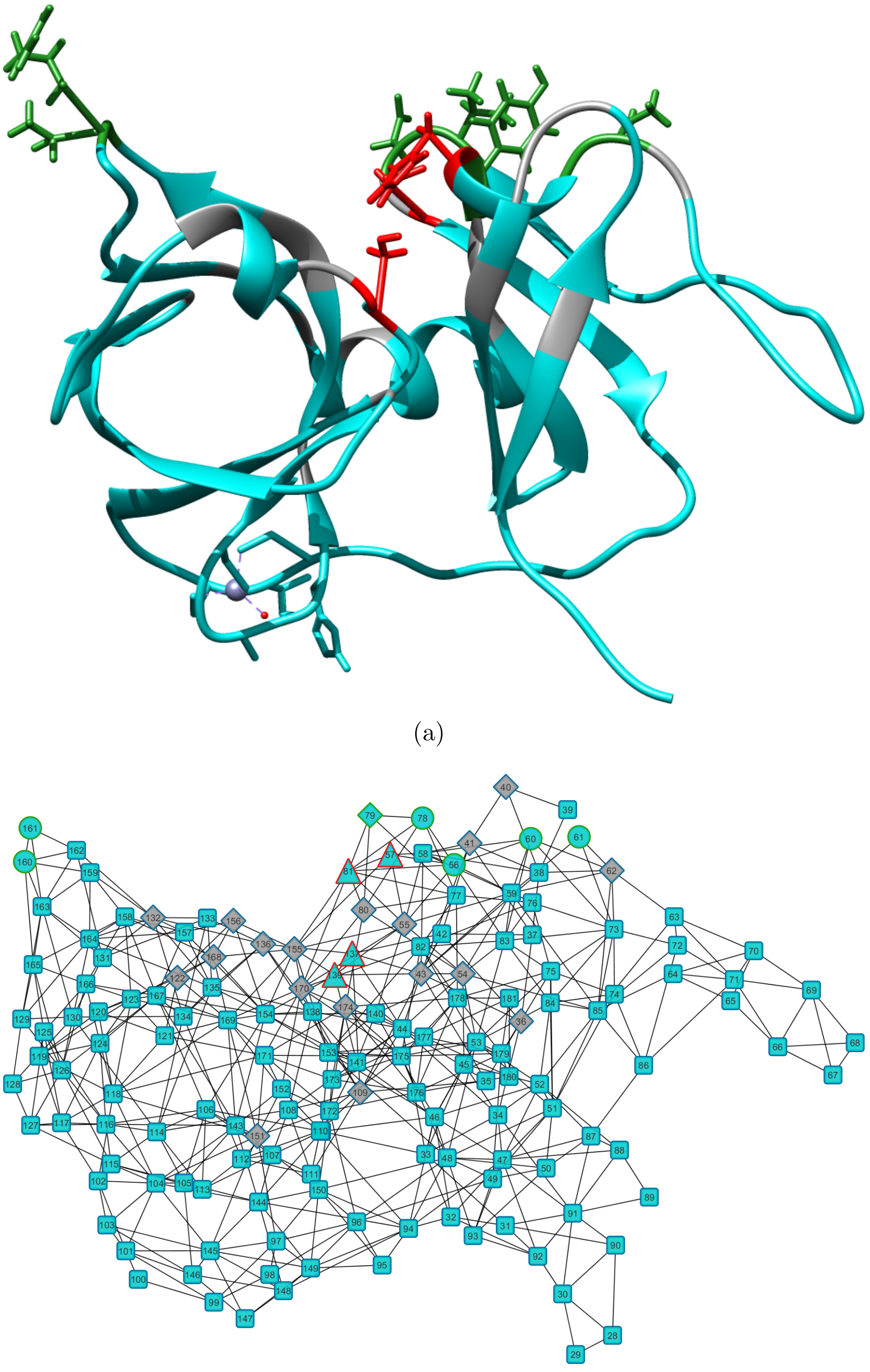


(a) (b)

Figure B.3: Summary of functional similarities between PSC and IBD associated loci for (a) the network from Figure B.2 and (b) the network that additionally contains all genes within 0.1cM of the pleiotropic and previously known PSC loci. The genes are grouped based on their association with PSC (red), IBD (green), and both PSC and IBD (violet), and each node is labeled with the number of associated genes. An edge indicates that the corresponding genes are connected by similarity edges and is labeled by the numbed of such connections. The functional similarity edges connecting genes in the same group are represented as self-loops.

Table B.1: Fold change (FC) for resistance mutations in HCV NS3/4A protease against linear and macrocyclic protease inhibitors. The FC values have been collected from the following publications: 1) Shimakami *et al.* (2011); 2) Welsch *et al.* (2012b); 3) Welsch *et al.* (2012a); 4) Welsch *et al.* (2008); 5) Jiang *et al.* (2013); 6) Lawitz *et al.* (2013); 7) McPhee *et al.* (2013); 8) Dvory-Sobol *et al.* (2012); 9) McPhee *et al.* (2012); 10) unpublished data from Christoph Welsch (University Hospital Frankfurt).

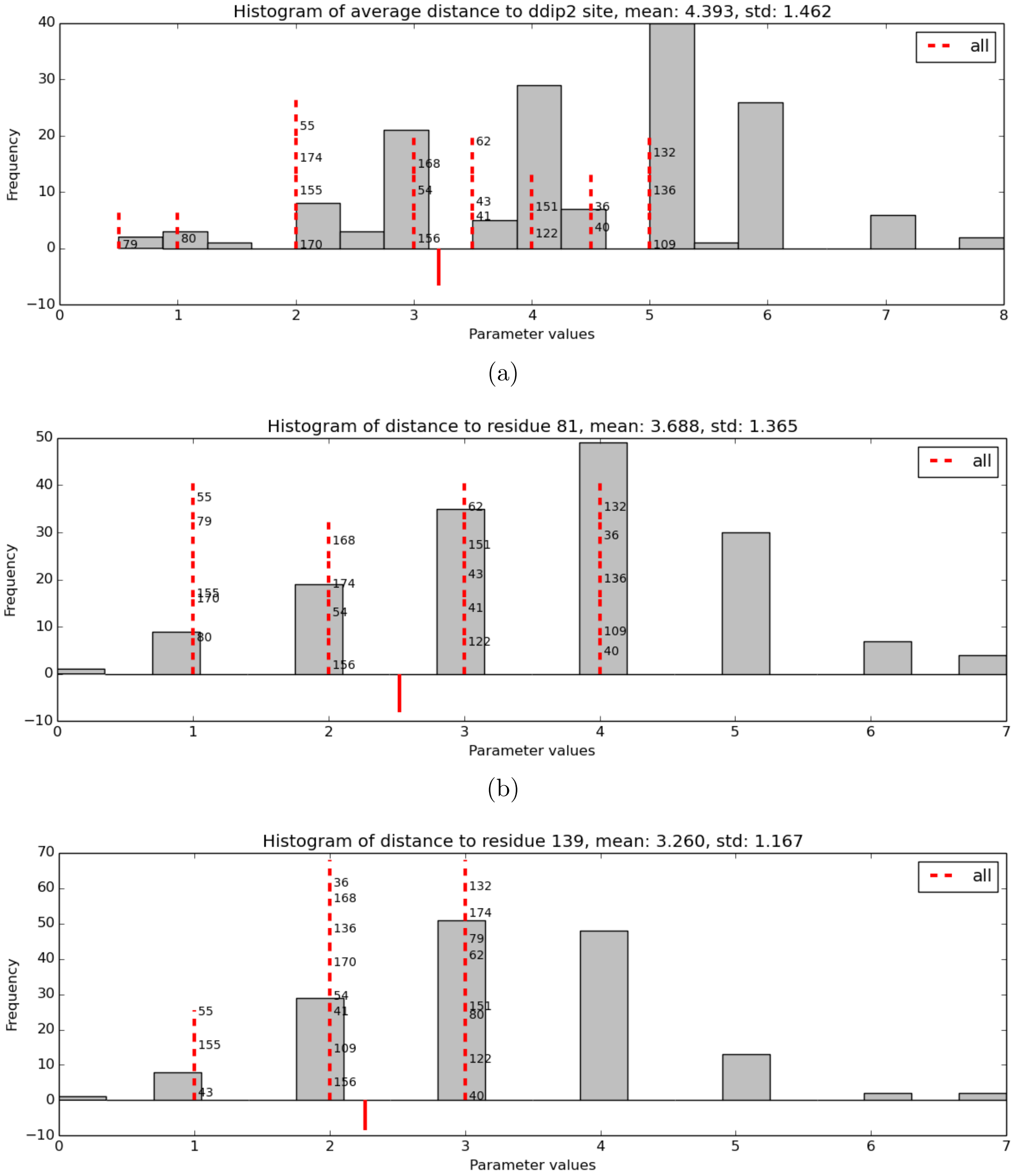
Drug class		Linear	Linear	unk	Macrocyclic	Macrocyclic	Macrocyclic	Macrocyclic	Macrocyclic	
Compound		Telaprevir	Boceprevir	GS-9451	Vaniprevir	Ciluprevir	Danoprevir	Simeprevir	Asunaprevir	
		(VX-950)	(SCH 503034)		(MK7009)	(BILN2061)	(ITMN191)	(TMC435)	(BMS-650032)	
Residue	RAV					FC				Publication
V36	A	21.58	3.6			1.2	1.8		3	1,5,9,10
	G	28.08	7.9			1	2.3			1,5,10
	L	2.2	3			1.2	1.3		2	1,5,9
	M	7	3		1.8	1	1.8	2.1	2	1,5,9
	C	7.8	1.8				1.4			5
	I	0.3								5
T40	A								1	9
Q41	R		1.5		2.4	6.2	3			1
	H	3.5	1.2							3
F43	S	18.78	6.9				44			1,10
	C									
	L								4	9
T54	A	12.28	5.5		1.1	0.7	1.1	1	0.4	1,4,5,9,10
	S	8.22							1	4,9,10
V55	A		1.6							2
	I	1.24	0.8						3	2,9,10
R62	K								1	9
D79	E								1	9
Q80	R	1.09				9.3				1,10
	K								3	9
	L								1	9
R109	K	3.86	1.2			0.9	0.8			1,5,10
S122	G								1	9
	N								1	9
	R								3	9
I132	V	2.4	1.1							3
K136	R	0.9	0.5							3
S138	T									
V151	A	0.9								5
F154	Y									3
R155	K	16.59	3.7	150	538	510	316	18	21	1,5,8,9,10
	T		5.2		312	460	45			1
	Q	4.1	0.4			267				1,10
	G	7.4	3.3		181.8	580	19			5,6
	M	5.6	2.9			30	1.9			5
	S	4.1	2.1		72.8	418	7.9			5,6
	I	24	7.7			26	1.3			5
	N				39.8					6
A156	S	9.6	7		17.5		18	0.4		1,5
	T	62	13.7		250	706	18	33		1,5
	V	62	40		2041		6.1			5
	F	62								5
	N	93								5
D168	A		1.1		900	453	330		23	1
	E	4.1		82			56		58	1,8,10
	H						160			1
	I									
	V	1.52	1	1000	1700	733	166		373	1,8,10
	G	2.2	0.6	85	55.2		31		14	3,6,8,9
	N						20			1
	Y								622	9
	T								205	7
I170	A		2.2		1.4					5
	T	4.61	2.1						5	9,10
	V								1	9
N174	Y								1	9



(a)

(b)

Figure B.4: Visualization of (a) the HCV NS3/4A protease structure as ribbon in UCSF Chimera (PDB identifier 3KF2) and (b) the corresponding RIN in Cytoscape with focus on the functional and phenotypic residues. The catalytic residues are represented as red sticks in the 3D structure and as red bordered triangles in the network, while the *ddip* residues are shown as green sticks and green bordered circles. The phenotypic residues and nodes are colored in gray.

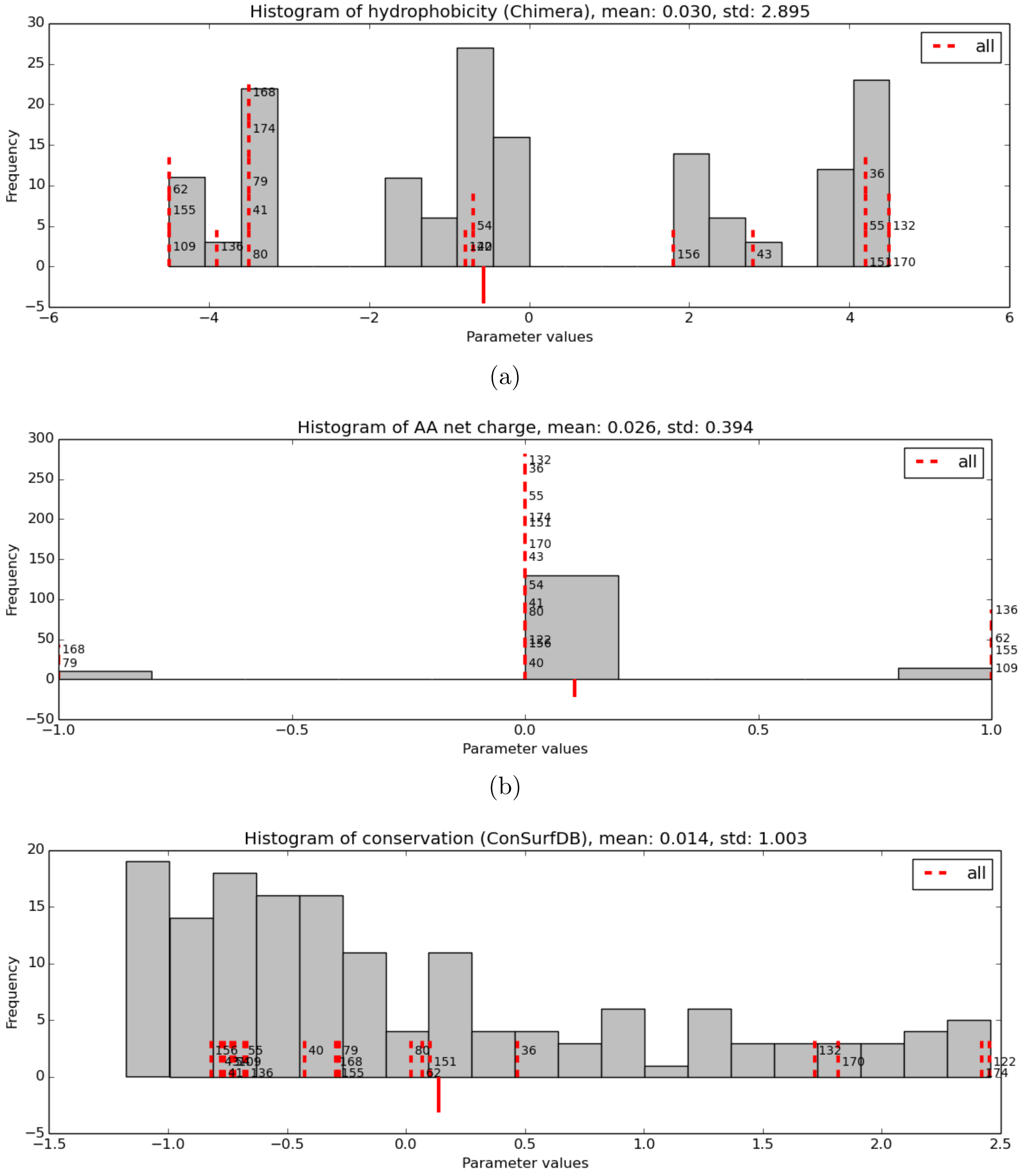


(a)

(b)

(c)

Figure B.5: Distribution of residue values for selected topological properties: (a) distance to *ddip2* site, (b) distance to catalytic residue 81, and (c) distance to catalytic residue 139. The values of the phenotypic residues are highlighted by vertical dotted lines in red and the vertical solid line indicates the average value. All vertical lines have the same size (proportional to the plot height).

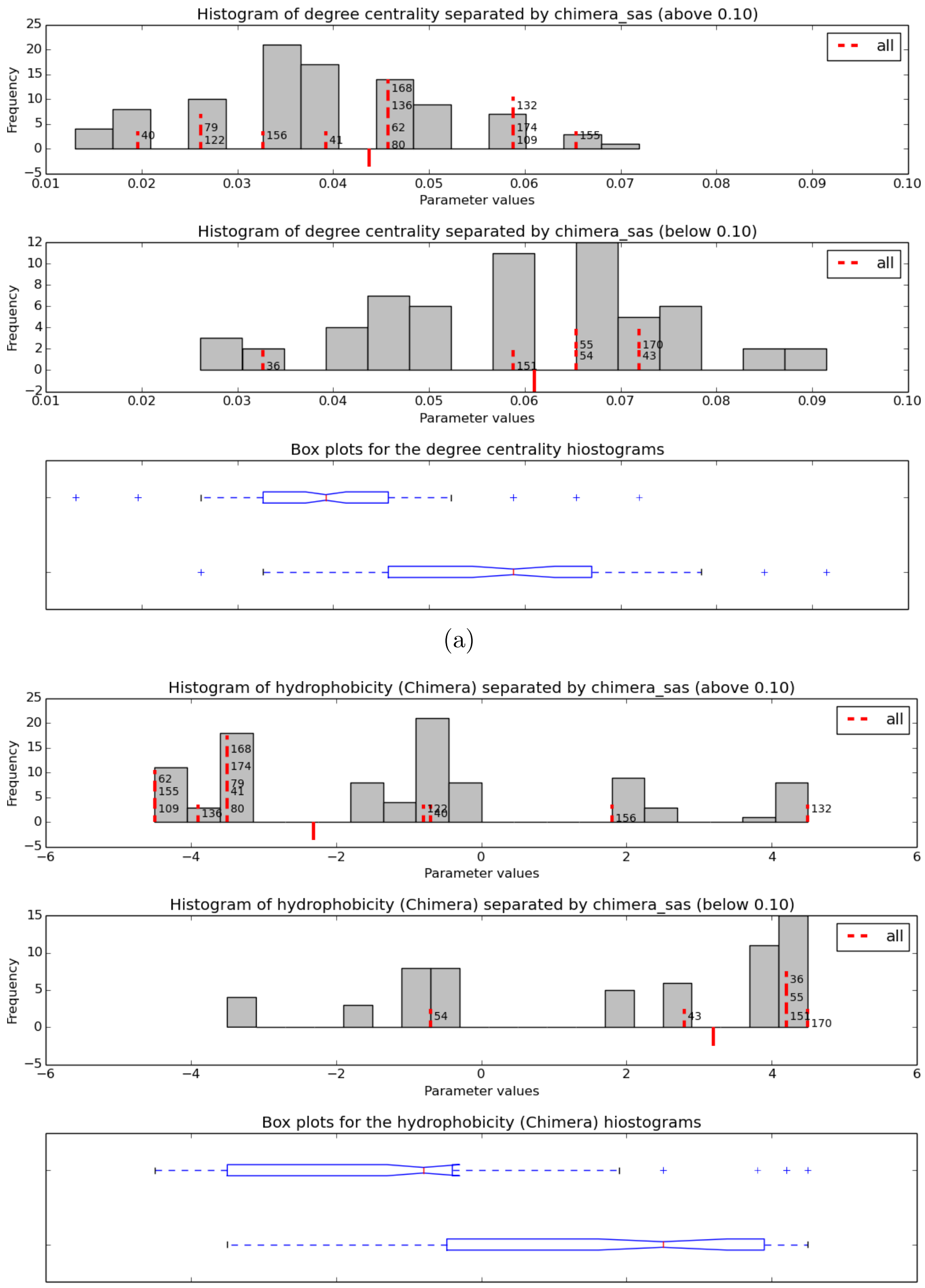


(a)

(b)

(c)

Figure B.6: Distribution of residue values for selected physico-chemical properties: (a) hydrophobicity (from UCSF Chimera), (b) net charge (AA index KLEP840101), and (c) conservation (from ConSurfDB). The values of the phenotypic residues are highlighted by vertical dotted lines in red and the vertical solid line indicates the average value. All vertical lines have the same size (proportional to the plot height).



(a)

(b)

Figure B.7: Histogram and box-and-whisker plot of selected properties for exposed (top) and buried (middle) residues: (a) degree centrality, (b) hydrophobicity. The values of the phenotypic residues are highlighted by vertical dotted lines in red and the vertical solid line indicates the average value. All vertical lines have the same size (proportional to the plot height).

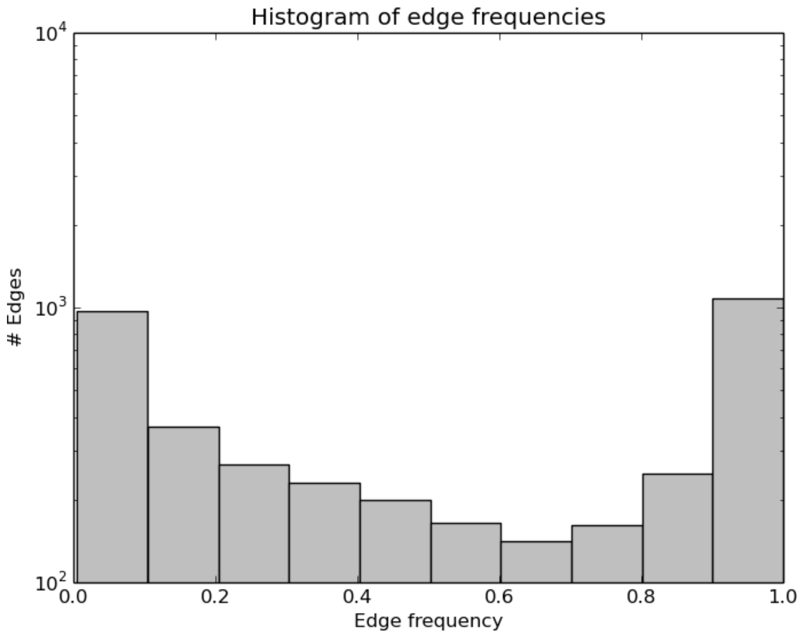


Figure B.8: Histogram of edge frequencies in the wild-type dRIN of DJ-1 (last 10 ns only).

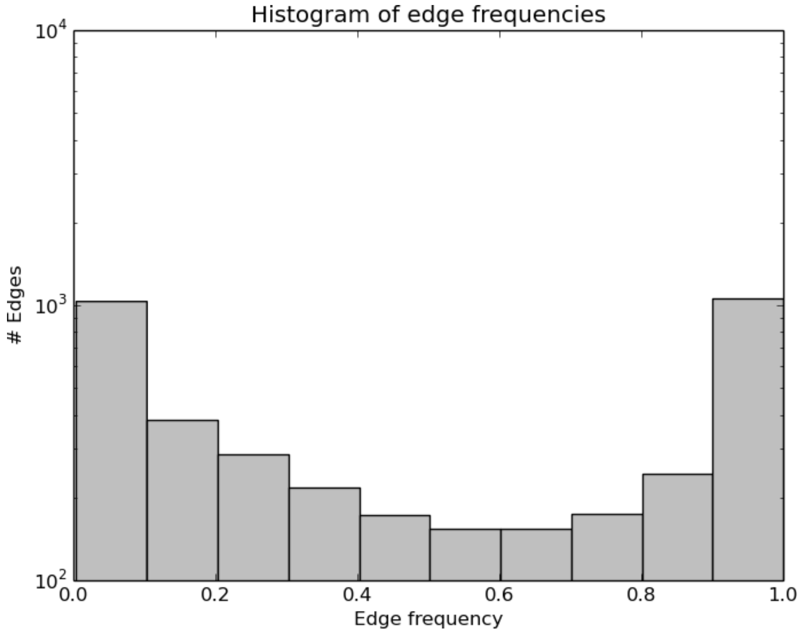


Figure B.9: Histogram of edge frequencies in the L166P mutant dRIN of DJ-1 (last 10 ns only).

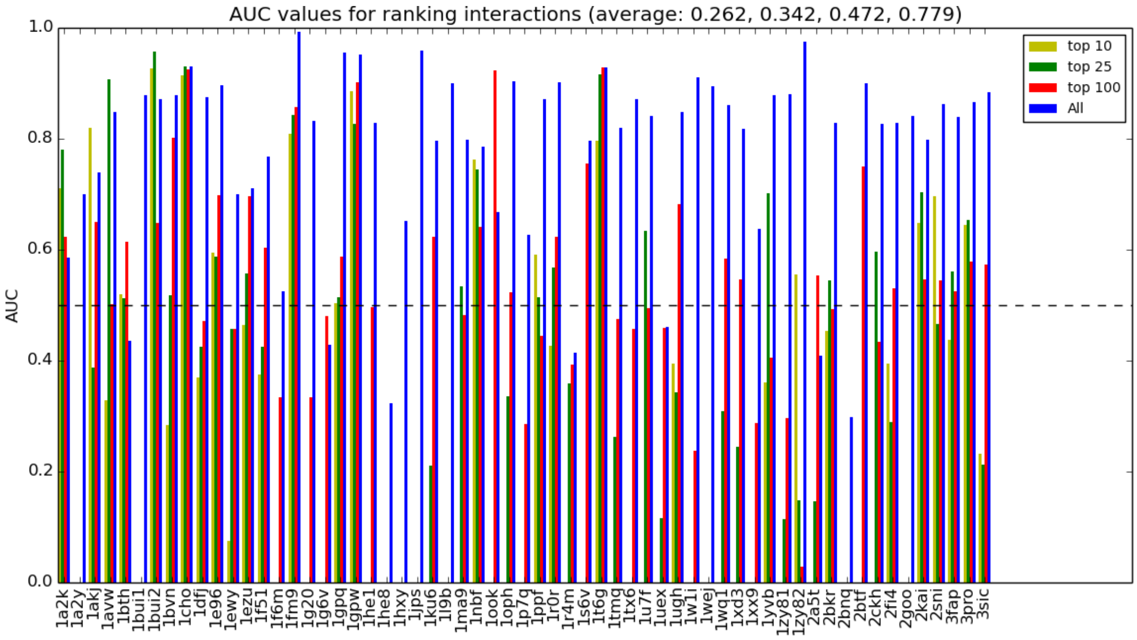


Figure B.10: Performance per structure for ranking interactions in the DOCKGROUND set using the *top 10, 25, 100* and *All* decoys.

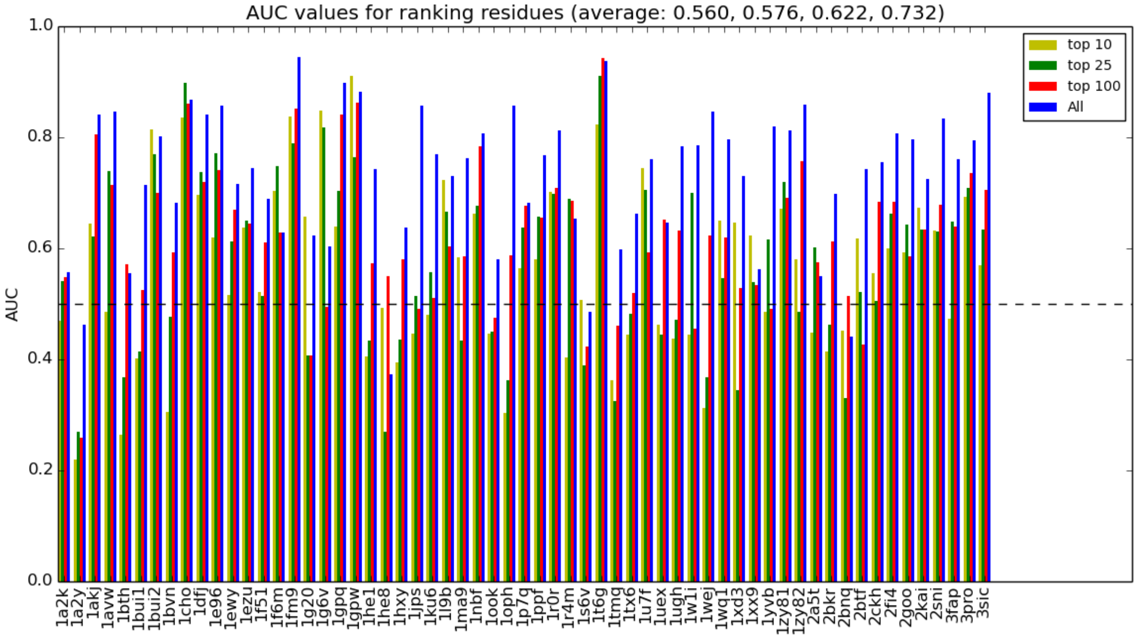


Figure B.11: Performance per structure for ranking residues in the DOCKGROUND set using the *top 10, 25, 100* and *All* decoys.

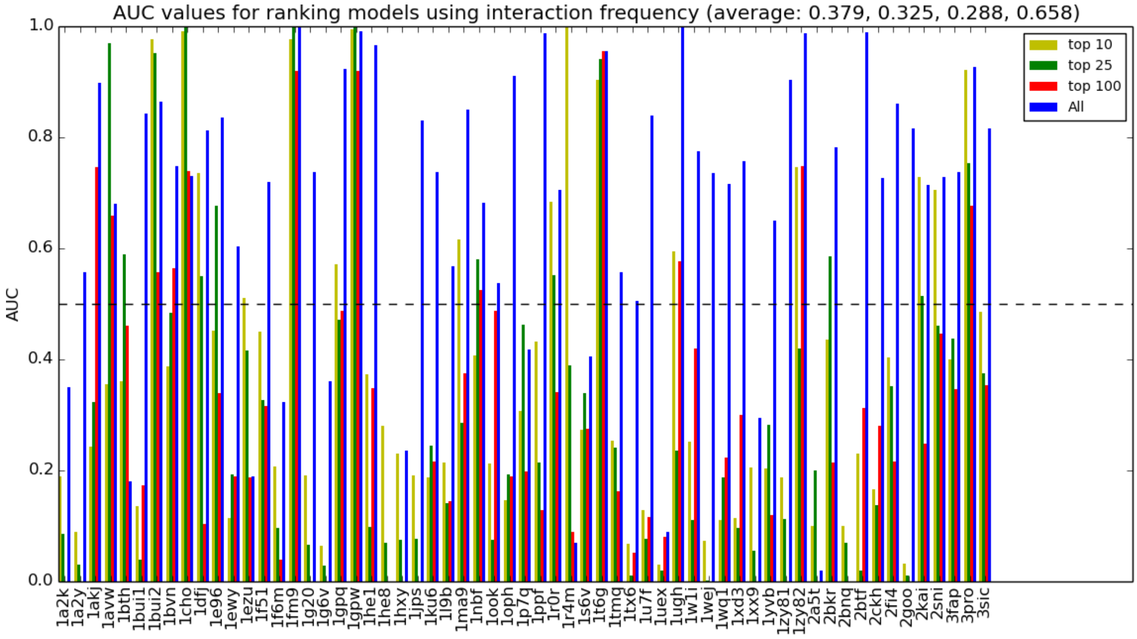
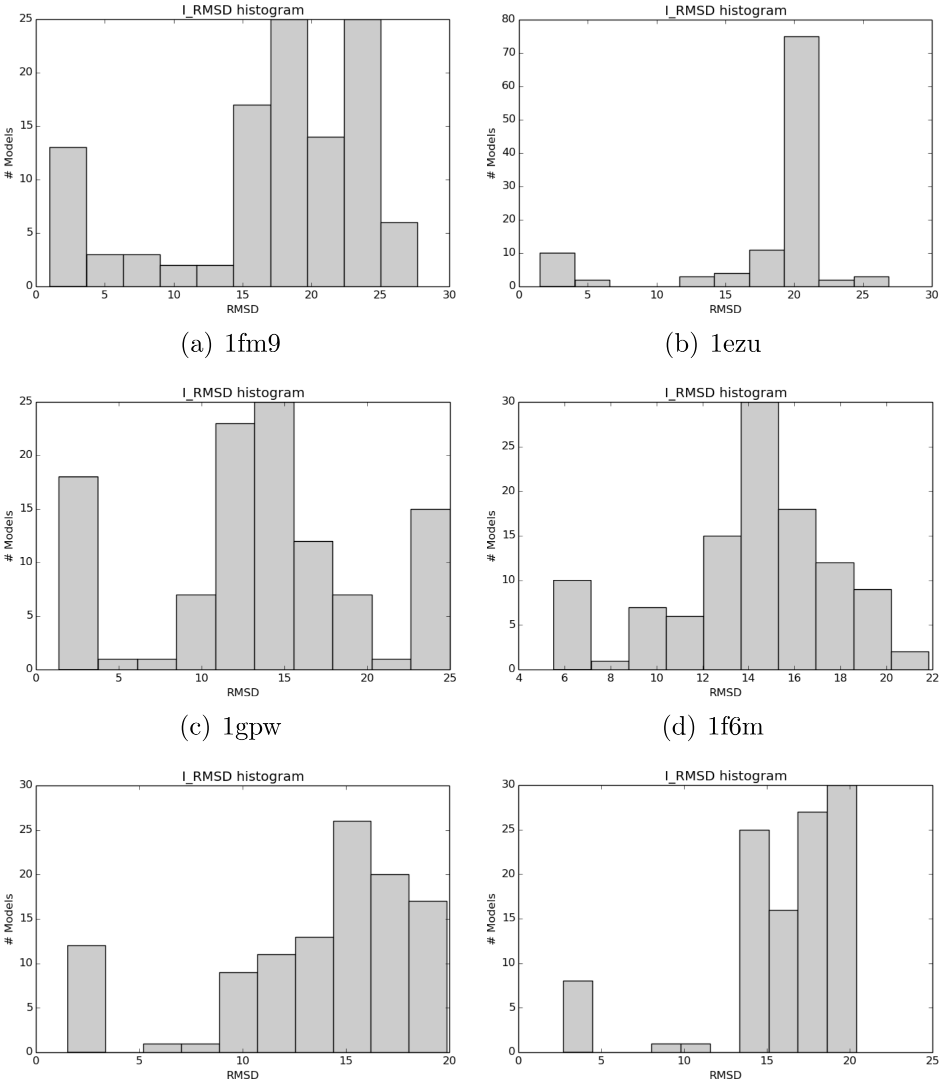


Figure B.12: Performance per structure for ranking models in the DOCKGROUND set using the *top 10, 25, 100* and *All* decoys.



(a) 1fm9

(b) 1ezu

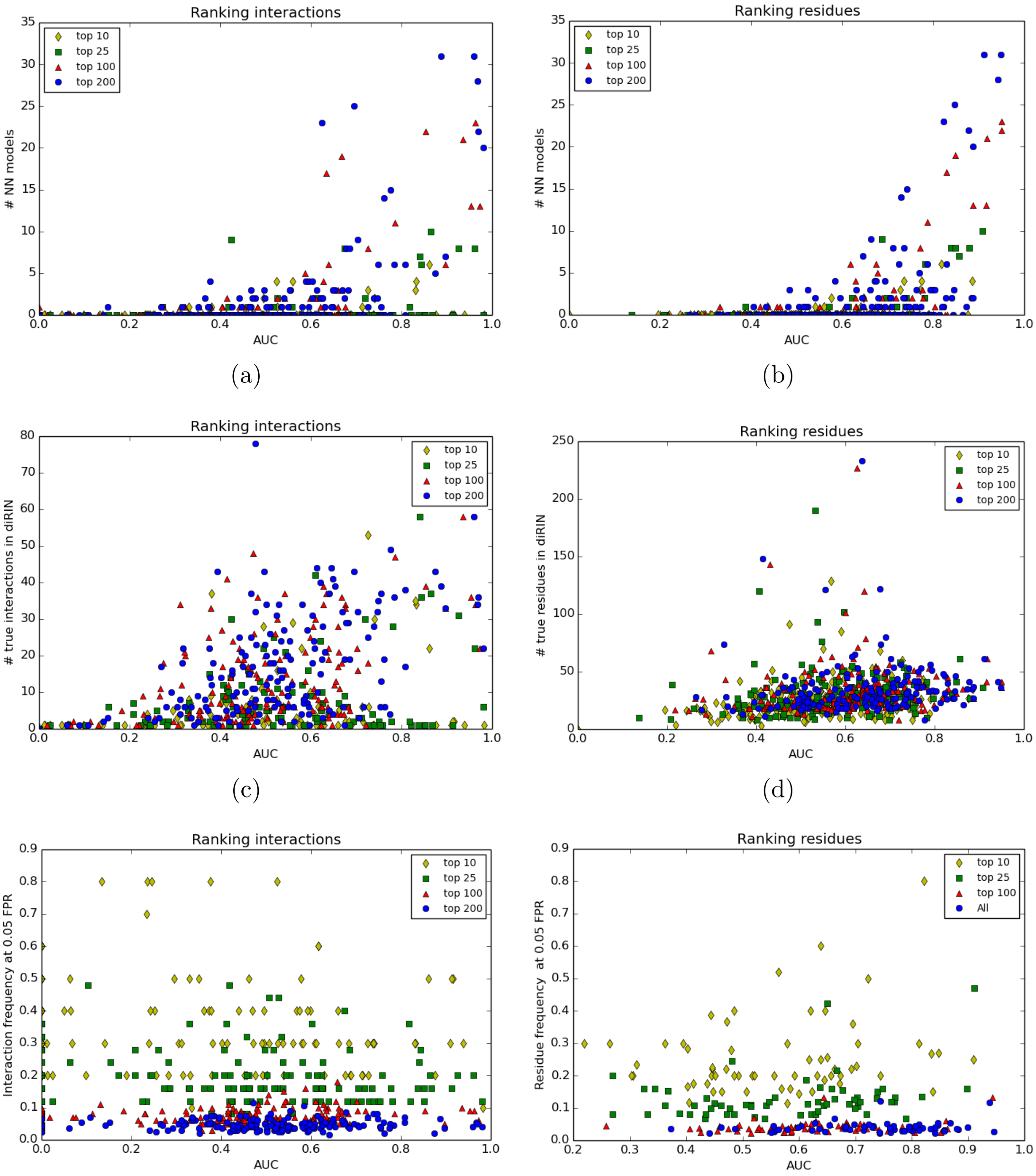
(c) 1gpw

(d) 1f6m

(e) 1ugh

(f) 1g6v

Figure B.13: I_{RMSD} distribution for selected DOCKGROUND targets.



(a) (b)

(c) (d)

(e) (f)

Figure B.14: Dependencies between performance for ranking interactions or residues and the number of NN models per target (a) and (b), the number of true (target) interface interactions (c) and residues (d) in the diRIN, and the interaction (e) and residue (f) frequency scores at 0.05 FPR (top 5 %) in the PatchDock set.

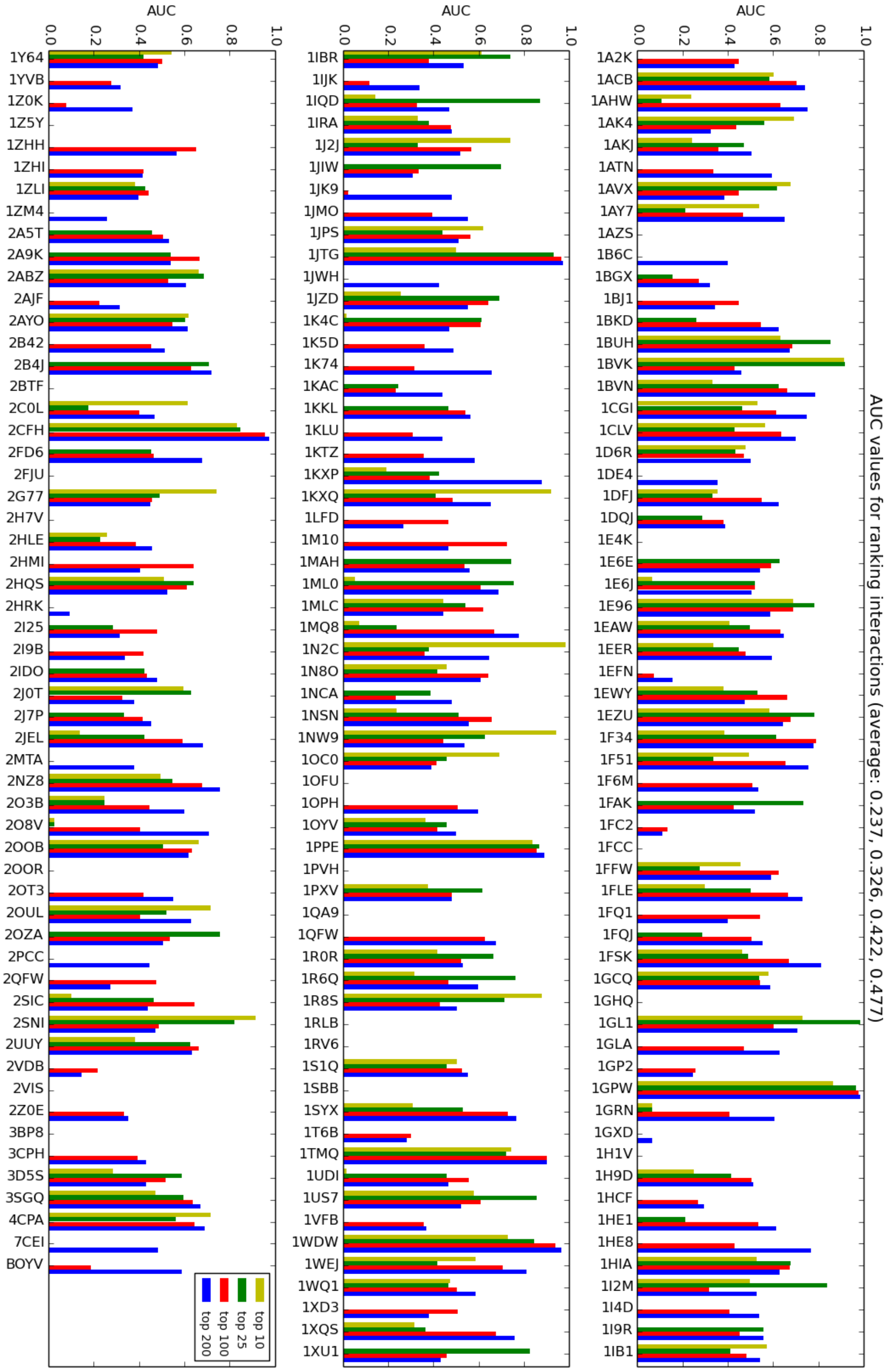


Figure B.15: Performance per structure for ranking interactions in the PatchDock set using the *top 10, 25, 100, 200* decoys.

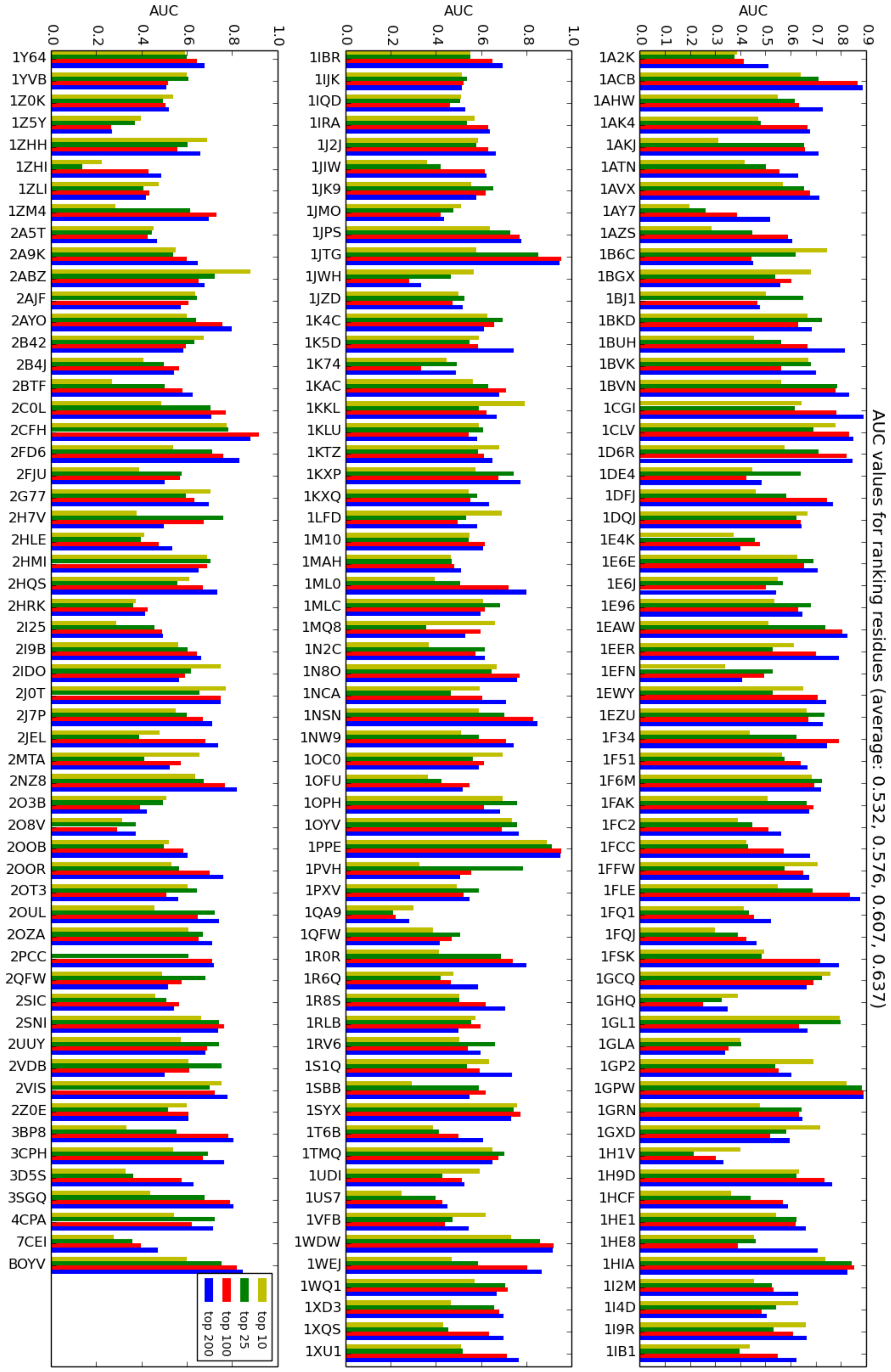


Figure B.16: Performance per structure for ranking residues in the PatchDock set using the *top 10, 25, 100, 200* decoys.

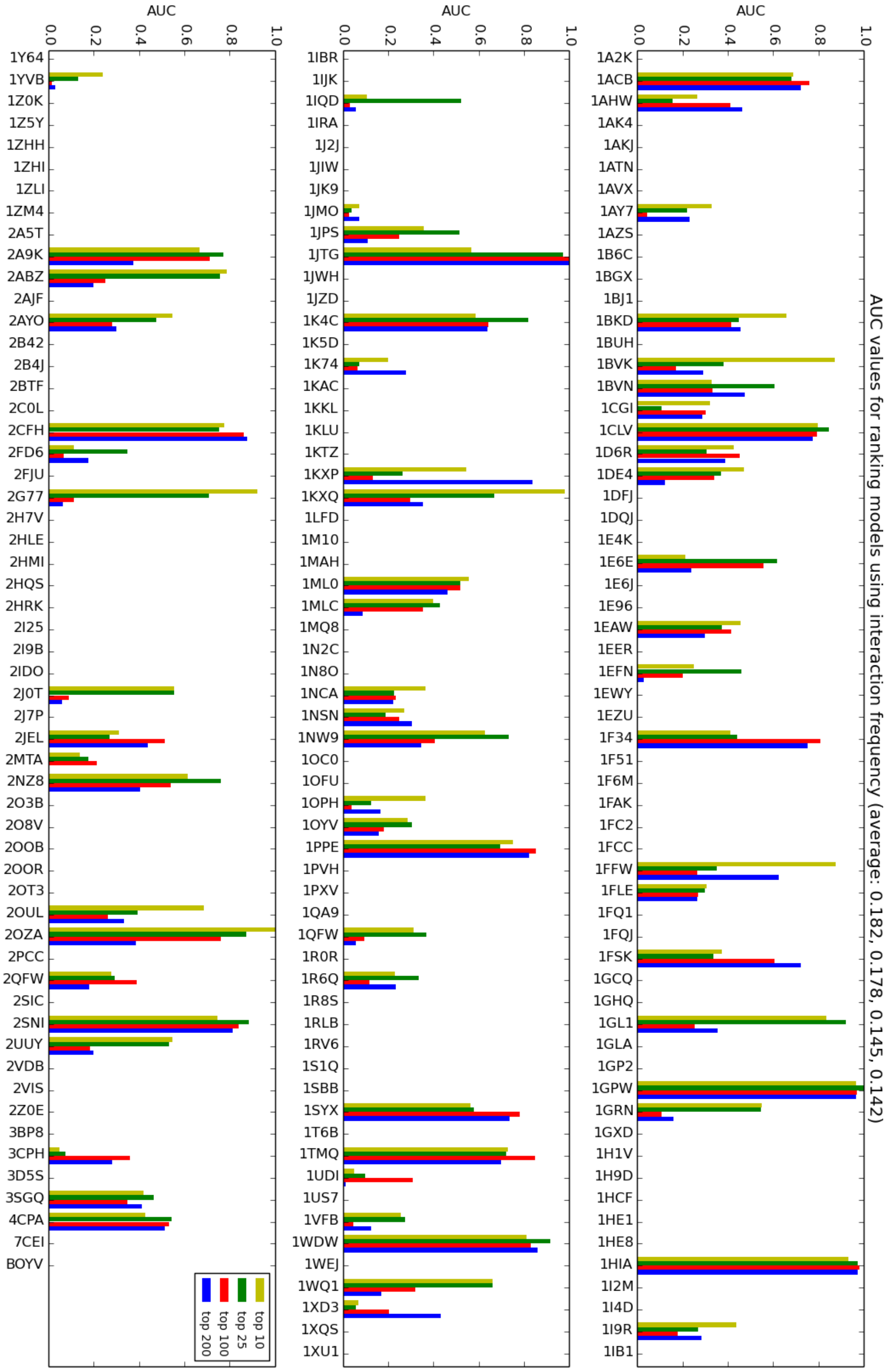


Figure B.17: Performance per structure for ranking models in the PatchDock set using the *top 10, 25, 100, 200* decoys.

APPENDIX C

List of own publications

1. **Doncheva, N.T.**, Klein, K., Domingues, F.S., Albrecht, M. (2011). Analyzing and visualizing residue networks of protein structures. *Trends in Biochemical Sciences*, 36(4):179-182, doi:10.1016/j.tibs.2011.01.002

2. **Doncheva, N.T.***, Kacprowski, T.*, Albrecht, M. (2012). Recent approaches to the prioritization of candidate disease genes. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4(5):429-442, doi:10.1002/wsbm.1177, (* equally contributing first authors)

3. **Doncheva, N.T.**, Assenov, Y., Domingues, F.S., Albrecht, M. (2012). Topological analysis and interactive visualization of biological networks and protein structures. *Nature Protocols*, 7:670-685, doi:10.1038/nprot.2012.004

4. Liu, J.Z.*, Hov, J.R.*, Folseraas, T*., Ellinghaus, E.*, Rushbrook, S.M., **Doncheva, N.T.**, Andreassen, O.A., Weersma, R.K., Weismüller, T.J., Eksteen, B., Invernizzi, P., Hirschfield, G.M., Gotthardt, D.N, Pares, A. *et al.* (2013). Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nature Genetics*, 45(6):670-677, doi:10.1038/ng.2616, (* equally contributing first authors)

5. Kacprowski, T., **Doncheva, N.T.**, Albrecht, M. (2013). NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, 29(11):1471-1473, doi:10.1093/bioinformatics/btt164

6. Zanon, A., Rakovic, A., Blankenburg, H., **Doncheva, N.T.**, Schwienbacher, C., Serafin, A., Alexa, A., Weichenberger, C.X., Albrecht, M., Klein, C. Hicks, A.A., Pramstaller, P.P., Domingues, F.S., Pichler, I. (2013). Profiling of parkin-binding partners using tandem affinity purification. *PLoS ONE*, 8(11): e78648, doi:10.1371/journal.pone.0078648

7. Ellinghaus, D., Zhang, H., Zeissig, S., Lipinski, S., Till, A., Jiang, T., Stade, B., Bromberg, Y., Ellinghaus, E., Keller, A., Rivas, M.A., Skieceviciene, J., **Doncheva, N.T.**, Liu, X. *et al.* (2013). Association between variants and Crohn's disease based on exome sequencing and functional studies. *Gastroenterology*, 145(2):339-347, doi:10.1053/j.gastro.2013.04.040

8. **Doncheva, N.T.**, Klein, K., Morris, J.H., Wybrow, M., Domingues, F.S., and Albrecht, M. (2014). Integrative visual analysis of protein sequence mutations. *BMC Proceedings*, 8(Suppl 2):S2, doi:10.1186/1753-6561-8-S2-S2

9. Morris, J.H., Lotia, S., Wu, A., **Doncheva, N.T.**, Albrecht, M., Ferrin, T.E. (2015). setsApp for Cytoscape: Set operations for Cytoscape Nodes and Edges. *F1000Research*, 3:149, doi:10.12688/f1000research.4392.2

10. Andreassen, O.A., Desikan, R.S., Wang, Y., Thompson, W.K., Schork, A.J., Zuber, V., **Doncheva, N.T.**, Ellinghaus, E., Albrecht, M., Mattingsdal, M., Franke, A., Lie, B.A., Mills, I., Aukrust, P. *et al.* (2015). Abundant genetic overlap between blood lipids and immune-mediated diseases indicates shared molecular genetic mechanisms. *PLoS ONE*, 10(4): e0123057, doi:10.1371/journal.pone.0123057