Network-based inference of protein function and disease-gene association

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Dr. rer. nat.
im Fach Informatik

eingereicht an der
Mathematisch-Naturwissenschaftliche Fakultät II
Humboldt-Universität zu Berlin

von
M.Sc. Samira Jaeger

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftliche Fakultät II:
Prof. Dr. Elmar Kulke

Gutachter:
1. Ulf Leser
2. Miguel Andrade-Navarro
3. Oliver Kohlbacher

eingereicht am: 09.09.2011
Tag der mündlichen Prüfung: 16.12.2011

Abstract

Protein interactions are essential to many aspects of cellular function. On the one hand, they reflect direct functional relationships, i.e., if two proteins interact with each other they are often involved in the same biological process or pathway. On the other hand, alterations in protein interactions, e.g., caused by mutations in their interfaces, perturb natural cellular processes and contribute to diseases. In this thesis we analyze both the functional and the pathological aspect of protein interactions to infer novel protein function for uncharacterized proteins and to associate yet uncharacterized proteins with disease phenotypes, respectively.

The first part of this thesis addresses the functional characterization of proteins. Knowing a protein's function is fundamental to understand the molecular and biochemical processes that sustain health or cause disease. Different experimental and computational approaches have been developed in the past to investigate the basic characteristics of proteins systematically. Yet, a substantial fraction of proteins remains uncharacterized, particularly in human. In this work, we present a novel approach to predict protein function from protein interaction networks of multiple species. The key to our method is to study proteins within modules defined by evolutionary conserved processes, combining comparative cross-species genomics with functional linkage in interaction networks. We show that integrating different evidence of functional similarity allows to infer novel functions with high precision and a very good coverage. For instance, when considering the combination of human, fly and yeast, we achieve a precision of 84% to 87%. Overall, our method generates novel functional knowledge for every species included in the analysis at varying, yet always high levels of precision. For human we predict 27,100 novel annotations with an estimated precision of 83%.

In the second part, we investigate the role of proteins in human diseases as for many genetic diseases it is not known which gene products are involved in their pathogenesis. Elucidating the underlying pathological mechanisms is important for understanding the onset of diseases and for developing diagnostic and therapeutic approaches. We introduce a network-based framework for identifying yet uncharacterized disease-related gene products by combining protein interaction data and protein function with network centrality analysis. Given a disease, we first extract all genes associated with this disease. We then compile a disease-specific network by integrating directly and indirectly linked gene products using protein interaction and functional information. Proteins in this network are ranked based on their network centrality. We demonstrate that using indirect interactions significantly improves disease gene identification, i.e., the cross-validation recovery rate increases by up to 20%. Predicted functions, in turn, enhance the ranking of disease-relevant proteins. However, the functional enrichment integrates many global "hub" proteins which feature a high centrality but are mostly unspecific for a disease. To adjust the ranking for a bias toward hub proteins in disease networks, we introduce a novel normalization procedure which decreases the fraction of highly ranked hub proteins (by 23%) while increasing the fraction highly ranked disease proteins at the same time (by 22%). Finally, we use our framework to detect novel surface membrane factors that are involved in a cascade of events contributing to HIV-1 infection. Their involvement includes serving as co-receptors for cell entry, mediating trans-infection or activating immune cells to inducing viral production from latently infected cells.

Zusammenfassung

Proteininteraktionen sind entscheidend für verschiedene Aspekte zellulärer Funktion. Interaktionen reflektieren einerseits direkte funktionale Beziehungen zwischen Proteinen, anderseits tragen Veränderungen in spezifischen Interaktionsmustern zur Entstehung von Krankheiten bei. In dieser Arbeit werden sowohl die funktionalen als auch die pathologischen Aspekte von Proteininteraktionen analysiert, um Funktionen für bisher nicht charakterisierte Proteine vorherzusagen und Proteine mit Krankheitsphänotypen zu assoziieren.

Der erste Teil der Arbeit befasst sich mit der funktionalen Charakterisierung von Proteinen. Die Funktionsweise von Proteinen ist von grundlegender Bedeutung, um die molekularen und biochemischen Prozesse, die Gesundheit oder Krankheiten verursachen, zu verstehen. Verschiedene experimentelle und computergestützte Methoden wurden in den letzten Jahren entwickelt, die die funktionalen Eigenschaften von Proteinen untersuchen. Dennoch bleibt ein wesentlicher Teil der Proteine, insbesondere menschliche, uncharakterisiert. In dieser Arbeit wird eine neue Methode zur Vorhersage von Proteinfunktionen vorgestellt, die auf Proteininteraktionsnetzwerken verschiedener Spezies beruht. Dieser Ansatz analysiert Proteine innerhalb von funktionalen Modulen, die über evolutionär konservierte Prozesse definiert werden. In konservierten funktionalen Modulen werden neue Proteinfunktionen gemeinsam über Orthologie-Beziehungen und Interaktionspartner vorhergesagt. Die Evaluierung dieser Methode zeigt, dass die Integration verschiedener funktionaler Ähnlichkeiten die Vorhersage von neuen Proteinfunktionen mit hoher Genauigkeit und sehr guter Abdeckung ermöglicht. Der Vergleich der Interaktionsnetzwerke von Mensch, Fliege und Hefe resultiert beispielsweise in einer Vorhersagegenauigkeit von 84% bis 87%. Insgesamt generiert unsere Methode neue funktionale Annotationen für verschiedene Spezies mit variierender aber hoher Präzision. Für den Menschen werden 27.100 neue Annotationen mit einer geschätzten Genauigkeit von 83% vorhergesagt.

Im zweiten Teil der Arbeit wird der Einfluss von Proteinen auf die Pathogenese menschlicher Krankheiten untersucht. Die Aufklärung der zugrunde liegenden Mechanismen ist wichtig, um die Entstehung von Krankheiten zu verstehen und diagnostische und therapeutische Ansätze zu entwickeln. Wir stellen einen netzwerkbasierten Ansatz für die Identifizierung krankheitsrelevanter Genprodukte vor, der auf der Kombination von Proteininteraktionsdaten, Proteinfunktionen und Netzwerk-Zentralitätsanalyse basiert. Gegeben eine Krankheit, werden zunächst alle Gene extrahiert, die bereits mit dieser Krankheit assoziiert sind. Anschließend werden krankheitsspezifische Netzwerke durch die Integration von direkt und indirekt interagierender Genprodukte und funktionalen Informationen generiert. Proteine in diesen Netzwerken werden dann anhand ihrer Zentralität sortiert. Es wird gezeigt, dass das Einbeziehen indirekter Interaktionen die Identifizierung von Krankheitsgenen deutlich (um bis zu 20%) verbessert. Die Verwendung von vorhergesagten Proteinfunktionen wiederum verbessert das Ranking von krankheitsrelevanten Proteinen. So konstruierte Netzwerke enthalten häufig globale Hub-Proteine, die eine hohe Zentralität aufweisen, jedoch unspezifisch für eine Krankheit sind. Aus diesem Grund wurde eine Methode zur Normalisierung des Rankings entwickelt, mit Hilfe derer der Anteil hoch gerankter Hub-Proteine um 23% reduziert wird und gleichzeitig der Anteil hoch gerankter Krankheitsproteine um 22% erhöht wird. Unsere Methode verwenden wir außerdem, um bisher unbekannte rezeptor-ähnliche Faktoren zu identifizieren, die maßgeblich an HIV-1 Infektion beteiligt sind.

Acknowledgements

This PhD thesis would not have been possible without the encouragement, assistance and support of many different people.

First of all, I would like to thank my supervisor Prof. Ulf Leser. I am very happy that I had the opportunity to do my PhD in his research group, Knowledge Management in Bioinformatics, at Humboldt-Universität zu Berlin. His encouraging and dedicated guidance during the last four years have provided a good basis for the present thesis. It has been a great pleasure to work with him.

Thanks to the Land Berlin and the Elsa-Neumann-grant, Deutsche Forschungsgemeinschaft (DFG) and EMBO for funding my research.

During the time I have been working at the Humboldt-Universität zu Berlin I have met many colleagues with whom I spent many pleasant moments. I would like to thank all current and past members of WBI who contributed to such an enjoyable working environment. I will miss coffee breaks, lunch-time discussions and cakes. Thanks to Karin, Astrid, Stefan, Philippe, Silke, Johannes, Andre, Illes and Björn for proof-reading parts of my thesis, for constructive criticism and helpful suggestions. Special thanks to the scientific quartet for stimulating discussions and Sebastian who worked with me quite some time. Thanks to Roger for providing me with cookies whenever I needed some.

I am grateful to many other people I met at conferences and research stays who gave me advice and motivation for my research. In particular, Gökhan Ertaylan and David van Dijk from University of Amsterdam with whom I worked together in a joint project for a while.

Mein besonderer Dank gilt all den Menschen, die sich bisher weniger für die “Proteine und Krankheiten dieser Welt” begeistern konnten, die dennoch immer ein offenes Ohr hatten und mich auf andere Art und Weise während meiner Doktorandenzeit begleitet und unterstützt haben. Besonderer Dank gilt meiner Familie, insbesondere meiner Großmutter, die maßgeblich für meinen Werdegang in der Bioinformatik verantwortlich ist. Besten Dank auch an meine Freunde, in der Nähe und in der Ferne.

Contents

1 Introduction

The past decade has seen a revolution in genomic sequencing technologies, in particular, so-called next-generation sequencing delivers fast and accurate data about genome and more recently also about metagenomic projects (Schloss and Handelsman, 2005; Metzker, 2010). As of March 2011, 1609 bacterial, 85 archaeal, and 299 eukaryotic genomes have been completely sequenced[1], while several other genomes are just about to be finished. Transferring this wealth of data into biological knowledge is a fundamental challenge in the post-genomic era.

The completion of a new genome is commonly followed by a process known as genome annotation to predict, among others, its protein coding regions and to associate biological information to them (Stein, 2001). Elucidating the functional role of each individual gene product in development, physiology, and pathology is one of the major challenges in molecular biology and bioinformatics. It is fundamental to understand biological processes, cellular mechanisms, evolutionary changes and the onset of diseases (Eisenberg *et al.*, 2000; Frishman, 2007).

Traditionally, biochemical experiments, such as functional assays, knock-out experiments or targeted mutations, have been used to determine the biological function of single genes (Whisstock and Lesk, 2003). These *in vivo* approaches, largely based on the one-gene one-function concept (Vidal and Furlong, 2004), led to impressive discoveries. For instance, knock-out studies in mice advanced molecular biology in particular by enhancing the understanding of genes in higher organisms (Kühn *et al.*, 1995). Comparative genomics allows for transferring functional knowledge derived from such experiments to those human genes which are direct counterparts of the investigated genes in mouse (Pennacchio, 2003). Other model organisms, such as yeast and fly, are also widely used for studying biological phenomena in species that are more difficult to analyze directly.

Despite of technical advances in so-called high-throughput methods, such as DNA microarrays (Schena *et al.*, 1995), RNA interference (RNAi) (Kamath and Ahringer, 2003) and large-scale systematic deletions (Que and Winzeler, 2002), many fundamental biological questions remain unanswered for several reasons. First, experimental characterization of proteins cannot keep up with the pace at which sequence data is produced (Frishman, 2007). Second, even detailed biochemical studies often cannot determine the full repertoire of biochemical activities within cells (Whisstock and Lesk, 2003). Third, conclusions from *in vitro* experiments might be limited as particularly eukaryotic proteins cannot be investigated in conditions close to their natural environment. Thus, even for well-known model organisms, such as yeast, a substantial fraction of proteins remains

[1] `http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi`

functionally uncharacterized (Sharan *et al.*, 2007).

An important aspect of proteins is their role in human diseases (Goh *et al.*, 2007). For many human diseases it is not yet known which genes are involved in their pathogenesis. As of May 2011, more than 7,000 Mendelian disorders are documented in OMIM (McKusick, 2007). However, for approximately 4,000 of them the molecular basis is still unknown. Elucidating the underlying pathological mechanisms is crucial for understanding the onset of diseases and for the development of specific diagnostic and therapeutic approaches. Traditional gene-mapping approaches, such as linkage analysis and association studies, are able to associate chromosomal regions, so-called linkage intervals, with a disease (Botstein and Risch, 2003). Yet, knowing the genomic region is often not sufficient to detect the associated gene(s). These regions are often large, typically comprising several megabases (Jorde, 2000). Investigating all candidates in the intervals experimentally is time-consuming and expensive. Furthermore, many genetic diseases are rare, which leads to a lack of samples and thus makes robust association studies impossible. The discovery process is even more complicated for diseases without confirmed or with multiple associated genomic regions. Finally, pleiotropy of genes (i.e. the ability of some genes to produce multiple phenotypes) and the heterogeneity of multifactorial diseases pose limitations to traditional gene-mapping approaches (Giallourakis *et al.*, 2005). For instance, type II diabetes (T2D), characterized by insulin resistance and dysfunction of β-cells, is a common multifactorial disease in which genetic alterations as well as environmental factors contribute to the onset of the disease (Stumvoll *et al.*, 2005). To date, more than 40 loci have been confidently associated with T2D (McCarthy, 2011), but the individual genes that mediate susceptibility to T2D have yet to be determined (Voight *et al.*, 2010).

Cellular function but also malfunctioning of proteins mostly emerges from the complex molecular interplay between proteins, metabolites, functional RNAs and genes (Barabási *et al.*, 2011). For instance, the tumor suppressor protein p53 mediates its natural function, namely cell cycle regulation, through several target proteins (Vogelstein *et al.*, 2000). Protein p53 is activated upon intra- and extracellular stimuli, such as DNA damage, activated oncogenes or oxidative stress. The activation induces the transcription of p53-regulated genes, e.g., p21 or Bax, through which cell cycle arrest, cellular senescence, apoptosis and DNA repair are mediated, depending on the physiological circumstances and cell type (see Figure 1.1). Mutations in p53 disrupt the complex network of stress response pathways leading to uncontrolled proliferation of damaged cells and eventually to various types of cancer (Hollstein *et al.*, 1991). This emphasizes that the relationship between genotypes and phenotypes is mostly determined by complex mechanisms which cannot be discerned by studying the respective gene in isolation. Albeit the function of a single gene might present a molecular description of cellular phenotypes, it is often not sufficient to provide mechanistic explanations on the particular process. The question of how a single genotype gives rise to distinct phenotypes remains a major challenge since Mendel's wrinkled peas (Mendel, 1866) and Morgan's white-eyed fruit flies (Morgan, 1910).

To understand the relationships between genotype, environment and phenotype, one has to consider the complex and nonsequential interaction patterns formed between

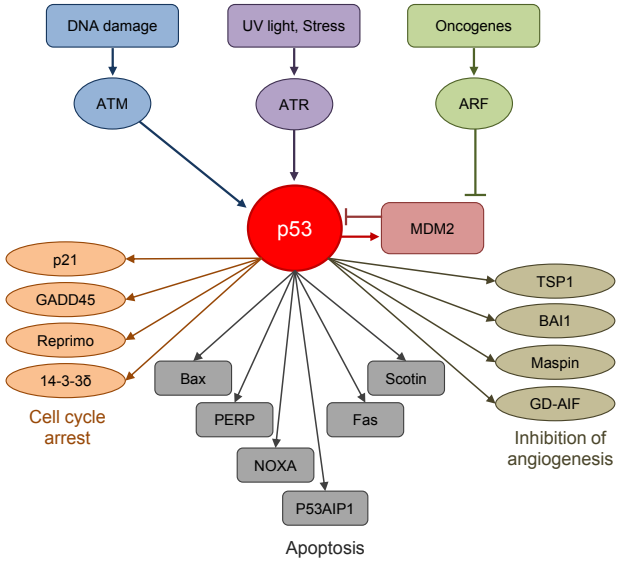


Figure 1.1: The p53 network. p53 is the central component within the complex network of stress response pathways (adapted from Vogelstein *et al.* (2000)). The activation of the network upon DNA damage, stress or activated oncogenes induces the modification of p53 and its negative regulator MDM2. Activated p53 initiates the expression of several target genes, such as p21, Bax or Fas, to mediate various functions including cell cycle arrest, DNA repair, apoptosis, and senescence.

the different sets of cellular entities. Advanced experimental techniques, such as DNA and protein microarrays, high-throughput localization studies and protein interaction mapping approaches, assist in determining how and when these molecules interact with each other. Several types of interaction networks, such as metabolic, signaling, protein interaction, and transcription-regulatory networks, emerge from the variety of these interactions (Barabási and Oltvai, 2004). Systematic studies of these networks for elucidating their basic function, structure and dynamics have become one of the key topics in systems biology and bioinformatics (Zhang, 2009).

In this work, we analyze cellular function in both physiological and pathological contexts by using one of the most commonly studied types of biological networks, i.e., protein-protein interaction networks. Protein interaction networks represent proteins that interact physically with each other. Such interactions are crucial to many aspects of cellular function, such as signal transduction, gene regulation, cell cycle control and metabolism (Piehler, 2005). Numerous experimental techniques have been developed for detecting protein interactions and their characteristics, both in small- and in large-scale (Phizicky and Fields, 1995).

Despite being still incomplete and error-prone, protein interaction networks have become particularly important for functional analysis, especially in human. On the one hand, protein interactions are direct and robust manifestations of functional relationships, i.e., if two proteins interact with each other they are likely to be involved in the same biological process or pathway (Sharan *et al.*, 2007). On the other hand, alterations in protein interactions disturb cellular processes and contribute to many diseases,

such as cancer (Ideker and Sharan, 2008). Mutations in protein interaction interfaces are often associated with loss of function or gain of function (Schuster-Böckler and Bateman, 2008). For instance, the cancer-predisposing mutation Tyr42Cys in *BRCA2* compromises its interaction with replication protein A, a protein involved in DNA repair, replication and recombination (Wong *et al.*, 2003). A lack of this interaction is presumed to inhibit the recruitment of double stranded break repair proteins and eventually leads to an accumulation of carcinogenic DNA changes.

Both the functional and the pathological aspect of protein interaction networks will be considered in this work to derive novel protein function for uncharacterized proteins and to associate yet uncharacterized proteins with disease phenotypes, respectively. The specific contributions to both problems are outlined in the following section.

1.1 Contribution

The central theme of this dissertation is the study of protein interaction networks with respect to two closely related problems: (1) protein function prediction and (2) inference of disease-gene associations.

As the first main contribution we present a method for predicting protein function from protein interaction networks. The proposed approach compares protein interaction networks across multiple species to detect evolutionarily and functionally conserved subgraphs, so-called conserved and connected subgraphs (CCS). Within each CCS we infer novel protein functions from orthology relationships across species and along conserved interactions of neighboring proteins within a species. Specific contributions to the objective of protein function prediction are summarized below:

- We develop a framework for integrating various small- and large-scale protein interaction data sets from six public databases into a meta-database called PiPa. This framework allows to combine heterogeneous data sets to provide comprehensive protein interaction networks as basis for this thesis.
- We introduce the idea of identifying functional modules in protein interaction networks by exploiting subgraphs that are evolutionary conserved across multiple species.
- As protein interaction data are known to be inherently noisy and incomplete, we implement a strategy to account for data quality as well as evolutionary variation by using two different definitions for identifying conserved interactions: a strict and a relaxed definition.
- We eventually integrate three different sources of evidences, namely evolutionary conservation of functional modules, orthology relationships, and direct and indirect protein-protein interactions into a single, comprehensive prediction method which yields high-quality predictions with very good coverage.

In comparison to three related methods, CCS-based function prediction clearly outperforms Neighbor Counting and χ^2. A comparable or even better performance is achieved

when comparing against FS-Weighted Averaging. We further contribute to the field of protein function prediction with a comprehensive survey on the different methodologies for protein function prediction, providing insights on current progress and limitations.

As second main contribution we present a linkage interval-independent, network-based algorithm to identify disease-related genes. We introduce a network biology framework that integrates protein interaction, protein function, and network centrality analysis. To detect disease-related genes with a particular disease, we first extract all genes that are known to be involved in this disease. We compile a disease-specific network by integrating directly and indirectly linked gene products based on protein-protein interaction and functional information. Proteins in this network are ranked based on their network centrality. Specific contributions to disease gene identification are summarized below:

- In our approach, we integrate genes indirectly linked to other disease genes. Thus, we uncover susceptibility genes that are not directly linked but that are part of the same pathway. This leads to more comprehensive disease networks and significantly increases cross-validation re-discovery rates by up to 20%.
- The extension by indirect interaction partners might lower the precision since larger networks naturally integrate many global "hub" proteins that get high centrality ranks but are mostly unspecific for a particular disease. To cater for this effect, we introduce a novel normalization procedure. Adjusting the centrality scores decreases the fraction of highly ranked hub proteins (by 23%) while increasing the fraction of highly ranked disease proteins at the same time (by 22%).
- In contrast to previous approaches, we also include predicted functional information to address the problem that yet uncharacterized proteins can neither be captured nor sensibly ranked by previous methods, which in turn prevents the detection of truly novel disease-gene associations.

In contrast to most previous works, our algorithm is particularly applicable for complex diseases without associated or with multiple causative genomic regions. Furthermore, the benchmark comparison with two state-of-the-art approaches demonstrates that our disease-specific framework significantly outperforms PRINCE (Vanunu *et al.*, 2010). In comparison to RWR (Köhler *et al.*, 2008), we achieve comparable results.

Another important feature of our method is its generality. Albeit we developed the framework for finding novel genes/proteins associated with genetic disorders, it can be used to address various biological questions, e.g., detecting further members of cellular processes, pathways or other definable mechanisms. In a comprehensive case study, we apply our framework successfully to identify novel surface membrane factors that contribute to HIV-1 infection.

1.2 Outline of this Thesis

Chapter 2 provides background information relevant throughout this work. We briefly review proteins, their basic biochemistry as well as their role in human diseases. Next, we introduce properties and types of physical protein-protein interactions, and give an

overview on protein interaction networks including their properties and significance for bioinformatics and experimental research.

Chapter 3 presents a comprehensive overview on protein function prediction; starting with a general introduction to protein function, followed by a survey on computational approaches for protein function prediction.

Chapter 4 describes our novel approach for protein function prediction, namely CCS-based function prediction, that combines link-based and module-based prediction with orthology. We depict an algorithm to analyze proteins within modules that are defined by evolutionary conserved processes. We also discuss related work on network-based protein function prediction.

Chapter 5 presents the systematic evaluation of the proposed protein function prediction method. We apply our strategy to different sets of species and use leave-one-out cross-validation to assess its performance in terms of precision and recall. We consider different evaluation settings and discuss inherent properties of our method. In addition, we benchmark our approach against two baselines and three related prediction methods.

Chapter 6 first gives a short introduction into the field of disease-gene association. We review the broad range of methods available for disease gene identification. In the main part of the chapter, we present our network-based approach for identifying disease-causing proteins in a genome-wide setting. The chapter is completed with a thorough survey of related work along with a discussion of the various methodological differences.

Chapter 7 presents the systematic evaluation of the developed disease gene identification approach. We verify whether disease proteins are central in their disease-specific networks and study the ability of our method to identify novel disease-related protein using leave-one-out cross-validation across all known disease proteins. We compare the performance of our method with two other published methods. In addition, we apply our method in case studies to elucidate genes associated with two types of cancer, namely classical Hodgkin Lymphoma and colorectal cancer, as well as to identify surface membrane factors contributing to HIV-1 infection.

Chapter 8 summarizes the thesis, its main contributions and concludes with an outlook to future work.

Appendix A provides information on databases and terminologies that are used in the experiments throughout this work.

Appendix B provides additional results discussed in the main part of the thesis.

1.3 Own prior work and contributions

Chapter 4 of this thesis describes the function prediction approach initially proposed in Jaeger and Leser (2007) and further extended in Jaeger *et al.* (2010a). Chapter 5 presents the evaluation of this method and is mainly based on Jaeger *et al.* (2010a). The contributions described in both chapters can be attributed to the authors as follows: Leser conceived and supervised the project. Jaeger proposed and implemented the distinct methods for identifying conserved protein interaction subgraphs and for predicting protein functions. All evaluations were performed by Jaeger. Sers assessed the

manual verification of function predictions in context of colorectal cancer described in Section 5.5. Leser, Jaeger and Sers contributed to the manuscript.

Chapter 6 presents the framework for genome-wide disease gene identification which has been applied by Jaeger *et al.* (2010b) for finding novel surface membrane factors of HIV-1 infection as described in Section 7.5 of Chapter 7. Experiments in this study were conceived and designed by Jaeger, Ertaylan and van Dijk. The respective data were analyzed by Jaeger, Ertaylan and van Dijk. All experiments were performed by Jaeger and Ertaylan, and both authors wrote the manuscript. Leser and Sloot critically revised the manuscript and supervised the work.

2 Biological Background

This chapter provides background information on proteins, protein-protein interactions and protein interaction networks relevant throughout this work.

Section 2.1 briefly reviews the history of protein research and introduces their basic biochemistry including structural and functional characteristics. We discuss the role of proteins in human diseases regarding alterations which impact their natural function and which may lead to cell malfunction and, eventually, to a disease.

Section 2.2 focuses one of the most important types of biomolecular relationships among proteins: protein-protein interactions. We introduce specific properties of different types of protein-protein interactions, and summarize the techniques that can be used to identify them experimentally. Furthermore, we discuss strengths and limitations of the individual techniques that are reflected in the resulting quality and coverage of the data. We complete this section with a survey on common protein interaction databases.

Section 2.3 discusses protein interaction networks. We give an overview on their properties and their significance for bioinformatics and experimental research.

2.1 Proteins

2.1.1 Historical background

The true nature of proteins and the origin of their basic biochemistry had not been understood until the late 18^{th} century when proteins were recognized as a distinct type of biological molecule. Systematic protein research started in the early 19^{th} century by studying their chemical composition. In 1838, Gerhard J. Mulder firstly described the chemical composition of the nitrogen-containing substances fibrin, white of egg, blood serum and wheat albumin (Tanford and Reynolds, 2001). Mulder hypothesized from his experiments that proteins are composed from one fundamental substance (*Grundstoff*). Based on this assumption Jöns J. Berzelius proposed the term 'protein', derived from the Greek word $\pi\rho\omega\tau\epsilon\iota o\varsigma$ (*proteios*) meaning 'primary', 'in the lead' or 'standing in front', to describe this type of molecule.

The central role of proteins in living organisms was only fully acknowledged in 1926 when James B. Summer demonstrated that the enzyme urease is a protein (Sumner, 1926), a controversial assumption at the time (Perrett, 2007). Ever since then, proteins have been subjects of experimental studies in molecular biology. Sequencing the B chain of insulin (Sanger and Tuppy, 1951b,a) and elucidating the structures of myoglobin (Kendrew *et al.*, 1958) and hemoglobin (Perutz, 1960) at atomic resolution led to the modern age of protein research.

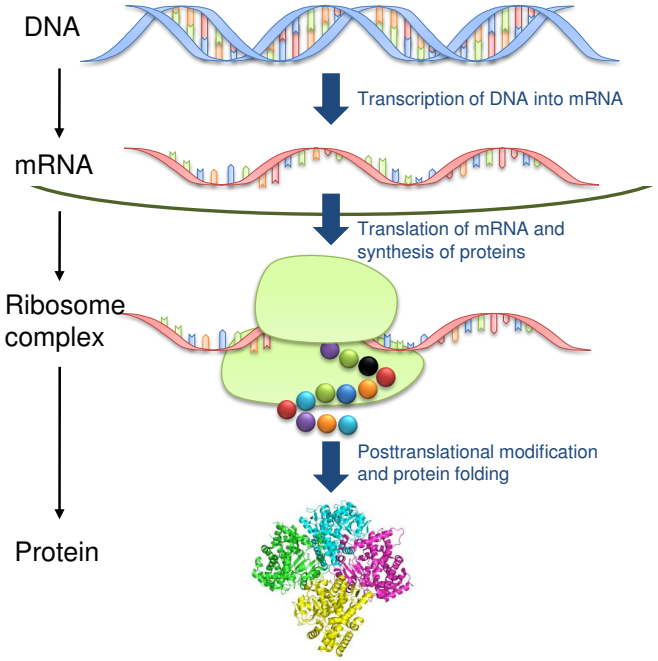


Figure 2.1: Basic principles of protein biosynthesis. The central dogma of molecular biology describes the conversion of a gene to protein via the transcription and translation phases.

Nowadays proteins are known to be one of the most important macromolecules in living organisms. They form the basic modules of cells and participate in virtually all cellular processes. Proteins are amazingly versatile molecules, capable of catalyzing an extraordinary range of biochemical reactions, functioning as antibodies in the immune system, providing structural stability to the cell, actively transporting molecules, controlling cell growth and differentiation, and regulating gene function (Lodish *et al.*, 2007). Although this tremendous functional scope is common knowledge these days, it has taken over 200 years and numerous controversial discussions, disputes and advanced technologies to move from the concept of a single unique 'protein' to our present understanding of thousands of distinct proteins in an organism (Perrett, 2007).

2.1.2 Protein composition and structure

Proteins are macromolecules that are manufactured by transcribing their coding genes into mRNA, which is then translated into a polypeptide, as shown in Figure 2.1.

The main building blocks of proteins are amino acids whose linear arrangement is defined by the nucleotide sequences of the genes encoding a protein. There are 22 proteinogenic amino acids that can be incorporated into proteins. Twenty of them are directly encoded in the universal genetic code whereas two, selenocysteine and pyrrolysine, originate from unique synthesis mechanisms (Ambrogelly *et al.*, 2007). All amino acids can be found in all eukaryotes, except for pyrrolysine which is currently only known

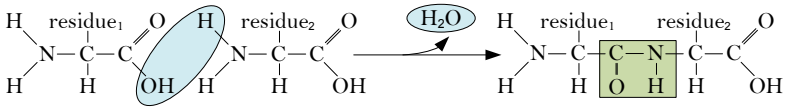


Figure 2.2: Formation of a dipeptide from two amino acids. Amino acids are linked to each other by a peptide bond that is formed through a condensation reaction that includes the removal of a water molecule.

for certain species of methanogenic archaea and one bacterium.

Amino acids have a common basic structure. They contain an amino group (NH_2), a carboxyl group (COOH) and a variable but specific residue (or side chain). The residues differ in properties such as size, form, charge, hydropathy and chemical reactivity, giving each amino acid its distinct biochemical characteristics. Individual amino acids are linked by peptide bonds (see Figure 2.2) to form one or more linear polypeptide chains which in turn constitute the backbone of a protein. The specific combination of residues and their distinct biochemical properties characterize the structure and function of each protein while the exponential number of combinations of amino acids accounts for the vast functional diversity of the proteins.

Proteins have highly variable sequence lengths and molecular weights. This variety partly reflects the diversity of the functional roles for proteins within different organisms (Lipman *et al.*, 2002). Proteins in prokaryotes, for instance, have on average shorter sequences than proteins in eukaryotes (Galperin *et al.*, 1999) reflecting the greater complexity of eukaryotic cells, e.g., multi-domain and multifunctional units (Brocchieri and Karlin, 2005).

However, a minimal number of amino acids is necessary to form a functional protein that fulfills its designated biological functions. Approximately 40 to 50 residues are thought to be the lower limit for a functional domain. Protein sequences range from this lower limit up to several thousands of residues in multifunctional and structural proteins. The median protein length in human measures around 375 residues (Brocchieri and Karlin, 2005) whereas the largest known human protein, Titin, a component of the contractile apparatus in muscle cells, consists of 34,350 amino acids[2] and 350 protein domains.

Protein structure

Proteins fold into specific spatial conformations. The folding of the linear strand of amino acids into a fully functional protein is one of the most complex challenges within the cellular protein factory and crucial for the functionality of each protein. An unique protein conformation arises from non-covalent interactions, such as hydrogen bonding, ionic interactions, Van Der Waals forces, and hydrophobic packing, between the amino acids of a sequence (Lodish *et al.*, 2007). The structural organization of a protein is commonly described on four different hierarchical levels:

[2] `http://www.uniprot.org/uniprot/Q8WZ42`

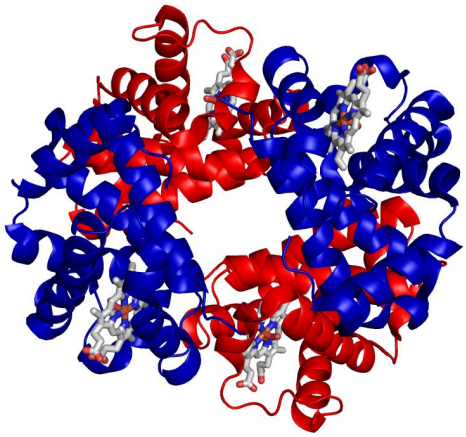


Figure 2.3: Quaternary structure of the human hemoglobin A. The model shows the assembly of the two α (red) and the two β (blue) subunits into a functional complex together with the iron-containing heme groups (illustrated with POLYVIEW-3D, Porollo *et al.* (2004)).

- The linear arrangement, or sequence, of amino acids in a polypeptide chain constitutes the *primary structure* of a protein.
- The *secondary structure* refers to intra- and intermolecular hydrogen bondings between amino acids of the linear sequence. Common secondary structures include α-helix, β-sheet, β-turn and random-coil structures which might occur separately or jointly within a protein.
- The *tertiary structure* describes the stable spatial conformation of local secondary structures and non-covalent interactions between specific amino acid residues. The tertiary structure presents the highest level of structural organization.
- Proteins with more than one polypeptide chain are only functional if their different subunits assemble to a larger complex. Depending on the protein, subunits might be identical, homologous (with similar functions) or completely distinct contributing to disparate tasks. The *quaternary structure* defines the spatial conformation of the distinct non-covalently linked subunits within such a multimer. Figure 2.3 shows the tertiary and quaternary structure of the human hemoglobin A which is assembled from two α- and the two β-globins. Other classical examples with a quaternary structure are actin, immunoglobin, ribosome and proteasome.

2.1.3 Protein function and their role in diseases

Protein structure and function are intrinsically tied to each other as a protein's function is largely determined by its three-dimensional conformation. Functionally, proteins are versatile macromolecules that evolved to carry out a wide range of functions (Lodish *et al.*, 2007). According to their different cellular roles, proteins can be classified into distinct functional classes:

- *Enzymes* present the largest class of proteins. They catalyze and accelerate the

rates of biochemical reactions that take place in a cell. Enzymes are typically named based on the reaction they facilitate. For instance, the enzyme tripeptide aminopeptidase is a hydrolase that cleaves off the amino-terminal amino acid from a polypeptide.

- *Regulatory proteins* or messenger proteins regulate the ability of other proteins to perform their biological functions. They transmit signals to coordinate biological processes between different cells, tissues, and organs. A classical regulatory protein is insulin – a hormone that regulates the glucose metabolism.
- *Transport proteins* serve as carriers that bind and transfer small molecules within cells and throughout the organism. Two different types of transport proteins can be distinguished: (i) those that transport molecules within cells or organisms, such as hemoglobin that transports oxygen from lungs to tissues, and (ii) membrane-bound proteins that serve as gateways for shuttling molecules, such as glucose, vitamins and amino acids, across otherwise impermeable cell membranes.
- *Storage proteins* function as biological reservoir for small molecules, e.g., metal ions and amino acids, which are mobilized and utilized for maintenance and growth of organisms. For instance, ferritin stores iron, an important component of heme which in turn is essential for binding oxygen by hemoglobin. Others encapsulate small molecules to protect cells, for instance, from metabolites that might be toxic when being released in the wrong cell compartment.
- *Contractile and motile proteins* endow cells with unique capacities for special forms of movement. Cell division, muscle contraction and cell motility present basic ways in which cells achieve motion. Prominent examples include actin and myosin as important contractile muscle proteins or tubulin, a major component of microtubules which facilitate cell division. Another class of proteins involved in motion are so-called motor proteins that control the movement of vesicles, granules, and organelles.
- *Structural proteins* are, in terms of molecular weight, the heaviest class of proteins. These fibrous molecules, typically insoluble, provide strength, structure and support for cells. α-keratins are the crucial proteins in skin, hair, and fingernails. Another example is collagen, a major component of bone, connective tissue, tendons, and cartilage.
- *Scaffold proteins* act as adaptors by linking various proteins to form scaffolds upon which certain protein or protein-DNA complexes are assembled. Scaffold proteins are crucial for regulating signaling pathways by tethering signaling components, localizing these components to specific compartments of the cell, regulating signal transduction by coordinating feedback signals and insulating correct signaling proteins from competing proteins. Prominent scaffold proteins include, for instance, KSR and MEKK1 in the MAPK pathway, HOMER in calcium signaling and DLG1 in T-cell receptor signaling.
- *Protective and exploitive proteins* are essential elements for cell defense and protection. Classical members of this class are immunoglobulins (or antibodies), critical

components of the immune system that locate and indirectly neutralize molecules that are not intrinsic to the host system. Other important examples are blood clotting proteins, e.g., thrombin and fibrinogen, that help to prevent severe loss of blood upon damage of the circulatory system.

- *Transcription factors* are proteins involved in the regulation of gene expression. They recognize and bind specific DNA sequences (motifs), thereby attracting other transcription factors to create a complex which eventually induces the recruitment of RNA polymerase to specific genes. The most common transcription factors include TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH.

It should be emphasized that numerous proteins, particularly in higher eukaryotes, posses multiple different functions rather than only a single one. An intriguing class of such multifunctional proteins are so-called moonlighting proteins that perform multiple autonomous but often unrelated functions without separating these functions into distinct protein domains (Huberts and van der Klei, 2010). Moonlighting proteins contribute to basic cellular functions, such as metabolism, angiogenesis, cell motility, DNA synthesis or repair, as well as in physiological functions and biochemical pathways that are involved in cancer and other diseases. Other striking examples are enzymes, which in addition to their catalytic function are involved in completely unrelated processes, such as autophagy, protein transport or DNA maintenance (Huberts and van der Klei, 2010).

Proteins and their role in diseases

A particular important aspect of proteins is their role in human diseases. Diseases are pathological conditions that impair the normal state of an organism by altering or destroying its vital functions (Merskey, 1986). Abnormal functioning is caused by inherited genetical defects or variations, spontaneous mutations, internal dysfunctions and environmental influences, such as stress, infection or other external factors, that directly or indirectly affect genes and their products (Mackenbach, 2006). Even slight alterations, for instance, in a single gene, might yield an aberrant protein, which may lead to cell malfunction and, eventually, to a disease. Furthermore, many known variations do not necessarily cause a disease but might increase the risk of developing a particular disease.

Disease-related alterations, e.g., mutations or dysregulations may affect proteins in various ways and on several functional levels. However, most alterations will eventually perturb the cellular machinery and its biological processes by impairing the natural function of a protein. Protein function can be severely disrupted by aberrations that affect either the specific protein expression, post-translational modification patterns, the folding into a stable tertiary structure or the combination of such events.

Protein expression

The expression of biologically active proteins is determined by the expression of their encoding genes which is regulated in many different ways. Precise expression control

is vital for cells to synthesize gene products whenever they are needed and to adapt to environmental changes, external signals or damages to the cell (Perdew *et al.*, 2006).

Gene expression is mostly controlled at the level of the transcription initiation and transcription rate but also through microRNA. Transcriptional activity is responsible for the steady state levels of mRNA of the regulated gene, which in turn correlates with protein levels for most genes. Modifications in the regulatory sequences, chromatin structure and proteins that trigger the transcription of a gene, might alter the cellular concentration of particular proteins which in turn perturbs the sensible balance within a cell. Aberrant expression patterns in central regulatory proteins, such as transcription factors that control cell proliferation and differentiation, are known to be a major cause of cancer (Delgado and León, 2006). In particular, (proto-)oncogenes and tumor suppressor proteins that regulate the cell cycle or promote apoptosis are typically over- and underexpressed, respectively, in various types of cancer (Weinberg, 1996; Croce, 2008).

Post-translational modification

Nascent proteins emerging from the translational machinery are often subjected to covalent chemical modifications that alter their amino acid residues. Post-translational modification is a common biological mechanism contributing to the vast diversity in protein structure, function and dynamics (Seo and Lee, 2004; Walsh, 2006). Various biochemical modifications, such as phosphorylation, glycosylation and proteolysis, increase the diversity of functional groups beyond the inherent properties of proteinogenic amino acids and extend the functional and structural repertoire encoded in a genome.

Amino acid substitutions and other sequence variations might disrupt designated post-translational modification sites in proteins. This may have severe functional consequences including conformational changes, alterations in subcellular locations, modulation of enzyme activity and abnormal interaction patterns (Walsh, 2006). Aberrant post-translational modifications are, for instance, involved in the pathogenesis of Huntington's disease (Wang *et al.*, 2010), Alzheimer's disease (Gong *et al.*, 2005) and different types of cancer (Krueger and Srivastava, 2006; Radivojac *et al.*, 2008; Reis *et al.*, 2010). However, also imbalances and alterations in the close proximity of modification sites have been found to be causative for human diseases (Baenziger, 2003; Li *et al.*, 2010).

Protein folding

The cellular function of proteins depends primarily on their tertiary structure. Alterations in the protein sequence, either emerging from inherited or spontaneous variations or aberrant amino acid modifications, may interfere with the folding process and result in incorrectly folded proteins. Misfolding of proteins might have serious implications ranging from functional insufficiency and loss-of-function to perturbation of cellular pathways to aggregation of abnormally folded proteins causing cell damage (Dobson, 2003).

Different diseases have been associated with protein misfolding (Chiti and Dobson, 2006; Gregersen, 2006), often classified into two types: loss-of-function pathogenesis caused by protein degradation and gain-of-function pathogenesis induced by protein

accumulation (Winklhofer *et al.*, 2008).

- In the first case, aberrant proteins are prematurely eliminated by the degradation systems, which results in loss-of-function pathogenesis and protein deficiency diseases (Gregersen, 2006). Cystic fibrosis, Marfan syndrome and some types of cancer, are characterized by the absence of central proteins that have been recognized as misfolded and thus degraded by the proteasome. For instance, the loss-of-function of the crucial tumor suppressor p53 induced by misfolding is thought to be a frequent cause of cancer (Nigro *et al.*, 1989; Lubin *et al.*, 2010).
- Aberrant proteins, which circumvent the cellular surveillance and accumulate to intractable aggregates, induce toxic gain-of-function pathogenesis and amyloidosis (Merlini and Bellotti, 2003; Aigelsreiter *et al.*, 2007). Large quantities of accumulated proteins in the intra- or extracellular space may damage and destroy cells through mechanisms which just have started to be elucidated (Selkoe, 2003). Alzheimer's disease, Parkinson's disease and Type II diabetes, are directly associated with the deposition of such aggregates in tissues, including brain, heart and spleen (Jaikaran and Clark, 2001; Shah *et al.*, 2006; Irvine *et al.*, 2008).

2.2 Protein-Protein Interactions

Once it was widely presumed that proteins are rather isolated entities acting mostly independently of their surroundings. Proteins were assumed to diffuse freely within cells while biochemical reactions result from random encounterings between two proteins. Today it is widely established that this picture is far too simplistic to explain the complex mechanisms that coalesce within living systems.

Specific proteins have evolved to bind every conceivable molecule – from small simple ions to large complex molecules like fats, sugars, (ribo-)nucleic acids, and other proteins (Lodish *et al.*, 2007). They mediate their function within complex networks of highly connected macromolecules rather than in isolation (see Figure 2.4). Enzymes, for instance, bind substrates to catalyze biochemical reactions, antibodies attach to viruses and bacteria to inactivate them directly or target them for degradation, α- and β-hemoglobin chains assemble into heterotetramers for transporting oxygen from lungs to tissues, and transcription factors bind the DNA to induce transcription.

One of the most important types of biomolecular relationships are protein-protein interactions[3]. Virtually all cellular mechanisms rely on the physical binding of two or more proteins to accomplish a particular task. To understand these processes and the importance of protein-protein interactions on a molecular and biophysical level, one needs to identify the different types of interactions, discern the extent to which they take place in the cell, and determine their consequences (Piehler, 2005).

Protein-protein interactions are commonly defined as physical contacts involving molecular docking between two or more gene products (Rivas and Fontanillo, 2010). From

[3]In this work we are primarily focusing on protein-protein interactions. Terms like 'protein interaction' or 'interaction' will refer in the following chapters to protein-protein interactions only. When talking about other biological relationships we will point this out.

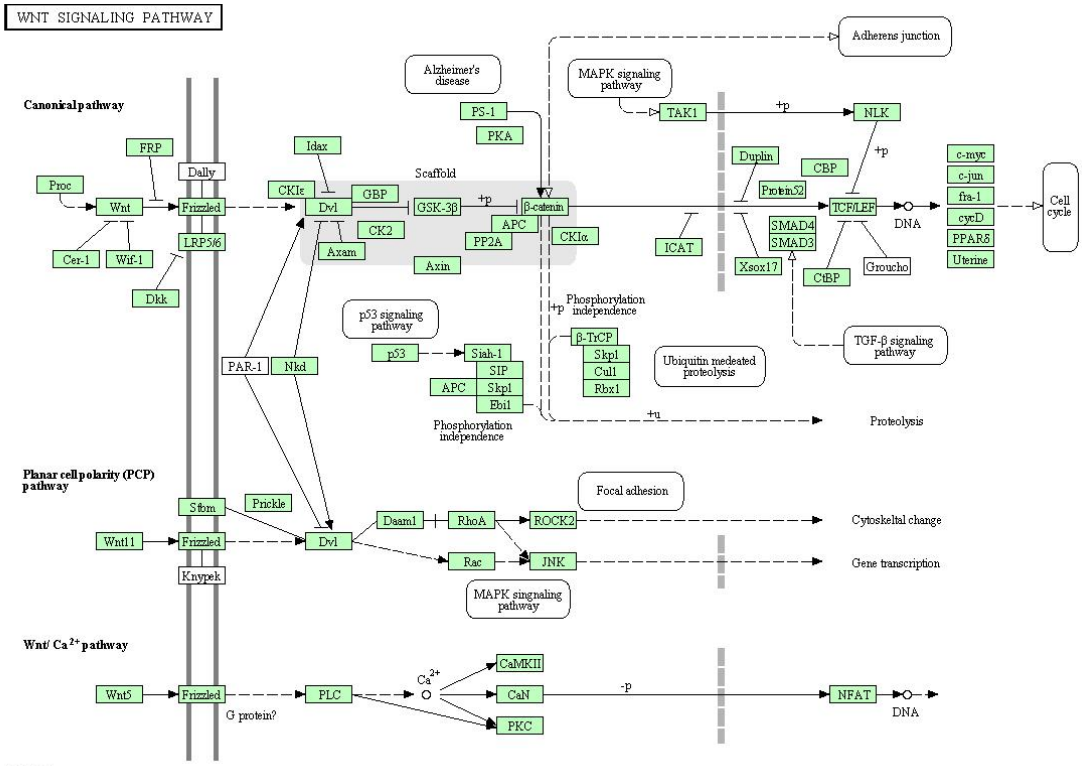


Figure 2.4: The human Wnt signaling pathway. Members of the Wnt pathway form a family of highly conserved, secreted signaling molecules that regulate cell-to-cell interactions during embryogenesis. Mutations in Wnt pathway components lead to specific developmental effects. Various human diseases, including cancer, are caused by abnormal Wnt signaling (hsa04310 retrieved from KEGG (Kanehisa *et al.*, 2010)).

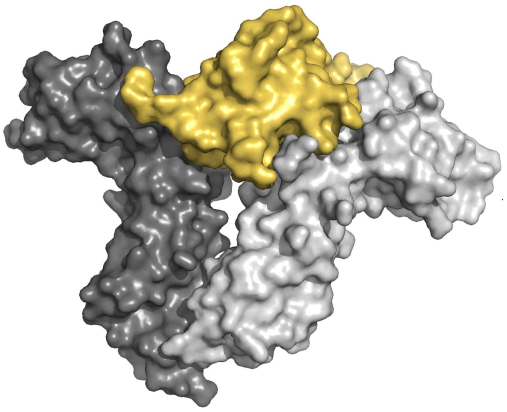


Figure 2.5: Molecular docking. Binding of the human growth hormone (yellow) to the extracellular portion of its homodimeric receptor (light and dark gray, taken from Ofran and Rost (2007)).

the physical point of view, any two proteins can interact – but on what conditions and at which strength? An important aspect for the formation of an interaction is the biological context. Whether two proteins do physically interact with each other depends on the cell type, cell cycle phase and state, environmental conditions, developmental stage, post-translational modifications and the presence of cofactors and other binding partners (Rivas and Fontanillo, 2010).

Protein-protein interactions are non-covalent interactions of two proteins primarily driven by hydrophobic effects, hydrogen bonds and electrostatic interactions (Nussinov and Tsai, 2005). Protein interactions differ based on their diverse structural and functional characteristics. Several types of interfaces facilitate the specific binding of proteins to each other. The most common way for proteins to interact is through the precise matching of two complementary, rigid protein surfaces. These interactions often target just a single interaction partner from the different proteins found in a cell. A second type of interaction is established among two α-helices, one from each protein, that pair together to form a coiled-coil. Finally, proteins may interact by linking the rigid surface on one protein to an extended loop of the polypeptide on a second protein (Lodish *et al.*, 2007).

Protein interactions differ in their strength, specificity and the type of their interacting subunits (Shoemaker and Panchenko, 2007a). *Strength* depicts whether an interaction is permanent or transient. Permanent interactions are usually associated with proteins that are part of multi-subunit protein complexes whereas transient interactions are temporary and typically require specific conditions for stimulating this interaction. Transient interactions are believed to regulate the majority of cellular processes (Perkins *et al.*, 2010). *Specificity* refers to the selective binding of interaction partners. Highly specific interactions are those where a protein only binds one or few proteins out of the different ones it may encounter. Non-specific interactions, on the other hand, include bindings that a protein experiences during its life cycle when being translated, folded, modified, quality checked or degraded. All proteins, for instance, interact with the ribosome, many of them contact chaperones and the degradation machinery. The *type of interacting subunits* specifies whether an interaction forms a hetero-oligomer with several different subunits or a homo-oligomer with only one type of protein subunit.

2.2.1 Identification of protein interactions

Detecting all possible physical interactions within an organism – the interactome (Cusick *et al.*, 2005) – is an essential step toward deciphering the complex molecular relationships in living systems. Different experimental and computational methodologies have been developed to identify the specific mechanisms of protein recognition at the molecular level and to elucidate the global picture of protein interactions in the cell. We briefly introduce (1) two established experimental methods, (2) literature curation and (3) *in silico* techniques for discovering protein interactions and discuss their methodical capabilities and limitations.

[4] *In vivo* methods refer to experiments performed in living cells while *in vitro* methods are carried out in a controlled environment.

Table 2.1: Experimental methods for detecting protein interactions and their characteristics. The table summarizes for each technique whether it is suitable for large-scale analysis (+ vs. −), whether it is an *in vivo* or *in vitro* system[4], the type of interaction it detects (binary vs. complex) and the type of interaction characterization. (Table adapted from Shoemaker and Panchenko (2007a))

Method	Large-scale approach	Cell assay	Type of interaction	Type of characterization
Yeast two-hybrid	+	*in vivo*	binary	Identification
Tandem affinity purification–MS	+	*in vitro*	complex	Identification
Protein microarrays	+	*in vitro*	complex	Identification
Phage display	+	*in vitro*	complex	Identification
Co-immunoprecipitation	−	*in vivo*	complex	Identification
Surface plasmon resonance	−	*in vitro*	complex	Kinetic, dynamic characterization
Electron microscopy	−	*in vitro*	complex	Structural and biological characterization
Fluorescence Resonance Energy Transfer (FRET)	−	*in vivo*	binary	Biological characterization
X-ray Crystallography, NMR spectroscopy	−	*in vitro*	complex	Structural and biological characterization

2.2.1.1 Experimental detection methods

Experimental elucidation of interactions between gene products is done either at small- or large-scale (Rivas and Fontanillo, 2010). Experiments detecting less than 100 protein interactions are commonly considered to be small-scale while the others are denoted as large-scale (Patil *et al.*, 2011). Methods that identify direct physical interactions among protein pairs are called *binary* methods. Approaches that determine physical interactions between a group of proteins, without distinguishing between direct and indirect interactions, are *co-complex* methods.

Numerous experimental methods have been developed for protein interaction detection, see Table 2.1 and Phizicky and Fields (1995) for a review. Traditionally, protein interactions have been detected by genetic, biochemical or biophysical techniques, such as X-ray crystallography or fluorescence resonance energy transfer (FRET). Such small-scale studies focus on individual proteins for generating specific interaction maps (Finley and Brent, 1994; Mayes *et al.*, 1999; Goehler *et al.*, 2004). However, the increasing availability of fully sequenced genomes and the speed at which proteins are discovered increased the interest in techniques that screen large sets of candidates systematically. Two widely established large-scale methodologies are the yeast two-hybrid (Y2H) system (Fields and Song, 1989) and tandem affinity purification coupled to mass spectrometry (TAP-MS) (Rigaut *et al.*, 1999); the former system is a binary and the latter a co-complex method. Both methodologies have been used for large-scale experiments in different model organisms, including yeast, fly, worm and human. The majority of interaction data currently available in the databases IntAct and MINT, for instance,

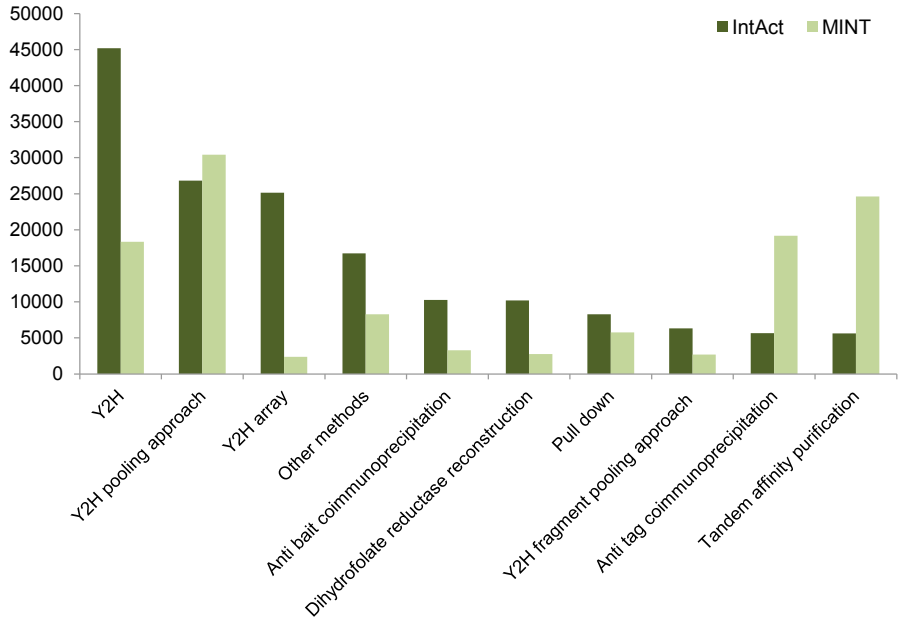


Figure 2.6: Overview on the number of protein interactions per detection method as provided in the public databases IntAct and MINT (March 2011).

is derived from Y2H and its variants. A general overview on the number of protein interactions per detection method is shown in Figure 2.6.

In the following, we briefly introduce Y2H and TAP-MS as the work presented in this thesis largely relies on protein interaction data derived from such experiments. We will highlight the systematic and methodological limitations inherent to each method. These effects have to kept in mind as the amount of experimental errors inevitably affects the outcomes of further analysis.

Yeast two-hybrid assay (Y2H) The Y2H assay determines whether two proteins physically interact with each other by using the principle of transcriptional activation. Genetically modified yeast strains are used to express two fusion proteins (two hybrids), which, if they interact, induce the expression of a reporter gene. Fusion proteins are created by linking proteins to separable protein domains of transcription factors. One protein, the bait, is fused to the DNA-binding domain that is capable to bind the promoter of a reporter gene. A potential binding partner, the prey, is linked to the activator domain that activates transcription by facilitating the binding of the RNA polymerase to the promotor. If both proteins interact, their complex forms an intact, functional transcriptional activator which mediates the transcription of the reporter gene (see Figure 2.7). Reporter genes encode proteins whose function provides a simple readout, such as *LacZ* from *E. coli* which causes a colorimetric reaction within the cell (Brueckner *et al.*, 2009).

Large-scale library screens can be performed by using a cDNA library instead of a single prey protein. Y2H has been extensively applied in several large-scale screens (Uetz *et al.*, 2000; Ito *et al.*, 2001; Rual *et al.*, 2005; Stelzl *et al.*, 2005) and for individual

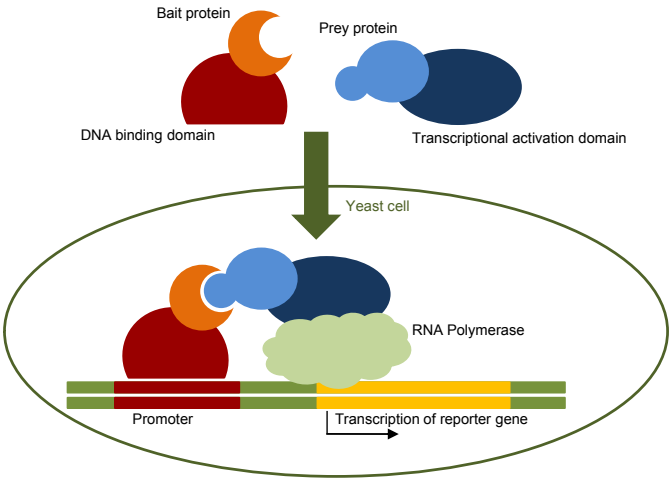


Figure 2.7: The yeast two-hybrid system for detecting binary protein-protein interactions (adapted from Alberts (1998)). A target protein, the bait, is fused to a DNA-binding domain that localizes it to the promoter region of a reporter gene. A potential binding partner, the prey, is linked to an activator domain. The interaction of both fusion proteins forms an intact, functional transcriptional activator which triggers the expression of the reporter gene.

experiments (Finley and Brent, 1994; Mayes *et al.*, 1999; Davy *et al.*, 2001).

Overall, Y2H is an established *in vivo* technique, well-suited for large-scale analysis. It allows to detect both transient and stable interactions, independently of endogenous protein expression. Albeit yeast cells are utilized for expressing fusion proteins, Y2H is not restricted to interactions between yeast proteins; in principle, the genetic code of any fusion protein may be introduced into the yeast cell. The major drawback of the yeast two-hybrid assay is its poor reliability. Y2H is performed in the nucleus, hence many proteins are not analyzed in their native compartment. Thus, two proteins may interact in the experiment although they would not do so in their natural environment (Koegl and Uetz, 2007). In turn, essential post-translational modifications of non-yeast proteins may not be carried out, or the fusion process might interfere with the true interactions between proteins. In consequence, Y2H data are associated with a large number of false positive and false negative interactions. Early estimates on distinct data sets indicated that only 30–50% of the detected interactions are biologically meaningful. More recent quality assessments suggested that Y2H data contain less false positives as previously presumed. Nevertheless, Y2H screens are still far from being reliable and the rate of interactions not detectable by Y2H remains substantial (Yu *et al.*, 2008).

Tandem affinity purification mass spectrometry (TAP-MS) In this technique, individual proteins are first fused to a protein fragment (the 'tag') which is used as an anchor for biochemical purification of protein complexes. The modified proteins are expressed and purified from cell extracts using the tag. Other proteins bound to the tagged protein are co-purified and subsequently identified by mass spectrometry (see Figure 2.8).

In contrast to Y2H assays, data derived from co-complex approaches, such as TAP-

MS or co-immunoprecipitation, cannot be directly translated into binary interactions. Co-complex methods only identify proteins involved in a given complex rather than the direct interactions between them. Different models are employed to translate the group-based observations into pairwise interactions. The matrix model assumes that all proteins of a purified complex interact whereas the spokes model infers only interactions between the tagged protein and each co-purified protein. The latter one is often used, as it yields a smaller number of false positives (Hakes *et al.*, 2007). Bader and Hogue (2002) estimated, for instance, that the number of false positives is three times larger in the matrix model.

Genome-wide TAP-based studies have been successfully performed for yeast (Krogan *et al.*, 2006; Gavin *et al.*, 2006), and for a smaller number of proteins in human (Ewing *et al.*, 2007) and *E. coli* (Butland *et al.*, 2005). Contrary to Y2H, TAP-MS detects protein complexes and interactions within the native cellular environment and is able to capture several members of a complex. In turn, protein complexes that are not present under the given conditions might be missed, loosely associated proteins of a complex might be washed of during purification and the tagging of a protein may interfere with the complex formation. Accordingly, the coverage of TAP-MS is limited as a large fraction of interactions, e.g., transient interactions, might be missed. Yet, false positive and false negative rates are much lower than for other experimental techniques (Kemmeren *et al.*, 2002; von Mering *et al.*, 2002), including Y2H, as interaction information are obtained under more natural physiological conditions than those induced by Y2H. However, both methods detect rather complementary types of interaction and only the combination of different approaches with bioinformatic tools will eventually yield a more complete characterization of physiologically relevant protein interactions in a given cell or organism (Brueckner *et al.*, 2009).

Literature curation

Protein interaction data, retrieved from small- and large-scale experiments, are commonly published in the scientific literature. To make this knowledge available to the scientific community, interaction data have to be curated and archived in specialized databases.

Literature curation translates information on physical interactions between proteins from free-text publications into a structured format (Chatr-aryamontri *et al.*, 2007). Curators read through the literature, identifying and extracting all significant information: the organism being studied, the gene product annotated, the proteins that interact, the type of experiment performed, and an identifier (typically the PubMed ID) as the source of information. This allows for quality control of the data.

However, the volume and growth of biomedical literature makes it hard to curate all newly published information (Hunter and Cohen, 2006; Chatr-aryamontri *et al.*, 2007). In addition, relevant data may be missed by oversight, an intrinsic weakness of purely human curation, and literature curation is 'hypothesis-driven' with prior assumptions of what could be learned. Accordingly, literature-curated data are often biased toward better-characterized genes and proteins (Cusick *et al.*, 2009). In consequence, only a

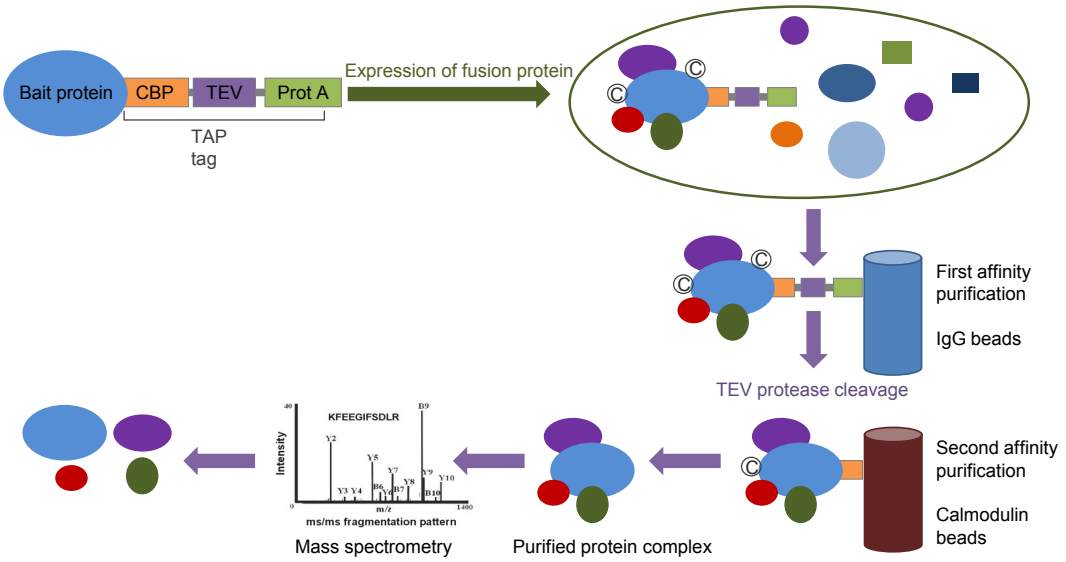


Figure 2.8: TAP-MS procedure for characterizing protein complexes. A target protein, the bait, is fused to a protein fragment - the TAP tag - comprising a protein A-IgG binding domain (ProtA), a calmodulin binding peptide (CBP) and a TEV protease cleavage site. The modified protein is then expressed in cells where it may carry out its natural function participating in one or more protein complexes. Protein complexes are purified from cell extracts by two subsequent affinity chromatographies using the TAP tag. Co-purified proteins bound to the bait are then identified by standard mass spectrometry.

small fraction of all published interactions has been captured in the interaction databases so far.

Computational detection approaches

As discussed above, experimental detection methods and literature curation have several limitations and do not yet come close to elucidate full interactomes. Thus, several computational methods have been proposed for predicting protein-protein interactions *in silico* based on various evidence. A complete review of the available approaches including their strengths and limitations is beyond the scope of this section. We shall briefly introduce established concepts for predicting protein interactions and refer interested readers to extensive reviews (Valencia and Pazos, 2002; Shoemaker and Panchenko, 2007b; Liu *et al.*, 2008) as predicted protein interactions are not used in this work.

A common approach for inferring novel protein interactions is based on the analysis of protein domains to determine which domains participate in an interaction. Given a set of protein domains that interact frequently in known interactions, novel interactions can be predicted between proteins containing the same domain pairs (Deng *et al.*, 2002; Chen and Liu, 2005; Jothi *et al.*, 2006). Another established methodology relies on the concept of 'interologs', which refers to pairs of homologous proteins interacting in different organisms (Matthews *et al.*, 2001). Novel protein interactions are thus inferred by identifying evolutionarily conserved protein interactions in related genomes (Sharan

et al., 2005; Wiles *et al.*, 2010). Additional methods employ:

- phylogenetic profiles (Pellegrini *et al.*, 1999; Goh and Cohen, 2002),
- gene fusion events (Rosetta stone) (Marcotte *et al.*, 1999b; Enright *et al.*, 1999),
- co-localization information, such as gene neighborhood or gene cluster (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999),
- patterns of co-occurrence or co-expression (Jansen *et al.*, 2002; Ge *et al.*, 2001),
- sequence and structural similarities between interacting proteins (Comeau *et al.*, 2004; Sikić *et al.*, 2009).

2.2.2 Quality and coverage of interaction data

Over the last decade there has been an increase in the quantity and quality of interaction data (Yu *et al.*, 2008). Yet, current data sets are still limited in both terms. Hence, critical evaluations are essential to quantify the reliability of any specific data set to identify interactions that actually occur in a cell at a given state.

When evaluating the reliability of interaction data, quality and quantity have to be considered together. Reference sets of trusted interactions from manually curated protein complexes, e.g., MIPS and CORUM (Mewes *et al.*, 2002; Ruepp *et al.*, 2010), are used as benchmarks to assess the quality of experiments and prediction methods by determining the proportion of reported interactions that is reproducible. However, this comparison is not always fair as interactions are often derived under different conditions. Further, the evaluation depends largely on the reference sets which are naturally incomplete and might be biased themselves.

2.2.2.1 Data quality

The different approaches for identifying protein interactions have different strengths and limitations, and the systematic methodological differences are often reflected in the quality of the resulting data sets.

- *Low-throughput interaction data* largely result from small-scale experiments where individual proteins are studied after careful selection by biologists (Yoon *et al.*, 2003). These data are presumed to be highly reliable with a low number of false positives since these types of experiments are performed under precise control. Despite their high quality, interaction data from small-scale experiments are limited to the specific processes studied. Thus, many laborious small-scale studies are necessary to obtain a global picture of a cell.
- *High-throughput interaction data* derived from large-scale studies are often unreliable and still incomplete. A large number of interactions from such screens has been shown to be false, i.e., not occurring in the cell. For instance, 50–70% of the interactions detected for yeast are assumed to be false positives (Sprinzak *et al.*, 2003; Bork *et al.*, 2004). In addition, most studies have not reached saturation while others are limited or biased toward particular proteins (von Mering *et al.*,

2002). For instance, proteins with high abundance are often favored which implies that a large fraction of interactions for proteins with lower abundance remains undiscovered. Other methods are biased toward particular cellular localizations of the interacting proteins. The incompleteness and unreliability of data sets is evidenced in the small overlap between such sets (von Mering *et al.*, 2002; Rivas and Fontanillo, 2010). Albeit recent data sets have improved in quality and coverage (Gavin *et al.*, 2006; Krogan *et al.*, 2006), missing interactions and incorrectly identified interactions remain a major problem.

- *Literature-curated interaction data* derived from manual curation of small-scale experiments are commonly presumed to be more reliable than high-throughput data (Reguly *et al.*, 2006; Gandhi *et al.*, 2006). However, recent assessments of different data sets question the superior reliability of literature-curated data indicating that these data might contain a larger fraction of false positives than previously anticipated (Ramirez *et al.*, 2007; Mackay *et al.*, 2007; Cusick *et al.*, 2009; Wu *et al.*, 2009; Venkatesan *et al.*, 2009).
  In fact, large-scale curation of primary literature is ambitious. One of the main challenges is the lack of formal representation of interactions in published manuscripts (Cusick *et al.*, 2009). Interaction data are typically reported either as free text or in tables of variable format, often lacking key pieces of information that are important for a detailed understanding of the experiments (Orchard *et al.*, 2007). Inconsistencies and missing information hinder the curation process considerably and lead to misinterpretation and time-consuming, error-prone attempts to derive missing evidence by other means. For instance, protein and gene names are often mistaken as their synonyms are hard to trace back to their canonical names, information on the species of each interactor are hidden or missing and standardized descriptions of the detection methods are absent (Turinsky *et al.*, 2010). To address this issue, the minimal information required for reporting a molecular interaction experiment (MIMIx) was proposed to define a community-wide consensus on what information is required to appropriately describe a molecular interaction (Orchard *et al.*, 2007). Submitting species, protein names, identifiers and methodological descriptions in a standardized format upon publication, as already required for sequences, microarray data and protein structures, will facilitate the accurate extraction of relevant information for curators and further improve the curation process and the reliability of literature-curated data (Lehne and Schlitt, 2009).

Given the varying data quality, different concepts have been proposed to increase the confidence in experimental interaction data especially from high-throughput experiments. These concepts include filtering for interactions that have been observed in multiple experiments or in multiple species, and assigning weights to interactions depending on their detection method (Suthram *et al.*, 2006; Braun *et al.*, 2009). Other measures assess the degree to which interacting proteins are associated with the same functional categories and cellular locations to estimate confidence values.

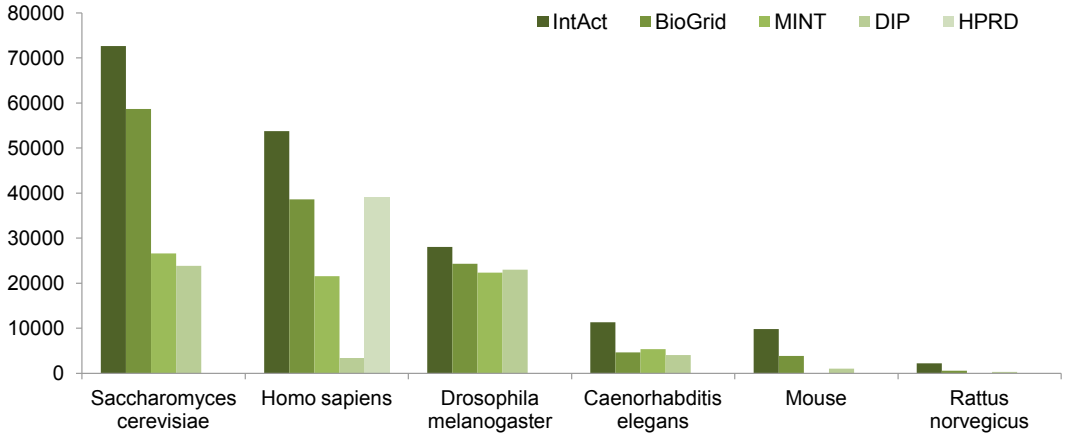


Figure 2.9: Number of protein interactions per species in the different databases (March 2011).

2.2.2.2 Coverage of interaction data

The majority of known protein interactions accounts so far for human and yeast, as illustrated in Figure 2.9. Assessing the current coverage requires an estimate of the complete set of interactions. Table 2.2 shows the anticipated number of protein interactions for human and yeast in comparison to the number of interactions discovered so far. Different attempts for appraising the potential interactome sizes yield clear discrepancies in the expected number of interactions, even for the well-studied model organism yeast. Empirical estimates for yeast range from 13,500 up to 75,000 interactions between the 5,800 proteins of yeast. The expected number of interactions in human also varies greatly, from 130,000 up to 650,000. These estimates deviate, in particular for yeast, significantly from the number of currently known interactions in the public databases. There are two possible explanations:

- A large number of experimentally determined interactions in the database are false positives (Rivas and Fontanillo, 2010).
- Another explanation might be the fact that the potential size of the yeast interactome has been under estimated.

2.2.3 Interaction databases and repositories

Several publicly available databases and repositories have been designed to collect, store and organize protein interaction data on a large scale as well as across studies and methods.

Interaction data are accumulated in three types of databases:

1. *Primary databases* – containing only experimentally proven interactions.
2. *Meta-databases* – integrating interactions from several primary databases.

Table 2.2: Expected number of protein interactions vs. known number of protein interactions for human and yeast. The number of known protein interactions has been accumulated from the in-house interaction databases PiPa which is compiled from integrating six different public protein interaction databases (see Section 5.1).

Species	Known interactions	Potential interactome size	
Human	81,868	130,000	Venkatesan *et al.* (2009)
		154,000 to 369,000	Hart *et al.* (2006)
		650,000	Stumpf *et al.* (2008)
Yeast	70,990	13,500 to 22,500	Yu *et al.* (2008)
		16,000 to 26,000	Grigoriev (2003)
		25,000 to 35,000	Stumpf *et al.* (2008); Blow (2009)
		75,000	Hart *et al.* (2006)

3. P*rediction databases* – combining predicted and experimentally detected interactions.

Primary databases focus on experimentally verified protein interactions from small- and large-scale studies. The major primary protein interaction databases are the Biological General Repository for Interaction Datasets – BioGRID (Stark *et al.*, 2006), the Database of Interacting Proteins – DIP (Salwinski *et al.*, 2004), the IntAct molecular interaction database – IntAct (Aranda *et al.*, 2010), the Molecular Interaction Database – MINT (Ceol *et al.*, 2010), the Human Protein Reference Database – HPRD (Prasad *et al.*, 2009) and the Biomolecular Interaction Network Database – BIND[5] (Bader *et al.*, 2003). Specific characteristic of each database, including size, species and references, are shown in Table 2.3.

Each database captures the interacting proteins and their species, the original publication and experimental method(s) that verified the individual interaction. Although these databases document only experimentally derived interactions, they differ greatly in scope and content. For instance, IntAct focuses primarily on high-throughput screens, e.g., Y2H and TAP-MS (see Figure 2.6) whereas HPRD focuses entirely on human interactions taking also small-scale studies into account (Lehne and Schlitt, 2009). Given the different foci, the interaction data in such databases have little overlap amongst each other. In fact, when comparing interaction data of the six resources there are only three human interactions that are contained in all six of them whereas the number of interactions exclusively reported in each database is much larger, e.g., 19,659 in HPRD (Rivas and Fontanillo, 2010).

Due to the heterogeneity and complementarity in the databases, data sets from different databases are often combined to generate more comprehensive data sets (Chaurasia *et al.*, 2007; Chatr-Aryamontri *et al.*, 2008). However, integrating interaction data from distinct databases is demanding. For instance, different gene or protein identifiers are used, even within the same database (Lehne and Schlitt, 2009). Moreover, interaction

[5]Note, BIND is now part of the Biomolecular Object Network Databank (BOND) and was subsequently acquired by the company Thompson Reuters.

data are often provided in many different formats capturing various details. To overcome this problem the International Molecular Exchange (IMEx) consortium defined an XML-based proteomics standard, the proteomics standards initiative – molecular interaction (PSI-MI, Hermjakob *et al.* (2004b)), to enable data exchange for improving data quality and curation. DIP, IntAct, MINT and BioGRID are current members of the IMEx consortium.

To obtain interaction data from the different sources, *meta-databases* have evolved to extract and integrate protein interaction data consistently into single formats (see Figure 2.10). To date, APID (Prieto and Rivas, 2006), PIANA (Aragues *et al.*, 2006) and PiPa (in-house database, *to appear*) represent the most comprehensive meta-databases. A number of other meta-databases exist, but these focus on specific species (Chaurasia *et al.*, 2007; Goll *et al.*, 2008) or incorporate other types of interactions, e.g., computationally predicted ones (Jensen *et al.*, 2009; Chen *et al.*, 2009b). Although these meta-databases provide access to more comprehensive datasets, they do have certain restrictions:

- First, they often do not allow the complete download of their data. In most cases their content is only accessible over the web on a per-protein basis, which renders them useless for systematic large-scale analyses. Even if the database content is available for download, it often excludes certain sources due to licensing issues.
- Second, many systems employ complex and task-specific data selection procedures, leading to an incomplete coverage of the integrated sources.
- Finally, meta-databases are often less up to date since the update intervals of these systems often are irregular and not adjustable by users.

In this work, we used PiPa to integrate interaction data from multiple sources into one centralized database (see Section 5.1). In contrast to most meta-databases in this field, PiPa does not perform any semantic integration itself; instead, data from the sources are integrated as such into the system (for instance, no duplicate detection is performed), leaving the decision onto which form of aggregation or quality filtering to perform to the user. Moreover, PiPa features a graphical administration tool for monitoring databases and for triggering updates allowing the user to control data import and update cycles.

Prediction databases provide relationships between proteins inferred by a variety of *in silico* methods combined with experimentally verified interactions. STRING (Jensen *et al.*, 2009) is one of the most established databases dedicated to combine known and predicted functional associations, including direct physical and indirect functional interactions. Information from numerous sources are weighted and integrated, including experimental repositories, computational prediction methods and public text collections, to map all interaction evidence onto a large set of genomes and proteins. POINT (Huang *et al.*, 2004) and OPHID (Brown and Jurisica, 2005), on the other hand, have been specifically designed to extend the human interactome using model organism data. Human protein interactions in these databases are inferred primarily based on available orthologous interactome datasets, domain-domain co-occurrence, co-expression and functional similarity of proteins.

Table 2.3: List of the major primary protein interaction databases and their characteristics (March 2011). The table summarizes for each data source the species covered, the size of the database in terms of proteins and interactions, the type of experiments included – low-throughput (LT) or high-throughput (HT) – as well as the number of experiments and publications (if available). Missing numbers indicate that no information on the number of experiments or publications could be obtained from the respective database.

			Database size		Type of		
Database	Reference	Species	# Proteins	#Interactions	Experiments	# Experiments	# Publications
BioGrid	Stark *et al.* (2006)	No restriction	31,423	136,544	LT, HT	–	19,588
DIP	Salwinski *et al.* (2004)	No restriction	23,201	71,276	LT, HT	16,640	1,602
IntAct	Aranda *et al.* (2010)	No restriction	56,037	247,309	LT, HT	12,844	–
MINT	Ceol *et al.* (2010)	No restriction	31,859	90,280	LT, HT	$>$ 19,000	4,532
HPRD	Prasad *et al.* (2009)	Human	30,047	39,194	LT, HT	–	–
MPPI	Pagel *et al.* (2005)	Mammalian	745	700	LT	–	–
BIND	Bader *et al.* (2003)	No restriction	31,972	58,266	LT, HT	–	–

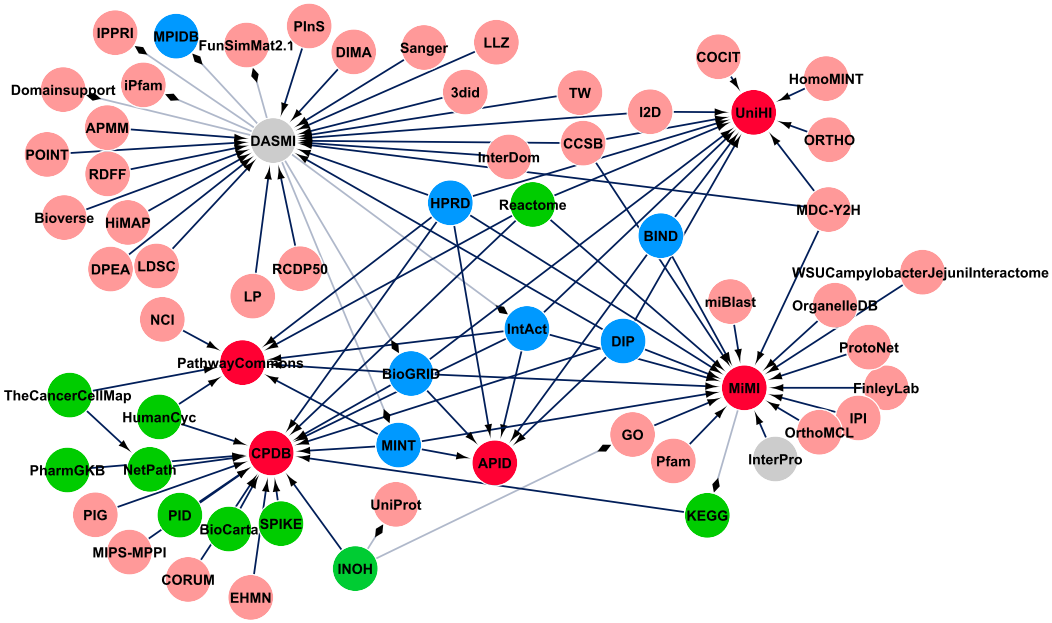


Figure 2.10: Overview on meta-databases and their data exchange patterns (from Klingström and Plewczynski (2010)). Primary protein interaction databases are shown in blue, pathway databases are represented in green and meta-databases are indicated in red. Three distinctive approaches to data meta-mining are illustrated. (1) APID, MiMI and UniHI unify data from source interaction databases to generate a centralized repository. (2) ConsensusPathDB (CPDB) and Pathways Commons provide a similar service aimed at integrating different pathway databases. (3) DASMI, in turn, maps to several databases rather than integrating them.

2.3 Protein-Protein Interaction Networks

Binary biomolecular interaction data can be represented as biological networks (see Figure 2.11). Analyzing such biomolecular networks has become one of the key topics in systems biology and bioinformatics (Zhang, 2009). Understanding biological networks at a systems-wide level allows to elucidate basic function, structure and dynamics of the network as well as the underlying essential mechanisms in living systems (Cusick *et al.*, 2005).

Systematic studies of protein interaction networks are particularly important for deciphering the relationships between network structure and function (Yook *et al.*, 2004; Pandey *et al.*, 2010), discovering novel protein function (Sharan *et al.*, 2007), identifying functionally coherent modules (Spirin and Mirny, 2003; Dittrich *et al.*, 2008) and conserved molecular interaction patterns (Sharan *et al.*, 2005; Jaeger and Leser, 2007). In addition, interaction networks have become essential tools for associating proteins with distinct phenotypes and diseases (Goh *et al.*, 2007; Ideker and Sharan, 2008), as well as for studying pharmacological drug-target relationships (Berger and Iyengar, 2009).

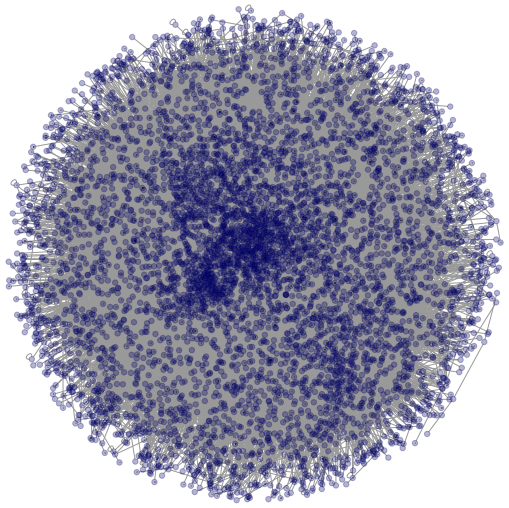


Figure 2.11: Visualization of the yeast protein interaction network used in this study.

2.3.1 Basic network nomenclature

Biological networks, in general, present abstractions of complex biological systems (Barabási and Oltvai, 2004). Nodes within a network commonly represent biomolecules, e.g., genes, proteins or metabolites. Edges indicate physical or functional interactions, including genetic interactions, protein-protein interactions, transcriptional binding, biochemical reactions, and many others. The probability or strength of an interaction can be modeled by assigning weights to the edges.

Depending on the nature of the interactions, edges can be directed or undirected. In directed networks, interactions between two molecules have well-defined directions, which represent, for instance, the flow of material from a substrate to a product in a metabolic reaction, or the flow of information from a transcription factor to the gene it regulates. In undirected networks, such as protein interaction networks, links only present mutual binding relationships.

To comply with the common nomenclature from biology and computer science, we use the term 'network' for discussing biological aspects and 'graph' when referring to algorithmic concepts; however, in principle, both terms can be used synonymously.

Basic graph concepts

For computational analysis, biological networks are commonly modeled as graphs, which allows to analyze the underlying data using graph-theoretical methods. Formally, a graph $G = (V, E)$ comprises two types of elements, namely nodes V and edges E where an edge $e = (u, v)$ connects two vertices u and v.

Different data structures are available for representing graphs. The most common are adjacency (incidence) lists and adjacency (incidence) matrices. The choice of the data structure depends mainly on its intended application. Adjacency lists, for instance, are preferred for presenting sparse graphs.

2.3.2 Properties of protein interaction networks

An important advance toward understanding biological systems is the realization that biological networks follow basic principles which determine at least partly their structural and functional properties (Yook *et al.*, 2004). Several graph-theoretical concepts have been established to study the organization of biological networks on different levels. For instance, elementary graph measures, such as degree distribution, clustering coefficient and centrality, can be used to describe global and local network characteristics that provide insights into network evolution, stability and dynamics (Tyson *et al.*, 2001).

A number of topological features in biological networks overlap significantly with that in other complex systems, such as the world wide web or social networks. Despite the difference between these networks, their architecture is determined by few principles which allow to employ the knowledge from large and well-mapped non-biological systems to characterize the interwoven relationships in a cell. For instance, small-world property, scale-free degree distribution and hierarchical modularity are typical features that also characterize biological networks (Almaas, 2007). In the following, we will introduce selected topological features that are important for protein interaction networks and for this thesis.

2.3.2.1 Degree distribution

One of the basic characteristics of a node v is its degree k, generally defined by the number of links it has to other nodes in the network. The node degree is commonly used to determine the degree distribution, $P(k)$, which yields the probability that an arbitrary node has exactly k links. $P(k)$ is computed by determining the number of nodes $N(k)$ with $k = 1, 2, \ldots, n$ links and dividing it by the total number of nodes N in the network. The degree distribution is an elementary measure to characterize the topology of a network and allows to distinguish between different classes of networks, such as random networks, hierarchical networks and, most importantly for us, scale-free networks (Barabási and Oltvai, 2004).

2.3.2.2 Scale-free topology

Several types of biological networks approximate a scale-free topology which is characterized by a power-law degree distribution. This means, that the probability $P(v)$ of a node v having k links is $P(k) \sim k^{-\gamma}$, where γ is the degree exponent that ranges between 2 and 3 in most networks (Barabasi and Albert, 1999). An important characteristic of such networks is their non-uniformity, i.e., most nodes have only few links while few nodes, so-called hubs, have many links. In particular, the absence of a typical node degree (or scale) that can be used to describe the nodes within a network characterizes the scale-free topology.

Two mechanisms are responsible for the emergence of scale-free topologies: growth and preferential attachment. Growth implies that networks evolve through the successive addition of new nodes whereas preferential attachment means that new nodes are attached preferentially to nodes that are already highly connected. Hubs are generated jointly

through a 'rich-gets-richer' mechanism as strongly connected nodes will acquire new links at a higher rate than less connected nodes which turns them into hubs (Barabási and Oltvai, 2004).

Scale-free organization is a typical feature of protein interaction networks. Most proteins participate in only few interactions while few proteins participate in dozens. Growth and preferential attachment in interaction networks are thought to have a common evolutionary origin: gene duplication (Pastor-Satorras *et al.*, 2003). The duplication of a gene yields two identical gene products. This induces growth as new nodes are added to the network. Further, duplicated gene products will interact with the same proteins as their ancestor proteins due to their structural similarity. Thus, proteins interacting with duplicated proteins acquire additional links whereas highly connected proteins are more likely to establish new interactions to duplicated proteins than less connected proteins.

Power-law distribution and scale-freeness seem to be common characteristics of biological networks. Yet, it is important to point out that other distributions with similar qualitative features, i.e., the existence of hub nodes, might explain their topological properties and their dynamical behavior (Lima-Mendez and van Helden, 2009). Such distributions include, for instance, generalized Pareto law, truncated power-law, stretched exponential distribution, geometric distribution, or a combination of the above (Khanin and Wit, 2006).

2.3.2.3 Network centrality

Network centrality is a quantitative measure that determines the relative position of a node in a network, which in turn can be used to assess its relative importance in the global network organization. Centrality represents a node property that quantifies the structural impact of a particular node on the processes within a network. Yet, the definition of 'central' varies depending on the context. In biological networks, for instance, centrality analysis is frequently used to identify interesting molecules within a network that are essential for biological processes, metabolic pathways or diseases. Network centrality, such as proximity and connectivity, offer effective means for identifying proteins that are either essential for viability or implicated in human diseases (Estrada, 2006; Lage *et al.*, 2007; Navlakha and Kingsford, 2010). Highly central proteins in interaction networks, for instance, are thought to be of essential functional and evolutionary importance as the deletion of such nodes is associated with lethality (Jeong *et al.*, 2001; Fell and Wagner, 2000).

The notion of centrality can be formally defined as a function C that assigns numerical values $C(v)$ to every node $v \in V$ in a given graph G. When considering the ranking of the nodes in G, a node u is defined to be more central, i.e., important, than another node v if and only if $C(u) > C(v)$ (Koschützki and Schreiber, 2008). Diverse node and edge characteristics can be used to determine centrality. Some emerged from biological sciences while others have been transferred from different fields, such as social network analysis. Traditional measures for network analysis, e.g., degree, closeness, betweenness and eigenvector centrality, are defined as follows (Junker *et al.*, 2006):

- *Degree centrality* is based on the connectivity of a node v. Given the degree of v, its degree centrality can be calculated as

$$C_D(v) = degree(v). \tag{2.1}$$

  C_D determines the amount of direct influence a particular node has on the network.

- *Closeness centrality* measures the centrality of a node v according to its distance to all other nodes in the network. The distance between two nodes $\delta(v, u)$ is defined as the length of the shortest path between them. Closeness is also often defined as inverse distance of v to every other node in the network:

$$C_C(v) = \frac{1}{\sum_{u \in V \setminus v} \delta(v, u)}. \tag{2.2}$$

  The rational behind C_C is that important nodes are typically close to many other nodes in the network to enable a quick communication.

- *Betweenness centrality* is based on the fraction of shortest paths going through a node v. It is formally defined as the number of shortest paths between two nodes, s and t, that pass v, $\sigma_{st}(v)$, divided by the total number of shortest paths from s to t, σ_{st}:

$$C_B(v) = \sum_{s \in V \setminus v} \sum_{t \in V \setminus v} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.3}$$

  C_B measures the control of a node over the flow of information within a network. Nodes that occur on many shortest paths between other nodes, i.e., as bridges between pairs of nodes, are considered to be more central.

- *Eigenvector centrality* measures the relative importance of a node within a network based on the assumption that not all relationships are equally important. The eigenvector centrality of a node v is proportional to the sum of the centralities of its neighbors u.

$$C_E(v) = \frac{1}{\lambda} \sum_{u \in N(v)} C_E(u) \tag{2.4}$$

  where $N(v)$ denotes the set of neighbors of v and λ is a constant. Links to highly central nodes contribute more to the eigenvector centrality of a node than links to less central nodes. A common variation of this measure is the PageRank centrality (Brin and Page, 1998).

Table 2.4 illustrates the different ranking outcomes for an example graph in Figure 2.12. The most important nodes according to degree centrality are F and K. The former node has the highest eigenvector centrality while the latter one is most important with respect to betweenness centrality. In contrast, the highest closeness centrality is assigned to L. The choice of centrality measure depends mainly on the biological network and the underlying question that is studied. Some measures can only be applied to undirected networks while others perform better on denser networks (Koschützki and Schreiber,

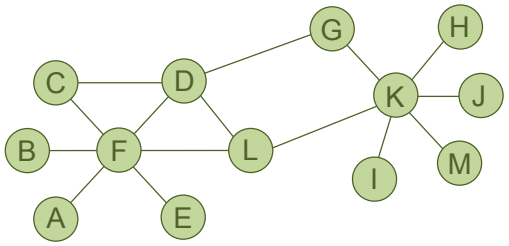


Figure 2.12: Example graph to illustrate the differences in the outcomes of the individual centrality measures. As specified in Table 2.4, the most important nodes according to degree centrality are *F* and *K*. The former node has the highest eigenvector centrality while the latter one is most important with respect to betweenness centrality. In contrast, the highest closeness centrality is assigned to *L*.

2008). In this work, we investigate the four discussed centrality measures (see Section 7.2). Eventually, we apply betweenness centrality to determine the importance of a node in a network as we are particularly interested in identifying bottlenecks, i.e., nodes controlling the flow of information within networks.

Table 2.4: Ranking of the vertices of the example graph in Figure 2.12 according to degree, closeness, betweenness and eigenvector centrality. The five most central nodes are listed for each measure.

	Centrality measure			
Rank	Degree	Closeness	Betweenness	Eigenvector
1	*F/K*	*L*	*K*	*F*
2	*D*	*F*	*F*	*D*
3	*L*	*K*	*L*	*L*
4	*C/G*	*D*	*D*	*K*
5	*A*	*G*	*G*	*C*

2.3.2.4 Modular organization of networks

Modular organization is another important topological feature of biological networks. Complex networks often exhibit modular structures defined by groups of nodes, so-called modules (or cluster), which are more densely connected within each other than across these modules (Newman, 2006). Structural modularity can be divided into two types: classical and hierarchical modularity. The first concept indicates that most nodes within such modules have approximately the same number of links. The latter implicates a power-law distribution of the node links as, for instance, observed in most biological networks.

Modularity in biological networks is presumed to reflect the modular organization of complex cellular function (Hartwell *et al.*, 1999). Highly interlinked subnetworks are believed to represent coherent functional units, e.g., cellular components and their interactions, that accomplish particular functions (Ravasz and Barabási, 2003). In protein

interaction networks, two types of functional modules can be distinguished: protein complexes and dynamic modules (Spirin and Mirny, 2003).

- *Protein complexes* comprise groups of proteins that interact with each other at the same time and place, e.g., transcription factor complexes or RNA splicing machinery.
- *Dynamic modules*, on the contrary, involve proteins that participate in a particular cellular process through consecutive interactions without being co-localized in time and space, such as the CDK/cyclin module responsible for cell proliferation and MAP signaling cascades.

Classifying functional modules into complexes and dynamic modules requires temporal and spatial information which are rarely captured in static protein interaction data. Thus, further evidence, e.g., co-expression data, are typically integrated to distinguish dynamic modules from static complexes (Lin *et al.*, 2010). In this work, we will focus on protein complexes as functional modules in protein interaction networks.

2.3.2.5 Identification of functional modules

The existence of topological modules in cellular networks is commonly reflected by a high clustering coefficient. The clustering coefficient C_v of a node v quantifies its tendency to cluster based on the connectivity of its neighbors. C_v is defined as follows:

$$C_v = \frac{2E_v}{N_v(N_v - 1)} \tag{2.5}$$

where N_v denotes the number of neighbors in the direct neighborhood of v and E_v equals the number of interactions between these neighbors. C_v equals 1 if all neighbors of v interact with each other, and 0 if there are no interactions between the neighbors of v. Based on C_v, the average cluster coefficient $\bar{C}$ of a network can be computed to determine the overall tendency of the nodes to form cluster:

$$\bar{C} = \frac{1}{|V|} \sum_{v \in V} C_v. \tag{2.6}$$

Identifying biologically relevant modules in interaction networks is far more challenging as the concept of modules does not imply clear boundaries between the modules or distinctive module sizes. Distinct methods have been proposed for decomposing networks into functional modules by exploiting either the network topology alone or in combination with functional genomic data (Sharan *et al.*, 2007).

Cluster analysis is a common methodology for extracting functional modules from interaction networks by dividing proteins into groups based on common properties. Distance-based clustering approaches consider different distance measures while graph-based techniques exploit the network topology. Common clustering techniques include the identification of k-cores (Bader and Hogue, 2003), restricted neighborhood search clustering (King *et al.*, 2004), and Markov clustering (Pereira-Leal *et al.*, 2004). More

advanced techniques, such as co-clustering, integrate the similarity of gene expression patterns and network topology into a combined distance measure that is used for hierarchical clustering (Hanisch *et al.*, 2002). Network alignment can also be used to identify functional modules through evolutionary conserved modules. These methods combine the interaction topology and protein similarity to detect protein complexes and pathways that are evolutionarily conserved across different species (Jaeger and Leser, 2007; Kalaev *et al.*, 2008).

The detected functional modules can be validated experimentally or by comparison to well-known manually curated protein complexes and modules (Mewes *et al.*, 2002; Ruepp *et al.*, 2010). However, it is difficult to assess to which extent extracted clusters reflect true modules within an organism.

2.4 Evolution of protein interaction networks

The evolution of biological networks contributed largely to the diversification of living organisms. On an evolutionary time-scale protein interaction networks evolved through two fundamental mechanisms: (i) gene duplications and (ii) gain and loss of interactions through mutations (Berg *et al.*, 2004). The first mechanism contributes primarily to network growth while the latter one accounts for functional divergence.

The duplication of a single gene generates a pair of genes whose products have initially identical binding partners. Duplication events are followed either by gene silencing in which one of the duplicates is immediately inactivated upon formation or by functional divergence of the duplicates (Berg *et al.*, 2004). About 90% of the duplicated genes in yeast are silenced directly after duplication indicating that gene duplication itself does not govern network evolution (Wagner, 2003). Yet, gene duplications occur frequently in eukaryotes at high rates, which accounts for the fact that up to 50% of a eukaryotic genome may consist of duplicate genes (Lynch and Conery, 2000).

Functional divergence after duplication, i.e., acquiring (partially) new function, results from changes in the interaction patterns of the duplicated proteins. Point mutations in their genes affect the interface of the interacting proteins leading to gain and loss of protein interactions. Although duplicated proteins may share interaction partners, the fraction of duplicates without common interaction partners is significantly higher (Makino and Gojobori, 2007). Empirical studies in yeast show that evolutionary rates of duplicates are considerably accelerated shortly after duplication due to their differentiation (Lynch, 2007). In consequence, the number of shared interactions between duplicates decreases according to their evolutionary distance (Wagner, 2001). Different studies indicated that the prevalence of degenerative mutations, i.e., mutational loss of interactions after gene duplications contributes most to the diversification (Wagner, 2003). Moreover, interactions are often lost asymmetrically, where one of the duplicates loses most of its original interactions while the other retains them.

The evolutionary rate of proteins depends on their interaction strength, i.e., transient and stable interactions (see Section 2.2). Proteins involved in the formation of stable complexes have been shown to evolve at similar rates (Fraser *et al.*, 2002). Residues in

their interfaces evolve at slower rate, and appear to co-evolve. This means, substitutions in one protein induce complementary alterations in its interaction partner to preserve the functionality of the interaction (Mintseris and Weng, 2005). In contrast, proteins participating in transient interactions show little evidence of co-evolution and thus are presumed to evolve at different rates.

Overall, gain and loss of protein interactions is the primary evolutionary force which shapes the structure of interaction networks while gene duplications affect, in first place, its size (Berg *et al.*, 2004).

3 Approaches to Protein Function Prediction

The following chapter provides a comprehensive overview on approaches to protein function prediction. We start with a general introduction on protein function emphasizing its importance for the post-genomic era. Subsequently, we briefly discuss traditional experimental methods for elucidating protein function and highlight their limitations for characterizing human proteins. Section 3.2 surveys established approaches that have been developed to circumvent technical and ethical drawbacks of experimental methods by computational means. We explain the most important concepts behind common sequence-, structure and genome-based function prediction methods. Section 3.3 focuses on the principles of network-based function prediction which is a central theme in this thesis. We summarize the benefits and limitations of the distinct methodologies to categorize our proposed protein function prediction approach (see Chapter 4) within the scope of network-based prediction methods.

3.1 Protein function

The large number of genome sequencing projects provide a wealth of knowledge on hundreds of organisms. The interpretation of this wealth of data is a fundamental challenge of the post-genomic era. Completing a new genome is commonly followed by a process known as genome annotation to predict, among others, its protein coding regions and to associate biological information to them (Stein, 2001). Elucidating the functional role of each individual gene product is one of the major challenges in molecular biology and bioinformatics, fundamental to understand biological processes, cellular mechanisms, evolutionary changes and the onset of diseases (Eisenberg *et al.*, 2000; Frishman, 2007).

Traditionally, protein function has been determined for single proteins, one at a time, using classical biochemical and molecular biological experiments. Function derived from, e.g., knock-out experiments, targeted mutations and functional assays (Whisstock and Lesk, 2003), has been commonly reported in the biomedical literature, which in turn is assessed by database curators. Manual curation of such experimental data provides comprehensive and accurate knowledge for genes/proteins (Dimmer *et al.*, 2008) which is widely considered as gold standard for functional annotation.

However, experimental characterization of protein function cannot compete with the pace at which genomic data is being produced (Frishman, 2007). Performing functional assays for each uncharacterized gene in every genome is technically and ethically impossible. This has several reasons:

- Even detailed biochemical studies often cannot identify the full repertoire of functional activities (Whisstock and Lesk, 2003).
- Conclusions from *in vitro* experiments might be limited as particularly eukaryotic proteins cannot be studied in conditions close to their natural environment.
- Knock-out experiments in human beings are prohibited for the obvious ethical reasons.

Annotation of protein function becomes more and more a bottleneck in the progress of biomolecular sciences. The gap between available sequence data and functionally characterized proteins is widening (Frishman, 2007). Even for the best-studied model organisms, such as yeast and fly, a substantial fraction of proteins is still uncharacterized (Sharan *et al.*, 2007). In attempt to close that gap, numerous high-throughput methods have been developed to study the basic properties of gene products systematically. Techniques such as DNA microarrays (Schena *et al.*, 1995; Lockhart *et al.*, 1996), yeast two-hybrid systems (Fields and Song, 1989), RNA interference (RNAi) (Fire *et al.*, 1998; Kamath and Ahringer, 2003) and large-scale systematic deletions (Que and Winzeler, 2002) generated a variety of data sets. However, the huge amount of data, accumulated over the last years, rendered biological discovery via manual analysis impossible (Baumgartner *et al.*, 2007; Dimmer *et al.*, 2008).

Facing these circumstances, scientists turn increasingly toward advanced *in silico* methods for annotating the vast amount of biological data. Numerous approaches have been developed exploiting the different biological data for assigning functions to uncharacterized proteins. Note that today, functional annotation of newly sequenced genomes relies primarily on computational methods (Friedberg, 2006; Pandey *et al.*, 2006; Frishman, 2007; Sleator and Walsh, 2010).

In the following sections, we present distinct computational methodologies for predicting protein function from various types of input data. Before we introduce the different approaches, we will first define biological function and the means of describing it by using standardized machine readable ontologies, such as the Gene Ontology.

Definition of protein function

Function is a highly context-sensitive concept covering all functional activities a gene product may be involved in (Sleator and Walsh, 2010). When speaking of function, one might refer to the molecular, biochemical, cellular, developmental or physiological characteristics of a protein. For instance, the function of a protein kinase, in a biochemical aspect, involves the phosphorylation of the hydroxyl group of specific substrates. In a physiological aspect, the kinase is part of distinct signaling pathways, where proteins both phosphorylate, and are phosphorylated by, interaction partners. A mutation in this kinase might implicate a disease, so yet another functional aspect is a phenotypic one. Clearly, the exact meaning of function depends on the biological context in which it is used (Rost *et al.*, 2003; Friedberg, 2006).

Because of its various facets, the "functions" under study need to be clearly defined to be subject of computational studies. Specifying function in a concise manner is

difficult as it should reflect the complexity of the concept. In first place, functional information is typically not available in machine-readable format but described in the scientific literature using natural language. However, for studying and inferring function computationally, function needs to be presented in a controlled and well-defined format.

To this end, different vocabularies and annotation schemes have been devised to standardize the description of protein function, typically in a hierarchical fashion starting with generic function and progressing toward more specific function. The first systematic scheme, the Enzyme Classification (EC), was proposed in 1992 to classify enzymes based on their enzymatic activity using a four-level hierarchy (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, 1992). Several other functional classification systems emerged (Ruepp *et al.*, 2004; Keseler *et al.*, 2009), often in context with individual species or protein families (see Rison *et al.* (2000); Ouzounis *et al.* (2003) for an overview).

Gene Ontology

In this work, protein function is defined by the Gene Ontology (GO) (Ashburner *et al.*, 2000), the most widely adopted vocabulary for representing function in a systematic manner. GO consists of two components: the ontology itself, defined by concepts and relationships between concepts (GO ontology); and the associations between gene products and concepts (GO annotations). GO covers three major aspects of function, each structured as an independent subontology:

- *Molecular function* describes the fundamental biochemical activities of a gene product at the molecular level.
- *Biological process* describes the series of molecular events or functions that are crucial for the functioning of cells, tissues, organs, and organisms.
- *Cellular component* characterizes the compartments of a cell or its extracellular environment.

Currently, there are about 32,000 concepts defined in GO but more will be included as the ontology continues to mature, see Table 3.1.

Table 3.1: Gene Ontology statistics for its three categories, molecular function (MF), biological process (BP), and cellular component (CC). Data have been retrieved from the Gene Ontology website (January 2011) and its archives (January 2005 and April 2008).

Date	Molecular Function	Biological Process	Cellular Component
2005 (Jan)	6,962	8,924	1,397
2008 (Apr)	8,260	14,659	2,064
2011 (Jan)	8,933	20,188	2,796

Each subontology is modeled as a directed acyclic graph (DAG) where nodes represent GO terms and edges denote the different relationships between them (see Figure 3.1). Initially, two relationship types have been used to link terms: *is a* and *part of*. GO_A

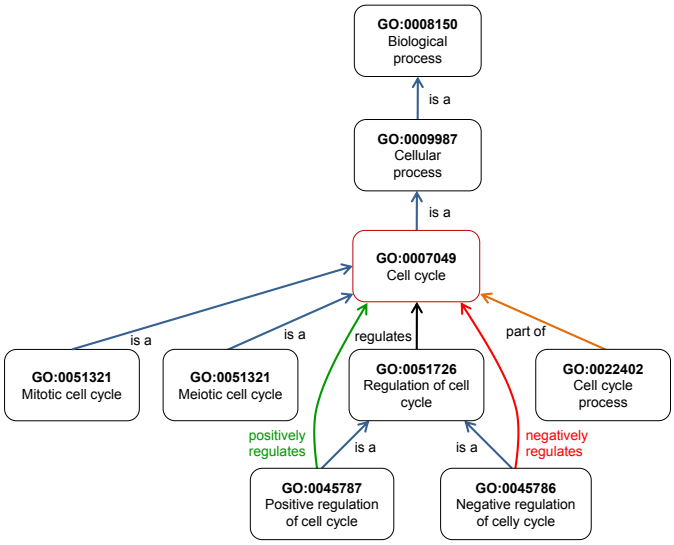


Figure 3.1: Example of the Gene Ontology. Visualization of a small excerpt of the GO subontology *biological process*, i.e., showing the GO term *cell cycle*, including its parent terms, children and the different types of relationships between them.

is a GO_B means that GO_A is a subtype of GO_B, e.g., *mitotic cell cycle* is a subtype of *cell cycle* which in turn is a subtype of *cellular process* (see Figure 3.1). The transitivity of this relation implies that *mitotic cell cycle* is also a subtype of *cellular process*. *Part of* indicates part-whole relationships where a relation is only added if a concept GO_B is necessarily part of another concept GO_A. For instance, whenever *cell cycle process* exists, it is part of *cell cycle*. Hence, the presence of the first term implies the presence of latter one. The *part of* relation has been recently extended by three other types of relationships to distinguish gene products that play more regulatory than direct roles in biological processes (Gene Ontology Consortium, 2010). *Regulates* and its sub-relations *positively regulates* and *negatively regulates* are similarly used to specifically mean necessarily-regulates.

Associating gene products with GO annotations can be performed either manually by database curators or automatically through prediction methods. Each association includes an evidence code referencing the type of information the annotation is based upon (Rhee *et al.*, 2008). Such evidence codes can be broadly divided into four categories: experimental, computational analysis, author statements, and curatorial statements[6]. Out of the many different codes, only one is not assigned by curators but automated methods. Annotations without curatorial judgment are associated with the 'inferred from electronic annotation' (IEA) evidence code.

Annotations derived manually from direct experimental evidence are generally thought to be of higher quality than those inferred from computational or indirect evidence. However, over 98% (September 2009) of the annotations in GO are automatically assigned and have not been curated yet (Gene Ontology Consortium, 2010). To ensure a high

[6] `http://www.geneontology.org/GO.evidence.shtml`

quality basis for function prediction, we only use annotations with curatorial judgment throughout this work and disregard annotations with evidence code IEA.

In summary, GO circumvents the primary shortcomings of natural language descriptions, namely ambiguity and lack of structure, by defining a set of terms and relationships in a controlled and structured manner (Azuaje *et al.*, 2006). This enables the analysis of gene products based on their annotations for inferring functional relationships and common characteristics beyond the traditional sequence-based approaches as we will demonstrate in the following chapters.

Other functional categories

Other functional classification schemes, such as the Yeast Proteome Database (YPD, Costanzo *et al.* (2001)) and the Functional Catalogue (FunCat, Ruepp *et al.* (2004)), have been widely used in the past for manual and automatic genome annotation as well as for systematic analyses of large-scale transcriptome and proteome data. YPD, for instance, covers, similar to GO, three categories of yeast protein function: biochemical function, cellular role and subcellular localization. However, the different categories have only 57, 41 and 22 members, respectively. FunCat, in turn, represents a hierarchically structured, species-independent classification system with 28 categories describing general protein features such as metabolism, cellular transport and transcription. The distinct main categories cover more than 1,300 subcategories which enable a more detailed functional characterization of proteins than provided by YPD.

The size and complexity of ontologies often influences the performance evaluation of computational methods. Evaluating, for instance, function prediction methods on small ontologies increases the likelihood to predict correct terms purely by chance as compared to evaluations using GO in which methods have to choose between up to 20,188 functional categories. Previously reported results on the YPD scheme dropped significantly when applying the same methods to GO (Chua *et al.*, 2007; Jaeger *et al.*, 2010a) (see Section 5.4).

3.2 Computational approaches for protein function prediction

Different data mining and machine learning techniques have been employed to systematically exploit genomic and large-scale experimental data that depict distinct functional protein characteristics. Early approaches used mostly protein sequences as basis. Subsequent methods utilized other types of biological data, including protein structure, phylogenetic and gene expression data, protein complexes and interaction networks.

Function prediction methods can be loosely divided into sequence-, structure- and genome-based approaches (see Figure 3.2). Independently of the data used, most methods rely on the identification, characterization, and quantification of similarities between a protein of interest and proteins for which functional information is available. The challenge of each method is to capture the true relation between the respective protein information and its functional characteristics. This can be typically achieved by (1) inferring relationships from characterized proteins that permit the transfer of functional

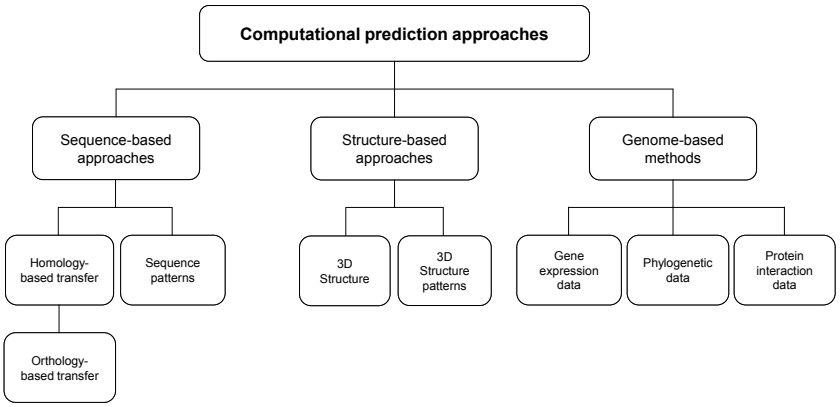


Figure 3.2: Overview on the different protein function prediction approaches (from Sleator and Walsh (2010)). The large arsenal of methods can be loosely divided into sequence-, structure- and genome context-based approaches.

annotations and (2) elucidating the correlation between the detected similarities and the actual level of functional relatedness (Loewenstein *et al.*, 2009).

In the following, we briefly discuss the most important principles behind the different strategies. We refer the reader to Pandey *et al.* (2006) for a more comprehensive survey on automated protein function prediction. In Section 3.3 we provide a detailed overview on methods which rely on protein interaction data since our method is based the same data source for inferring function (see Chapter 4).

3.2.1 Sequence-based approaches

Protein sequences provide the most fundamental information about proteins as their amino acid residues define the structural and functional characteristics of proteins. There are two basic approaches for predicting protein function from amino acid sequences alone: overall sequence similarity and sequence signature patterns.

3.2.1.1 Sequence similarity

Sequence similarity results either from convergence (similarity without common evolutionary history) or descent from a common ancestor, also known as homology. Although similarity due to convergence, often limited to small gene regions, can be useful for some functional predictions (Henikoff *et al.*, 1997), similarity-driven prediction methods are usually based on significant similarities originating from homology.

Homology-based transfer Inferring protein function from homology is based on the assumption that highly similar sequences evolved from a common ancestor and thus have similar, if not identical, functional properties (Whisstock and Lesk, 2003). Homologous sequences can be retrieved from databases using BLAST (Altschul *et al.*, 1997) and function is transferred from the highest scoring homolog(s) to the protein of interest (Tomb *et al.*, 1997).

Albeit the concept is straightforward, homology-based function prediction has severe limitations and systematic errors associated with this paradigm have become increasingly apparent in the databases (Valencia, 2005). Gene duplication, domain shuffling, moonlighting proteins, evolutionary divergence in distantly related species and propagation of incorrect annotations contribute primarily to erroneous function prediction (Friedberg, 2006; Punta and Ofran, 2008).

Orthology-based transfer An important aspect in refining sequence-based function prediction is the distinction between orthologous and paralogous sequences: orthologs originate from a common ancestor through speciation events, while paralogs result from gene duplications within the same genome (Fitch, 1970). Both concepts are well-established, and have been further extended to describe more complex events associated with extensive gene duplications commonly observed in eukaryotic species (Dolinski and Botstein, 2007). Paralogs can be further classified into out- and in-paralogs, denoting genes that have been duplicated either before or after the speciation event, respectively (Sonnhammer and Koonin, 2002). Orthologs and in-paralogs are more likely to retain equivalent or similar function over evolutionary time, while (functionally redundant) out-paralogs have diverged in their functions, e.g., through point mutations and domain recombinations (Li *et al.*, 2003b; Koonin, 2005).

Identifying orthologs is particularly challenging for higher eukaryotes due to their larger genome size, the presence of large protein families, the complexity of protein domain architectures and extensive gene duplications (Dolinski and Botstein, 2007). To address these difficulties, distinct strategies have been employed to distinguish orthologs and in-paralogs from out-paralogs using, for instance, phylogeny, evolutionary distance metrics and bi-directional best hits followed by sequence clustering (Alexeyenko *et al.*, 2006). Depending on their concepts these methods differ in their ability to distinguish orthologs from paralogs which results either in small but functionally pure groups or larger groups that may include out-paralogs.

Deriving functional annotation from the closest ortholog(s) improves the reliability of function assignment considerably (Gabaldón and Huynen, 2004). Yet, most methods are still limited in their predictive power as sequence clustering classifies levels of similarity rather than accurately infers evolutionary relationships (Eisen *et al.*, 1998). Moreover, the coverage provided by identifiable orthologs tends to be smaller than the one achieved by homology detection (Lee *et al.*, 2007).

3.2.1.2 Sequence patterns

Proteins also can be classified by considering only locally conserved sequence patterns, instead of complete sequences. Proteins with related functions but diverged sequences usually share one or more sequence patterns that determine their structure and function (Punta and Ofran, 2008). Such patterns may suffice to preserve the function of the protein even if the sequence evolved further. Also, non-homologous proteins might have acquired the same functional motif convergently (Friedberg, 2006).

Several computational tools extract common distinctive features, i.e., motifs, domains or patterns, from functionally related sequences and provide them in large repositories, such as Pfam (Finn *et al.*, 2010). Newly sequenced proteins can be compared against these resources and, if well-characterized motifs are found, the proteins can be associated to the corresponding family. Functional annotation are then implied by the presence of a particular domain based on curated mappings between domains and GO terms, e.g., Interpro2GO (Camon *et al.*, 2005). More advanced methods use, for instance, classification models (Hayete and Bienkowska, 2005), rule-based predictors (Schug *et al.*, 2002) or probabilistic approaches (Forslund and Sonnhammer, 2008), to predict protein function at the domain level.

3.2.2 Structure-based approaches

The structure of a protein is usually more informative than the underlying amino acid sequence as it is more conserved, particularly in distantly-related proteins (Whisstock and Lesk, 2003). Thus, structural information allows to elucidate functional relationships which could not have been detected even with the most sensitive sequence analysis methods (Skolnick *et al.*, 2000).

Structural data can be utilized in various ways (Watson *et al.*, 2009). Similar to sequence-based methods, two methodological concepts can be distinguished: global and local structural similarity.

- *Global similarity:* Global methods proceed by searching for structurally similar proteins associated with function. Structural alignment tools (Kolodny *et al.*, 2005) compare newly determined structures against structural classification databases or the Protein Data Bank (PDB, Berman (2008)). Proteins with significant structural similarity are likely to share similar or identical functions as structural similarity is a strong indicator for similar function (Shapiro and Harris, 2000).
- *Local motifs:* Proteins with low structural similarity or novel structures are often neglected when considering global similarity for function prediction (Shapiro and Harris, 2000). For such proteins functional information can be deduced by focusing on local structures (Friedberg, 2006). Structurally defined motifs, commonly derived from functionally related proteins, describe conserved functional aspects, such as potential binding or active sites (Punta and Ofran, 2008). Different databases have evolved for searching and recognizing structural features in a protein of interest. Functional knowledge associated with such features can be integrated into functional predictions (Jones and Thornton, 2004; Polacco and Babbitt, 2006).

Functional inference from structure is a promising approach, yet, with a limited scope as only ∼64,500 experimentally solved structures are currently available in the PDB (March 2011). On the other hand, particularly alignment-based methods suffer from analogous limitations as their sequence-based counterparts. Similar structure, for instance, does not always imply similar function and vice versa (Punta and Ofran, 2008).

3.2.3 Genome-based approaches

Proteins without structural or sequence similarity but with related functions are presumed to share other features that indicate functional association. Non-homology-based methods use, for example, subcellular localization (Jensen *et al.*, 2002) while genome-based methods exploit complementary data, such as phylogenetic, gene expression or protein interaction data (Galperin and Koonin, 2000).

3.2.3.1 Gene expression-based prediction

Genes with common biological function tend to exhibit similar expression patterns across different experimental conditions (Eisen *et al.*, 1998; Quackenbush, 2006). Based on this assumption, two techniques are commonly applied for function prediction: clustering (D'haeseleer, 2005) and classification (Asyali *et al.*, 2006). The former technique clusters genes into different functional classes using similarity (or distance) measures defined on the expression behavior, while the latter considers function prediction as a classification problem. Once a group of co-expressed genes has been identified, functional annotation can be inferred using the 'guilt-by-association' principle (Walker *et al.*, 1999). Genes, co-expressed with genes involved in particular cellular processes, are assigned to the same processes using, for instance, the most common annotation or the annotation of the nearest neighbor in the respective cluster (Li *et al.*, 2006; Miozzi *et al.*, 2008).

3.2.3.2 Phylogenetic data

Protein function changes as a result of evolution. Hence, several approaches attempt to reconstruct the evolutionary history of gene products to facilitate the inference of protein function. For analyzing protein function from an evolutionary perspective, evolutionary knowledge is commonly exploited in terms of phylogenetic profiles and phylogenetic trees (Bittar and Sonderegger, 2009).

- *Phylogenetic profiles* represent the evolutionary history of a protein by indicating whether it is present or absent in a set of genomes. Proteins with highly similar profiles are expected to be functionally related (Pellegrini *et al.*, 1999). Thus, function can be inferred by matching the phylogenetic profile of a protein of interest to those with known function.
- *Phylogenetic trees* can be used to encode evolutionary information (Sjölander, 2004). In general, a phylogenetic tree is constructed from homologous sequences. The tree is overlaid with annotations and its topology is used to distinguish orthologs from paralogs. Protein function is then inferred based on the orthologs identified by this process (Brown and Sjölander, 2006).

Function prediction based on phylogenetic profiles (Date and Marcotte, 2005) or trees (Engelhardt *et al.*, 2009) has been shown to perform significantly better than homology-based approaches (Marcotte *et al.*, 1999a). However, phylogenomic inference is not often used in practice, most likely due to the preceding construction of phylogenetic trees. This process is more complicated than simple database searches as it requires

more expertise and computational resources making it impractical for high-throughput applications (Brown and Sjölander, 2006).

3.3 Network-based function prediction

An important source for functional information is provided by protein interaction data. Protein interactions depict, in contrast to sequences, a complementary type of function describing the role of a protein within cells rather than its specific biochemical activity. Further, interaction partners often share similar function. Therefore, protein interaction data are ideally suited to form the basis for function prediction methods (Sharan *et al.*, 2007).

A wide range of methods has been developed for studying protein interaction networks in order to predict protein function (Sharan *et al.*, 2007). Most of them rely on the concept of guilt-by-association, where proteins are annotated based on the function of their interaction partners. Network-based prediction approaches can be categorized into two main classes (see Figure 3.3):

1. *Direct prediction methods* infer novel functions for a protein by transferring known functions from directly or indirectly interacting proteins. This may be achieved by studying the set of neighbors (Schwikowski *et al.*, 2000), considering the position of the protein within its neighborhood (Huynen *et al.*, 2003), or looking at the position of the protein in the entire interaction network (Vazquez *et al.*, 2003; Karaoz *et al.*, 2004).
2. *Module-based methods* assign functions to proteins by first computing clusters (or modules) within the protein network (Bader *et al.*, 2003). Based on the hypothesis that cellular function is organized in a highly modular manner (Hartwell *et al.*, 1999), all members of a cluster are assigned annotations that are enriched within the module (Sharan *et al.*, 2005).

In the following, we will provide a short overview on the basic concepts of direct and module-based methods. A more detailed description of selected approaches can be found in the Related Work of Chapter 4.

3.3.1 Direct prediction methods

The rational behind direct prediction methods is the correlation between network and functional distance: the closer two proteins are in a network, the more similar are their functional annotations (Sharan *et al.*, 2007). Direct methods differ primarily in whether they utilize local or global properties of the interactome to discern and exploit this correlation.

3.3.1.1 Local approaches

Local approaches utilize the close neighborhood of a protein to transfer the most predominant function among these neighbors to the protein of interest (Schwikowski *et al.*,

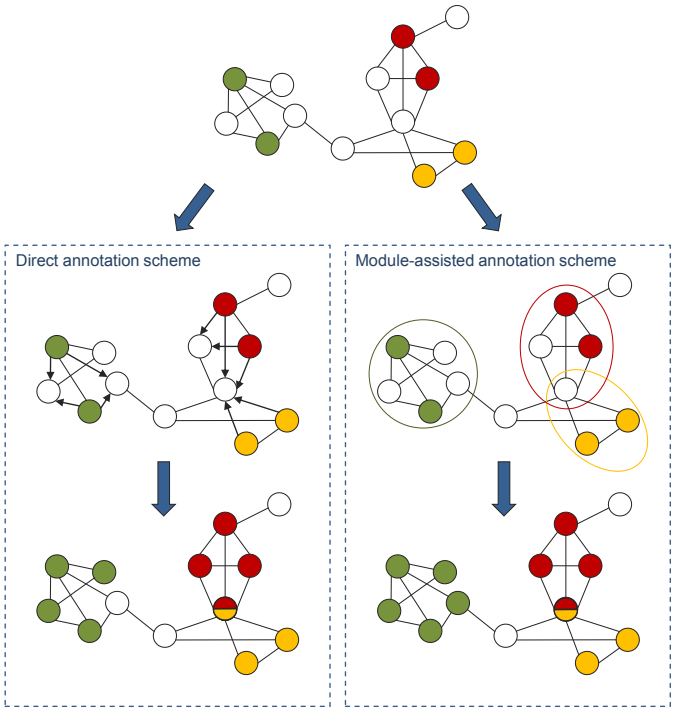


Figure 3.3: Direct versus module-based approaches for functional annotation. The scheme illustrates the principles behind the two basic network-based approaches for function prediction (adapted from Sharan *et al.* (2007)). Proteins with known functions are indicated by different colors and proteins without functions are white. Direct prediction methods (left) infer functions for uncharacterized proteins based on their direct and indirect neighbors. Module-based methods (right), first identify modules based on their density and associate proteins then with functions that are prevalent in the module.

2000). More advanced methods move beyond the direct neighborhood and consider also the local network topology, i.e., by assigning different weights to direct and distant neighbors (Hishigaki *et al.*, 2001; Chua *et al.*, 2006).

Though the neighborhood approach is straightforward, it suffers from several limitations. For instance, the predictive power of local methods is often limited as interaction and/or annotation data are often sparse. The presence of contradictory annotations among neighbors hinders the derivation of coherent predictions and thus compromises the quality of predictions.

3.3.1.2 Global approaches

Global approaches are commonly based on the same concepts as local methods. However, they also take the global network topology into account, usually by computationally more expensive and less intuitive transfer algorithms which employ graph theory and iterative stochastic approaches. Such methods aim to optimize, either directly or indirectly, an objective function which is defined on the entire network determining properties the network should possess once all its proteins have been characterized (Pandey *et al.*,

2006).

Cut-based approaches, for instance, minimize the number of protein interactions across different functional classes or maximize the functional similarity among neighboring proteins in a network (Vazquez *et al.*, 2003; Karaoz *et al.*, 2004). *Flow-based approaches* simulate the spread of function throughout a network by measuring the amount of functional flow each uncharacterized protein receives during the simulation (Nabieva *et al.*, 2005) while *probabilistic approaches* determine the likelihood that a protein performs a particular function depending on the number of neighbors featuring this function and those that do not (Deng *et al.*, 2003).

Global approaches circumvent a number of limitations associated with neighborhood-based methods. Although these techniques are presumed to provide a substantial advantage over the simple guilt-by-association rule, several studies suggest that global methods do not significantly improve on the simpler local approaches (Murali *et al.*, 2006; Chua *et al.*, 2006). These inconsistent observations emphasize the need for common annotation benchmarks and evaluation methodologies to assess the growing number of functional prediction systems.

3.3.2 Module-based prediction methods

This class of prediction methods is based on the hypothesis that cellular functions are organized in a highly modular manner (see Section 2.3.2.4). Module-based methods identify first modules (or cluster) within an interaction network that are likely to represent functional units and associate then functions to the proteins within the module.

Module-based methods differ mostly in their module detection approach; identifying functional units is often their primary intention rather than functional assignment. Most methods use the assumption that proteins within modules are more densely connected than proteins in different modules for identifying modules (Spirin and Mirny, 2003). As discussed in Section 2.3.2.5, graph clustering, hierarchical clustering and decomposing protein interaction networks according to topological properties or evolutionary conservation are common approaches for module detection. A systematic evaluation of several network clustering methods, namely *NetworkBlast* (Kalaev *et al.*, 2008), *CFinder* (Adamcsek *et al.*, 2006), *MCL* (Enright *et al.*, 2002), *DPClus* (Altaf-Ul-Amin *et al.*, 2006), *MCODE* (Bader and Hogue, 2003) and *SpectralMode* (Newman, 2006) revealed substantial differences (Song and Singh, 2009). *NetworkBlast* and *CFinder*, for instance, have been shown to discover high quality modules within dense and well-studied interactomes while *MCL* is more applicable for very sparse and incomplete interaction networks. Overall, there is no clustering approach which is able to consistently outperform the other methods.

Once a module is identified, simple strategies are usually used for predicting function within the module. For example, functional annotations shared by the majority of the module's proteins can be transferred to the uncharacterized proteins in the module. Alternatively, the overrepresentation of a function can be considered using, for instance, hypergeometric distribution. Functions enriched in a module, i.e., having a p-value below some threshold, are then associated with all members in the module (Sharan *et al.*, 2005).

3.4 Conclusion

Direct and module-based function prediction methods have their benefits and their drawbacks as summarized in Table 3.2. To the best of our knowledge, no systematic comparison of network-based function prediction including both methodologies has been performed yet.

Interaction-based prediction methods provide a better coverage than module-based approaches but are more sensitive to the high level of false-positives and false negatives in current interaction data sets (von Mering *et al.*, 2002; Hart *et al.*, 2006). Module-based methods are more robust to missing or wrong interactions, but are often only able to predict function within dense regions of a species network disregarding, for instance, chain-like pathways. This largely reduces their coverage (Bader *et al.*, 2003; Spirin and Mirny, 2003). Module-based methods have been shown to be less accurate than, for example, simple guilt-by-association approaches but their performance improves in networks with less functional coverage (Sharan *et al.*, 2007; Song and Singh, 2009). Local prediction methods are often limited to proteins that interact with characterized proteins while module-based methods are also able to predict novel functions for proteins interacting with proteins of unknown function. Furthermore, both methods in first place only work within a species, which disregards the wealth of information that might be available in evolutionary related species (this is particularly true for humans). This limitation can be circumvented by including annotations of orthology protein relationships underpinned by interologs as we will demonstrate in Chapter 4.

Table 3.2: Overview on benefits and limitations of direct and module-based prediction methods.

Approach	Benefits	Drawbacks
Direct methods	+ High prediction accuracy + High prediction coverage	− Sensitive to FP and FN interactions − Lower performance in networks with low functional coverage − Cross-species information not used
Module-based methods	+ Robust to FP and FN interactions + Performance increase in networks with low functional coverage	− Reduced prediction coverage − Lower prediction accuracy − Cross-species information not used

4 CCS-based Protein Function Prediction

As we have emphasized in the previous chapter, knowing a protein's function is fundamental for understanding the molecular and biochemical processes that sustain health or cause disease. In this chapter, we describe a novel algorithm for protein function prediction based on protein interaction networks that combines the benefits of direct and module-based prediction approaches with orthology to overcome the respective limitations of the individual approaches. The key of our method is to analyze proteins within functional modules that are defined by evolutionarily conserved processes, combining comparative cross-species genomics and functional linkage within species-specific networks. To this end, we first compare protein interaction networks to identify interactions that are highly conserved within a given set of species. These so-called interologs are then assembled to *conserved and connected subgraphs* (CCS).

The underlying assumption of exploiting CCS for function prediction is that proteins and their interactions, forming the topological structure of a CCS, are involved in the same biological context as complex cellular function is assumed to be carried out in a highly modular manner (Hartwell *et al.*, 1999). This functional modularity is thought to be reflected in modular network structures where proteins are grouped according to their biological function as discussed in Section 2.3.2. On the other hand, evolutionary conservation of interaction patterns across several species indicates functional conservation of the underlying processes since proteins and interactions evolve together to preserve their functionality (Mintseris and Weng, 2005). Thus, CCS are presumed to represent functional modules or complexes which are biologically coherent and meaningful. For a given protein, we then predict functions from proteins in the same CCS using both directly interacting proteins within the same species as well as orthology relationships across species.

This chapter is organized as follows: We explain in Section 4.1 how putative functional modules can be identified from a set of protein interaction networks by employing a two-step algorithm. We then describe how conserved functional modules and complexes can be assessed to verify whether they represent biological meaningful and functionally coherent modules. In Section 4.2, we propose a novel approach for functional inference which is based on the previously detected conserved and connected subgraphs. We end the chapter in Section 4.4 with a discussion of related work in the field of network-based function prediction.

4.1 Network Comparison

The key to our prediction approach is to study protein function within evolutionarily conserved subnetworks. To this end, we first compare protein interaction networks across

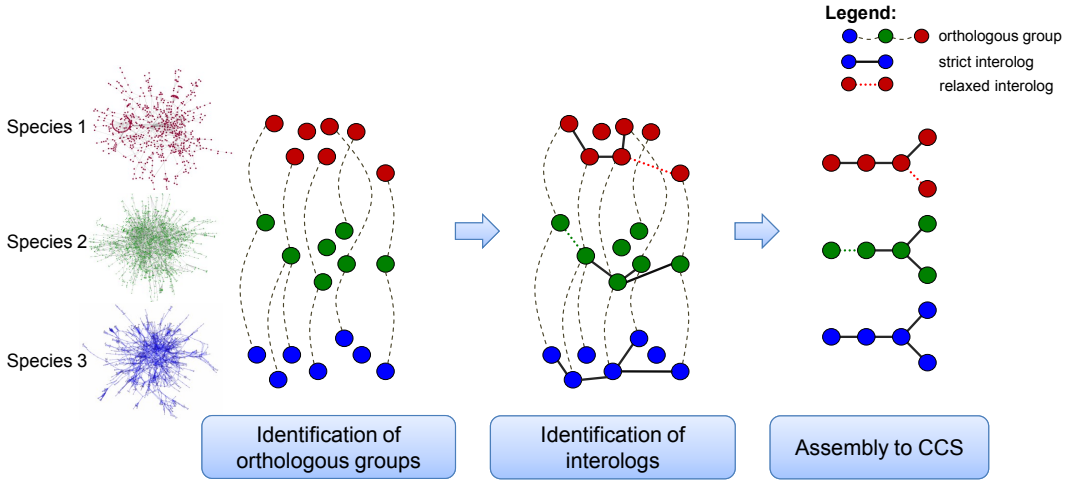


Figure 4.1: Illustration of the detection of CCS. Protein interaction networks are compared across different species to identify evolutionarily conserved and connected subgraphs (CCS). First, orthology relationships across multiple species are determined by using OrthoMCL. Second, all pairs of conserved interactions (interologs) are identified between the orthologs. Adjacent interologs are then assembled to CCS.

different species to detect subgraphs that are evolutionarily conserved. Such subgraphs are presumed to represent functional modules. The proposed algorithm follows a three-step strategy, as illustrated in Figure 4.1 (from left to right):

- First, we identify orthology relationships, i.e., proteins with high sequence homology, across multiple species.
- Second, all pairs of conserved interactions (interologs) are detected between the orthologs within the species.
- Third, adjacent interologs are assembled into maximal conserved and connected subgraphs (CCS).

In the following, we will explain each of the individual steps in detail. Lastly, we present a GO-based scoring scheme which allows to study detected CCS according to their functional coherence.

4.1.1 Identification of orthologous proteins

Orthology is the backbone of our network comparison methodology. As explained in Section 3.2.1.1, orthology detection is particularly challenging for higher eukaryotes. Therefore, we employ an established approach, called OrthoMCL (Li *et al.*, 2003b), for identifying putative orthology relationships across multiple species.

OrthoMCL

This approach has been selected for several reasons:

- First, OrthoMCL discriminates orthologs and in-paralogs from functionally unrelated (out-)paralogs at a reasonable balance of specificity and sensitivity (Chen *et al.*, 2007a), i.e., identifying functionally pure groups while eliminating out-paralogs.
- Second, OrthoMCL is a robust method for excluding out-paralogs that occur when comparing distantly related species to more closely related species, although it does not completely avoid such erroneous assignments (Alexeyenko *et al.*, 2006).

OrthoMCL (Li *et al.*, 2003b) is based on sequence similarity and a Markov clustering (MCL) algorithm to classify protein sequences into families (Enright *et al.*, 2002). Given a set of species, it first performs all-against-all BLASTP comparisons to identify putative orthologous relationships by finding reciprocal best similarity pairs. Subsequently, potential in-paralogs are detected for each putative ortholog as sequences that are (reciprocally) more similar to each other within the same genome than either is to any sequence from another genome. Putative orthologous and paralogous relationships are converted into a graph structure where nodes represent proteins and weighted edges describe their relationships. This graph is modeled as a symmetric similarity matrix to which the MCL algorithm is applied. MCL employs flow simulation to separate diverged paralogs, distant orthologs erroneously assigned by (weak) reciprocal best hits, and sequences with distinct domain structures. An important parameter for the clustering is the markov inflation index that controls the cluster granularity (increasing this index increases cluster tightness, and the number of clusters). We used a comparatively strict inflation value of 4 (default 1.5) to obtain functionally coherent groups. For the BLAST search we used the default E-value cut-off of e^{-5}. Cluster with sequences from at least two species form the final output, and each cluster either represents one-to-one, one-to-many or many-to-many orthology relationships.

We applied OrthoMCL to the distinct interaction data sets described in Section 5.1. Accordingly, the number of orthologous protein cluster differs depending on the number of species being compared as well as on their evolutionary distance and their current interactome coverage. For instance, 3,882 orthologous protein groups are identified between mouse and human of which 2,801 involve at least one protein of each species. These 2,801 groups cover about 91% and 29% of the mouse and the human proteins, respectively, whereas the significant difference in the coverage originates from the different sizes of the protein sets, e.g., 3,701 proteins in mouse vs. 14,218 proteins in human. In contrast, the comparison between human, fly and yeast results in 6,190 cluster of which only 1,114 contain at least one protein of each species (see Table Appendix B, Table B.1).

4.1.2 Detection and assembly of conserved interactions

Once orthology relationships have been identified, we proceed to detect evolutionary conserved and connected subgraphs (CCS) within a set of k species. For each species m we generate a specific protein interaction network by integrating interaction data from various public databases (see Section 5.1) which is represented as a graph G_m with

$m \in \{1, \ldots, k\}$. Given k species-specific protein interaction networks $\mathbb{G} = \{G_1, \ldots, G_k\}$, we aim to find all maximal connected subgraphs $C_s \subseteq G_1, C_s \subseteq G_2, \ldots, C_s \subseteq G_k$ as CCS.

To this end, we consider all orthologous protein groups that comprise at least one protein of each species under consideration and use an adaption of an algorithm for frequent subgraph discovery (Koyutürk *et al.*, 2004). This includes (1) identifying conserved binary interactions, so-called interologs (Matthews *et al.*, 2001), and (2) assembling them into CCS. Both steps are described in detail below, pseudo-code is given in Algorithms 1 and 2.

First, we identify all interactions (interologs) that are conserved across the different species (see Algorithm 1). Note, interaction implies here that at least one protein from each orthologous group interacts with a protein from the other group. For identifying interologs we use two different definitions depending on the number of species that are involved: a strict and a relaxed definition.

- When comparing only two species, we use the classical, strict definition considering only those interactions as interologs that are present in both species.
- Requiring perfect conservation is too strict when studying more than two species, especially due to the incompleteness (Hart *et al.*, 2006) and noise of interaction data (von Mering *et al.*, 2002), evolutionary variation and experimental errors. Therefore, we relax our strict demand on interologs when comparing multiple species and consider each interaction as interolog that is present in at least 50% of the species.

As indicated in Algorithm 1, identifying interologs depends primarily on the number of orthologous protein groups, $|\mathbb{O}|$, and thus has a complexity of $O(k \cdot |\mathbb{O}|^2)$.

To assemble interologs into CCS, one interolog is chosen as subgraph seed and all interologs adjacent to this subgraph are added recursively (see Algorithm 2). If a subgraph cannot be further extended we store this maximal and connected subgraph as CCS. The complexity of assembling interologs is $O(|\mathbb{I}|)$ while the overall complexity of detecting CCS is $O(k \cdot |\mathbb{O}|^2 + |\mathbb{I}|)$.

From an abstract point of view, we are able to identify all CCS, $\mathbb{C} = \{C_1, C_2, \ldots, C_n\}$, that are either perfectly or approximately conserved across a given set of k protein interaction networks $\mathbb{G} = \{G_1, \ldots, G_k\}$. Each CCS, $C_S = (O, I)$, is specified by its set of orthologous proteins (O) and the set of interologs (I). CCS differ in their orthologous proteins and interologs depending on the species they belong to. We refer to a CCS within a particular species m as C_S^m with $m \in \{1, \ldots, k\}$. Accordingly, we denote its species-specific set of orthologs and interologs with O_m and I_m, respectively. An individual orthologous group is indicated by $o_i \in O$ while its species-specific form is denoted by o_i^m. Similarly, interologs between orthologous groups are denoted by $l = (o_i, o_j)$ and $l_m = (o_i^m, o_j^m)$, respectively.

Algorithm 1 Identification of interologs across k interaction networks $\mathbb{G}$

Input: Set of orthologs $\mathbb{O} = \{o_1, \ldots, o_n\}$; Set of interaction networks $\mathbb{G} = \{G_1, \ldots, G_k\}$
Output: Set of interologs $\mathbb{I}$

$\mathbb{I} = \emptyset$
for all $(o_i, o_j) \in \mathbb{O}^2 \quad \wedge \quad i \geq j$ **do**
  count $\leftarrow 0$ {counts the presence of an interaction across $G_1, \ldots, G_k$}
  for all $G_m \in \mathbb{G}$ **do**
    if $(o_i^m, o_j^m) \in G_m$ **then**
      count $\leftarrow$ count + 1
    end if
  end for
  if $|\mathbb{G}| = 2 \wedge$ count $= 2$ **then**
    $\mathbb{I} \leftarrow \mathbb{I} \cup (o_i, o_j)$
  end if
  if $|\mathbb{G}| > 2 \wedge$ count $\geq |\mathbb{G}| * 0.5$ **then**
    $\mathbb{I} \leftarrow \mathbb{I} \cup (o_i, o_j)$
  end if
end for
return $\mathbb{I}$

Algorithm 2 Assembly of interologs $\mathbb{I}$ into CCS

Input: Set of interologs $\mathbb{I}$
Output: Set of CCS $\mathbb{C}$

$\mathbb{C} = \emptyset$
while $\mathbb{I} \neq \emptyset$ **do**
  $I \leftarrow$ {any interolog $I \in \mathbb{I}$}
  $C_S \leftarrow \{I\}$
  $\mathbb{I} \leftarrow \mathbb{I} \setminus \{I\}$
  for all J adjacent to $I \in C_S$ **do**
    $C_S \leftarrow C_S \cup \{J\}$
    $\mathbb{I} \leftarrow \mathbb{I} \setminus \{J\}$
  end for
  $\mathbb{C} \leftarrow \mathbb{C} \cup \{C_S\}$
end while
return $\mathbb{C}$

4.1.3 Functional coherence of CCS

In the previous section we elaborated on the detection of putative functional modules through evolutionarily conserved proteins and interaction patterns. As we identify modules entirely on the conserved topology of the interaction networks, it is important to study whether structural conservation correlates with functional conservation within such modules.

To this end we verify whether CCS present biologically meaningful and functionally coherent modules that can be exploited for function prediction later on. We employ a GO-based scoring scheme to assess the functional coherence within CCS based on the functional annotations of the participating proteins. This measure will be used later to exclude CCS from function prediction that are too heterogeneous due to the noise and incompleteness in the existing interaction and annotation data sets (see Section 4.2.4).

We use an information content based measure to first determine the similarity of two GO terms which is then extended to determine the functional similarity of two proteins annotated with several GO terms. Finally, we compute for each CCS its average functional similarity across the species (Sim_{ortho} – similarity between orthologs) and within a species (Sim_{neigh} – similarity between neighboring proteins).

4.1.3.1 Semantic similarity between GO terms

We use the approach proposed by Lin (1998) to define semantic similarity between two GO terms. Following Lin's definition, the information content of a GO term t is defined as follows:

$$IC(t) = -\log\left(\frac{freq(t)}{freq(root)}\right), \tag{4.1}$$

where the frequency of a term is defined as the number of times a term or any of its descendants occurs. Thus, less frequent terms and terms with few occurring descendants are considered more informative.

Based on this measure, the semantic similarity between two terms is defined as ratio of the information content of their most informative common ancestor and the information content of both concepts (Lin, 1998). The information content of the most informative common ancestor is given by:

$$shareIC(t_1, t_2) = \max\left\{IC(t) | t \in CA(t_1, t_2)\right\}, \tag{4.2}$$

where $CA(t_1, t_2)$ is the set of all common ancestors between terms t_1 and t_2. The similarity between two terms is then defined as:

$$sim(t_1, t_2) = \frac{2 * shareIC(t_1, t_2)}{IC(t_1) + IC(t_2)}. \tag{4.3}$$

Note that $sim(t_1, t_2) \in [0, 1]$ by definition.

4.1.3.2 Semantic similarity between proteins

The semantic similarity between proteins is determined based on the pairwise similarity of their associated GO terms. Since proteins are often annotated with more than one term, the similarity of a protein p_1 to a protein p_2 is defined as average similarity of its terms $t(p_1)$ to the most similar terms in $t(p_2)$ (where $t(p)$ is the set of terms associated

with protein p):

$$Sim(p_1, p_2) = \frac{\sum_{t_1 \in t(p_1)} max\{sim(t_1, t_2) | t_2 \in t(p_2)\}}{|t(p_1)|}. \quad (4.4)$$

It should be noted that $Sim(p_1, p_2)$ is an asymmetric measure as $Sim(p_1, p_2)$ does not necessarily equal $Sim(p_2, p_1)$. The reason is that $t_2 \in t(p_2)$ being the most similar term to $t_1 \in t(p_1)$ does not imply that t_1 is the most similar term for t_2.

We use the definition of Couto *et al.* (2007) to determine the GO similarity between two proteins as the average similarity of their GO terms:

$$GO_{Sim}(p_1, p_2) = \frac{Sim(p_1, p_2) + Sim(p_2, p_1)}{2}. \quad (4.5)$$

4.1.3.3 Functional similarity within CCS

Finally, we determine the functional similarity within CCS. Given the two sources of conservation in CCS, namely orthologs and interologs, we measure functional similarity separately between orthologs across the species (Sim_{ortho}) and between interacting proteins within a species (Sim_{neigh}).

Orthology-based similarity – Sim_{ortho} For functional similarity between orthologs we first compute all pairwise similarities between the proteins within an orthologous group $o_i \in O$. Subsequently, we add all pairwise protein similarities and divide the sum by the number of protein comparisons $n = \frac{k*(k-1)}{2}$ within a group o_i to obtain the average score for o_i:

$$GO_{Sim}(o_i) = \frac{\sum_{m,p(m<p)}^{k} GO_{Sim}(o_i^m, o_i^p)}{n}. \quad (4.6)$$

The individual similarity scores of each group are added and divided by the number of orthologous groups ($|O|$) in the CCS:

$$Sim_{ortho}(C_S) = \frac{\sum_{o_i \in O} GO_{Sim}(o_i)}{|O|}. \quad (4.7)$$

Interaction-based similarity – Sim_{neigh} This measure determines the functional similarity between the interaction partners of one species within a CCS. To compute Sim_{neigh} for species m we first determine the functional similarities between all interacting orthologs of m within the CCS which are then added and divided by the number of edges ($|I|$) in the CCS:

$$Sim_{neigh}(C_S^m) = \frac{\sum_{(i,j) \in I, i<j} GO_{Sim}(o_i^m, o_j^m)}{|I|}. \quad (4.8)$$

Sim_{ortho} and Sim_{neigh} range between 0 and 1, where 1 indicates functional equality and 0 indicates maximal functional distance. CCS lacking protein annotations result in a decreased semantic similarity due to missing annotations. Both measures will be used to filter CCS that do not provide a reliable functional basis for inferring novel functions.

4.2 Prediction of Functional Annotation

This section introduces our novel algorithm for predicting protein function (see Figure 4.2) that uses interaction data from multiple species and combines three different sources of evidence for functional similarity:

- evolutionary conservation of functional modules in protein interaction networks,
- orthology relationships, and
- direct and indirect protein-protein interactions.

While we introduced the first evidence as a filter in Section 4.1, we will discuss the two remaining types of evidence, namely orthology relationships and conserved neighborhood, in the following. First, we present each evidence individually and demonstrate how these evidence can be combined to form a function prediction algorithm.

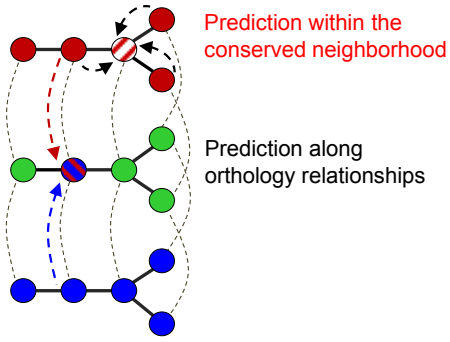


Figure 4.2: CCS-based function prediction. The three complementary approaches, namely orthology relationships, evolutionary conserved functional modules, as well as direct and indirect protein-protein interactions, are integrated into a single prediction strategy.

Additionally, we discuss two strategies to further increase the accuracy of our method: CCS filtering and CCS pre-processing. The first technique filters functionally incoherent CCS by using the GO-based evaluation scheme described in Section 4.1.3 while the latter accounts for large CCS which are, due to their sheer size, usually functionally heterogeneous. We complete this section with a description of the evaluation procedure that shall be used for validation, see Section 4.3.

4.2.1 Prediction using orthology relationships

To predict function from orthology in CCS, we determine orthologous groups that differ significantly in their individual functional similarity from the similarity score of the CCS

by computing the standardized z-score (Freedman *et al.*, 1998). The z-score specifies the difference between the similarity within an orthologous group o and the similarity of the CCS normalized by the standard deviation of the orthologous similarities in the CCS (std_{ccs}):

$$z\text{-}score(o) = \frac{GO_{Sim}(o) - Sim_{ortho}(\text{CCS})}{std_{ccs}} \tag{4.9}$$

Based on the z-score we derive a p-value to determine whether an observed difference is significant. P-values smaller than the significance level of $\alpha = 0.01$ are considered to be significant.

Protein groups that differ significantly (p-value < 0.01) from the otherwise functionally coherent CCS are likely to lack functional annotations. In such groups we transfer all known protein annotations to poorly annotated or uncharacterized orthologs. Note that an orthologous protein group might consist of more than one protein per species (orthologs and in-paralogs). Although all proteins within such a group should be functionally highly similar, this is, probably due to missing or wrong annotations, not always reflected in the data. Therefore, we define the consensus annotation of all proteins of one species in an orthologous group to be the set of all GO terms that are associated to more than half of the annotated proteins of that species in that group. When considering more than two species we combine the species-specific sets of consensus annotations and transfer them to the other proteins in the same group.

4.2.2 Prediction using neighboring proteins

To infer protein function from direct links between proteins we consider the functional annotations of the neighbors of a protein. Given a protein in a CCS, we decide for each GO term associated to any of its neighbors whether it can be also annotated to the protein itself.

Let A be the set of terms annotated to at least one neighbor of a target protein u, and let $N(u)$ be the set of direct neighbors interacting with u. We first determine the functional similarity between u and each of its neighbors $v \in N(u)$ (see Eq. 4.5). We transfer $g \in A$ to u if the number of proteins in $N(u)$ annotated to g, with a functional similarity to u higher than a given threshold t, exceeds a threshold f. Both thresholds have been optimized towards precision using manual grid search. Finally, we set $t = 0.7$ and $f = 0.5$.

This method has the major flaw that for candidate proteins without annotation, we cannot compute the semantic similarity to its neighbors and thus cannot predict novel function. Therefore, we also consider the pairwise functional relations between its interaction partners, assuming that a high functional similarity between indirectly linked interaction partners of the protein has to be reflected in the protein itself. Again, if their pairwise similarity exceeds the threshold t we predict their common GO annotations for our target protein.

4.2.3 Combined CCS-based function prediction

Finally, we integrate the three complementary approaches, namely orthology relationships, evolutionary conserved functional modules, as well as direct and indirect protein-protein interactions, into a single prediction strategy. Any protein that is only weakly and incompletely characterized or not annotated at all is a candidate for CCS-based function prediction. For each candidate we infer novel protein function (a) within functionally coherent CCS by exploiting its (b) orthology relationship across other species as well as (c) the information shared by its neighboring proteins as explained in Sections 4.2.1 and 4.2.2, respectively.

4.2.4 Filtering for candidate CCS

CCS are presumed to comply with functional modules whose proteins participate in the same biological processes and pathways. However, not all detected CCS are qualified candidates for function prediction due to the noise and incompleteness within the existing interaction and annotation data sets. Therefore, we first filter CCS that are simply too small or too heterogeneous to be used for function prediction.

In detail, we only process CCS further which contain more than two proteins as smaller CCS are unlikely to present biologically meaningful processes. Furthermore, we only consider CCS whose similarity score exceeds a given threshold. We determine for each CCS its average functional similarity within a species (see Eq. 4.8) and across the species (see Eq. 4.7), and apply three different thresholds (low: 0.3, medium: 0.5, high: 0.7) to Sim_{ortho} and Sim_{neigh} to study the performance of our method for different levels of functional coherence. This scheme is applied separately for each subontology of GO (molecular function, biological process, cellular component).

4.2.5 Processing large CCS

Comparing evolutionarily close species (such as human and mouse) might result in very large CCS with up to several hundreds of proteins. However, biological processes typically involve only between 5 and 25 proteins (Spirin and Mirny, 2003). Consequently, large CCS often encompass various functions. For instance, the largest CCS between human, fly, worm and yeast illustrated in Figure 4.3 clearly contains several highly conserved clusters, probably forming discrete protein complexes. Functional analysis of its proteins reveals that the CCS encompasses at least four different biochemical activities, e.g., protein degradation, translation, signaling and protein transport, indicating a reduced functional homogeneity. Our results confirm this fact, as large CCS always get low coherence scores (see Section 5.3.4.1).

To adequately treat such CCS, we split CCS with more than 25 proteins into smaller, overlapping sub-subgraphs. Sub-subgraphs are built by considering each protein of the CCS as seed of a new, smaller CCS. We add all direct neighbors of this seed to the new CCS as exemplified in Figure 4.4. Subgraphs with less than three proteins are removed. We then consider each of these subgraphs as an independent CCS.

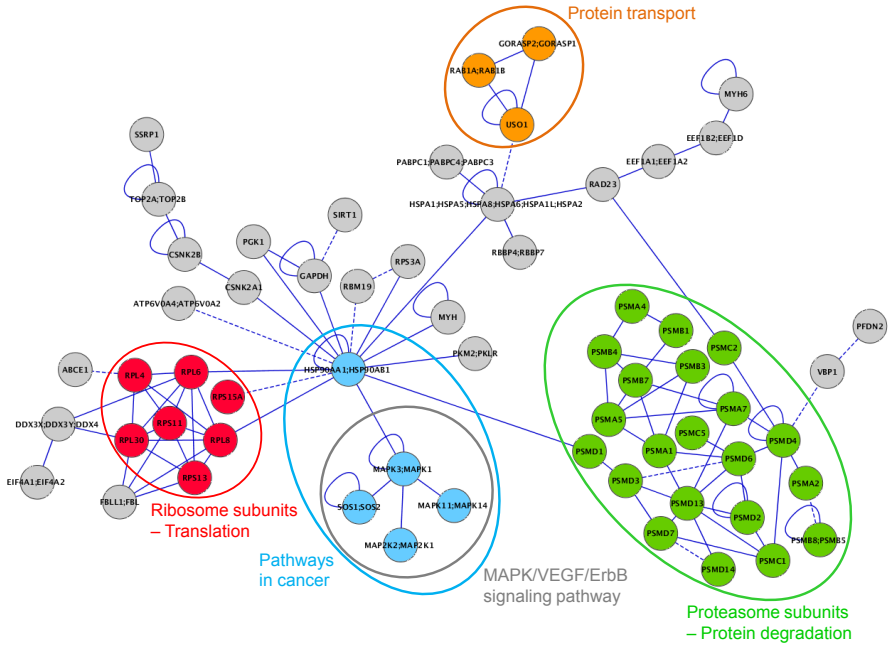


Figure 4.3: Different biological subprocesses within the largest CCS from human, fly, worm and yeast. This CCS consists of 61 proteins and 108 interologs and encompasses different biochemical activities, such as protein degradation, translation, signaling and protein transport.

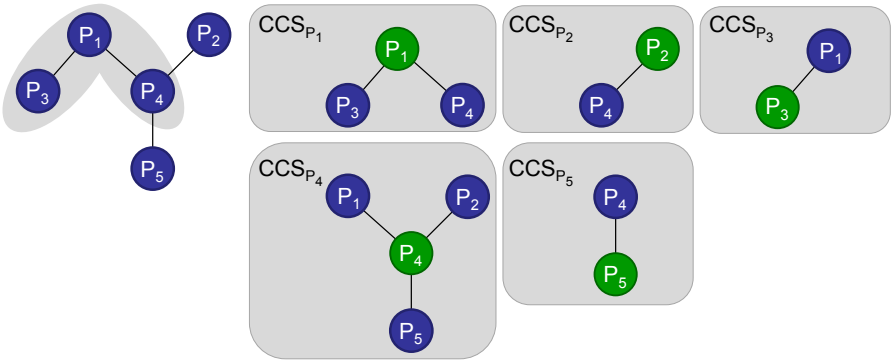


Figure 4.4: Processing large CCS for function prediction. CCS with more than 25 proteins are split into smaller, overlapping sub-subgraphs by considering each protein of the CCS as seed (green node) of a new, smaller CCS. All direct neighbors of this seed are added to the new CCS. Sub-subgraphs with less than three proteins are removed. For example, P_1 is used as seed and its direct neighbors P_3 and P_4 are added to form the new sub-subgraph CCS_{P_1}. Splitting the entire CCS results in five independent sub-subgraphs, but only CCS_{P_1} and CCS_{P_4} are considered further for function prediction while the others are pruned (those having less than three proteins).

4.3 Evaluation methods

To assess the performance of our CCS-based function prediction approach we use *precision* (P) and *recall* (R). Both concepts present well-established measures for evaluating

prediction algorithms. Precision indicates the fraction of correctly predicted annotations, true positives (TP), amongst all predictions, both true positives and false positives (FP):

$$P = \frac{\mathrm{TP}}{\mathrm{TP} \ + \ \mathrm{FP}} \tag{4.10}$$

Recall, on the other hand, depicts the fraction of correctly identified predictions amongst all known functions, true positives and false negatives (FN):

$$R = \frac{\mathrm{TP}}{\mathrm{TP} \ + \ \mathrm{FN}} \tag{4.11}$$

We assess the performance of CCS-based function prediction according to the following criteria (which are explained in detail below):

- First, leave-one-out cross-validation is used to estimate the expected precision and recall of each single and the combined function prediction methods.
- Second, we evaluate our approach by comparing it with two baselines, namely orthology and neighbor baseline.
- Third, we validate our approach against two classical prediction methods: *Neighbor counting* (Schwikowski *et al.*, 2000) and χ^2 (Hishigaki *et al.*, 2001). We also compare it with *FS-Weighted Averaging* (Chua *et al.*, 2006), a method that considers indirect functional associations and topological weights.

Cross-validation and the two baselines are defined below while the detailed description of the three related prediction strategies is provided in the Related Work of this chapter (see Section 4.4.1).

4.3.1 Cross-validation

For cross-validation we blind the known annotations for each protein before applying our algorithm. Predicted terms are then compared to the held out annotations. We count a GO term as correctly predicted if the proposed term is an ancestor of the original term on the path to the root or the term itself. Otherwise, the prediction is considered to be incorrect (false positive).

Precision and recall are determined for proteins within CCS that exceed a given similarity threshold. Note, for all methods involving CCS, we give recall values on the basis of all annotations of proteins within qualifying CCS. We call this measure per-protein recall. It must be distinguished from the traditional per-species recall (Eq. 4.11) which is also used frequently, but which punishes all methods that first filter proteins. When determining the per-protein recall (R_{pp}) we consider only proteins p that are part of a CCS:

$$R_{pp} = \frac{\sum\limits_{p \ \in \ \mathrm{CCS}} \mathrm{TP}_p}{\sum\limits_{p \ \in \ \mathrm{CCS}} \mathrm{TP}_p + \ \mathrm{FN}_p} \tag{4.12}$$

To also give an idea of the per-species performance, we always complement precision and recall values with coverage, which simply counts the total number of predictions.

4.3.2 Baselines

For evaluation we also defined an orthology and a neighbor baseline. The orthology baseline considers only OrthoMCL orthology relationships ignoring structural network conservation. We randomly select 500 orthologous protein groups, remove annotations from one protein in the group and predict their functions using only its orthologs. The neighbor baseline takes only direct interaction partners into account, independent of evolutionary and structural network conservation. For each species we randomly choose one third of the proteins from the corresponding interaction network and exploit their direct neighbors for deriving novel functions. We repeat this procedure 100 times for each baseline and compute average precision and recall including their standard deviation across all runs.

4.3.3 Further evaluations

We shall use the cross-validation setting described above to assess further features of our approach according to the following aspects (see Section 5.3.5):

- First, we study CCS-based function prediction with respect to the three GO subontologies: molecular function, biological process and cellular component, and determine subontology-specific precision and recall. Further, we examine the average depth of predicted terms in the GO hierarchy (see Section 5.3.5.4).
- Second, we assess whether specific GO branches are better predictable than others and if those correlate with evolutionarily conserved functions and processes. To this end, we determine for each GO term a term-specific precision and recall (see Section 5.3.5.4).
- Third, we analyze how CCS-based function prediction performs on proteins without any or with only very little functional information by considering all novel predictions for these proteins which are counted as false positives in the cross-validation (see Section 5.3.5.5).
- Fourth, we study whether there is a difference in the prediction performance between more general genes, such as housekeeping genes, or specific genes. Therefore, we extract tissue-specific and housekeeping genes from microarray studies. Human proteins are then classified according to this list (if possible). Protein-specific precision and recall is determined and compared between the two groups (see Section 5.3.5.6).

Finally, we discuss predicted functions for selected proteins that are highly relevant for colorectal cancer (see Section 5.5). Specifically, we study the gene products MLH1, PMS2 and EPHB4, which receive 14, 16, and 15 novel annotations using our method.

4.4 Related Work

In Chapter 3 we provided a broad overview on the different approaches to protein function prediction. In this section, we review prediction strategies that utilize protein interaction data for functional annotation (see classification scheme in Section 3.3). We describe direct, local and global, as well as module-based methods and discuss their distinctive features with respect to the CCS-based prediction approach.

4.4.1 Direct local prediction approaches

Direct local prediction methods utilize the close interaction neighborhood of a protein to infer protein function.

4.4.1.1 Neighbor counting

Schwikowski *et al.* (2000) proposed the first local prediction method, known as *majority-vote* or *neighbor-counting*, based on the most common function(s) annotated to the direct interaction partners of a protein. Each uncharacterized protein is associated with the $k \leq 3$ most frequent functions of its direct neighbors.

Cross-validation has been performed on yeast interaction data annotated with functional categories from YPD which covers three categories of yeast protein function: biochemical function, cellular role and subcellular localization. During evaluation only cellular role and yeast proteins with at least one annotated interaction partner have been considered. In this setting, *majority-vote* achieves a prediction precision of 72%.

The concept of *majority-vote* is simple but has several drawbacks compared to CCS-based function prediction. First, the poor reliability of protein interaction data (see Section 2.2.2) is not accounted for. Thus, protein function might be derived from false positive interactions without biological relevance. Further, missing protein interactions largely reduce the coverage of this approach. We address both limitations by exploiting only evolutionarily conserved interactions for function prediction. Using interologs excludes spurious interactions and thus increases the quality of the underlying data which in turn improves prediction precision (see Section 5.3.3). On the other hand, missing interactions are implicitly inferred by using the relaxed interolog definition when considering more than two species. Second, *majority-vote* can only derive function for proteins with annotated interaction partners. However, large fractions of uncharacterized proteins interact with proteins of unknown function. We circumvent this restriction by including cross-species information. We further increase the coverage of our method by considering the indirect relationships between the direct interaction partners of a protein. The last drawback concerns the bias of *majority-vote* toward more general functions. Considering the most frequent functions in a neighborhood favors functions that are either less specific or more broadly annotated within the network. Thus, more general functions tend to be associated with proteins (Pandey *et al.*, 2006). In contrast, we consider each function as potential prediction as long as it is supported by sufficient biological evidence.

4.4.1.2 χ^2 approach

The χ^2 algorithm extended the majority approach by moving beyond direct neighborhood and by considering the background frequency of a functional annotation (Hishigaki *et al.*, 2001). This approach analyzes the k–neighborhood of a protein, i.e., all neighbors that can be reached via k links, and computes a χ^2-score for each function. The χ^2-score for a protein u and a function j is determined based on the number of neighbors associated with j, $n_u(j)$, and the expected number of neighbors annotated with j, $e_u(j)$. $e_u(j)$ is defined as $|(N(u))|\, f(j)$ where $N(u)$ corresponds to the neighbors of u and $f(j)$ denotes the frequency of j among all proteins, i.e., the background frequency of j. Given these parameters, the χ^2-score can be calculated as follows:

$$\chi^2_u(j) = \frac{(n_u(j) - e_u(j))^2}{e_u(j)}. \tag{4.13}$$

Those functions with the best χ^2-score are assigned to the protein of interest. Cross-validation has been performed on interaction data from yeast and the functional categories from YPD. The χ^2 approach predicts subcellular localization, cellular role and biochemical function with precision of 72.7%, 63.6% and 52.7%, respectively, when considering either the $k = 1$– or $k = 2$–neighborhood.

Using χ^2 statistics alleviates shortcomings of the simple *majority-vote* and improves the statistical significance of predictions considerably (Pandey *et al.*, 2006; Chua *et al.*, 2007). However, the network topology is not taken into account during the annotation process. Equal weights are assigned to direct and distant neighbors, while in practice immediate neighbors are more likely to share the same function with the protein in question.

χ^2 statistics suffers mostly from the same limitations as *majority-vote*. Contrary to CCS-based function prediction, the quality of experimental interaction data is not taken into account. This reduces, on the one hand, the level of accuracy as function prediction is based on false positive interactions. On the other hand, proteins without available interaction data are neglected which limits the overall coverage of the method. Another advantage of our method is the usage of functional information of established model organisms such as yeast or fly (see Section 5.3.5.3). This allows to infer function for proteins with uncharacterized interaction partners that are disregarded otherwise. Similar to CCS-based function prediction, the χ^2 approach also uses indirect relationships by considering the k–neighborhood of a protein. However, except for biochemical function, the prediction performance decreases significantly when going beyond direct interaction partners. This indicates that noise and redundancies impact the function prediction, if too many neighbors are considered without differentiating between direct and indirect neighbors. By contrast, we only consider the shared functions between indirect interaction partners assuming that these also pertain to the protein under consideration.

4.4.1.3 Functional Similarity Weighted Averaging

A further extension of the simple neighborhood approach has been proposed by Chua *et al.* (2006). In addition to the indirect neighborhood, their method considers the relationship between functional similarity and network distance. Focusing on direct and indirect neighbors (1- and 2-neighborhood), they assign weights to each neighboring protein according to their functional similarity with the target protein, considering the local network topology as well as the reliability of experimental sources. The *Functional Similarity Weighted Averaging* (FS-WA) method predicts functions for proteins based on their weighted frequencies in the neighboring proteins (Chua *et al.*, 2006, 2007).

The performance of FS-WA has been measured in two ways. First, cross-validation has been performed on yeast interaction data and functional categories from FunCat and YPD (see Section 3.1). In their study, the authors do not report specific numbers for precision and recall but only show precision vs. recall graphs for varying FS-Weight thresholds. When using FunCat, the precision ranges from 46% to 85% having a recall of 40% to 10%, respectively. Slightly better results have been achieved for the three YPD categories (Chua *et al.*, 2006). The second evaluation setting uses interaction data from seven different species, including yeast, human, and functional annotations from GO. For yeast, precision ranges from 58% to 92% in a recall interval of 40% to 10%. In contrast, a precision of only 10% to 35% can be achieved for human proteins.

FS-WA differs in several aspects from our CCS-based approach. Similar to the previously described methods, FS-WA considers only one species and does not incorporate cross-species information, i.e., function from orthologs. This reduces the functional coverage of the method, especially when studying species whose proteins are only sparsely annotated. The varying quality of protein interactions is considered by integrating the reliability of the different experimental sources into the similarity weight. In contrast, we use evolutionary conservation to account for the data quality. This presumably reduces the amount of false positive interactions but also accounts for missing data and variations. The latter aspect is not covered by the reliability measure of FS-WA, thus incomplete interaction data remain a problem.

4.4.2 Direct global prediction approaches

Interaction networks are commonly modeled as graphs suggesting the application of graph-theoretic algorithms for their functional analysis. Three main strategies have been followed: cut- and flow-based as well as probabilistic approaches. Although direct local- and module-based methods are more relevant to this work, we also present a global approach to exemplify the basic principles of such methods.

4.4.2.1 Cut-based prediction

A number of approaches utilize the concept of graph cuts, i.e., partitioning the vertices of a graph into disjoint subsets, when considering the entire network, including its topology and functional annotations. Function prediction is formulated as a global optimization

problem to maximize the number of interactions that connect proteins with the same function.

Vazquez *et al.* (2003) assign functional classes to unannotated proteins by minimizing the number of protein interactions across different functional classes. A scoring function measures the number of interacting proteins associated with the same function. This score is associated with any given functional assignment determined for the set of uncharacterized proteins. The contribution to the overall score is computed from the number of protein neighbors annotated with that function. A function σ_u is assigned to each unclassified protein u by minimizing the following scoring function ϵ:

$$\epsilon = - \sum_{(u,v) \in E'} \delta(\sigma_u, \sigma_v) - \sum_u h_u(\sigma_u), \tag{4.14}$$

where E' corresponds to the set of edges incident to two unannotated proteins, $\delta(\sigma_u, \sigma_v)$ is the discrete δ function which equals 1 if $\sigma_u = \sigma_v$ and 0 otherwise, and $h_u(\sigma_u)$ denotes the number of interaction partners of u associated with σ_u. The first term of the optimization problem concerns the unclassified proteins whereas the second one accounts for interactions between unannotated and previously annotated proteins. Simulated annealing is employed to minimize all scoring functions simultaneously.

The approach of Vazquez *et al.* (2003) has been assessed on interaction data of yeast and functional categories from MIPS. For evaluation the function of 40% of the proteins in the network has been removed. The estimated precision for sparsely connected proteins, i.e., with one or two neighbors, is 30% and increases up to 74% for proteins with eight or more interaction partners. In average the prediction precision varies between 60 and 70% for proteins with more than one interaction partner. Leave-one-out cross-validation yields a precision of approx. 80%.

One of the main differences between CCS- and cut-based function prediction is the extent to which the network topology is exploited. Vazquez *et al.* (2003) follow a global approach while we integrate a local method into a module-based approach. Another important aspect is the varying reliability of protein interaction data. In contrast to CCS-based function prediction, the quality of the protein interaction data is not taken into account. Although this method has been shown to perform robustly in the presence of noisy and incomplete interaction data, high levels of error inevitably compromise the quality of predictions. Spurious and missing interactions interfere with the optimization principle of the assignment procedure in which unclassified proteins with unclassified interaction partners must be associated with functions that are consistent with those assigned to their partners. Last, functional information from other species is not incorporated into cut-based function prediction.

4.4.3 Module-based methods

Module-based methods exploit functionally coherent groups of proteins to derive novel functions for uncharacterized members within these groups. As discussed in Section 3.3.2 several methods have been proposed for finding modules. Methods developed specifi-

cally for function prediction differ in their module detection strategy but use a common prediction approach.

One of the first approaches in this category has been developed by Sharan *et al.* (2005). They proposed a framework for analyzing protein interaction networks by integrating interaction data with sequence information to create a network alignment graph. Each node in this graph corresponds to a group of potentially orthologous proteins $(p_1, ..., p_k)$, one of each species, whose sequences are sufficiently similar (BLAST e-value $< 10^{-7}$). Two protein groups $(p_1, ..., p_k)$ and $(q_1, ..., q_k)$ are connected by an edge if and only if one of the following conditions is satisfied with respect to the pairs (p_i, q_i):

1. One protein pair (p_i, q_i) interacts directly with each other while the remaining ones interact indirectly at distance two (through a third protein);
2. All protein pairs interact indirectly with each other at distance two;
3. At least $\max\{2, k-1\}$ of the protein pairs interact directly with each other.

A heuristic search is conducted over the network to detect conserved network structures, i.e., linear paths of interacting proteins and densely connected protein cluster. High-scoring subnetwork seeds are identified within the alignment graph and then extended by using a probabilistic model for scoring candidate subnetworks. The significance of conserved subnetworks is then evaluated by comparing their probability scores against randomized data sets.

Protein function is inferred within the conserved subnetwork whenever (i) the subnetwork is significantly enriched for a particular GO function, p-value < 0.01, and (ii) at least half of its proteins are associated with this function. Cross-validation has been performed for proteins involved in conserved subnetwork between worm, fly and yeast. A prediction specificity of 58%, 60% and 63% could be achieved for GO biological process depending on the species. On the same species combination we achieve an overall precision of 82%, 81% and 79% (see Chapter 5, Table 5.11).

Sharan *et al.* (2005) follow a similar strategy as we do. Evolutionary modules are identified across different species to predict protein function. However, there are several differences between both approaches. For instance, the detection of putative orthology relationships is simplified by only considering sequence similarities between proteins. In contrast, we apply OrthoMCL to discriminate between in- and out-paralogs which is crucial for function prediction (see Section 4.1.1). In addition, reliability estimates for protein interactions are included into the search heuristic and subnetworks with more reliable interactions receive higher scores. Another major difference in their approach is the limitation of the size of detectable network structures to 15 proteins. Contrary to our approach, larger conserved subnetworks are disregarded although those represent strong indicators for conserved functions (see Section 5.3.4.1).

For function prediction, Sharan *et al.* (2005) use the hypergeometric distribution to detect and then transfer significantly enriched functions to all proteins within a conserved subnetwork. By this means, cross-species function prediction is achieved implicitly. Further, these functions are then associated with each member of the module. In contrast, we transfer such function more accurately along defined orthology relationships. Note

that the rationale behind transferring overrepresented protein functions to all proteins within a module might be only valid for GO's biological process. For molecular function and cellular component this concept might be too general. For instance, when considering proteins within a pathway these proteins will be involved in the same biological process but they will also exhibit specific molecular functions. To capture this functional diversity more differentiated approaches are needed to ensure high quality predictions. For this purpose, we consider either the orthologous partner proteins (including in-paralogs) or the direct neighbors and their indirect relationships for retrieving novel functions.

5 Evaluation of CCS-based Protein Function Prediction

In this chapter, we present the evaluation of the CCS-based protein function prediction approach outlined in Chapter 4. We apply our strategy to different sets of species, ranging from species pairs to groups of up to six species. Throughout this evaluation, we will focus on selected species combinations covering different interactome sizes and evolutionary distances to discuss results and properties of our function prediction technique.

Chapter 5 is organized as follows. First, we describe in Section 5.1 the protein interaction data that provide the basis for this evaluation. We characterize the protein interaction networks of the different species by describing distinctive features as well as functional coverage. Section 5.2 describes the outcomes of the CCS assembly across different species and discusses the impact of using either the strict or the relaxed interolog definition. Section 5.3 proceeds with a detailed evaluation of our function prediction method.

We show that combining different sources of evidence for functional similarity between proteins reaches very high prediction precision, especially for multiple species (three and four). For instance, for the combination of human, fly and yeast we achieve a precision of 87%, 84% and 87%, respectively. Furthermore, we predict many novel functions for uncharacterized or only weakly characterized proteins. When integrating novel predictions from different species combinations, our method produces 27,100 novel annotations for human with an estimated precision of 83%, and 10,586 for mouse with 80% precision. However, also weakly characterized proteins of the well-studied species yeast and fly, whose functional annotations are still incomplete, benefit from our method. These results underline the importance and power of combining different function prediction methods into ensembles and of studying different species for deriving novel functions.

We also demonstrate that our predictions are rather specific, which is reflected in a mean GO-depth of 8 for humans and 7 for mice. Systematic evaluations regarding the different GO subontologies and their specific GO branches reveal branches and processes that are more precisely predictable than others. Essential processes associated with housekeeping genes, such as *protein biosynthesis* and *transcription*, yield significant higher prediction precision than more (species-)specific function like *mating behavior* or *response to drug*. Further investigations show that our method performs better on evolutionary conserved genes, such as housekeeping genes. Yet, our method is not limited to well-studied housekeeping proteins as we also predict function for (tissue-)specific proteins with comparable precision.

In Section 5.4 we report on the performance comparison against three recent function prediction methods from the literature, namely Neighbor Counting (Schwikowski *et al.*,

2000), χ^2 (Hishigaki *et al.*, 2001), and FS-Weighted Averaging (Chua *et al.*, 2006). We show that our CCS-based method performs significantly better than those methods in almost all settings we studied, especially in terms of precision.

We complete this chapter with an extensive literature evaluation by manually verifying a number of predictions in the context of colon cancer to confirm our estimated precision values, see Section 5.5. We study the gene products of *MLH1*, *PMS2* and *EPHB4*, which receive 14, 16, and 15 novel annotations through our method, and discuss their relevance with respect to their known cellular function and their role in colorectal cancer. Detailed literature analysis indicates that at least 73% of the novel functions actually are true predictions.

5.1 Protein interaction data

We use interaction data of the well-studied model organisms *S. cerevisiae*, *D. melanogaster* and *C. elegans*, and the mammals *R. norvegicus*, *M. musculus* and *H. sapiens*. Baker's yeast, fruit fly and the nematode worm are widely used model organisms for studying biological phenomena in species that are more difficult to analyze directly. However, higher eukaryotes often possess evolutionarily more evolved features, e.g., the complex immune system, that are unlikely to have direct counterparts in those simple model organisms. Thus, genetically closer models, i.e., mammalian models are often employed to study and develop novel therapeutic strategies in human (Taketo, 2006; Craig *et al.*, 2006).

For each selected species we obtained protein interaction data from the major public databases: DIP, IntAct, MPPI (Pagel *et al.*, 2005), HPRD, MINT and BioGRID (except for BIND[7]). We integrated the interaction data from multiple sources into one centralized database, called PiPa, as the performance of computational analysis methods often depends largely on the completeness of their input data.

PiPa

PiPa is a system for integrating protein interaction (*Pi*Pa) and pathway data (Pi*Pa*) automatically into a homogeneously structured MySQL database. The integration process involves three stages:

- First, PiPa downloads source files from the respective databases, typically in PSI-MI 2.5 XML format.
- Second, each source file is parsed to extract and insert essential information about the proteins and their interactions into the database. As most database identifiers differ between the distinct resources, we map the various protein identifiers to their

[7]Interaction data of BIND have not been considered as the database is no longer maintained, i.e., active curation ended in 2005. Since then a significant amount of interaction information has been corrupted mostly as database identifiers become out of date. Gene and protein identifiers referenced in BIND, for instance, slowly deteriorate as they refer to retired or altered protein identifiers (Isserlin *et al.*, 2011).

unique UniProt IDs. This allows for an unambiguous integration of the different data sets. Protein interactions are then stored as undirected links between proteins while additional information, such as the type of experiment used to confirm the interaction as well as references to PubMed, are associated with each entry in the database, if such information is available.

- In the last step, proteins are annotated with important attributes such as sequence or chromosomal location as well as further information describing their function, domains in their primary sequence, and associations with diseases (see Figure 5.1). In particular, functional information are complemented by using UniProt (UniProt Consortium, 2010) and EntrezGene (Wheeler *et al.*, 2008) as well as species-specific databases, such as FlyBase (FlyBase Consortium, 2003), MGD (Bult *et al.*, 2008), RGD (Twigger *et al.*, 2007), SGD (Hong *et al.*, 2008) and WormBase (Bieri *et al.*, 2007) (see Appendix A for a listing of the different data sources).

In addition to the MySQL database, PiPa features a graphical administration tool to monitor the databases, to trigger updates and to compute statistics on the included data sources.

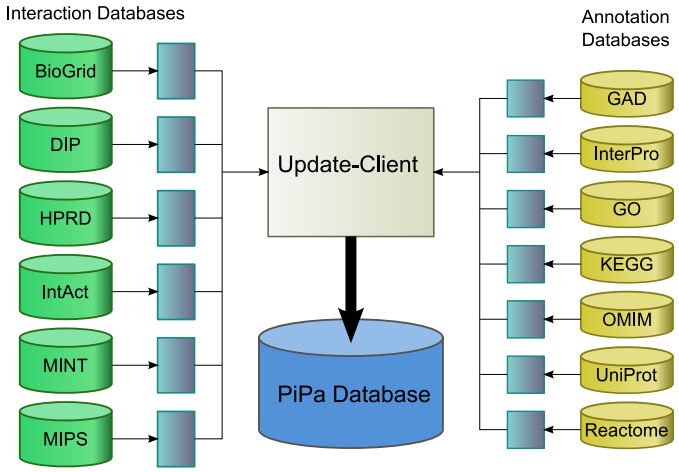


Figure 5.1: Overview of data sources integrated in PiPa.

Protein interaction networks

By means of PiPa we integrated various data sets from multiple sources to generate comprehensive species-specific protein interaction networks. We apply the spokes model when incorporating interaction data from co-complex methods to avoid the inclusion of more false positive interactions (see Section 2.2.1.1). The characteristics of the resulting protein interaction networks are summarized in Table 5.1.

The distinct protein interaction networks differ significantly in their size and complexity. The largest network is assembled for human with 14,218 proteins and 81,868 protein interactions while the yeast network exhibits the highest density with approximately 20.3 interactions per protein and 70,990 interactions in total. The smallest set

Table 5.1: Characteristics of the integrated species-specific protein interaction networks. For each species the total number of proteins and protein interactions as well as the average (and median) number of protein interactions per protein is specified.

Species	Acronym	#Proteins	#Interactions	Interactions per protein
R. norvegicus	*rno*	1,396	1,661	2.3 (1)
M. musculus	*mmu*	3,701	5,582	2.9 (2)
H. sapiens	*hsa*	14,218	81,868	10.3 (4)
D. melanogaster	*dme*	8,272	27,646	5.6 (2)
C. elegans	*cel*	4,364	10,216	3.7 (2)
S. cerevisiae	*sce*	5,845	70,990	20.3 (9)

of interactions is obtained for rat and mouse. The majority of interactions for both species has been obtained from the MIPS–MPPI database which focuses primarily on (manually curated) mammalian interaction data from small-scale experiments. On this account only few interaction data are available yet. Early interaction studies in human have been performed in small-scale to analyze signal transduction pathways or diseases, such as the TGF pathway or Huntington disease (Goehler *et al.*, 2004). However, given the potential of protein interaction for elucidating function and disease mechanisms, more and more large-scale studies have been conducted since then. The variety of small-scale experiments and high-throughput studies contributed to the increasing number of interactions for human (Rivas and Fontanillo, 2010).

For each of the species-specific networks we determined the distribution of interactions per protein as shown in Figure 5.2. The log-log plot emphasizes that most proteins participate in few interactions while some proteins participate in several hundreds of interactions. The human and the yeast network, for instance, comprise 15 proteins with more than 300 interactions, including Ubiquitin-60S ribosomal protein L40 with 345 interactions in human and 713 in yeast, respectively. This analysis demonstrates that, independent of the network size, the generated protein interaction networks exhibit a scale-free topology (see Section 2.3.2.2).

Table 5.2 summarizes the functional coverage of proteins within the protein interaction networks. Proteins of rat and mouse are functionally well-characterized with on average 15.3 and 12.8 GO annotations per protein, respectively. As mentioned above, protein interactions of both species have been mainly detected in specific small-scale experiments. These studies often yield supplementary information regarding, for instance, function, interaction partners and disease phenotypes that help to characterize proteins. Thus, such proteins are often functionally better characterized than proteins analyzed in high-throughput studies. Yeast proteins, for instance, are associated with about six annotations on average (and a median of five). However, in comparison to the other species, yeast proteins have a fairly high functional coverage, mainly due to yeast's role as model organism in various research areas. Human, fly and worm proteins are mostly poorly annotated.

We also investigated whether the number of GO terms annotated to a protein is related to the number of its protein interactions. Figure 5.3 shows that the amount of functional

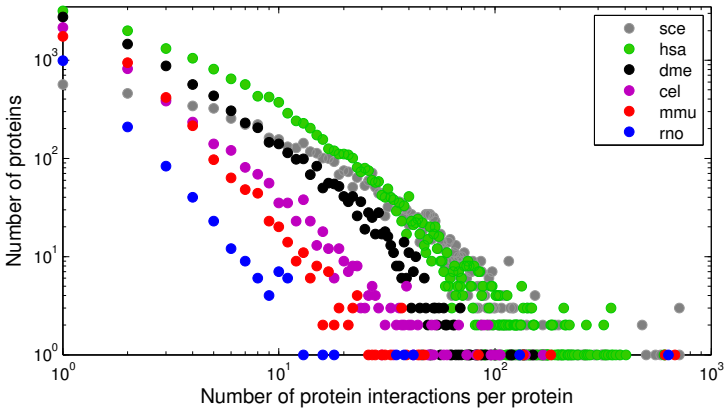


Figure 5.2: Degree distribution in the protein interaction networks. The log-log plot of the node degrees, i.e., number of interactions per protein, in the different interaction networks approximates a power-law distribution indicating that the generated protein interaction networks are scale-free, as discussed in Section 2.3.2.2.

information on a protein does not correlate with its number of interactions. The lack of correlation can be explained by the fact that functional characterization of proteins is more challenging than retrieving protein interactions in large-scale. Yet, a higher correlation is expected as proteins with a large number of interactions are assumed to be multifunctional, i.e., being involved in diverse biological processes.

To complete the overview on the functional coverage of the interaction data, we determined the fraction of proteins with and without function for each species and subontology. Figure 5.4 mostly reflects the conclusions from Table 5.2. The majority of rat, mouse and yeast proteins is associated with at least one function while a large number of proteins in human, fly and worm remains uncharacterized. However, contrary to Table 5.2, Figure 5.4 indicates that yeast proteins feature the best functional coverage as only a small fraction of proteins is completely uncharacterized. Overall, there is only one protein in fly which is not annotated in neither of the three subontologies.

Table 5.2: Functional annotation data for proteins within the protein interaction networks. For each species the average (and median) number of GO terms per protein as well as the average (and median) depths of the respective GO terms is specified in general and per GO subontology: molecular function (MF), biological process (BP) and cellular component (CC).

Species	GO per protein	Depths	MF	Depths	BP	Depths	CC	Depths
rno	15.3 (10)	6.4 (6)	3.1 (2)	4.2 (4)	7.8 (3)	7.1 (7)	4.3 (3)	6.8 (7)
mmu	12.8 (8)	6.5 (6)	2.9 (2)	3.8 (4)	6.7 (3)	7.5 (7)	3.2 (2)	6.9 (7)
hsa	5.8 (3)	6.4 (6)	1.5 (1)	4.0 (4)	2.8 (1)	7.3 (7)	1.5 (1)	7.2 (7)
dme	4.7 (3)	6.1 (6)	1.4 (1)	3.6 (2)	2.4 (1)	7.2 (7)	0.9 (0)	7.3 (7)
cel	3.5 (1)	4.8 (5)	0.6 (0)	2.8 (2)	2.6 (1)	4.9 (5)	0.4 (0)	6.9 (7)
sce	6.0 (5)	6.2 (6)	1.8 (2)	3.6 (3)	2.2 (2)	7.2 (8)	1.8 (2)	7.4 (7)

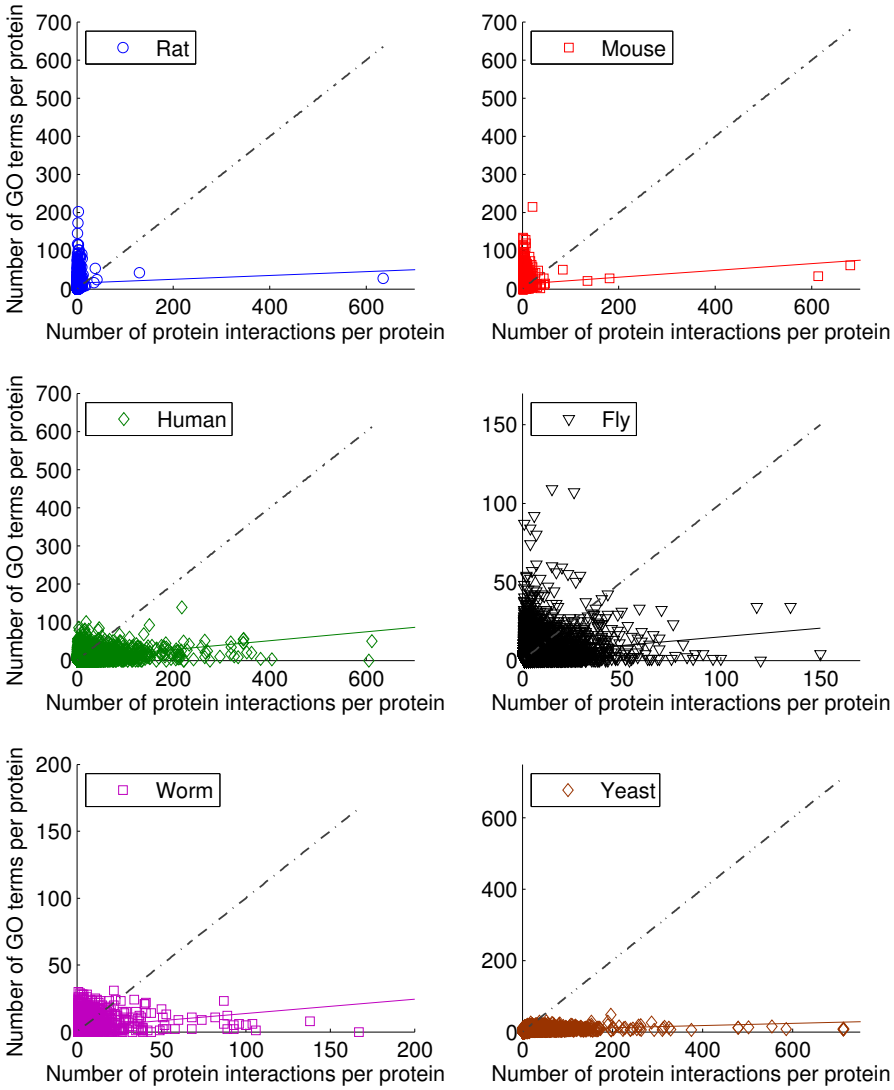


Figure 5.3: Functional characterization of a protein with respect to its number of protein interactions. For each protein in a species, the number of protein interactions is plotted against its overall number of GO terms. Dotted lines represent the angle bisector and thus a perfect linear correlation between both data sets while solid lines represent the regression of the data indicating the actual correlation.

5.2 Network comparison

After constructing protein interaction networks for rat (*rno*), mouse (*mmu*), human (*hsa*), fly (*dme*), worm (*cel*) and yeast (*sce*) (see Table 5.1) we performed network comparisons across different species combinations, ranging from pairs to groups of up to six species, to identify evolutionarily and functionally conserved subgraphs. Conserved sub-networks are assembled by combining conserved interactions, called interologs, using

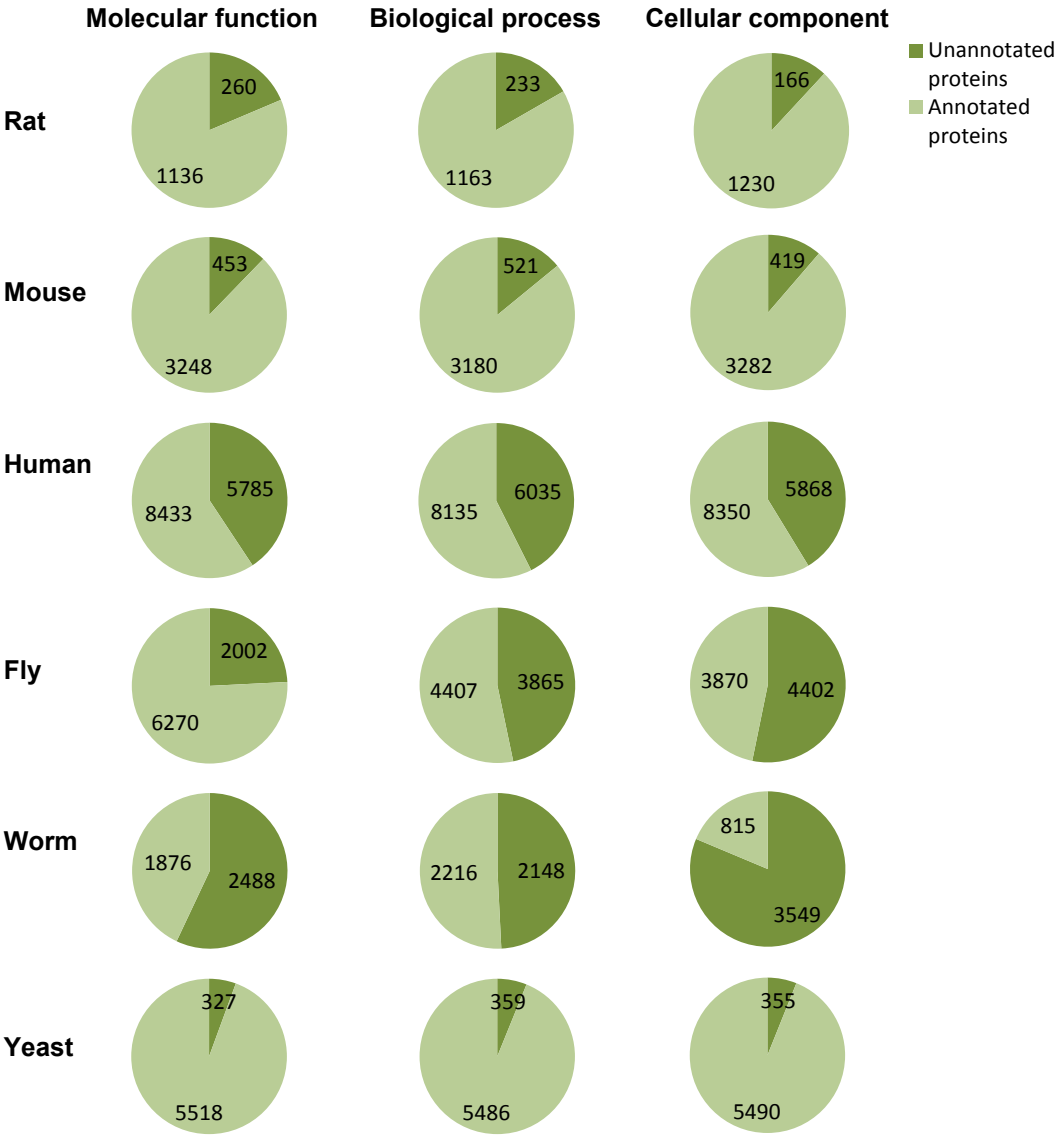


Figure 5.4: Functional coverage of proteins in the interaction data. For each species, the fraction of proteins with and without function is specified with respect to the three subontologies: molecular function, biological process and cellular component.

two different definitions of interologs as described in Section 4.1.2.

1. In the classical, strict definition an interolog is defined as an interaction present in any species under consideration.
2. The relaxed definition defines an interolog as interaction which is present in at least 50% of the species being compared.

Overall, we computed CCS for 15 combinations of two species, 20 combinations with three, 15 with four species, six with five species, and one with six species, for both in-

Table 5.3: Overview on the outcomes of the selected network comparisons. For each combination the number of orthologous groups, interologs and CCS from strict and relaxed definition is presented as well as the size of the largest CCS.

Species	Criteria	# OrthoMCL groups	# Interologs	# CCS (≥3)	Largest CCS Proteins (Edges)
mmu sce	strict	551	126	48 (17)	13 (26)
	relaxed		4762	2 (1)	539 (4761)
hsa dme sce	strict	1114	119	65 (12)	13 (12)
	relaxed		959	127 (23)	344 (727)
hsa dme cel sce	strict	552	22	19 (3)	3 (2)
	relaxed		477	67 (12)	200 (372)
mmu hsa dme sce	strict	395	16	11 (3)	4 (3)
	relaxed		433	53 (14)	146 (324)
rno hsa dme cel sce	strict	206	0	–	–
	relaxed		35	19 (2)	11 (12)

terolog definitions. In the following, we will focus on five selected species combinations covering different interactome sizes and evolutionary distances. By means of these examples we discuss results and properties of our function prediction approach. An overview on the respective network comparison of *mmu-sce*, *hsa-dme-sce*, *hsa-dme-cel-sce*, *mmu-hsa-dme-sce* and *rno-hsa-dme-cel-sce* is presented in Table 5.3. For each combination we summarized results for using strict and relaxed interolog definition. Complete results are given in Appendix B, Table B.1.

As expected, the number of orthologous protein groups, interologs and identified CCS differ significantly depending on the number of compared species, their evolutionary distance as well as their current interactome coverage. Strict comparison of mouse and yeast, for instance, results in 17 CCS (out of 48) qualifying for function prediction. In contrast, (strict) multiple comparisons do rarely result in any or only very few CCS with two or more proteins (see Table B.1). While *hsa-dme-sce* results in 12 qualifying CCS (out of 65), we yield only few rather small CCS (three to four proteins) for *hsa-dme-cel-sce* and *mmu-hsa-dme-sce*. In turn, only few predictions can be derived for such a small number of CCS (see Section 5.3.5.1).

The usage of the relaxed interolog definition considerably increases the number and size of the detected CCS (see Table 5.3 and B.1). For instance, we identify 53 CCS for *mmu-hsa-dme-sce* of which 14 comprise more than two proteins. These CCS are shown in Figure 5.5. On the contrary, we only found 11 CCS using the strict interolog definition of which only three were larger than two proteins. Similarly, species combinations without any qualifying CCS from strict comparison do now result in a reasonable number of CCS. Figure 5.6 exemplarily illustrates the increasing number and size of CCS for combinations of three species.

Strikingly, even combinations with five and six species yield qualifying CCS. For instance, *mmu-hsa-dme-cel-sce* produces nine CCS with more than two proteins. A com-

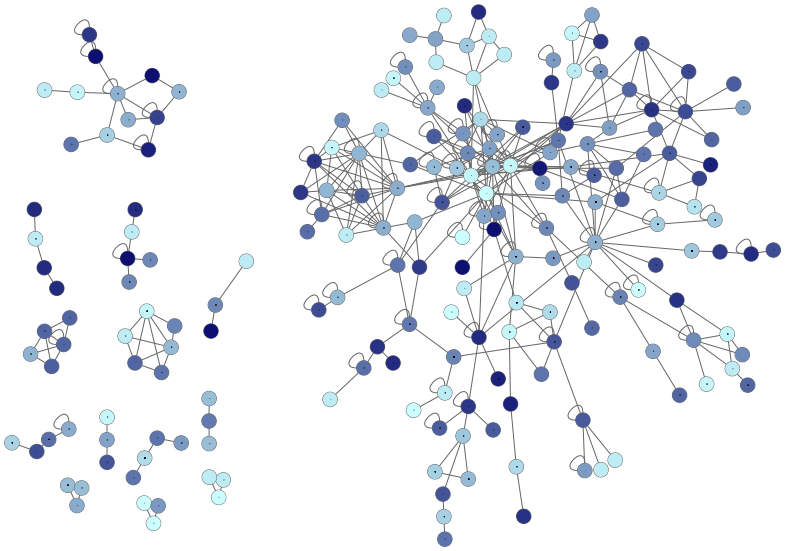


Figure 5.5: CCS conserved among mouse, human, fly and yeast. 14 (out of 53) subgraphs with more than two proteins are approximately conserved between the four species. The largest CCS comprises 146 proteins and 324 interologs. Proteins are colored according to their functional coverage in human within molecular function; from no function (light blue) to functionally well-characterized (dark blue).

bination of all six species yields still two CCS (out of 15) with the largest comprising five proteins (see Table B.1). These observations emphasize that being less strict allows for a significant improvement of the coverage of our network comparison method. The impact of the relaxed interolog definition on CCS-based function prediction will be discussed in Section 5.3.5.1.

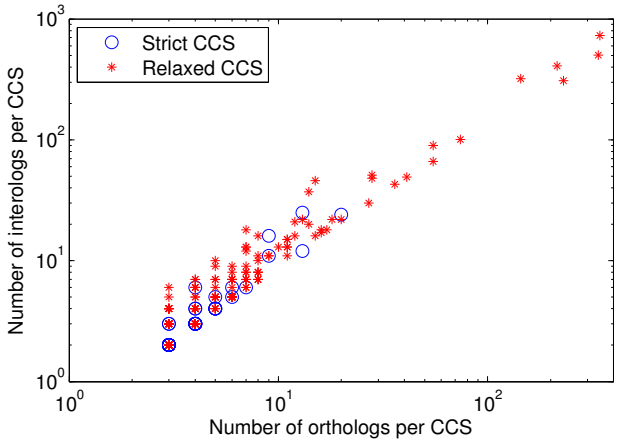


Figure 5.6: Comparison of the number of qualifying CCS from strict and relaxed network comparison for combinations of three species. Strict CCS are represented by circles while relaxed CCS are presented by stars. The size of each CCS can be determined from x- and y-axis which show the number of orthologs (x-axis) and interologs (y-axis), respectively.

We also tested the effect of applying the relaxed definition of interologs to species pairs. This leads to very few (often only one) yet very large CCS (e.g. 539 proteins and 4,761 interologs for *mmu-sce*), as it only creates the union of interactions between orthologous proteins of the two species. However, this does not reflect evolutionary conservation of protein interactions and therefore misses the important signals of functional conservation. For this reason, we use the classical, strict definition for pairs of species and the relaxed definition when comparing multiple species.

5.3 Protein function prediction

We use orthology relationships, functionally conserved modules as well as direct and indirect protein interactions to predict functional annotations for proteins in qualifying CCS. Qualifying CCS are those (a) with more than two proteins and (b) exceeding a given similarity threshold to ensure evolutionary and biological conservation. Precision and per-protein recall, as defined in Section 4.3, are computed for the following thresholds: low (0.3), medium (0.5) and high (0.7) functional coherence. Keep in mind that, as always when comparing to an incomplete gold standard, cross-validation inherently considers any new annotations as false, although new annotations are the primary target of function prediction. This penalizes any novel prediction the algorithm derives, even if it is correct. In consequence, the observed number of false positives is higher than in reality. Thus, precision and recall documented throughout the evaluation present lower bounds as both are typically underestimated.

We evaluated our approach in several ways. First, we compared our combined strategy to baseline methods which disregard conservation in networks. Second, we compared it to the results obtained from using orthology and PPI neighborhood within CCS in isolation. Third, we performed a comparison to three recent function prediction methods from the literature. Further, we assess several important features of our function prediction method and discuss their impact on the prediction performance.

5.3.1 Baselines

We first show the performance of our two baseline methods, orthology and link-based, for function prediction (see Section 4.3.2). Precision for predictions based solely on orthology relationships varies between 22% and 70% (see Table 5.4). Precision rises the more species are considered. Recall is roughly the same for two and three species combinations (34% to 61%), but decreases steeply with the number of species being compared reaching, for instance, 8% to 16% for *rno-hsa-dme-cel-sce*. Precision of the link-based baseline ranges from 23% to 48%. Contrary to the orthology baseline, recall is mostly low, varying between 4% and 41% (see Table 5.5).

5.3.2 Orthology Relationships in CCS

We use orthology relationships underpinned by interologs to infer novel functions from multiple species. Considering only orthology relationships for transferring functions to

Table 5.4: Baseline for utilizing OrthoMCL orthology relationships for function prediction. Precision (P) and recall (R) are estimated from randomly sampling 1/3 of the OrthoMCL groups from a species combinations and predicting function along orthology within the groups. Results are averaged across 100 runs.

Species	# Terms	∅ P (± Std)	∅ R (± Std)
mmu	103573	0.52 (0.01)	0.61 (0.01)
sce	161855	0.38 (0.01)	0.38 (0.01)
hsa	18643	0.45 (0.02)	0.55 (0.01)
dme	17496	0.43 (0.02)	0.44 (0.01)
sce	12169	0.69 (0.01)	0.34 (0.01)
mmu	5654	0.62 (0.03)	0.29 (0.02)
hsa	6400	0.52 (0.03)	0.28 (0.02)
dme	6411	0.48 (0.04)	0.34 (0.02)
sce	4634	0.70 (0.03)	0.30 (0.02)
hsa	3318	0.42 (0.03)	0.09 (0.01)
dme	2289	0.51 (0.04)	0.11 (0.01)
cel	5643	0.22 (0.03)	0.17 (0.01)
sce	1686	0.60 (0.03)	0.09 (0.01)
rno	662	0.63 (0.06)	0.08 (0.01)
mmu	1215	0.49 (0.08)	0.08 (0.02)
hsa	968	0.50 (0.05)	0.10 (0.02)
dme	2010	0.25 (0.04)	0.16 (0.02)
cel	666	0.68 (0.05)	0.08 (0.01)

Table 5.5: Baseline for link-based function prediction within species-specific PPI networks without utilizing interologs. Precision (P) and recall (R) are estimated from sampling randomly 1/3 of the proteins of each interaction network independently of any species combination. Results are averaged across 100 runs.

Species	# Terms	∅ P (± Std)	∅ R (± Std)
rno	3153	0.44 (0.04)	0.04 (0.007)
mmu	22901	0.31 (0.01)	0.10 (0.005)
hsa	141129	0.27 (0.02)	0.29 (0.006)
dme	30445	0.23 (0.01)	0.12 (0.005)
cel	8960	0.30 (0.02)	0.14 (0.011)
sce	58653	0.48 (0.01)	0.41 (0.008)

proteins within CCS results in predictions with medium to high precision. Table 5.6 shows precision and per-protein recall estimated for the selected examples. Precision reaches 85% to 93% for yeast proteins and 76% to 81% for mouse proteins when comparing *mmu-hsa-dme-sce*. Precision values increase considerably with higher coherence thresholds for CCS (see Section 5.3.4.2), but this improvement comes at the cost of lower coverage, that is, less predictions. Contrary to the coverage, the per-protein recall increases with the functional coherence. This behavior differs from the standard recall which typically decreases when filtering for certain criteria, i.e., less proteins are taken

into account at higher thresholds. Yet, annotations of proteins passing the thresholds are recovered with higher coverage. Particularly low numbers of predictions are obtained for comparisons involving species with low interaction coverage. Low coverage limits the number of qualifying CCS which, in turn, limits the number of function predictions. This is especially prominent for *rno-hsa-dme-cel-sce*, where the network comparison results in two qualifying CCS yielding only 5 predictions for *rno*, *hsa* and *sce* – but with a precision of 100%.

Besides the coherence threshold, also the number of species being compared has a strong impact on performance. Higher precision is achieved when analyzing more species. For instance, we achieve an average precision of 71% for *mmu-hsa-dme-sce* at the lowest threshold of 0.3 which is almost as good as precision for *mmu-sce* with 76% at the highest coherence threshold of 0.7. This shows that using more species implicitly selects functions that are conserved more strongly, which underlines the impact of evolutionary conservation for protein function (see Section 5.3.4.2). This fact also shows up when comparing to the orthology baseline (Table 5.4): Precision and per-protein recall using orthology within CCS are much higher, in particular for medium and high functional conservation, but the overall coverage is much lower. This means that CCS strongly restrict the number of proteins for which predictions are made, but this restriction is done in a very sensible way removing mostly false positive predictions.

Table 5.6: Prediction results from exploiting only orthology relationships within CCS derived by exact (pairs) and approximative (multiple) network comparisons. Precision (P) and per-protein recall (R_{pp}) are estimated for low (0.3), medium (0.5) and high (0.7) functional similarity/conservation thresholds. Missing numbers (−) indicate combinations where no CCS is homogeneously enough to pass the respective similarity threshold.

	0.3			0.5			0.7		
Species	# Terms	P	R_{pp}	# Terms	P	R_{pp}	# Terms	P	R_{pp}
mmu	7488	0.41	0.55	4742	0.50	0.63	763	0.71	0.75
sce	5368	0.46	0.47	3344	0.57	0.49	478	0.81	0.74
hsa	26156	0.54	0.32	698	0.55	0.32	52	0.71	0.34
dme	20926	0.59	0.41	695	0.59	0.39	67	0.70	0.80
sce	14619	0.81	0.34	462	0.81	0.34	57	0.83	0.34
mmu	8486	0.76	0.27	486	0.78	0.56	207	0.79	0.57
hsa	9712	0.63	0.24	447	0.75	0.56	181	0.76	0.59
dme	8403	0.64	0.30	382	0.93	0.68	148	0.99	0.66
sce	5931	0.85	0.25	413	0.90	0.64	168	0.93	0.87
hsa	791	0.57	0.15	8	0	0	0	–	–
dme	390	0.74	0.13	0	–	–	0	–	–
cel	853	0.25	0.30	21	0	0	0	–	–
sce	293	0.92	0.11	0	–	–	0	–	–
rno	171	0.90	0.18	5	1.00	0.09	0	–	–
hsa	271	0.69	0.19	5	1.00	0.18	0	–	–
dme	344	0.51	0.22	12	0.54	0.71	0	–	–
cel	187	0.78	0.13	7	0.71	0.62	0	–	–
sce	189	0.46	0.21	5	1.00	0.36	0	–	–

5.3.3 Neighborhood in CCS

The second part of our approach infers protein function by considering the conserved interaction neighborhood within CCS. Table 5.7 shows precision and recall for inferring functions from interaction partners only. Compared to predicting function based on orthology within CCS, precision is higher, while per-protein recall roughly stays the same. At the same time, neighbor-based prediction has a considerable better coverage. For instance, when considering *rno-hsa-dme-cel-sce* about 1,250 predictions are retrieved at 0.7 for human with an estimated precision of 99%. Precision and per-protein recall correlate again with the functional coherence of CCS and the number of compared species (see Section 5.3.4.2), but the impact of multiple species is less pronounced. Compared to the link-based baseline (see Table 5.5), considering CCS also leads to a clear and significant increase in precision and per-protein recall at any threshold while coverage decreases. This effect can be explained by the fact that using interologs (strict or relaxed) instead of single interactions largely improves the reliability of protein interactions (Saeed and Deane, 2008), since false positive interactions are unlikely to be reproduced across multiple species.

Table 5.7: Precision (P) and per-protein recall (R_{pp}) for function prediction along interactions within CCS derived by exact (pairs) and approximative (multiple) network comparisons.

		0.3			0.5			0.7	
Species	# Terms	P	R_{pp}	# Terms	P	R_{pp}	# Terms	P	R_{pp}
mmu	2736	0.70	0.20	1739	0.79	0.26	993	0.81	0.30
sce	3342	0.86	0.22	3325	0.86	0.24	2565	0.88	0.42
hsa	43484	0.43	0.41	2009	0.79	0.47	881	0.85	0.49
dme	16393	0.56	0.28	3088	0.70	0.34	961	0.91	0.60
sce	17194	0.79	0.35	17056	0.79	0.50	2130	0.89	0.50
mmu	12072	0.55	0.29	4130	0.73	0.48	320	0.91	0.49
hsa	18395	0.44	0.32	672	0.81	0.65	564	0.91	0.69
dme	9335	0.54	0.26	1626	0.70	0.30	846	0.86	0.63
sce	8783	0.74	0.30	8725	0.74	0.31	901	0.87	0.54
hsa	20567	0.51	0.35	344	0.49	0.18	0	–	–
dme	8922	0.52	0.22	3981	0.67	0.34	230	0.83	0.34
cel	4284	0.56	0.24	4111	0.56	0.24	72	0.65	0.13
sce	8016	0.82	0.30	7134	0.83	0.31	689	0.91	0.34
rno	694	0.96	0.48	609	0.99	0.67	609	0.99	0.67
mmu	1261	0.99	0.71	1261	0.99	0.71	1248	0.99	0.72
hsa	412	0.63	0.21	219	0.74	0.62	0	–	–
dme	116	0.87	0.20	116	0.87	0.20	116	0.87	0.53
cel	433	0.97	0.42	433	0.97	0.42	433	0.97	0.59

5.3.4 Combining module, orthology and link-based PPI evidence

We hypothesized that the integration of orthology relationships, evolutionary conserved functional modules, and direct and indirect protein-protein interactions into a single prediction strategy will combine the strengths of the three individual methods. To this end, we unify predicted functions from each approach. Selected results from this combined strategy are shown in Table 5.8. As before, precision varies (from 42% to 99%) depending on the species combination and the threshold for functional coherence. Best results are obtained for *hsa-dme-sce* at a threshold of 0.7, with precision of 84%, 89% and 89%, respectively, as well as *rno-hsa-dme-cel-sce* with precision of 95%, 99%, 87% and 97%, respectively.

As mentioned before, one of the major drawbacks of using only CCS orthology relationships is the low number of predictions due to the restriction to orthologous proteins with at least one known function (see Table 5.6). In contrast to orthology-only, the combined approach creates much more predictions (2- to 10-times more). It generates hundreds or even thousands of predictions also for those cases where the orthology-only method could not predict any function.

Comparing the combined method and CCS link-based only (see Table 5.7) shows an increase within the amount of predictions (e.g. about 2-times for *mmu* from *mmu-sce*), although it is less steep than observed for orthology-only. This increase has mostly only minor influence on precision and recall. Precision reaches similar levels and the recall increases slightly. Note, for few combinations the combined method yields the same results as link-based-only because no predictions could be inferred through orthology relationships.

5.3.4.1 Impact of large CCS

One essential feature of our prediction strategy remains to be discussed: the processing of large CCS. Large CCS with more than 25 proteins become increasingly frequent when using the relaxed interolog definition or when studying closely related species (see Table 5.3). However, such CCS naturally encompass various biological functions. In consequence, their functional homogeneity is often too low which excludes the entire CCS from function prediction. The functional similarity, orthology- and interactor-based, for large CCS in *hsa-dme-sce*, *mmu-hsa-dme-sce* and *hsa-dme-cel-sce* is depicted in Table 5.9. These measures confirm that most large CCS receive low coherence scores while only few exhibit medium and none exhibits high functional similarity. Thus, considering such large CCS as a whole is insufficient. For this reason, we modified our approach for large CCS by breaking them up into sub-subgraphs (see Section 4.2.5). Applying this procedure to CCS from Table 5.9 yields 248, 104 and 131 sub-subgraphs, respectively. Assessing the functional coherence of the considerably smaller sub-subgraphs reveals an increase in their functional homogeneity. In consequence, a larger number of subgraphs exceeds the coherence threshold for medium and high functional similarity, see Figure 5.7, and thus can be subjected to our function prediction method.

The impact of splitting large CCS on precision and recall is exemplarily shown in

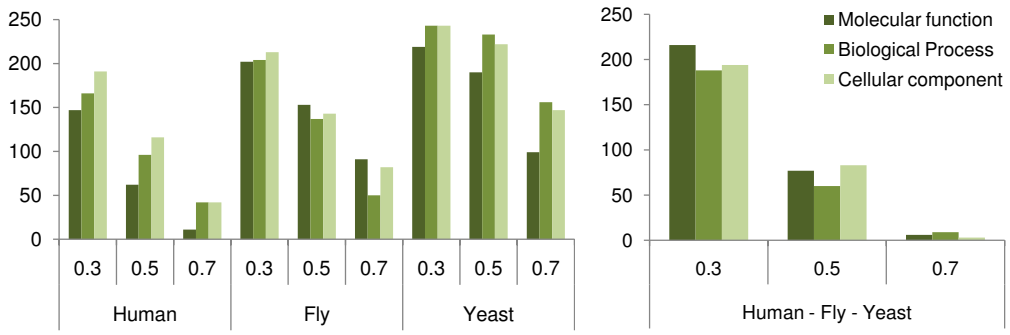
Table 5.8: Prediction results when combining CCS, orthology relationships, and neighboring proteins. Precision (P) and per-protein recall (R_{pp}) are estimated for low (0.3), medium (0.5) and high (0.7) functional similarity/conservation thresholds.

	0.3			0.5			0.7		
Species	# Terms	P	R_{pp}	# Terms	P	R_{pp}	# Terms	P	R_{pp}
mmu	8405	0.42	0.43	5091	0.50	0.55	1398	0.74	0.44
sce	7007	0.54	0.37	5061	0.64	0.34	2494	0.87	0.48
hsa	57020	0.44	0.54	2394	0.74	0.51	932	0.84	0.49
dme	28496	0.52	0.46	3537	0.66	0.37	1028	0.89	0.60
sce	24670	0.76	0.48	17332	0.79	0.36	2187	0.89	0.50
mmu	15945	0.59	0.37	4262	0.73	0.48	449	0.82	0.49
hsa	22742	0.47	0.42	831	0.75	0.62	594	0.89	0.67
dme	13590	0.54	0.38	1746	0.72	0.31	873	0.86	0.63
sce	11314	0.75	0.39	8876	0.74	0.31	947	0.85	0.54
hsa	21046	0.51	0.36	352	0.49	0.17	0	–	–
dme	9115	0.52	0.23	3981	0.67	0.34	230	0.83	0.34
cel	4903	0.53	0.25	4125	0.56	0.24	72	0.65	0.13
sce	8173	0.82	0.31	7134	0.83	0.31	689	0.91	0.34
rno	694	0.96	0.48	609	0.99	0.67	609	0.95	0.67
hsa	1261	0.99	0.71	1261	0.99	0.71	1248	0.99	0.74
dme	539	0.63	0.21	219	0.74	0.62	0	–	–
cel	412	0.87	0.20	116	0.87	0.20	116	0.87	0.53
sce	116	0.97	0.42	433	0.97	0.42	433	0.97	0.59

Table 5.9: Functional similarity of large CCS determined between orthologs across the species (Sim_{ortho}) and between interacting proteins within a species (Sim_{neigh}).

	Largest CCS		Sim_{ortho}			Sim_{neigh}			
Species	# Proteins	# Interologs	MF	BP	CC	Species	MF	BP	CC
hsa dme sce	344	727	0.44	0.40	0.41	*hsa*	0.35	0.42	0.44
						dme	0.52	0.46	0.50
						sce	0.50	0.66	0.65
mmu hsa dme sce	146	324	0.42	0.39	0.42	*mmu*	0.27	0.47	0.51
						hsa	0.35	0.42	0.46
						dme	0.44	0.44	0.46
						sce	0.54	0.63	0.62
hsa dme cel sce	200	372	0.36	0.28	0.28	*hsa*	0.37	0.46	0.46
						dme	0.55	0.46	0.53
						cel	0.21	0.62	0.09
						sce	0.49	0.68	0.65

Table 5.10 and B.2 for combinations yielding CCS with more than 25 proteins. As can be seen, processing large CCS creates significantly more predictions with mostly better precision. For example, when comparing *split* and *non-split* results from *hsa-dme-sce* the number of predictions multiplies almost tenfold for human proteins at a similar or



(a) Interactor-based similarity (b) Orthology-based similarity

Figure 5.7: Number of splitted sub-subgraphs of *hsa-dme-sce* in which (a) the interactor-based similarity or (b) the orthology-based similarity exceeds the functional coherence threshold of 0.3, 0.5 and 0.7.

even better precision. A similar increase in the number of predictions can be observed for fly and yeast proteins. The same holds for *mmu-hsa-dme-sce* and *hsa-dme-cel-sce* (see Table B.2). This underlines the importance of processing large CCS for function prediction with high coverage and more importantly a high level of precision.

The complete results of the combined CCS-based prediction approach for pairs and multiple species combinations from exact (pairs) and approximative (multiple) network comparisons, including processing large CCS, is shown in Table 5.11. Overall, the impact of our combined approach is positive, especially in terms of the number of predictions. Precision drops for some combinations compared to the single methods. However, the decrease of precision does not indicate a lower prediction quality. It rather indicates that the combined method yields many more novel predictions that can not be validated during cross-validation rather than successfully reproducing known function for well-characterized proteins (see Section 5.3.4.3 for a thorough discussion of selected pre-

Table 5.10: Impact of processing large CCS on function prediction in multiple species. CCS with more than 25 proteins are splitted into smaller, overlapping sub-subgraphs.

	0.3			0.5			0.7		
Species	# Terms	P	R_{pp}	# Terms	P	R_{pp}	# Terms	P	R_{pp}
Non-split									
hsa	57020	0.44	0.54	2394	0.74	0.51	932	0.84	0.49
dme	28496	0.52	0.46	3537	0.66	0.37	1028	0.89	0.60
sce	24670	0.76	0.48	17332	0.79	0.36	2187	0.89	0.50
Split									
hsa	51957	0.61	0.20	27868	0.69	0.23	9489	0.87	0.23
dme	33546	0.59	0.18	19674	0.72	0.20	6556	0.84	0.21
sce	35349	0.79	0.19	28936	0.81	0.18	18806	0.87	0.21

dictions). Precision is affected the least for the highest similarity threshold (0.7) fostering the most reliable precisions.

Table 5.11: Complete results of the combined CCS-based prediction approach for pairs of species and three, four, five and six species combinations. CCS are derived from exact (pairs) and approximative (multiple) network comparisons and CCS with more than 25 proteins are splitted into smaller, overlapping sub-subgraphs. Precision (P) and per-protein recall (R_{pp}) are estimated for low (0.3), medium (0.5) and high (0.7) functional similarity/conservation thresholds. Species combinations discussed in more detail throughout the chapter are highlighted in gray.

	0.3			0.5			0.7		
Species	# Terms	P	R_{pp}	# Terms	P	R_{pp}	# Terms	P	R_{pp}
rno	29626	0.67	0.19	15572	0.69	0.18	3462	0.72	0.27
hsa	52116	0.39	0.34	23227	0.41	0.53	5505	0.43	0.64
mmu	416	0.67	0.14	344	0.70	0.17	52	0.94	0.34
cel	199	0.28	0.18	180	0.24	0.24	48	0.75	0.22
rno	578	0.46	0.28	155	0.74	0.18	117	0.75	0.33
cel	164	0.24	0.18	28	0.50	0.11	28	0.50	0.11
hsa	44085	0.36	0.28	12055	0.52	0.36	2442	0.71	0.43
dme	27880	0.50	0.24	11916	0.71	0.21	4601	0.81	0.29
rno	9322	0.76	0.36	7116	0.84	0.38	5097	0.91	0.43
mmu	15485	0.64	0.41	13363	0.64	0.44	10627	0.66	0.51
rno	3021	0.42	0.36	954	0.62	0.53	189	0.86	0.26
sce	3476	0.36	0.45	1095	0.48	0.72	102	0.84	0.32
dme	11510	0.41	0.42	6464	0.51	0.47	1448	0.65	0.31
sce	8459	0.62	0.31	6098	0.69	0.31	3245	0.81	0.30
dme	808	0.55	0.22	572	0.62	0.20	357	0.66	0.21
cel	1211	0.32	0.25	660	0.48	0.24	348	0.55	0.22
rno	691	0.36	0.41	569	0.33	0.45	0	–	–
dme	225	0.44	0.36	214	0.46	0.35	59	0.69	0.17
mmu	145901	0.67	0.18	95592	0.69	0.20	35714	0.73	0.32
hsa	175748	0.45	0.23	103161	0.46	0.29	40609	0.50	0.41
mmu	8644	0.45	0.36	5834	0.49	0.40	795	0.68	0.55
dme	4536	0.58	0.35	3395	0.64	0.47	479	0.65	0.35
mmu	8405	0.42	0.43	5091	0.49	0.55	1398	0.74	0.44
sce	7006	0.54	0.37	5060	0.64	0.34	2493	0.87	0.48
cel	798	0.42	0.16	511	0.55	0.23	300	0.53	0.59
sce	1219	0.81	0.25	974	0.92	0.31	835	0.93	0.41
hsa	4516	0.64	0.26	1578	0.91	0.41	1129	0.97	0.51
cel	2173	0.52	0.17	1167	0.63	0.15	317	0.73	0.22
hsa	105760	0.48	0.21	54010	0.56	0.28	18290	0.69	0.33
sce	66586	0.62	0.22	47973	0.70	0.20	27455	0.84	0.24
hsa	27914	0.68	0.21	18581	0.77	0.25	9009	0.89	0.29

Continued on next page

Table 5.11 – (continued)

Species	0.3 # Terms	0.3 P	0.3 R_{pp}	0.5 # Terms	0.5 P	0.5 R_{pp}	0.7 # Terms	0.7 P	0.7 R_{pp}
cel	8206	0.53	0.12	5902	0.59	0.11	2128	0.71	0.11
sce	16345	0.82	0.18	15284	0.85	0.19	11574	0.89	0.25
rno	873	0.50	0.12	317	0.64	0.24	0	–	–
mmu	951	0.58	0.19	107	0.81	0.35	0	–	–
cel	643	0.41	0.19	417	0.45	0.20	0	–	–
rno	2925	0.68	0.08	1759	0.69	0.10	816	0.91	0.30
hsa	3116	0.66	0.19	1201	0.95	0.33	911	0.97	0.55
cel	1894	0.35	0.11	695	0.52	0.10	384	0.65	0.20
rno	761	0.75	0.29	117	0.85	0.14	33	0.97	0.17
mmu	1191	0.58	0.30	167	0.65	0.15	33	0.97	0.17
sce	748	0.35	0.15	494	0.25	0.10	32	0.69	0.20
mmu	10031	0.57	0.09	3317	0.77	0.15	737	0.84	0.39
hsa	8036	0.53	0.13	1633	0.73	0.25	569	0.81	0.38
cel	4935	0.45	0.13	2478	0.58	0.14	276	0.57	0.17
rno	36582	0.74	0.12	27423	0.79	0.14	9621	0.88	0.26
mmu	50380	0.72	0.13	35738	0.77	0.16	12861	0.83	0.28
hsa	63967	0.42	0.15	44705	0.44	0.23	20231	0.46	0.38
hsa	11720	0.54	0.25	4078	0.70	0.39	133	0.76	0.29
dme	7481	0.59	0.16	4316	0.67	0.21	818	0.79	0.20
cel	3744	0.55	0.15	2340	0.63	0.15	631	0.69	0.13
rno	6618	0.68	0.15	3969	0.65	0.12	1403	0.92	0.32
hsa	7977	0.58	0.25	3052	0.71	0.29	1053	0.99	0.56
dme	5953	0.42	0.13	1898	0.60	0.09	512	0.74	0.09
mmu	2843	0.54	0.13	1067	0.67	0.30	44	0.16	0.14
dme	1519	0.64	0.19	1025	0.65	0.22	99	0.57	0.20
cel	833	0.45	0.14	672	0.47	0.16	295	0.63	0.17
hsa	51957	0.61	0.20	27868	0.69	0.23	9489	0.87	0.23
dme	33546	0.59	0.18	19674	0.72	0.20	6556	0.84	0.21
sce	35345	0.79	0.19	28932	0.81	0.18	18806	0.87	0.21
mmu	1694	0.58	0.30	790	0.81	0.33	609	0.88	0.41
cel	546	0.55	0.12	493	0.58	0.13	34	0.47	0.10
sce	1543	0.85	0.25	1463	0.85	0.28	931	0.88	0.40
rno	2183	0.69	0.26	782	0.79	0.32	12	1.0	0.55
mmu	2382	0.65	0.24	1082	0.68	0.27	12	1.0	0.55
dme	1588	0.43	0.17	1143	0.48	0.15	67	0.64	0.41
rno	934	0.77	0.16	651	0.92	0.34	575	0.95	0.45
cel	467	0.63	0.15	418	0.62	0.16	269	0.69	0.29
sce	574	0.79	0.14	433	0.97	0.23	377	0.98	0.52
mmu	25502	0.58	0.17	15688	0.63	0.20	5799	0.68	0.23
hsa	25201	0.59	0.16	14428	0.67	0.24	4448	0.79	0.27
sce	22313	0.67	0.18	19599	0.69	0.17	9925	0.83	0.20

Continued on next page

Table 5.11 – (continued)

Species	0.3 # Terms	0.3 P	0.3 R_{pp}	0.5 # Terms	0.5 P	0.5 R_{pp}	0.7 # Terms	0.7 P	0.7 R_{pp}
mmu	27932	0.66	0.13	14354	0.73	0.14	2366	0.86	0.17
hsa	25215	0.56	0.17	9448	0.66	0.23	1369	0.80	0.39
dme	25211	0.54	0.14	13773	0.61	0.14	3349	0.76	0.16
mmu	8413	0.63	0.25	5502	0.72	0.25	2184	0.85	0.22
dme	6733	0.58	0.23	4369	0.65	0.31	1049	0.81	0.29
sce	7441	0.73	0.23	5615	0.82	0.22	2846	0.91	0.25
rno	3048	0.71	0.37	2150	0.74	0.61	1022	0.91	0.72
dme	3384	0.47	0.35	765	0.73	0.54	220	0.76	0.62
sce	1823	0.77	0.32	1042	0.80	0.38	823	0.83	0.66
dme	2289	0.66	0.18	1647	0.75	0.20	543	0.82	0.25
cel	1755	0.57	0.17	1345	0.62	0.18	416	0.81	0.23
sce	3422	0.75	0.20	3155	0.78	0.21	2629	0.79	0.25
rno	8008	0.68	0.21	5907	0.72	0.21	3192	0.79	0.31
hsa	9864	0.67	0.22	6128	0.75	0.52	2724	0.85	0.75
sce	9761	0.51	0.19	6584	0.57	0.17	2188	0.86	0.30
rno	875	0.66	0.20	654	0.71	0.28	414	0.76	0.28
dme	369	0.66	0.09	342	0.68	0.14	190	0.61	0.15
cel	325	0.25	0.15	293	0.26	0.28	64	0.56	0.18
rno	1928	0.70	0.24	1574	0.72	0.44	1044	0.93	0.66
dme	962	0.60	0.20	347	0.64	0.63	0	–	–
cel	880	0.65	0.24	737	0.66	0.23	35	0.83	0.60
sce	1044	0.82	0.2	889	0.84	0.31	749	0.85	0.61
mmu	22505	0.60	0.14	12508	0.70	0.17	4941	0.83	0.27
hsa	21552	0.61	0.14	10762	0.75	0.22	3757	0.87	0.32
dme	17023	0.59	0.14	9863	0.72	0.21	3411	0.85	0.35
sce	17111	0.75	0.16	14223	0.77	0.15	8624	0.85	0.23
rno	5487	0.71	0.21	5061	0.71	0.22	2551	0.77	0.30
hsa	6129	0.73	0.26	4258	0.82	0.49	3474	0.87	0.75
dme	6257	0.43	0.15	2631	0.61	0.24	512	0.73	0.57
sce	4176	0.76	0.17	3685	0.76	0.16	2131	0.82	0.33
rno	6709	0.57	0.15	5537	0.57	0.16	1839	0.54	0.14
mmu	9283	0.51	0.15	2773	0.60	0.17	1039	0.73	0.30
hsa	4167	0.58	0.08	2061	0.59	0.32	736	0.62	0.29
sce	5320	0.53	0.13	3382	0.63	0.13	979	0.78	0.20
mmu	9323	0.59	0.13	5401	0.67	0.15	2131	0.87	0.28
hsa	10493	0.56	0.15	5320	0.71	0.21	1751	0.80	0.27
cel	4303	0.48	0.09	3082	0.56	0.09	1023	0.70	0.09
sce	7932	0.76	0.15	7450	0.76	0.15	4349	0.84	0.21
rno	5348	0.68	0.19	4616	0.71	0.21	2986	0.80	0.28
hsa	7840	0.64	0.22	4677	0.81	0.36	1794	0.99	0.56
cel	2108	0.52	0.12	1373	0.66	0.12	1029	0.69	0.12
sce	3292	0.78	0.15	3074	0.79	0.14	1907	0.91	0.24

Continued on next page

Table 5.11 – (continued)

Species	0.3 # Terms	0.3 P	0.3 R_{pp}	0.5 # Terms	0.5 P	0.5 R_{pp}	0.7 # Terms	0.7 P	0.7 R_{pp}
mmu	12659	0.58	0.09	6792	0.68	0.11	1327	0.83	0.19
hsa	8374	0.62	0.13	2785	0.74	0.20	984	0.86	0.31
dme	9369	0.57	0.10	6864	0.58	0.11	1641	0.65	0.12
cel	5107	0.50	0.11	3263	0.55	0.11	1232	0.57	0.11
rno	266	0.54	0.10	131	0.46	0.12	0	–	–
mmu	304	0.50	0.13	7	1.0	0.07	0	–	–
cel	213	0.62	0.14	186	0.58	0.17	0	–	–
sce	167	0.44	0.05	60	0.48	0.02	0	–	–
rno	1071	0.59	0.26	854	0.58	0.41	83	0.96	0.22
mmu	1468	0.60	0.41	87	0.97	0.26	86	0.97	0.26
dme	1208	0.39	0.17	108	0.40	0.28	21	1.0	0.45
sce	589	0.45	0.12	172	0.65	0.09	69	0.67	0.16
rno	7771	0.67	0.09	5247	0.66	0.10	967	0.92	0.19
mmu	10170	0.63	0.09	4697	0.66	0.10	400	0.76	0.15
hsa	9807	0.51	0.11	2656	0.64	0.21	243	0.94	0.32
dme	6862	0.40	0.08	2940	0.47	0.07	540	0.80	0.10
rno	3749	0.62	0.06	2044	0.56	0.05	165	0.78	0.08
mmu	5804	0.61	0.07	1352	0.63	0.07	82	0.61	0.09
hsa	5272	0.48	0.09	605	0.62	0.15	5	1.0	0.38
cel	2456	0.44	0.09	1007	0.60	0.09	165	0.60	0.10
rno	983	0.67	0.13	415	0.72	0.17	42	0.74	0.28
mmu	1430	0.53	0.16	12	1.0	0.21	0	–	–
dme	659	0.62	0.13	456	0.64	0.10	0	–	–
cel	565	0.56	0.19	327	0.58	0.16	0	–	–
rno	2405	0.65	0.08	1444	0.79	0.12	781	0.95	0.38
hsa	2443	0.76	0.20	1566	0.94	0.42	1326	0.99	0.59
dme	2563	0.46	0.07	1178	0.57	0.07	161	0.61	0.09
cel	1728	0.50	0.12	1223	0.60	0.12	466	0.67	0.19
mmu	3190	0.61	0.29	1531	0.63	0.35	566	0.84	0.42
dme	2341	0.55	0.22	915	0.75	0.28	345	0.78	0.44
cel	1280	0.57	0.18	715	0.56	0.14	0	–	–
sce	2282	0.82	0.27	2171	0.82	0.27	848	0.92	0.36
hsa	23497	0.63	0.17	14722	0.69	0.20	6380	0.88	0.23
dme	12813	0.58	0.13	7484	0.74	0.18	4593	0.82	0.29
cel	7801	0.51	0.11	5043	0.57	0.09	1747	0.68	0.10
sce	13005	0.85	0.16	12135	0.86	0.16	9420	0.91	0.23
rno	350	0.58	0.06	271	0.65	0.16	0	–	–
mmu	837	0.39	0.08	5	1.0	0.04	0	–	–
hsa	143	0.57	0.10	15	0.47	0.50	0	–	–
dme	456	0.64	0.11	422	0.64	0.11	86	0.55	0.16
cel	376	0.54	0.20	287	0.56	0.24	265	0.55	0.25
rno	0	–	–	0	–	–	0	–	–

Continued on next page

Table 5.11 – (continued)

Species	0.3 # Terms	0.3 P	0.3 R_{pp}	0.5 # Terms	0.5 P	0.5 R_{pp}	0.7 # Terms	0.7 P	0.7 R_{pp}
mmu	0	–	–	0	–	–	0	–	–
dme	0	–	–	0	–	–	0	–	–
cel	0	–	–	0	–	–	0	–	–
sce	0	–	–	0	–	–	0	–	–
rno	694	0.96	0.48	609	0.99	0.67	609	0.99	0.67
hsa	1261	0.99	0.71	1261	0.99	0.71	1248	0.99	0.72
dme	412	0.63	0.21	219	0.74	0.62	0	–	–
cel	116	0.87	0.20	116	0.87	0.20	116	0.87	0.53
sce	433	0.97	0.42	433	0.97	0.42	433	0.97	0.59
rno	11	0.27	0.01	0	–	–	0	–	–
mmu	0	–	–	0	–	–	0	–	–
hsa	0	–	–	0	–	–	0	–	–
cel	19	0.42	0.02	19	0.42	0.02	0	–	–
sce	2	1.0	0.0	2	1.0	0.0	0	–	–
rno	651	0.94	0.40	543	0.97	0.73	543	0.97	0.73
mmu	1167	0.65	0.64	848	0.68	0.76	83	0.96	0.32
hsa	190	0.97	0.40	189	0.97	0.40	189	0.97	0.40
dme	550	0.36	0.16	202	0.64	0.56	20	1.0	0.57
sce	184	0.64	0.11	179	0.65	0.11	158	0.63	0.25
mmu	1524	0.80	0.39	1361	0.81	0.38	733	0.86	0.41
hsa	2121	0.70	0.46	1216	0.82	0.41	561	0.88	0.47
dme	896	0.70	0.23	650	0.72	0.27	346	0.80	0.44
cel	440	0.48	0.10	349	0.52	0.11	25	0.80	0.28
sce	1183	0.86	0.27	1130	0.87	0.27	899	0.91	0.41
rno	214	0.75	0.19	203	0.78	0.33	83	0.96	0.32
mmu	227	0.67	0.28	83	0.96	0.32	83	0.96	0.32
hsa	189	0.97	0.40	189	0.97	0.40	189	0.97	0.40
dme	429	0.38	0.14	45	0.93	0.37	20	1.0	0.57
cel	264	0.53	0.25	250	0.54	0.25	25	0.80	0.28
sce	99	0.75	0.10	99	0.75	0.10	80	0.74	0.33

5.3.4.2 Impact of functional and evolutionary conservation

As indicated in Table 5.11, the results of our prediction method vary depending on the amount of interaction data and functional annotations available for the species under considerations. They are mostly better for well-annotated species, such as yeast, rat or mouse. This is an inherent property of methods that transfer annotations, since better annotated species provide more source functions. This property underpins the importance of comparative genomics for elucidating the function of human proteins. Further, precision correlates with the functional coverage of a species (see Figure 5.8). Prediction methods perform better on well-studied organisms than on species that are functionally less well characterized, e.g., worm, as new findings in such species are always counted as false positives, independently of their real, biological relevance.

As briefly discussed for selected combinations in Section 5.3.2 and 5.3.3, prediction

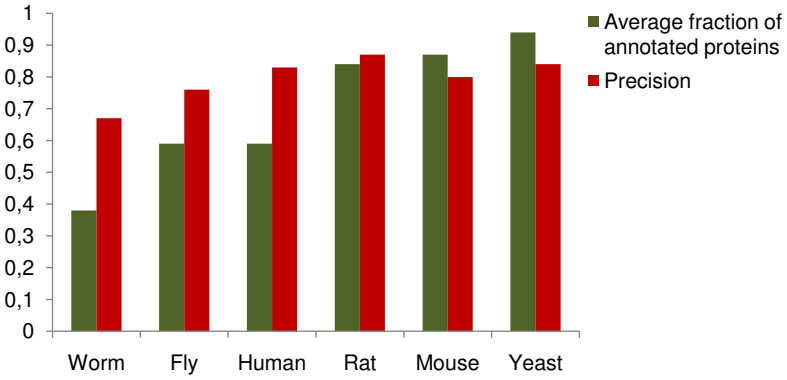
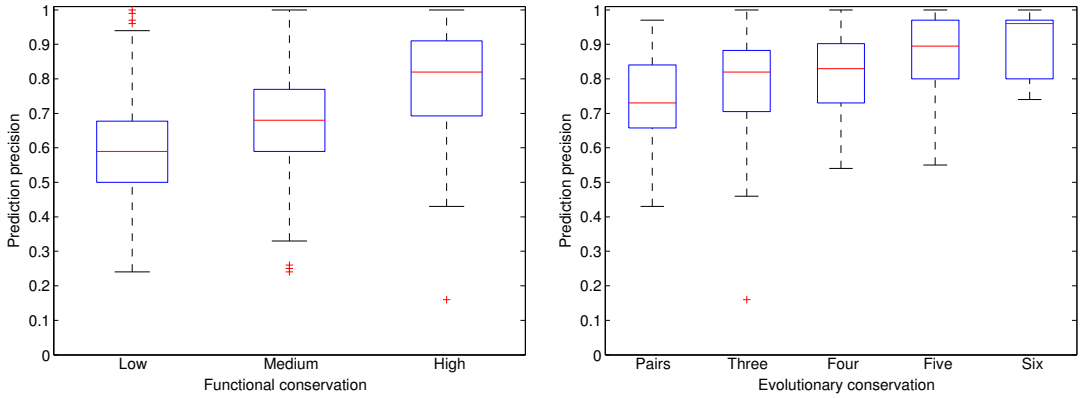


Figure 5.8: Correlation of prediction precision with functional coverage. Functional coverage is determined as the average fraction of proteins with at least one functional annotation across the three subontologies. Precision for each species is estimated across all species combinations.

precision correlates with functional conservation (see Figure 5.9(a)). The median prediction precision increases significantly from 58% (low similarity) to 82% (high similarity). Obviously, the functional conservation threshold is an important possibility to tune our method to the specific needs of an application. A similar correlation can be observed along the degree of evolutionary conservation of CCS (see Figure 5.9(b)). The more species we consider the higher is the evolutionary conservation of the respective CCS which, in turn, is reflected in an increased prediction precision. The median precision improves, for instance, from 73% for pairwise to 96% for multiple CCS of six species.



(a) Functional conservation

(b) Evolutionary conservation

Figure 5.9: Correlation of prediction precision with (a) functional and (b) evolutionary conservation, respectively. (a) Prediction precision is estimated across all species combination for each of the three similarity thresholds (low: 0.3, medium: 0.5, high: 0.7) indicating the level of functional conservation. (b) Precision with respect to the evolutionary conservation of CCS indicated by the number of involved species.

5.3.4.3 Novel protein function

Altogether, our method predicts thousands of protein functions for every species included in the analysis at varying, yet always high levels of precision. Table 5.12 presents for each species the total number of newly derived GO annotations and their estimated average precision and per-protein recall using low, medium and high coherence thresholds. For human, for instance, we predict 27,100 novel functions with a precision of at least 83% or 69,200 functions with a precision of 72%.

Assessing truly novel functions is challenging. Novel functions are typically verified by finding supporting evidence in the literature or databases. However, given the amount of predictions manual curation is impossible. For this reason, we compared GO annotations derived from the CCS-based function prediction method to function inferred by electronic annotation (IEA). Table 5.13 shows the fraction of CCS-based function predictions that have been associated with proteins by other computational methods. About 25% of our predictions for human within biological process have been also inferred automatically while only 7.5% of the predictions for yeast within the same ontology are supported by IEA annotations. However, these numbers do not directly confirm CCS-based prediction but indicate their relevancy and novelty. Therefore, we performed an extensive literature evaluation to verify the correctness of novel annotations of selected proteins (see Section 5.5).

Table 5.12: Overall function prediction statistics across all species combinations. Total number of newly derived GO annotations and their estimated average precision (P) and per-protein recall (R_{pp}) for each species using low, medium and high similarity thresholds.

	0.3			0.5			0.7		
Species	# Terms	P (± std)	R_{pp} (± Std)	# Terms	P (± Std)	R_{pp} (± Std)	# Terms	P (± Std)	R_{pp} (± Std)
rno	18580	0.65 (0.15)	0.21 (0.12)	12259	0.71 (0.14)	0.28 (0.18)	3357	0.87 (0.11)	0.36 (0.18)
mmu	54033	0.59 (0.09)	0.22 (0.14)	32337	0.73 (0.14)	0.25 (0.16)	10586	0.80 (0.17)	0.31 (0.13)
hsa	117014	0.62 (0.16)	0.23 (0.13)	69182	0.72 (0.16)	0.33 (0.13)	27099	0.83 (0.16)	0.43 (0.15)
dme	28885	0.53 (0.10)	0.19 (0.09)	13716	0.64 (0.11)	0.27 (0.16)	3281	0.76 (0.12)	0.30 (0.16)
cel	9432	0.49 (0.13)	0.15 (0.05)	5642	0.55 (0.11)	0.16 (0.06)	1757	0.67 (0.11)	0.22 (0.15)
sce	19035	0.71 (0.17)	0.20 (0.10)	13497	0.75 (0.16)	0.22 (0.14)	4953	0.84 (0.08)	0.32 (0.14)

Table 5.13: CCS-based function predictions compared to annotations inferred from electronic annotation.

Species	Molecular function	Biological Process	Cellular Component
rno	15.3%	6.2%	20.3%
mmu	22.3%	13.2%	20.8%
hsa	39.8%	24.4%	33.3%
dme	19.7%	9.2%	13.1%
cel	51.3%	14.7%	29.8%
sce	11.9%	7.5%	6.8%

5.3.5 Further evaluations

We use the cross-validation setting applied above to evaluate and discuss further properties of our methods according to the following aspects:

- First, we determine the effect of utilizing relaxed CCS on the performance of CCS-based function prediction (see Section 5.3.5.1).
- Second, we study the diversity of function predictions derived from the orthology- and link-base methods (see Section 5.3.5.2).
- Third, we examine whether our prediction strategy benefits from analyzing several species combinations (see Section 5.3.5.3).
- Fourth, we evaluate CCS-based function prediction according to the three GO subontologies: molecular function, biological process and cellular component and their specific GO branches (see Section 5.3.5.4). We determine the performance of our approach with respect to each subontology. Further, we study whether particular GO branches are better predictable than others and if those correlate with evolutionarily conserved function and processes.
- Fifth, we analyze how CCS-based function prediction performs on proteins without any or with only very little functional information by considering all novel predictions for these proteins which are typically counted as false positives in the cross-validation (see Section 5.3.5.5).
- Sixth, we assess whether there is a difference in the prediction performance between more general genes, such as housekeeping genes, or specific genes (see Section 5.3.5.6).
- Last, we study the impact of filtering for high density CCS on function prediction (see Section 5.3.5.7).

Note that in the following evaluations, based on exact or relaxed CCS, we apply the combined prediction strategy including the processing of large CCS with more than 25 proteins.

5.3.5.1 Strict vs. relaxed CCS

For comparing protein interaction networks we used two criteria to identify interologs (see Section 4.1.2): the strict and the relaxed definition. As reported in Section 5.2, we experimented with using the strict interolog definition for multiple species which often results in zero or only very few and small CCS. In contrast, applying the relaxed definition generally yields a considerable higher number of CCS (see Table 5.3 and B.1 for strict and relaxed results).

Table 5.14 illustrates the impact of relaxed CCS by comparing the number of predicted terms along with precision and recall for CCS identified by strict and relaxed network comparisons for: *hsa-dme-sce* and *mmu-hsa-dme-sce*. Utilizing strict CCS leads to a small but highly precise set of function predictions. For instance, a precision of 86% to 95% is achieved for CCS from *hsa-dme-sce*. However, being less strict leads to a significant improvement in the coverage of our prediction method, e.g., the number of predictions increases drastically (10- to 15-times), at comparable precision (see Table 5.14). The effect is most dominant for *rno-hsa-dme-cel-sce* where no (strict) CCS could be identified and consequently no predictions could be derived. The influence on prediction precision is mostly smaller and in some cases negative ($< 10\%$). Notably, especially the predictions with highest reliability (threshold 0.7) are the least affected in terms of precision and often there is an increase, e.g., from 86% to 87% and 73% to 80% for *hsa* and *dme* in *mmu-hsa-dme-sce*, respectively.

Altogether, the usage of the relaxed definition considerably increases the number of qualifying CCS for three or more species and has a significantly positive impact on function prediction.

5.3.5.2 Diversity of orthology- and link-based predictions

We combine orthology- and link-based function prediction within CCS to benefit from the strengths of both methods. To study whether predictions of the individual methods result in the same or complementary sets of predictions we determined the overlap of GO terms predicted by either strategy. For *hsa-dme-sce*, the respective numbers are shown as Venn diagrams in Figure 5.10. In general, the major fraction of unique predictions is derived from neighboring proteins. The overlap between both sets is comparably small and decreases when increasing the similarity threshold. This shows that both methods complement each other very well as they predict rather different sets of functions.

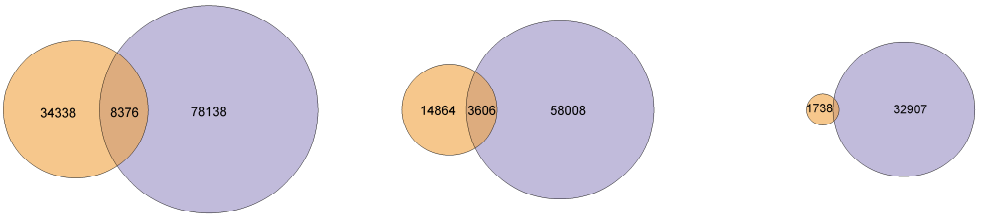
This behavior is also observable when predictions are analyzed separately per species (see Appendix B, Figure B.1). However, contrary to yeast proteins (see Figure B.1(c)), the amount of orthology and link-based predictions is less diverging for human and fly proteins (see Figure B.1(a) and Figure B.1(b)), particularly for low functional coherence, which can be explained by the much denser interaction and functional data available for yeast (see Table 5.1 and Table 5.2). This observation emphasizes that different species profit differently from our method. Especially less characterized species, such as human, benefit strongly from the functional knowledge of model organisms.

Further, we investigated the overlap per subontology. Figure 5.11 shows the fraction

Table 5.14: Impact of the strict and relaxed interolog definition on function prediction in multiple species.

Species	0.3 # Terms	0.3 P	0.3 R_{pp}	0.5 # Terms	0.5 P	0.5 R_{pp}	0.7 # Terms	0.7 P	0.7 R_{pp}
Strict									
hsa	7637	0.52	0.55	2166	0.83	0.68	1119	0.95	0.66
dm	3873	0.54	0.48	1423	0.76	0.50	355	0.86	0.48
sce	3143	0.79	0.42	2382	0.82	0.50	1317	0.92	0.50
Relaxed									
hsa	51957	0.61	0.20	27868	0.69	0.23	9489	0.87	0.23
dme	33546	0.59	0.18	19674	0.72	0.20	6556	0.84	0.21
sce	35349	0.79	0.19	28936	0.81	0.18	18806	0.87	0.21
Strict									
mmu	548	0.78	0.31	484	0.77	0.29	112	0.86	0.37
hsa	766	0.67	0.29	405	0.77	0.46	133	0.86	0.36
dme	447	0.58	0.24	338	0.67	0.39	44	0.73	0.59
sce	448	0.86	0.34	404	0.89	0.37	382	0.88	0.34
Relaxed									
mmu	22505	0.60	0.14	12508	0.70	0.17	4941	0.83	0.27
hsa	21552	0.61	0.14	10762	0.75	0.22	3757	0.87	0.32
dme	17023	0.59	0.14	9863	0.72	0.21	3411	0.85	0.35
sce	17112	0.75	0.23	14224	0.77	0.15	8625	0.85	0.23

of common predictions for molecular function, biological process and cellular component determined at a coherence threshold of 0.3. Regarding the ontology-specific Venn diagrams two further observations can be made. First, the total number of predictions differs significantly among the three ontologies (see Section 5.3.5.4). Second, in contrast to biological process and cellular component, the number of orthology and link-based predictions is quite similar for molecular function (with a fairly small overlap of 515). In this case the significantly smaller number of predictions from the neighbor-based approach might be explained by the following: Proteins within protein complexes/modules represented by CCS are very likely to participate in the same biological process within



(a) Threshold 0.3 (b) Threshold 0.5 (c) Threshold 0.7

Figure 5.10: Overlap between predicted functions derived from the orthology- (orange) and link-based (purple) method for CCS from *hsa-dme-sce*.

(a) Molecular function (b) Biological process (c) Cellular component

Figure 5.11: Overlap between predicted functions derived from the orthology- (orange) and link-based (purple) method for CCS from *hsa-dme-sce* for molecular function, biological process and cellular component. Note that the overlap between orthology- and link-based predictions for molecular function is 515.

the same cellular compartments. However, such proteins do not necessarily have the same but highly diverging biochemical functions. Thus, fewer functions are derived from neighbors in this ontology as molecular function is likely to be less coherent among interacting proteins.

5.3.5.3 Diversity of predictions from different species combinations

In addition to the overlap between the neighbor-based and orthology-based method, we examine whether our prediction strategy benefits from analyzing several species combinations. For this purpose, we assess the overlap between function predictions for human proteins derived from distinct species pairs by defining an overall and a per-protein overlap. The overall overlap, determined by dividing the number of overlapping predictions through the total number of predictions of a combination, represents the overall functional space commonly covered by two species. A low overlap indicates that the species cover complementary functional areas of GO while a high overlap implies a rather common functional basis. The per-protein overlap in turn measures the information gain for proteins when being considered by different species. This overlap is computed for human proteins that are considered by distinct species from different combinations. A low overlap indicates that common proteins receive rather distinct predictions. A high overlap denotes that predictions for the same proteins are quite similar.

Table 5.15 shows the overall and the per-protein overlap between predictions for human proteins inferred from different species pairs. The overall overlap in Table 5.15(a) varies significantly from 0.13 to 0.9 largely depending on the number of proteins within a species and their functional coverage. For instance, the overlap for worm is fairly high as its proteins are quite sparsely annotated. In turn, the overlap between rat and mouse is particularly high for rat. This might be explained by the low number of rat proteins in the data and the fact that both species are mammals, thus their underlying annotations are expected to be more similar. Overall, the data demonstrate that proteins from distinct species contribute differently to predictions for human proteins, i.e., covering

Table 5.15: Fraction of overlapping function predictions for human proteins derived from different species pairs. The overlap is defined as the number of overlapping predictions divided by the total number of predictions. Each cell value – i/j – specifies the overlap based on the total number of predictions of the two combinations. i is the overlap between the non-human species from row i and column j. j is the overlap between non-human species from column j and row i.

(a) Overall overlap between species pairs

	mmu hsa	*hsa dme*	*hsa cel*	*hsa sce*
rno hsa	0.85/0.40	0.42/0.47	0.22/0.69	0.49/0.39
mmu hsa	–	0.30/0.72	0.13/0.90	0.35/0.59
hsa dme	–	–	0.29/0.85	0.56/0.40
hsa cel	–	–	–	0.81/0.20

(b) Per-protein prediction overlap between species pairs

	mmu hsa	*hsa dme*	*hsa cel*	*hsa sce*
rno hsa	0.61/0.58	0.45/0.29	0.20/0.24	0.37/0.31
mmu hsa	–	0.38/0.43	0.36/0.41	0.47/0.35
hsa dme	–	–	0.61/0.60	0.53/0.43
hsa cel	–	–	–	0.46/0.43

different functional aspects.

The per-protein overlap, shown in Table 5.15(b), is mostly below 50% and strongly depends on the evolutionary distance between the species. For example, the overlap between human predictions derived from mouse and those derived from rat is much larger than the overlap between rat and worm. Likewise, predictions inferred from fly and worm are more similar than those obtained from rat as both species are closer related. An overlap of 0.31/0.37 between rat and yeast implies that predictions derived from both species are highly diverse and complementary to each other. The same holds for combinations of three and four species (data not shown). Yet, the diversity decreases the more species we combine as we focus indirectly on evolutionary conserved functions. This becomes more clear when studying predictions for highly conserved housekeeping proteins (see Section 5.3.5.4 and 5.3.5.6).

Our findings emphasize that not only predictions from the different approaches complement each other but also predictions inferred from different species combinations are rather complementary.

5.3.5.4 Performance on GO

Until now, we only focused on the prediction performance regarding the different species and approaches. We now evaluate our approach with respect to the different GO subontologies and specific GO branches. First, we study CCS-based function prediction separately for molecular function, biological process and cellular component, and determine subontology-specific precision and recall. Both subontology-specific precision

Table 5.16: Subontology-specific precision and per-protein recall determined separately for molecular function (MF), biological process (BP) and cellular component (CC) for proteins from *hsa-dme-sce*.

		0.3			0.5			0.7		
		# Terms	P	R	# Terms	P	R	# Terms	P	R
MF	*hsa*	5020	0.51	0.17	1924	0.51	0.21	348	0.55	0.33
	dme	3541	0.64	0.19	2360	0.74	0.22	1053	0.81	0.21
	sce	3492	0.77	0.19	2765	0.79	0.18	1343	0.83	0.19
BP	*hsa*	31318	0.62	0.18	18157	0.69	0.21	7768	0.89	0.23
	dme	18709	0.51	0.14	9863	0.64	0.15	2784	0.80	0.19
	sce	19683	0.74	0.16	15772	0.77	0.14	10875	0.84	0.31
CC	*hsa*	15619	0.62	0.28	1114	0.73	0.31	1373	0.82	0.23
	dme	11296	0.71	0.26	1181	0.82	0.30	2719	0.90	0.25
	sce	12174	0.86	0.27	10399	0.88	0.26	6588	0.91	0.27

and per-protein recall determined for *hsa-dme-sce* are shown in Table 5.16. Overall, our CCS-based method performs comparably consistent across the three subontologies, in particular at a threshold of 0.7. However, precision is generally higher for cellular component across the thresholds than for biological process and molecular function, 91%, 84% and 83%, respectively for yeast.

The largest number of predictions is obtained for biological process followed by cellular component and molecular function (see also Figure 5.11). In human, for instance, the number of predictions derived for biological process is two- to six-times larger than for cellular component and molecular function, respectively. When considering the depth of derived GO functions, annotations are predicted at a median level of 10 for cellular component, 9 for biological process and 4 for molecular function (at a threshold of 0.7). Note that the median depth of a GO term in GO is 5 for molecular function, 8 for biological process and 8 for cellular component. This demonstrates that we are able to infer very specific functions and not only general annotations (see Section 5.5 for a discussion of novel functions).

Compared to other methods presented in the literature, our method has also the important property that it is not limited to so-called "informative" GO terms (Zhou *et al.*, 2002). Many prediction methods use only GO terms that are associated to more than ten or 30 genes (Deng *et al.*, 2003; Chua *et al.*, 2007). Such an approach implicitly disregards more specific annotations, although those are the most valuable ones. For example, in 2007 82.5% of GO annotations in human were associated with less than ten genes (Tao *et al.*, 2007) leaving only 17.5% as annotation basis. GO-based methods have been shown to result in higher precisions when applied on a small number of frequently annotated GO terms. In contrast, we are able to generate accurate predictions also for rarely used GO terms.

Branches of GO Since our prediction strategy is mainly based on evolutionary and functional conservation, we study if this is reflected in the predictions. We assess whether

there are particular GO terms and branches that are better predictable than others and if those correlate with evolutionarily conserved functions and processes. To this end, we determine for each GO term a term-specific precision and recall.

Figure 5.12 presents predicted GO terms within molecular function. GO annotations are colored according to their precision from 0 (yellow) to 1 (green). The figure reveals branches with more precise predictions (green areas) and less precise predictions (yellow areas). Predictions with high precision include, for instance, *DNA* and *RNA polymerase activity*, *DNA-dependent ATPase activity*, and *H3 histone acetyltransferase activity*, while predictions with low precision comprise functions such as *juvenile hormone response element binding*, *retinal binding* and *androgen binding*.

Evolutionarily conserved functions Given the evolutionary framework of our method, we also assume that functions which are evolutionary highly conserved are more precisely predicted by our method. We use housekeeping genes and their annotated functions to validate this hypothesis. Housekeeping genes are known to be constitutively expressed to maintain cellular function essential for cell viability, such as transcription, translation and signaling (Eisenberg and Levanon, 2003; Zhu *et al.*, 2008). We define a set of *housekeeping functions* by extending a list generated by Ferrari and Aitken for classifying housekeeping genes (Ferrari and Aitken, 2006). This list comprises GO terms that are exclusively associated with at least one housekeeping gene. We extend this list by annotations that are known to be over-represented in this gene group (Tu *et al.*, 2006). The set of housekeeping-specific function is displayed in Table 5.17. Again, we determine GO term-specific precision and recall and compare the results for housekeeping-specific functions against non-housekeeping specific functions. Note that for this evaluation we only consider biological process since the majority of GO terms belong to this subontology.

Housekeeping functions yield an average precision of 93%. When considering predictions for human proteins from *hsa-dme-sce* we obtain average precisions of 62%, 69% and 89% at 0.3, 0.5, 0.7 (see Table 5.16). The housekeeping-specific precision is significantly higher, with p-values of $8.6 \cdot 10^{-15}$, $5.9 \cdot 10^{-11}$ and 0.0021, respectively, when comparing it to the average precisions at the three different threshold. However, not only those housekeeping-specific function but also other essential processes, such as *macromolecule biosynthetic process*, *ribosome localization*, *RNA catabolic process* and *mRNA export from nucleus*, are inferred with precisions between 89% and 100%. This shows, that evolutionary conserved processes are better predictable by our prediction method.

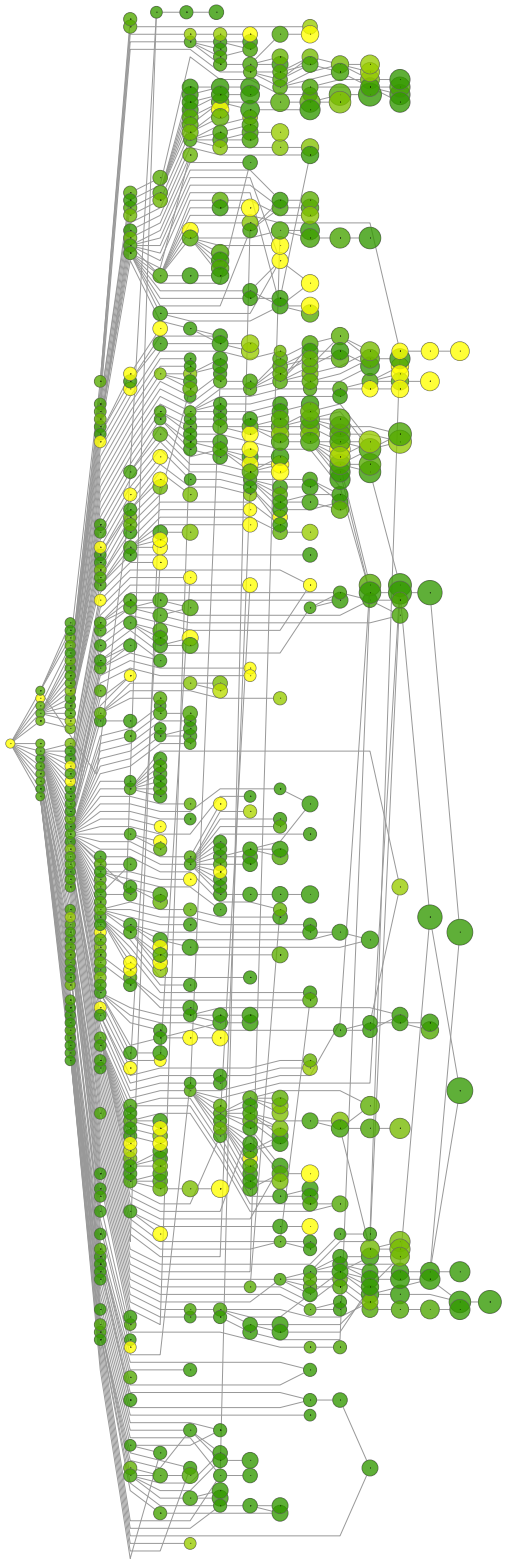


Figure 5.12: GO term specific prediction precision for function predictions in molecular function. Nodes represent GO terms which are colored according to their precision; from 0.0 (yellow) to 1.0 (green). The node size correlates with the depths of a GO term within the GO hierarchy.

Table 5.17: Functional GO annotation specific for human housekeeping genes.

GO Identifier	Description	Subontology	Reference
GO:0006412	Protein biosynthesis	BP	Ferrari and Aitken (2006)
GO:0003735	Structural constituent of ribosome	MF	Ferrari and Aitken (2006)
GO:0005840	Ribosome	CC	Ferrari and Aitken (2006)
GO:0022625	Cytosolic large ribos. subunit (sensu Euk.)	CC	Ferrari and Aitken (2006)
GO:0030529	Ribonucleoprotein complex	CC	Ferrari and Aitken (2006)
GO:0006414	Translational elongation	BP	Ferrari and Aitken (2006)
GO:0006350	Transcription	BP	Tu *et al.* (2006)
GO:0006091	Generation of precursor metabolites and energy	BP	Tu *et al.* (2006)
GO:0046907	Intracellular transport	BP	Tu *et al.* (2006)
GO:0015031	Protein transport	BP	Tu *et al.* (2006)
GO:0006996	Organelle organization and biogenese	BP	Tu *et al.* (2006)
GO:0007049	Cell cycle	BP	Tu *et al.* (2006)

5.3.5.5 Performance on weakly and uncharacterized proteins

An important goal of protein function prediction is to derive novel functions for proteins without any or with only very little functional information. We analyzed how our method performs on such proteins. We define any protein with an information content below a certain cutoff c as weakly annotated protein (WAP). The information content of a protein p is defined as the maximal information content of any of its GO terms assigned *a priori* in our data:

$$IC(p) = max\left\{IC(t_i)|t_i \in t(p)\right\}. \tag{5.1}$$

Recall the information content of a GO term correlates with its frequency and specificity. Thus, proteins with less frequent terms and terms with few occurring descendants are more informative. The cutoff for weak annotation is set for each species and ontology as 25 percentile of the $IC(p)$ distribution across the proteins within the respective species. Proteins with an information content below c are considered to be weakly annotated. For WAP analysis, we count annotations as new if they are more specific than the existing ones or if they belong to another sub-branch in the subontology. Note such annotations are counted as false positives in our evaluation as they cannot be validated from our gold standard data.

Figure 5.13 shows the number of predicted functions for proteins without any annotations within CCS from *hsa-dme-sce*. As expected, the highest number of proteins without any annotation can be found in human. Annotation coverage of fly is not as good as for yeast but still much better than in human. For example, CCS at threshold 0.3 contain 583 human proteins without any functional annotation in biological process. By means of our method, we predict 703 annotation for 191 of those proteins. Similarly, 96 fly proteins out of 313 are annotated with 235 GO annotation in biological process. In contrast, only very few yeast proteins are not annotated at all in any of the three ontologies. However, also well-studied species still contain many WAPs that benefit from our approach as illustrated in Figure 5.14. For instance, about 240 yeast proteins are

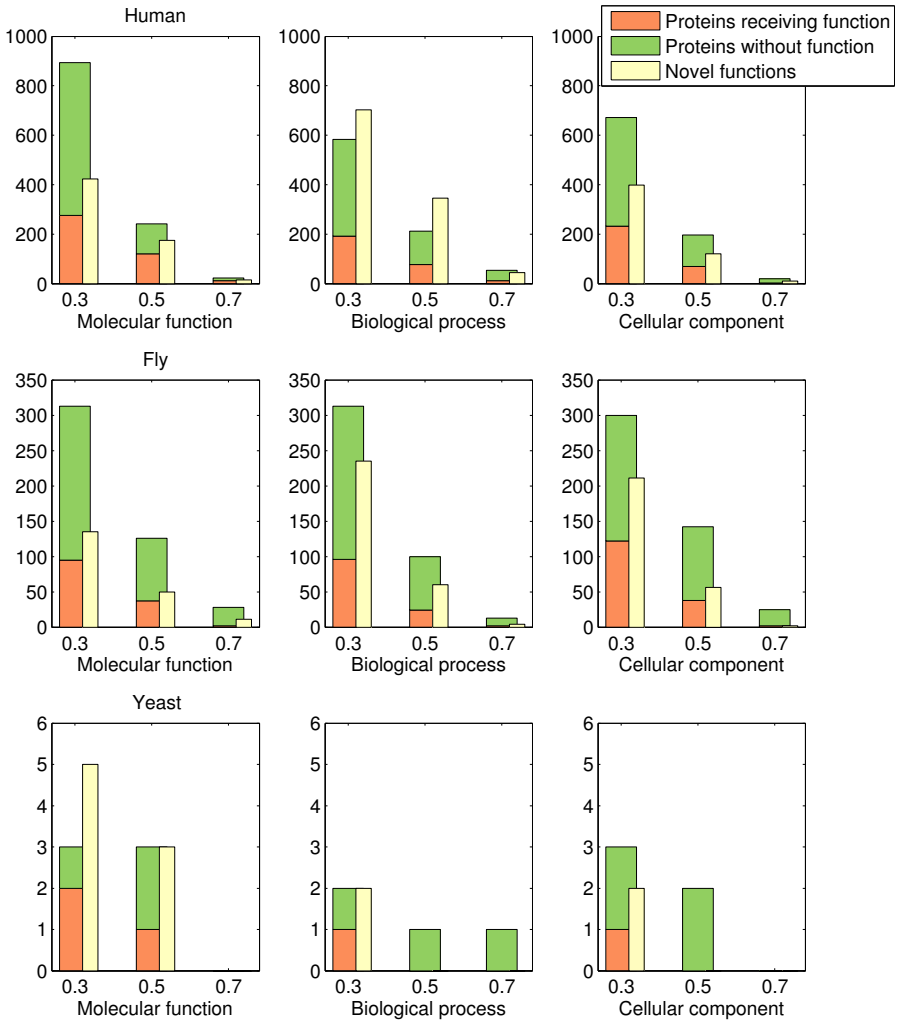


Figure 5.13: Number of predicted functions for proteins without annotations within CCS from *hsa-dme-sce*. For each subontology and similarity threshold the number of proteins without any annotation (olive), the number of proteins receiving new annotations (orange) and the total number of novel annotations are shown (yellow). Recall that a higher coherence threshold for CCS leads to less proteins being included in function predictions; thus, numbers generally decrease with higher thresholds.

only weakly characterized for cellular component and for more than a third of them we predict about 200 functions. Overall, the fraction of proteins receiving new annotations decreases for each species with increasing similarity threshold.

5.3.5.6 Performance on general and specific proteins

Additionally, we study whether there is a difference in the prediction performance between more general genes, such as housekeeping genes, or specific genes. In contrast to Section 5.3.5.4 in which we focused on functions specific to such genes, we now evalu-

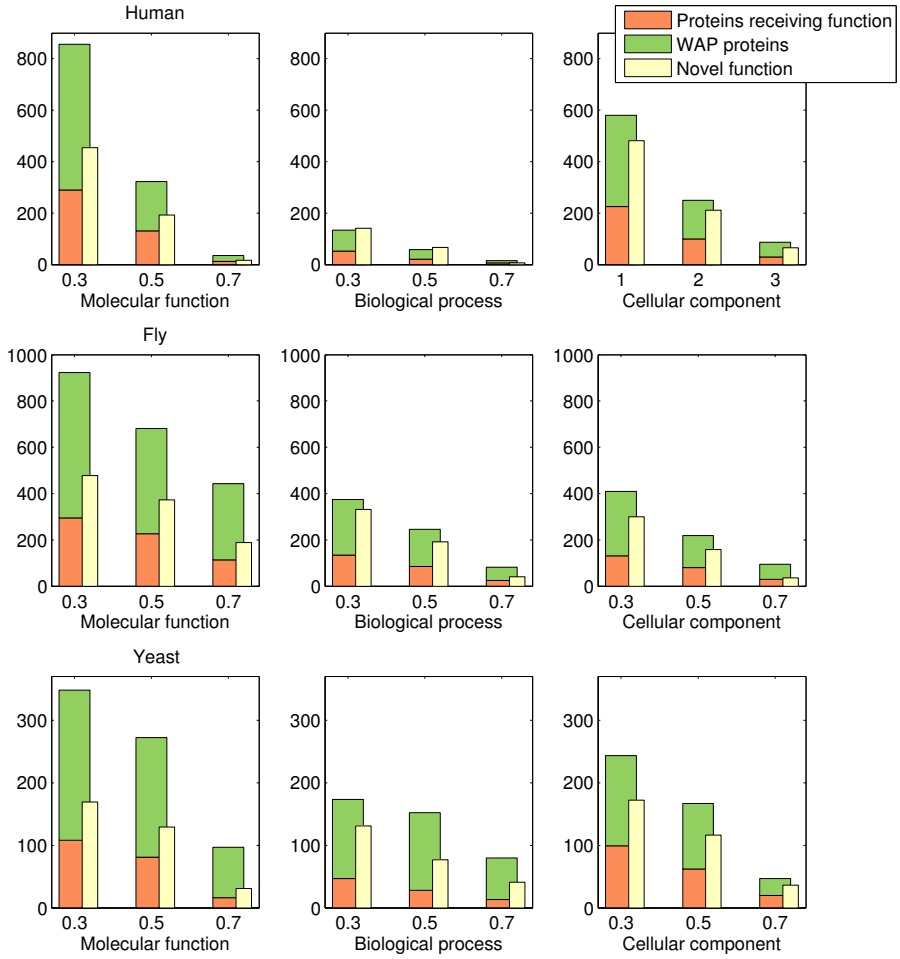
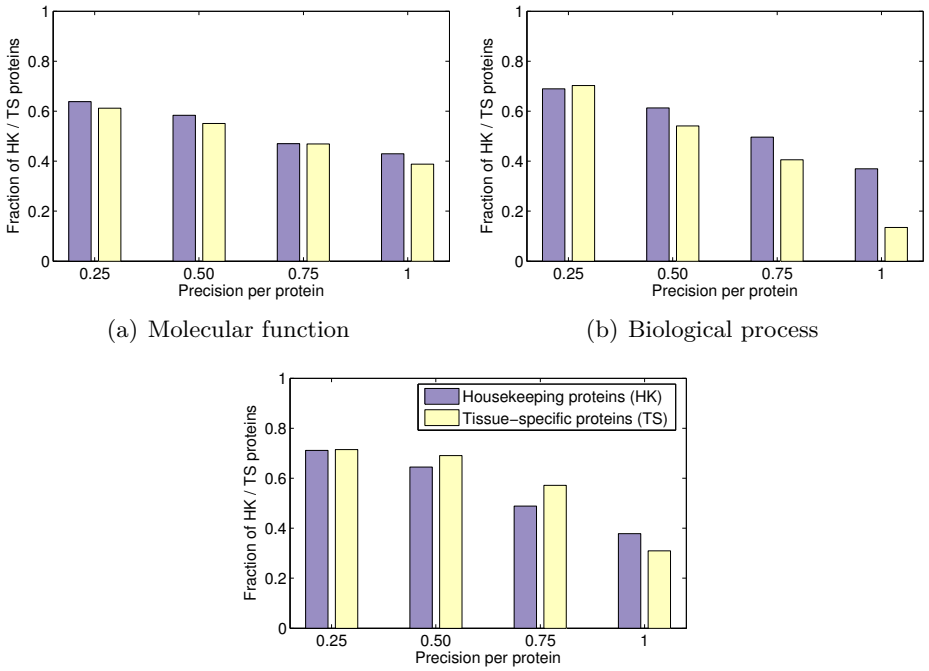


Figure 5.14: Function prediction for weakly annotated proteins (with low information content) within CCS from *hsa-dme-sce*. For each subontology and similarity threshold the number of weakly annotated proteins (olive), the number of proteins receiving new annotations (orange) and the total number of novel annotations are shown (yellow). Note only annotation that are more specific then existing ones are counted as novel.

ate our method on the protein level. We compiled a list of tissue-specific and housekeeping genes from microarray studies that focus on housekeeping (Warrington *et al.*, 2000; Hsiao *et al.*, 2001; Eisenberg and Levanon, 2003) and tissue-specific genes (Warrington *et al.*, 2000; Hsiao *et al.*, 2001; Ge *et al.*, 2005). This results in 1177 housekeeping and 1780 tissue-specific genes encoding human proteins in our data set. Gene lists are available as supplementary material on a CD (see `housekeeping_genes.tsv` and `tissuespecific_genes.tsv`).

We determine protein-specific precision and recall for human proteins from the species combination: *hsa-dme-sce*. Proteins are then classified as housekeeping or tissue-specific (if possible) and precision and recall are compared between both protein groups. As expected the overall number of housekeeping genes involved in CCS is significantly larger



(a) Molecular function

(b) Biological process

(c) Cellular component

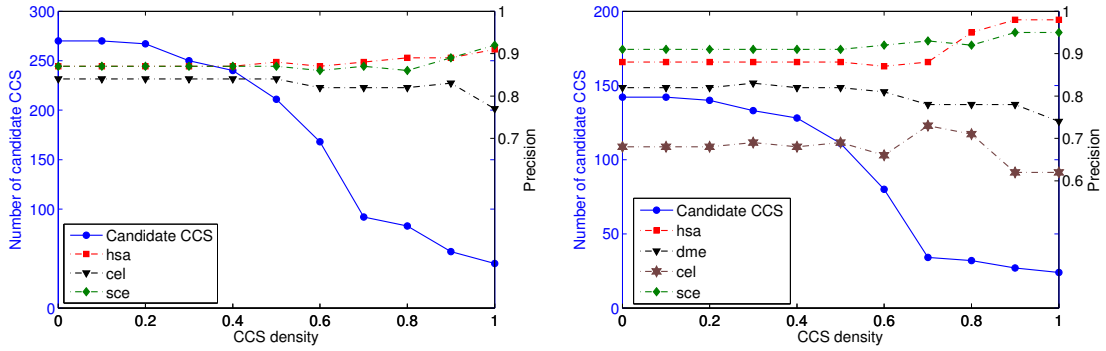
Figure 5.15: Comparison of the prediction performance on housekeeping and tissue-specific proteins of CCS from *hsa-dme-sce*. The fraction of proteins above a certain precision threshold is determined for each of the two protein groups.

than the number of tissue-specific genes as we focus on evolutionary conserved subgraphs in our studies. Housekeeping genes tend to be more conserved than tissue-specific genes as they evolve more slowly than tissue-specific genes (Zhang and Li, 2004).

The prediction performance of our method on the two sets of proteins has been assessed by determining the fraction of proteins above a certain precision within each set. Figure 5.15 shows the fraction of housekeeping and tissue-specific proteins having a protein-specific precision larger than 0.25, 0.50, 0.75 and 1.0. Except for cellular component, the fraction of housekeeping proteins above a certain precision is mostly higher than the fraction of tissue-specific proteins. The difference between both fractions correlates with the increasing precision but the discrepancy is not significant. As assumed these results indicate, that our method favors evolutionary conserved genes, such as housekeeping genes. Yet, we are not limited to them and do not only predict functions for well-studied housekeeping genes but also for (tissue-)specific proteins at a comparably high level of precision.

5.3.5.7 Module density

Our proposed method considers evolutionarily conserved subgraphs with high functional coherence as functional modules. However, our definition of functional modules differs



(a) Human-Fly-Yeast (b) Human-Fly-Worm-Yeast

Figure 5.16: Impact of CCS density on the number of qualifying CCS and the function prediction for (a) *hsa-dme-sce* and (b) *hsa-dme-cel-sce*. The number of CCS per density is displayed (left y-axis) as well as the influence on the prediction precision (right y-axis).

from the traditional ones that primarily consider dense complexes, with a high connectivity or clustering coefficient, as modules (Bader *et al.*, 2003; Spirin and Mirny, 2003; Altaf-Ul-Amin *et al.*, 2006). Therefore, we evaluated the effect of requiring CCS to be "module-like", i.e., to exhibit a certain density of interactions between its members. To study the impact of subgraph density on function prediction we performed an experiment where we only consider candidate CCS with a certain density D. CCS-density C_D is defined as:

$$C_D = \frac{2 * |I|}{|O|(|O| - 1)} \tag{5.2}$$

where I presents the interologs as edges and O denotes the orthologous proteins groups as nodes within a CCS. The influence of high density on the number of candidate CCS and on the prediction precision is demonstrated for *hsa-dme-sce* and *hsa-dme-cel-sce* in Figure 5.16. The number of candidate CCS decreases with an increasing density threshold, e.g., only 83 out of 270 CCS and 24 out of 142 CCS have a density above 0.7 in *hsa-dme-sce* and *hsa-dme-cel-sce*, respectively. On the other hand, the increasing density correlates mostly with an increasing prediction precision, e.g., in human from 88% without filtering up to 98% for a density of 1.0, see Figure 5.16(b). The effect is less pronounced in *hsa-dme-sce* as precision is fairly constant, see Figure 5.16(a). For worm and fly we observe a decrease in the precision for very high densities ($C_D \geq 0.9$). An optimistic hypothesis might be that among the predictions in very dense CCS are many that are incorrectly counted as false positives, due to the incompleteness of the gold standard used for evaluation.

However, filtering for highly connected CCS disregards pathways and modules that are less linked, most likely due to the incompleteness of the data. Thus, limiting CCS to highly connected subgraphs improves prediction precision further but at the expense of coverage. The number of predictions decreases significantly, for instance, for human from 6,380 without filtering to 1,583 for a density of 0.7 down to 382 for a density of 1.0.

5.4 Comparison to related methods

We evaluate our approach using two classical prediction methods, namely *Neighbor Counting* (*NC*, Schwikowski *et al.* (2000)) and χ^2 *statistics* (Hishigaki *et al.*, 2001). We also compare our approach with *FS-Weighted Averaging* (*FS-WA*, Chua *et al.* (2006)), a method that considers indirect functional associations and topological weights (see Section 4.4 for a detailed description of the individual methods). We did not compare to purely module-based prediction methods, as link-based techniques have been shown to outperform those (Sharan *et al.*, 2007; Song and Singh, 2009). Further, Chua *et al.* (2006) demonstrated earlier that the FS-Weighted Averaging significantly outperforms local and global network approaches, e.g., methods that are based on Markov random fields or functional flow (Deng *et al.*, 2003; Vazquez *et al.*, 2003).

For comparing the three approaches to our CCS-based approach we use a script provided by Chua *et al.* that implements the three methods. This script was also used to perform the comparison of those methods in (Chua *et al.*, 2006, 2007). To allow for a fair comparison with the CCS-based prediction we adjusted few parameters before applying the script:

- First, we do not limit predictable GO terms to informative annotations. Other methods restrict their data to frequently annotated functions (see Section 5.3.5.4) to obtain statistically sound conclusions (Deng *et al.*, 2003; Chua *et al.*, 2007) which, in turn, implicitly disregards more specific functions. As we consider both frequent and rare annotations in our approach, we do not exclude non-informative GO terms when comparing against the three methods.
- Second, we do not limit predictable GO terms to annotations above a certain GO level in the GO hierarchy as we do not exclude more general terms at lower levels from our approach, unlike Chua *et al.* (2006, 2007).

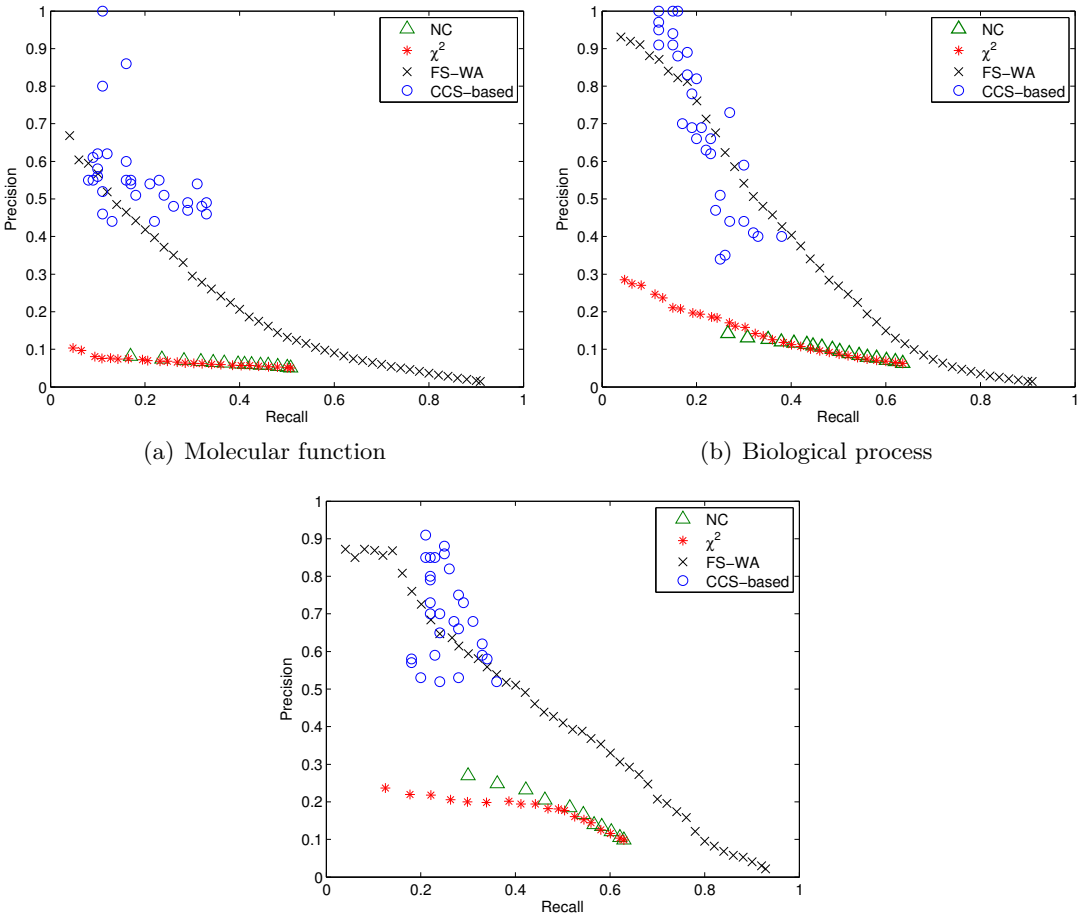
Further, we alter our original per-protein recall computation (see Eq. 4.12) to generate comparable precision – recall graphs. Recall is therefore based on proteins of all CCS rather than only on CCS passing a particular similarity threshold. The modified recall decreases, in contrast to the per-protein recall, with increasing similarity threshold as less proteins are considered during the evaluation. Precision and recall for CCS-based prediction are determined across *hsa-dme-sce* and *mmu-hsa-dme-sce* for varying similarity thresholds ranging from 0.1 to 1.0.

Figure 5.17 presents precision – recall graphs for predictions for human proteins separated by the three GO subontologies. Our combined CCS-based approach significantly outperforms *NC* and χ^2 statistics, especially in terms of precision. Precision and recall obtained from the latter two are very low and even below our baselines. This also holds for fly and yeast (see Appendix B, Figure B.2 and B.3). Furthermore, precision for *NC* and χ^2 is significantly lower than reported in the respective original publications (Schwikowski *et al.*, 2000; Hishigaki *et al.*, 2001). There are three explanations for this drop (from ∼70% to 15% precision):

- First, both methods have been previously evaluated on the functional classification scheme from YPD. This scheme covers, similar to GO, three categories of

yeast protein function: biochemical function, cellular role and subcellular localization (see Section 3.1). However, categories have only 57, 41 and 22 members, respectively. Compared to our evaluation using GO, in which methods have to chose between up to 20,188 functional categories (see Table 3.1), this increases the chances to predict correct terms purely by chance. Furthermore, yeast is a particularly well-studied organism, while we applied the method also to less-well covered species. A similar performance drop was observed by Chua *et al.* (2007), who also applied both methods to GO term prediction, with precision decreasing to 60% (NC) and 20% (χ^2) for yeast and to 20% (NC) and 16% (χ^2) for fly.

- The second point concerns the amount of interaction data. For example, results from Schwikowski *et al.* (2000) are based on only 2,709 interactions among 2,039



(a) Molecular function

(b) Biological process

(c) Cellular component

Figure 5.17: Performance comparison for human proteins. CCS-based precision and recall are compared against *Neighbor Counting* (NC), χ^2 *statistics* and *FS-Weighted Averaging* (FS-WA) for (a) molecular function, (b) biological process and (c) cellular component.

proteins. In contrast, we integrated six different databases, leading to, for instance, almost 71,000 interactions for 6,000 proteins in yeast. Thus, we cover many more proteins and interactions which also increases the probability of false positives.

- Third, many prediction methods, including the two studies compared to here, consider only annotated proteins with at least one annotated interaction partner for their studies (Schwikowski *et al.*, 2000; Hishigaki *et al.*, 2001; Deng *et al.*, 2003; Chua *et al.*, 2006). We did not exclude those proteins because we believe that especially weakly or uncharacterized proteins must be a primary target for function prediction. In our combined approach, such proteins often receive functions from orthologous proteins in other species, an option missing in *NC* and χ^2. However, functions predicted for non-annotated proteins are necessarily counted as false positives although these are truly novel findings. Thus, disregarding such proteins results in higher precisions.

When comparing *FS-WA* results with our approach, CCS-based function prediction performs comparably well or even better. Depending on species and subontology we achieve either higher precision at a similar recall or an improved precision and recall. For instance for human proteins, CCS-based prediction performs consistently well across the three subontologies. Comparing the performance on fly proteins shows that our method clearly outperforms *FS-WA*, achieving much higher precision at higher recall (see Figure B.2). A similar tendency is also observable for yeast. Precision is mostly similar but recall is higher (see Figure B.3). Notably, our method achieves significantly better results in species with less comprehensive interaction coverage, such as fly.

5.5 Predictions for Selected Human Proteins*

In the following, we discuss specific predictions for proteins that are relevant for colorectal cancer. Note, these predictions were counted as false positives in our evaluation because they are not contained in the Gene Ontology annotations at all or only marked as putative (mostly "inferred from electronic annotation", IEA). However, we show that many predictions already have strong experimental support in the literature. Thus, the group of novel predictions falls into two classes – those that, given the current literature, can be considered as true but have not yet made it into the annotation databases and those for which we could not find conclusive evidence in the literature. We argue that, given the large amount of predictions that fall in the first class, predictions from the second class should be considered as promising candidates for further studies.

We discuss predicted functions for the gene products of *MLH1*, *PMS2* and *EPHB4*, all of which have an established importance for colorectal cancer (Jiricny, 2006; Kumar *et al.*, 2009). Overall, literature curation largely confirms the predictions for these three genes by different experimental studies.

*Joint work with Christine T Sers (Institute of Pathology, Charite - Universitätsmedizin Berlin).

MLH1 and PMS2

The DNA mismatch repair protein MLH1 and the mismatch repair endonuclease PMS2 belong to the main components of the post-replicative DNA mismatch repair (MMR) system (see Figure 5.18) (Li, 2008). The MMR system is required for correcting base mismatches and insertion or deletion loops resulting from DNA replication, DNA damage, or from recombination events between non-identical sequences during meiosis (Jiricny, 2000). Curated annotation for *MLH1* and *PMS2* from UniProt and EntrezGene and newly inferred functions are listed in Table 5.18 and 5.19.

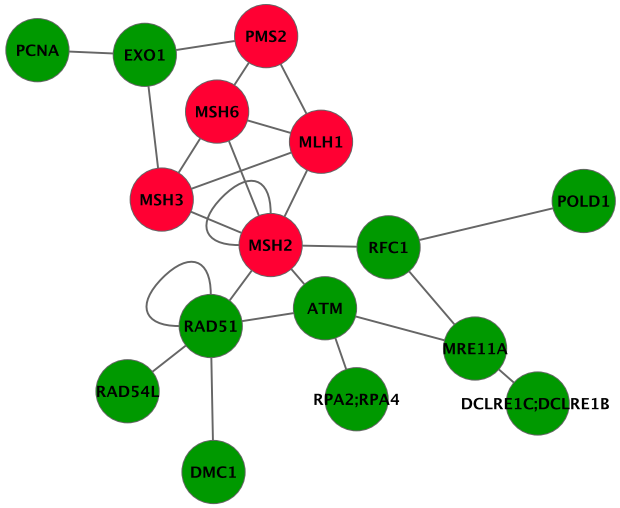


Figure 5.18: Components of the post-replicative DNA mismatch repair system (MMR). The illustrated subgraph is part of a larger CCS identified between human, fly and yeast, and clusters proteins that are involved in mismatch repair. Proteins associated to colorectal cancer are indicated in red.

The majority of our predictions (terms are set *italics* in the following) is directly related to the functionality of the MutLα complex which is formed by *MLH1* and *PMS2*. Rich supporting evidence can be found from the respective orthologs in yeast and mouse. For instance, *PMS1*, the *PMS2* ortholog in yeast, contributes to *dinucleotide insertion or deletion binding*, *loop DNA binding* (Habraken *et al.*, 1997). *Mlh1*, the mouse ortholog of *MLH1*, is annotated to *guanine/thymine mispair binding* (ichi Yoshioka *et al.*, 2006) and likely plays a role in the formation, stabilization and/or the resolution of Holliday junction intermediates (*four-way junction DNA binding*) (Baker *et al.*, 1996). High and low affinity ATP binding sites have been observed for *MLH1* and *PMS1* in yeast (Hall *et al.*, 2002) which supports the *ATP binding* and *ATPase activity* predictions for their human orthologs (Guarné *et al.*, 2001). Moreover, *PMS2* contains a conserved metal-binding motif constituting part of the active site for the endonuclease activity of the protein and might enable *magnesium ion binding* (Hsieh and Yamane, 2008). Considering *protein homodimerization activity*, the dysregulated gene expression of *PMS2*, either as a monomer or homodimer, can disrupt MMR function in mammalian cells (Gibson *et al.*, 2006). Note that although we support our predictions by literature evidence that

are mostly based on orthologs, our algorithm actually inferred them from conserved interaction partners as the orthologs in most cases do not carry the annotation we found in the literature.

Our algorithm also generates a number of predictions that are not as clearly supported by the existing literature, such as *guanine/thymine mispair binding*, *single guanine* and *thymine insertion binding* or *oxidized DNA binding*. Moreover, we associate both proteins to *base-excision repair* as well as *postreplication repair* and *MLH1* to *maintenance of DNA repeat elements*. These are interesting hypotheses supported by recent findings from Erdeniz *et al.* who suggested that the endonuclease activity of *PMS2* in MutLα is not only important in MMR-dependent mutation avoidance but also for suppression of homologous recombination, DNA damage signaling, and damage response functions (Erdeniz *et al.*, 2007). Association of yeast *PMS1* with *meiotic mismatch repair* and *DNA recombination* (Stone and Petes, 2006) further support these predictions. Regarding their cellular components both proteins are associated to the *MutLα complex* (Jiricny, 2006), an annotation predicted jointly from orthology and the CCS neighborhood. *MutLα complex* is a clearly sensible refinement of the existing annotation *nucleus* and only seven others genes are annotated to this term, which emphasizes the specificity of our method.

EPHB4

Ephrin type-B receptor 4 is a transmembrane receptor for the ephrin-B family. It belongs to the family of receptor tyrosine kinase (RTK) and is usually expressed in endothelial and neuronal cells. Known and predicted functional annotations are displayed in Table 5.20.

Several predicted functions, such as *protein*, *enzyme* and *ATP binding*, *SH3/SH2 adaptor* and *enzyme regulator activity* and *protein amino acid phosphorylation*, derived both from conserved interactions and orthology, are evidently consistent with the characteristics of receptor tyrosine kinases. Although those functions are well-known in literature, they are not yet curated and established in the corresponding databases.

Two functions inferred by orthology are *transmembrane-ephrin receptor activity* and *transmembrane receptor protein tyrosine kinase signaling pathways*. Both are supported by annotations from highly related receptors, such as *Ephb1* in mouse and *EPHB2* in human (Ikegaki *et al.*, 1995; Birgbauer *et al.*, 2001). Less evident predictions are, for instance, *cell-cell signaling* (Himanen and Nikolov, 2003), *cell migration* (Sturz *et al.*, 2004), *angiogenesis* and *behavior* (Pasquale, 2005). These functions were not predicted by orthology alone but only in combination with the conserved interaction neighborhood of *EPHB4*. *EPHB4* participates in the axon guidance pathway and in this context predictions like *axon guidance* or *axon guidance receptor activity* can be integrated (Brambilla and Klein, 1995; Dickson, 2002; Huot, 2004).

Table 5.18: Most specific existing and predicted functional annotation (per GO subontology) for *MLH1*. Predictions with supporting literature are marked as (+). The correctness of predictions without supporting literature (?) remains unclear.

Molecular function	Evidence	Biological process	Evidence	Cellular component	Evidence
Known functional annotations					
single-stranded DNA binding		mismatch repair		nucleus	
MutSα complex binding					
Predicted functional annotations					
ATPase activity	+ (Ban and Yang, 1998; Hall *et al.*, 2002)	base-excision repair	+ (Wu *et al.*, 2008)	MutLα complex	+ (Jiricny, 2006)
ATP binding	+ (Ban and Yang, 1998; Hall *et al.*, 2002)	postreplication repair	+ (Shcherbakova *et al.*, 2001)		
protein homodimerization activity	+ (Shcherbakova *et al.*, 2001)	maintenance of DNA repeat elements	+ (Lin and Wilson, 2009)		
four-way junction DNA binding	+(Baker *et al.*, 1996)				
guanine/thymine mispair binding	+ (ichi Yoshioka *et al.*, 2006)				
dinucleotide repeat insertion binding	?				
single guanine insertion binding	?				
purine-specific mismatch base pair DNA N-glycosylase activity	?				
single thymine insertion binding	?				
oxidized DNA binding	?				

Table 5.19: Most specific existing and predicted functional annotation (per GO subontology) for *PMS2*. Predictions with supporting literature are marked as (+). (−) indicates false predictions and for predictions without supporting literature (?) the correctness remains unclear.

Molecular function	Evidence	Biological process	Evidence	Cellular component	Evidence
Known functional annotations					
Single-stranded DNA binding		Mismatch repair		Nucleus	
Single base insertion or deletion binding					
MutSα complex binding					
Predicted functional annotations					
ATPase activity	+ (Ban and Yang, 1998; Guarné *et al.*, 2001; Hall *et al.*, 2002)	DNA recombination	+ (Stone and Petes, 2006; Erdeniz *et al.*, 2007)	MutLα complex	+ Jiricny (2006)
ATP binding	+ (Ban and Yang, 1998; Hall *et al.*, 2002)	Base-excision repair	+ (Wu *et al.*, 2008)		
Protein homodimerization activity	+ (Gibson *et al.*, 2006)	Meiotic mismatch repair	+ (Stone and Petes, 2006; Erdeniz *et al.*, 2007)		
Magnesium ion binding	+ (Hsieh and Yamane, 2008)				
Dinucleotide insertion or deletion binding	+ (Habraken *et al.*, 1997)				
Loop DNA binding	+ (Habraken *et al.*, 1997)				
Four-way junction DNA binding	?				
Guanine/thymine mispair binding	?				
Single guanine insertion binding	?				
Purine-specific mismatch base pair DNA N-glycosylase activity	?				
Single thymine insertion binding	?				
Oxidized purine DNA binding	?				

Table 5.20: Most specific existing and predicted functional annotation (per GO subontology) for *EPHB4*. Predictions with supporting literature are marked as (+). (–) indicates false predictions and for predictions without supporting literature (?) the correctness remains unclear.

Molecular Function	Evidence	Biological Process	Evidence	Cellular component
Known functional annotations				
Transmembrane receptor protein tyrosine kinase activity		Cell proliferation		Integral to plasma membrane
		Regulation of angiogenesis		Cell surface
Predicted functional annotations				
ATP binding	+ (Schlessinger, 2000; Pawson, 2002)	Protein amino acid phosphorylation	+ (Schlessinger, 2000; Pawson, 2002)	
Protein binding	+ (Schlessinger, 2000; Pawson, 2002)	behavior	+ (Pasquale, 2005)	
Enzyme regulator activity	+ (Schlessinger, 2000; Pawson, 2002)	Cell-cell signaling	+ (Himanen and Nikolov, 2003)	
Protein serine/threonine kinase activity	–	Cell migration	+ (Sturz *et al.*, 2004)	
Enzyme binding	+ (Schlessinger, 2000; Pawson, 2002)	Transmembrane receptor protein tyrosine kinase signaling pathway	+ (Tanaka *et al.*, 2004)	
Protein C-terminus binding	+ (Schlessinger, 2000)	Axon guidance	+ (Brambilla and Klein, 1995; Dickson, 2002; Huot, 2004)	
SH3/SH2 adaptor activity	+ (Schlessinger and Lemmon, 2003)			
Axon guidance receptor activity	+ (Brambilla and Klein, 1995; Dickson, 2002; Huot, 2004)			
Transmembrane-ephrin receptor activity	+(Ikegaki *et al.*, 1995; Birgbauer *et al.*, 2001)			

6 Disease Gene Identification

Once the function of a gene or gene product has been characterized, this knowledge can be utilized to study its (potential) role in diseases. This chapter focuses on the association of genes or proteins with particular diseases through interaction and functional data.

Section 6.1 introduces the relationships between genes and diseases such as genetic factors that cause a disease. We briefly review the broad range of methods available for disease gene identification – from traditional gene-mapping methods to advanced bioinformatic techniques that emerged to accelerate classical disease gene discovery. We present in Section 6.3 a novel network-based approach for identifying disease-causing proteins in a genome-wide setting. We give a detailed description of the method itself and the evaluation procedure that is used in Chapter 7. There, we show that our method yields promising results even without an associated locus. The chapter ends with a discussion of related work in Section 6.4.

6.1 Genes and Diseases

In Section 2.1.3 we elaborated on the role of proteins in human diseases with respect to alterations that impact their natural function and which may lead to cell malfunction and, eventually, to a disease. In the following, we focus on the genetic origins of changes in functional, structural and metabolic protein properties that account for the onset of disease but also for susceptibility to disease under particular circumstances.

Diseases are pathological conditions that impair the normal state of an organism by altering or destroying its vital functions. Abnormal functioning can be caused by inherited genetical defects, somatic or spontaneous mutations, internal dysfunctions and environmental influences, such as stress or infection (Mackenbach, 2006). Diseases rarely originate from abnormalities in single genes but rather reflect the perturbation of the complex intra- and intercellular network that links tissue and organ systems (Barabási *et al.*, 2011). Elucidating the underlying disease mechanisms is crucial for understanding the onset of diseases and the development of disease-specific diagnostic and therapeutic approaches.

Many human diseases have a strong genetic component. Diseases caused by abnormalities in an individual's genome are generally referred to as *genetic diseases* or *genetic disorders*[8]. More than 7,000 (classical) genetic diseases have been characterized (McKusick, 2007); from widely recognized disorders like Down syndrome, Spina bifida and Sickle cell anemia, to lesser known diseases, e.g., Tay-Sachs and Fray disease. Most diseases are neither purely genetically nor purely environmentally induced. Heredity can

[8]Both concepts are used synonymously in clinical settings.

predispose to diseases of primarily environmental origin while environment can influence diseases of mostly genetic origin (Porter, 1982).

Four different types of genetic diseases can be distinguished depending on the abnormalities contributing to the disease:

1. *Mendelian (or monogenic) diseases* are caused by mutations in a single gene (Peltonen and McKusick, 2001). These rather rare disorders can segregate as autosomal recessive, autosomal dominant, X-linked or Y-linked traits (Read and Strachan, 2003). Classical Mendelian diseases include Sickle cell anemia, Huntington's disease or Cystic fibrosis.
2. *Multifactorial (or polygenic) diseases* are caused by a combination of environmental factors and mutations in multiple genes, each contributing a small effect to the disease (Peltonen and McKusick, 2001). For instance, several genes that influence breast cancer susceptibility have been found on chromosomes 6, 11, 13, 14, 15, 17, and 22. The majority of human diseases belongs to this category, including several congenital defects and a number of adult-onset diseases, such as Alzheimer's or Parkinson's disease and autoimmune diseases (Hunter, 2005).
3. *Chromosomal diseases* are caused by abnormalities in the chromosome structure, namely, extra chromosomes (addition or duplication), missing chromosomes (deletion), or the relocation of parts of one chromosome onto another (translocation) (Gillberg, 1998). Down syndrome (or Trisomy 21), for instance, is a common chromosomal disorder caused by the duplication of chromosome 21 (Roizen and Patterson, 2003).
4. A relatively rare type of genetic disorders, *mitochondrial diseases*, is caused by mutations in the nonchromosomal DNA of mitochondria (Schaefer *et al.*, 2004). Mitochondrial defects have been implicated in a wide variety of degenerative diseases, such as aging, and cancer (Wallace, 1999; Schapira, 2006).

Determining the association between a disease and its causal genes, i.e., genes contributing to a disease when being mutated (further referred to as *disease genes*), remains a major challenge for molecular medicine. For many human diseases it is not yet known which genes are involved in their pathogenesis. For instance, currently more than 7,000 Mendelian and genetic disorders are documented in OMIM (McKusick, 2007) but for ∼4,000 of those the molecular cause is still unknown.

Identifying the origin of human genetic diseases is mainly based on finding statistical associations between genomic variations and clinical phenotypes (Cardon and Bell, 2001). Traditional gene-mapping approaches, such as genetic linkage analysis and gene association studies, are used to associate chromosomal regions with a disease (Botstein and Risch, 2003). Yet, knowing the associated genomic region is often not sufficient to detect the associated gene(s). Most efforts yield large genomic intervals of 0.5–10 centimorgan[9] with up to several hundreds of candidate genes (Jorde, 2000; Glazier *et al.*, 2002) as pedigrees are often too small and reproduction cycles are too long, particularly

[9] In humans one centimorgan corresponds to approx. one million base pairs on average (Lodish *et al.*, 2007).

Table 6.1: List of genomic regions associated with Alzheimer's disease (AD) in OMIM. AD is a genetically heterogeneous disorder. The different AD subtypes are linked to several chromosomal regions with different sizes (in mega base pairs, Mb) and a varying number of genes located in each region (determined by Biomart). Note that the prefix (#) represents phenotypes whose molecular basis are known while (%) denotes phenotypes or susceptibility loci whose molecular basis are unknown.

AD type	OMIM ID	Locus	Locus size in Mb	# of genes
AD 1	#104300	21q21	15.2	181
AD 2	#104310	19q13.2	4.4	185
AD 3	#607822	14q24.3	5.5	110
AD 4	#606889	1q31-q42	51	924
AD 5	%602096	12p11.23-q13.12	22.1	290
AD 6	%605526	10q24	7.7	232
AD 7	%606187	10p13	5	84
AD 8	%607116	20p	27.1	487
AD 9	%608907	19p13.2	5.7	274
AD 10	%609636	7q36	11.3	224
AD 11	%609790	9p22.1	1.4	25
AD 12	%611073	8p12-q22	76	1198
AD 13	%611152	1q21	10.9	545
AD 14	%611154	1q25	12.8	244
AD 15	%611155	3q22-q24	18.9	311
AD 16	%300756	Xq21.3	12	73

for rare diseases. Experimental studies for testing all candidates are time-consuming and costly. Often it is simply impossible to establish the true disease-gene relationship by inspecting genes within an interval.

Detecting genetic factors for diseases without confirmed or with multiple associated genomic regions is even more complicated than for monogenic diseases. Alzheimer's disease, for instance, is linked to more than 16 chromosomal regions containing up to 1,198 genes. However, only four loci are associated with a causal gene yet (see Table 6.1). The pleiotropy of genes, i.e., their ability to produce multiple phenotypes, and the heterogeneity of multifactorial diseases that do not obey the standard Mendelian patterns of inheritance pose limitations to traditional gene-mapping approaches (Giallourakis *et al.*, 2005). In addition, genetic factors often account only partially for complex phenotypes.

Thus, alternative techniques emerged for studying such disease phenotypes. These use single nucleotide polymorphisms (SNPs) (Marchini *et al.*, 2007), microarray expression analysis (Farber and Lusis, 2008), serial analysis of gene expression (SAGE) (Hene *et al.*, 2007) and more recently also copy number variations (CNVs) (Mardis, 2008) to analyze alterations in diseases. Genome-wide surveys, for instance, systematically assess the contribution of common SNPs or CNVs to complex diseases by discovering statistically significant associations between SNP genotypes or CNV measurements with gene expression phenotypes (Stranger *et al.*, 2007; McCarroll and Altshuler, 2007). However, such methods generate typically large sets of potential candidate genes for a given phenotype (Tiffin *et al.*, 2009).

6.1.1 Bioinformatic approaches to disease gene identification

To overcome experimental limitations and accelerate disease gene identification, the concept of *disease gene prioritization* emerged, i.e., integrating computational biology with the broad range of genomic data. Disease gene prioritization aims for identifying the most promising genes from large candidate sets obtained either from genome-wide association studies or from the (multiple) causative regions associated with a disease under consideration. The basic principles of this concept are similar to function prediction as inferring function of a gene or its implication in a disease are two closely related problems. However, associating genes with diseases is far more challenging as diseases often imply intricated mechanisms involving distinct molecular functions and pathways (Myers *et al.*, 2006).

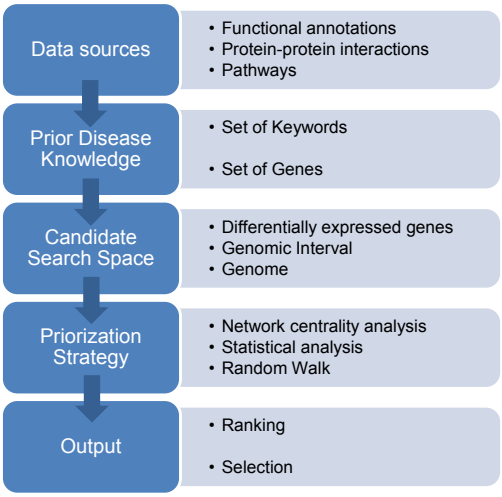


Figure 6.1: Basic work flow of disease gene prioritization.

One of the first approaches addressing this problem has been proposed by Perez-Iratxeta *et al.* (2002). The data-mining system for associating genes with genetically inherited diseases has later been implemented as a web application, namely G2D (Perez-Iratxeta *et al.*, 2005). Since then, various methodologies have been developed for identifying disease-related genes which will be discussed in detail in Section 6.4. These methods differ primarily in (i) the data sources they use, (ii) the included prior knowledge about a disease of interest, (iii) the candidate search space, (iv) the prioritization strategy and (v) the outcome they deliver (Tranchevent *et al.*, 2010) (see Figure 6.1):

- *Data sources:* Different experimental data can be used to represent gene characteristics that may correlate with disease phenotypes (Tiffin *et al.*, 2009). The most important ones are sequence features, gene expression data, pathway data, protein interactions, and functional annotations. Some methods exploit only a single data source while others integrate several complementary evidence (Aerts *et al.*, 2006; Franke *et al.*, 2006).

- *Prior disease knowledge:* Prior knowledge represents the current information about the disease under consideration. Such knowledge can, for instance, be defined by a set of keywords describing the different aspects of the disease or by a set of genes known to play a role in the disease. In the latter case, training sets are compiled from genes associated with the disease of interest. Alternatively, when no disease genes are available, proteins associated with pathways or processes perturbed by the disease can also be employed. The prior knowledge is used to deduce relationships between disease-causing and potentially related genes.
- *Candidate search space:* The candidate search space refers to the set of genes that represents the candidates for prioritization. Prioritization methods are often applied to genomic regions that have been associated with the disease, e.g., by linkage analysis. The average linkage interval in OMIM contains, for instance, 108.8 genes (Lage *et al.*, 2007). Otherwise, a list of differentially expressed genes can be used as candidates. In case no candidate set can be defined beforehand, e.g., due to the lack of associated genomic regions, the whole genome must be explored as candidate search space.
- *Prioritization strategy:* The core of each prioritization strategy is the algorithm for relating genes to a disease. Typically, several types of experimental evidence are first integrated and then different scoring methods, such as network centrality, order statistics or Bayesian predictor, are applied to score disease-gene associations (Ideker and Sharan, 2008). The common idea behind all scoring methods is the guilt-by-association principle: the most promising candidate(s) will be the one(s) that are most similar to the genes already associated with the disease (Tranchevent *et al.*, 2010).
- *Prioritization outcome:* Two types of prioritization outcomes can be distinguished: ranking or selection of candidate genes. In the ranking scenario, candidates are ranked according to their associated scores such that the highest scoring genes present the most promising candidates for further studies. A selection returns a subset of the original candidates comprising only the most promising candidate genes. A selection can be obtained either directly from a scoring method or from a ranking by using a threshold.

Leave-one-out cross-validation is generally used to evaluate prioritization methods. To this end, a known association of a gene with a disease is removed from the data to assess whether the algorithm recovers the hidden disease gene and at which rank. The set of potential candidates used in an evaluation differs depending on the search space. For instance, linkage interval dependent methods are commonly evaluated on artificial linkage intervals, i.e., defined as list of 100 – 110 genes located around the disease gene of interest according to their genomic distance on the chromosome (Lage *et al.*, 2007; Köhler *et al.*, 2008). Genome-wide methods, on the other hand, consider the entire genome for cross-validation. Others define candidates as a set of randomly selected genes and the blinded gene (Aerts *et al.*, 2006).

6.1.2 Using protein interaction data for disease gene association

Human diseases are often caused by perturbations in multiple genes. Mutations in genes with similar function often lead to the same or similar phenotype(s) which indicates that the underlying genes are likely to be functionally related. For instance, genes associated with the same disease share up to 80% of their functional annotations and protein domains (Turner *et al.*, 2003). Also genes from related biological pathway exhibit significant sequence similarity with other pathway members (Aerts *et al.*, 2006). Several functional characteristics correlate with disease phenotypes which can be exploited to identify novel genes for particular disease phenotypes (Tiffin *et al.*, 2009).

In particular, the increasing availability of protein interaction data provides valuable evidence through which disease-gene associations can be inferred (Navlakha and Kingsford, 2010). Because physically interacting proteins tend to be involved in the same cellular processes, interactions are direct and robust manifestations of functional relationships. In the context of understanding disease mechanisms at molecular level, several observations motivate the usage of protein interactions (Ideker and Sharan, 2008). Genes associated with a particular disease phenotype

- interact preferentially with genes known to be involved in the same disease (Ideker and Sharan, 2008),
- tend to exhibit a higher connectivity within the interaction network than non-disease gene products (Jonsson and Bates, 2006),
- occur in central network locations and
- often share topological network features with each other (Gandhi *et al.*, 2006; Xu and Li, 2006).

Another important aspect which is also exploited in this work is the concept of modularity (see Section 6.3.1). Several genetically heterogeneous hereditary diseases are known to be caused by mutations in gene products that participate in the same protein complexes. Such phenotypes might reflect underlying mechanisms in which the disease-related genes form some kind of functional module, e.g., a signaling pathway or multi-protein complex (Brunner and van Driel, 2004).

One example is Fanconi Anemia, a genetically heterogeneous disease associated with chromosomal instability, congenital abnormalities, progressive bone-marrow failure and cancer susceptibility. Fanconi Anemia originates from mutations in at least one of 13 distinct FANC genes whose products are believed to be involved in a common DNA repair signaling pathway – the Fanconi Anemia (FA) pathway (Kennedy and D'Andrea, 2005). As illustrated in Figure 6.2, these proteins cooperate closely with DNA repair proteins to prevent DNA from damage, induced through DNA interstrand cross-links and double-strand breaks, during replication (Patel and Joenje, 2007). Eight of the FANC proteins form a large core complex which is thought to play a central role in sensing and repairing DNA damage or in stabilizing chromosome structures. Mutations in any of these proteins disrupt the function of the FA pathway which in turn may result in chromosomal instability (Kennedy and D'Andrea, 2005).

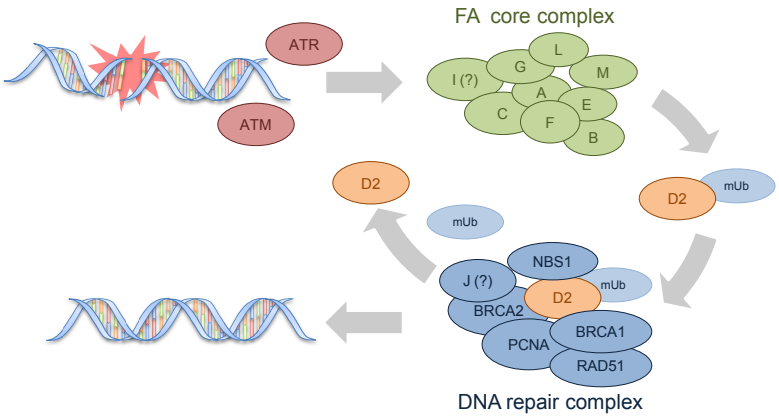


Figure 6.2: Schematic representation of the Fanconi Anemia DNA pathway. At least eight FANC proteins (A, B, C, E, F, G, M, L and possibly I) form the nuclear FA core complex which is activated upon DNA damage by DNA damage sensor proteins, such as ataxia telangiectasia mutated (ATM) or ataxia telangiectasia mutated and Rad3-related (ATR). The activation triggers the monoubiquitination of FANCD2 through the E3 ligase function of the FA complex. FANCD2 co-locates thereupon to the damage site and forms a DNA damage inducible foci with DNA repair proteins, e.g., BRCA1 and RAD51, which induces DNA repair. FANCD2 is deubiquitinated after DNA repair and the DNA replication fork proceeds (Patel and Joenje, 2007).

6.2 Overview

In this chapter, we present an interval-independent, network-based algorithm to identify disease-related genes. Our algorithm is particularly applicable for complex diseases without associated or with multiple causative genomic regions. For a given disease, we first extract all genes that are known to be associated with this disease (as seed genes). We compile a disease-specific network by integrating directly and indirectly linked gene products based on protein-protein interaction data and functional similarity. Proteins in this network are ranked based on network centrality. While the general approach is similar to those of other methods (see Related Work in Section 6.4), we use two distinctive features that improve our results considerably, in particular for diseases without associated loci.

- We consider genes indirectly linked to a seed gene. Thus, we uncover susceptibility genes that are not directly linked but that are part of the same pathway. This leads to more comprehensive disease networks and significantly increases recall. However, it also lowers precision, as larger networks naturally integrate many global "hub" proteins that also receive high centrality scores. The role of such hubs, i.e., proteins with an extremely high number of interaction partners, in diseases is controversial. Although hubs tend to be essential for many processes (He and Zhang, 2006; Zotenko *et al.*, 2008), they mostly are disease-unspecific (Goh *et al.*, 2007). Therefore, we developed a normalization procedure to down-rank such unspecific proteins.

- We use predicted functional information to overcome the incomplete functional coverage of the human genome (Chen *et al.*, 2009b). Most methods cannot consider genes that are functionally uncharacterized, which in turn prevents the detection of truly novel disease-gene associations. In contrast, we use predicted functions to increase the outreach of our networks and to assist the proper ranking of proteins without confirmed functional annotations. However, this also yields larger networks which makes an appropriate normalization even more important.

An important property of this approach is its generality. Although we introduced the framework for identifying novel proteins associated with genetic disorders, it can be used to address various biological questions, e.g., detecting further members of cellular processes, pathways or other definable mechanisms. For instance, in Section 7.5 we will report on how the framework can be employed to infer surface membrane factors that contribute to HIV-1 infection, a phenotype which clearly cannot be reduced to a genomic region.

6.3 Network-based disease gene identification

We developed a generic framework that infers novel disease-gene associations from disease-specific networks using network centrality analysis. The underlying assumption of our method is that the most central genes/proteins in a specific disease network are likely to be related to the disease (Özgür *et al.*, 2008; Chen *et al.*, 2009a).

The workflow comprises three steps as illustrated in Figure 6.3:

1. First, a *seed set* is defined from proteins that share specific characteristics of interest. This can be a set of proteins associated with a certain disease, involved in specific pathways, or transcripts that are differentially expressed in a condition of interest.
2. In the second step, a *disease-specific similarity network* is compiled. Starting from the seed set a graph is build by adding proteins based on their functional similarity to the seed set. In first place functional annotation and interaction data are used but other genomic data, such as expression data, sequences and phenotypes, can be integrated.
3. Finally, *network centrality analysis* is performed to rank proteins with respect to their relative importance within the network. The most central proteins are presumed to be of functional importance for the specific network.

In the following, we explain the details of the proposed framework with respect to the identification of novel disease genes. However, one should keep in mind that the framework is neither domain nor disease specific.

6.3.1 Building Disease Networks

Given a disease, we first map the genes that are associated with the disease in OMIM to their protein(s). These proteins are used as seeds for generating a disease network (Goh

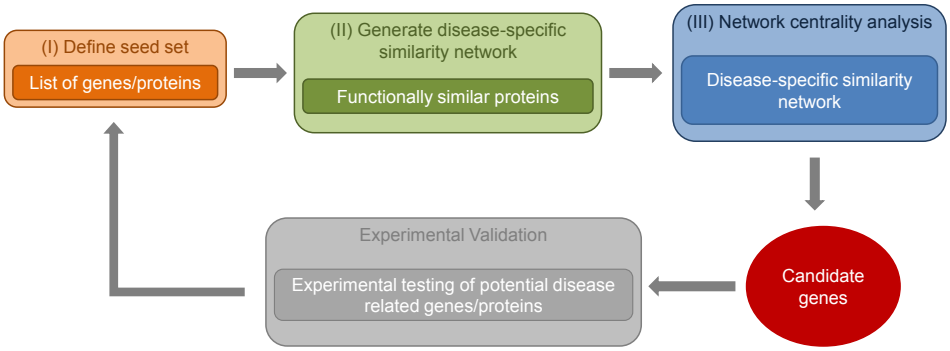


Figure 6.3: Conceptual framework for disease gene ranking. The method consists of three components. I) Definition of a *seed set* from genes/proteins sharing specific characteristics of interest; II) Generation of a disease-specific similarity network by including functionally related proteins (see Figure 6.4); III) Network centrality analysis to score candidate genes. The final step is the experimental validation of the identified candidate genes/proteins. Confirmed genes/proteins then can be included in the seed set and steps I-III can be repeated.

et al., 2007). The disease-specific network is initialized with the seeds and then extended by adding all proteins that interact either directly or indirectly with any seed protein or that are functionally similar to at least one seed (see Figure 6.4). We call the set of directly linked partners d_1 neighbors and the set of directly or indirectly linked partners (through one common interactor) d_2 neighbors ($d_1 \subseteq d_2$). Functional similarity between two proteins is determined by using the semantic similarity measure defined in Eq. 4.1.3.2 (Couto *et al.*, 2007) using only annotations from the GO subontology *biological process* (see Section 7.1). In principle, proteins are considered as functionally similar if their semantic similarity to a seed protein is above a pre-defined threshold. Thereby, we only consider close and significant biological relationships.

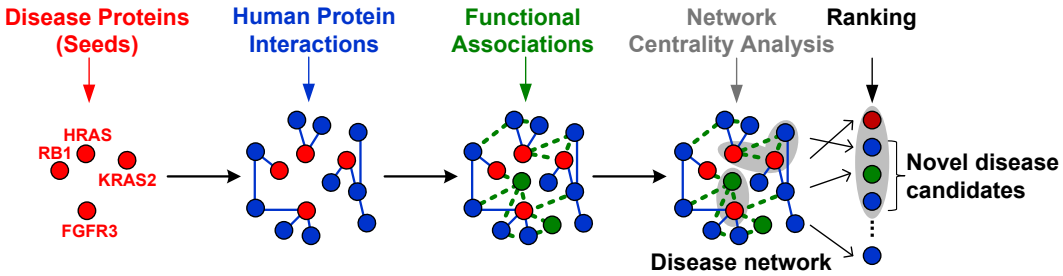


Figure 6.4: Illustration of the main steps in the prediction method. Starting from known disease proteins we add proteins that either 1) interact directly or indirectly with any of them (blue solid edges) or 2) that are functionally similar (green dashed edges) to at least one disease protein. This yields a disease-specific network. Proteins are then ranked according to their centrality within the network. Proteins in shaded areas represent highly central proteins and thus promising candidates.

Functional enrichment As the functional coverage of human proteins is limited, i.e., currently only a fraction of the genome is annotated with pathways, functions and phenotypes (Chen *et al.*, 2009a), we integrate predicted functions into the framework. We apply the network-based prediction method described in Chapter 4 to infer function (Jaeger *et al.*, 2010a). Predicted functions are used in the same way to infer functional relationships as original annotations. This improves in first place the ranking of disease proteins (see Section 7.2) but also increases the overall cross-validation recovery rate (see Section 7.3).

6.3.2 Disease Network Centrality Analysis

Once a disease-specific network has been generated, we apply network centrality analysis to identify the most relevant candidates for the disease. Different centrality measures have been proposed for analyzing various types of biological networks (Junker *et al.*, 2006; Koschützki and Schreiber, 2008). We investigated the following centrality measures (see Section 2.3.2.3 for definitions):

- Degree centrality
- Closeness centrality
- Betweenness centrality
- PageRank centrality.

We chose betweenness centrality for all further experiments because it (a) performs best on our data (see Section 7.2 for a comparison of the four measures) and (b) also showed favorable properties for generating new hypotheses on disease-gene associations by others (Özgür *et al.*, 2008). Accordingly, we rank all proteins with respect to their betweenness centrality within the network using the *igraph* library in R (Csardi and Nepusz, 2006).

6.3.2.1 Normalization for Hub Proteins

Betweenness centrality is applied to identify proteins that are central within disease-specific networks (*local hubs*). However, the ranking of disease-relevant elements becomes more difficult in large disease networks, for instance, when integrating d_2 neighbors or considering diseases with a large number of seed genes.

An important property of whole cell protein interaction networks is their *scale-free topology* (Albert, 2005), as discussed in Section 2.3.2.2. Thus, the more proteins are integrated in a disease network the higher is the likelihood of including *global hubs*, e.g., proteins with many interaction partners, independent of any disease context. These hubs affect the ranking since they often will be central due to their general high (but unspecific) connectivity rather than due to a particular relevance for a disease. However, hubs cannot be simply removed from a disease network because this would destroy their topology and might also affect disease-relevant local hubs (by means of missing links).

To account for these effects, we adjust the ranking for all proteins by considering their individual distribution across many disease networks. This is based on the assumption

that proteins that are involved in various disease networks are less disease-specific than those that occur only in particular networks. Highly ranked proteins, integrated in many disease networks, are likely to present global hubs that are not disease relevant.

We generate all disease networks for OMIM diseases and count for each protein P in how many disease networks it is involved. We define a normalized betweenness centrality score BC_N for P in a disease network D by normalizing the betweenness centrality score BC by the frequency of P across all disease networks:

$$BC_N(P|D) = \frac{BC(P|D)}{|\{k|P \in D_k\}|} \tag{6.1}$$

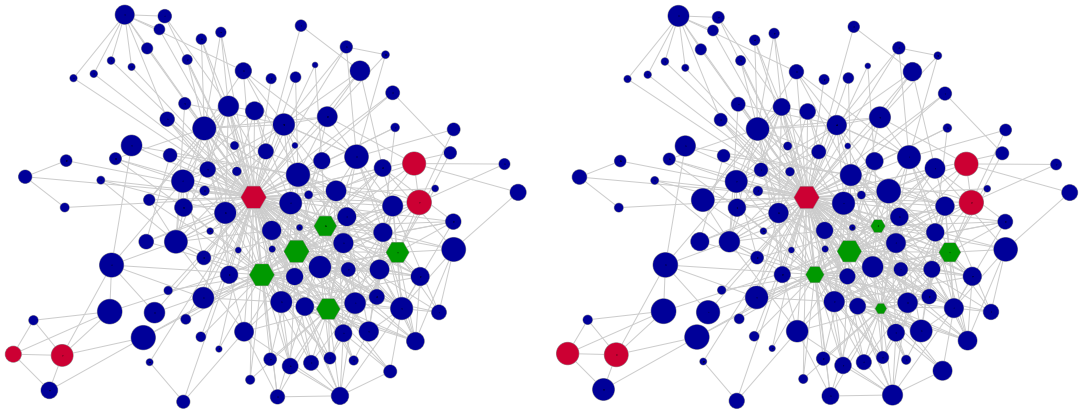
Proteins are then ordered according to their BC_N score for further analysis. Thus, proteins occurring in many disease networks (especially global hubs) are adjusted downwards.

The effect of the proposed hub normalization is exemplarily illustrated in Figure 6.5. The figure shows the prioritized d_1 disease networks, with and without hub correction, for *Familial Atypical Mycobacteriosis* (OMIM Id 209950), a tuberculosis-like disease caused by mycobacteria other than *Mycobacterium tuberculosis*. The mycobacteriosis network has been generated from five seeds proteins and comprises in total 119 proteins of which six are global hubs with more than 23 interactions (see Section 7.2). Proteins in the network are ranked according to their betweenness centrality whereas the rank of each protein is reflected in the node size, i.e., the larger the node the higher is its centrality and its rank.

Figure 6.5(a) indicates that most disease proteins are fairly central. Two seeds are ranked among the top five proteins while the remaining seeds are found among the top 53 proteins. However, also hub proteins are very central due to their high number of interactions which compromises the ranking of disease proteins. For instance, three hubs (of which one is a seed) are among the top 5 proteins. Yet, not all of them are disease relevant. Normalizing the centrality scores according to the protein frequencies estimated across all disease networks corrects most of the hubs downwards (see Figure 6.5(b)). In consequence, only one hub protein, the (hub-)seed, is found among the top five proteins. In turn, the ranking of true disease proteins improves considerably, e.g., the set of seed proteins can be found within the top 23 proteins. Figure 6.5(b) also demonstrates that our normalization effects mostly non-specific hubs as the rank of the hub-seed protein is not altered by the correction. Note that for more clarity we considered a fairly simple d_1 example disease network. The impact of hub proteins and the hub normalization on the ranking is much more pronounced in d_2 disease networks as we will show Section 7.2.

6.3.3 Evaluation methods

We shall evaluate our method in three ways. First, we verify whether (known) disease proteins are highly ranked in their disease-specific networks. Second, we assess the ability of our method to discover novel disease proteins by performing a leave-one-out-validation over all known disease proteins. For both cases, we study the top k% ranked proteins within a disease network for different values of k (from 1% to 100%). We compare the



(a) Ranking without hub normalization (b) Ranking with hub normalization

Figure 6.5: Effect of hub normalization on the protein ranking in a d_1 disease network generated for *Familial Atypical Mycobacteriosis*. Known disease proteins are shown in red. Hub proteins are represented as hexagons (green for non-seeds and red for seeds). The size of each node correlates with its betweenness centrality score and thus with its rank.

performance of our methods against two related methods, namely random walk with restart (RWR) applied by Köhler *et al.* (2008) and PRINCE (Vanunu *et al.*, 2010).

Centrality analysis and cross-validation are defined below while the detailed description of the two related strategies is provided in Section 6.4.2.

6.3.3.1 Centrality of disease proteins

We determine the amount of highly ranked disease proteins in a disease network by counting the number of seed proteins among the top k% ranked proteins of the network. Clearly, we expect the majority of seed proteins to be highly ranked in the prioritized list, since we build the disease networks around them which naturally puts them in a central position. However, not all seed proteins are central in their disease networks, and many non-seed proteins are highly ranked. We are especially interested in the latter since these present promising candidates for novel disease-gene relationships (Özgür *et al.*, 2008).

Note that this type of evaluation is often used for analyzing the performance of disease gene identification methods (see Section 6.4). Yet, this evaluation only reflects a method's ability to score and to rank candidates with respect to a particular disease rather than its predictive power as all disease genes remain in the data set. To assess whether a method is capable of *de novo* identification, disease gene associations have to be removed from the data (see below).

6.3.3.2 Cross-validation

For leave-one-out cross-validation, we consider all OMIM diseases with two or more known disease proteins, since our method requires at least one disease protein as seed. For each disease, we remove one associated protein from its set while using the remaining ones as seeds. We apply our method as described and count how often the blinded disease-associated protein is re-discovered. We consider only those proteins as re-discovered that rank among the top k% proteins of the prioritized list. We repeat this procedure for each seed per disease and determine an average relative recovery rate for different values of k (using macro average).

The inclusion of additional evidence leads to larger networks and thus to a higher number of potential candidates. Hence, the ratio between promising and false positive candidates decreases and the number of proteins in a top-k% list increases. To assess whether we truly gain additional information from our extended networks, we also study the absolute recovery rate by performing cross-validation as explained above using only the top 100 proteins within each network.

Further evaluations

We shall use the two evaluation settings described above to further assess the performance of our approach according to the following aspects:

- First, we measure the effect of integrating predicted protein functions into the framework. We compare the ranking of disease-related proteins in functionally enriched disease networks with their ranking in non-enriched networks. We will demonstrate, that predicted functions enhance the ranking of disease-related proteins (see Section 7.2).
- Second, we assess the impact of utilizing indirect interactions on finding disease-related gene products. To this end, we compare the recovery rates achieved when considering either direct (d_1) or also indirect interaction partners (d_2) against each other. We will show, that the inclusion of indirectly linked proteins significantly improves the cross-validation recovery rate (see Section 7.3).
- Third, we verify our hypothesis that the inclusion of indirect interaction partners also yields a higher number of (disease-unspecific) hub proteins which in turn compromises the ranking of proteins relevant to the disease of interest. In addition, we assess the impact of the hub normalization on the ranking and show that the proposed strategy is most effective for filtering proteins unrelated to a particular disease (see also Section 7.2).
- Fourth, we quantify the influence of the number of known disease-related genes on the performance of our method. Therefore, we analyze the cross-validation recovery rates with respect to the number of initial seed proteins (see Section 7.3.2).
- Fifth, we study whether the specific disease type influences the prediction quality of our method. To this end, we group OMIM diseases into 22 distinct disease classes according to a disease classification scheme proposed by Goh *et al.* (2007) and

consider the recovery rates with respect to each disease class (see Section 7.3.3).

- Sixth, we assess how much our method could benefit from utilizing information on genomic regions, i.e., disease loci. Related methods are mostly evaluated on artificial linkage intervals with around 100 to 110 genes including the target gene which is incomparable to evaluations considering an entire genome (see Section 6.4). Therefore, we perform a cross-validation in which we filter all proteins from the ranked candidate list that are not located on the same chromosome as the left-out protein. This mimics a scenario where the candidate search is restricted to a particular genomic region, i.e., a chromosome in this case (see Section 7.3.1).

Finally, we show in two biologically relevant use-cases that our approach is highly applicable for diseases with complex and incompletely known genetic background. First, we apply our method to investigate classical Hodgkin Lymphoma (cHL) and colon cancer (see Sections 7.3.4 and 7.3.5). Second, we utilize our method to study surface membrane factors that might contribute to HIV-1 infection, a phenotype which cannot be limited to particular chromosomal regions (see Section 7.5).

6.4 Related Work

In the following, we discuss related work in the field of computational disease gene identification focusing on (interaction) network-based approaches. We start with a classification of the different methodologies based on their underlying ideas. Representative methods will be briefly discussed regarding their main concepts and distinctive features with respect to our approach.

The currently existing prioritization strategies can be classified into three categories:

1. *Local methods* infer disease association for a gene product by investigating either its direct or indirect interaction partners or the shortest paths between the candidate and known disease genes (Oti *et al.*, 2006; George *et al.*, 2006).
2. *Global methods* model the flow of information within the cell to assess the connectivity and proximity between known disease genes and candidate genes (Franke *et al.*, 2006; Köhler *et al.*, 2008; Vanunu *et al.*, 2010).
3. *Disease module-based methods* associate proteins with diseases based on the hypothesis that common phenotypes are associated with dysfunction in proteins participating in the same complex or pathway. These methods first construct disease-specific networks around a set of genes related to the condition of interest which are assumed to present modular disease-machineries (Chen *et al.*, 2006; Gonzalez *et al.*, 2007; Özgür *et al.*, 2008). Different scoring functions are then used to score and rank proteins in such networks according to their relevance to the disease.

According to this classification scheme, we follow a module-based strategy, by generating disease-specific networks, and employing a global similarity measure for identifying disease-related genes within such networks.

6.4.1 Local prioritization methods

The rational behind local prioritization methods is based on the following assumption: if two gene products interact with each other, the known association of one protein with a disease indicates that its interaction partner is also associated with the same disease (Goh *et al.*, 2007; Oti and Brunner, 2007; Ideker and Sharan, 2008).

An early approach (Oti *et al.*, 2006) inferred disease-causing genes for heterogeneous diseases in which some causative genes have been already elucidated, while for others only locus information have been detected. Given a disease, they first determine the direct interaction partners for each associated causative gene product. Interaction partners that are located in a previously identified locus are then predicted to be candidate genes for the disease of interest. The work demonstrated that the systematic use of protein interaction data facilitates disease gene prediction. The likelihood for finding the correct disease gene product in a given locus ranges from 9% to 17% for different high-throughput data sets and reaches 58% when using HPRD, a data source known to be biased toward disease proteins. When disregarding locus information, the combined high-throughput set of interaction data yields a prediction accuracy of 0.7%. Although this accuracy is still higher than finding causative genes by chance in the genome (0.005%), the low performance emphasizes the large dependence of the approach on defined linkage intervals for achieving reasonable outcomes.

The method of Oti *et al.* (2006) has two major flaws compared to the prioritization method presented in this thesis. First, its dependence on defined genomic regions excludes diseases without associated causative loci. Second, their strategy considers only direct interaction partners disregarding indirect relationships between disease-related proteins. Further limiting aspects are sparseness and quality of interaction data (see Section 2.2.2) which are inherent to all methods that are largely based on this data type. Missing interactions hinder the disease gene identification as the large fraction of proteins without available interaction data are neglected. Spurious interactions, on the other hand, induce associations without biological relevance which reduces the level of accuracy. Navlakha and Kingsford (2010), for instance, extended the method of Oti *et al.* (2006) by requiring more corroborating interactions to support a disease-gene association, i.e., at least two or three causative proteins had to interact with the respective candidate to consider it as prediction. This modification increases the predictive confidence as spurious interactions can be excluded. However, the successively higher precision comes also at expensive of a lower recall.

6.4.2 Global prioritization methods

Global prioritization strategies are based on the same assumption as local methods but take the global structure of interaction networks into account by modeling, for instance, the flow of information to assess the connectivity between known disease genes and candidate genes. Two prominent global strategies are those proposed by Köhler *et al.* (2008) and Vanunu *et al.* (2010). The idea behind both methods is to identify disease-related proteins as those which are most often visited when iterating over a network.

Proteins interacting with several disease proteins will receive high weights, as well as those that may not directly interact with any disease protein but are in close network proximity.

6.4.2.1 Random walk with restart

Köhler *et al.* (2008) proposed a global distance measure based on random walks with restart (RWR) to rank candidates. RWR computes the similarity between two gene products i and j based on the probability that a random walk through the network starting in disease protein i ends in candidate j when taking all possible paths into account. Given a disease of interest and an associated linkage interval, RWR ranks each positional candidate based on its relative location to all related genes of the disease.

Cross-validation has been performed on manually selected disease families and on artificial linkage intervals to evaluate RWR. Disease families have been compiled on the basis of OMIM, domain knowledge and additional literature or database information. RWR performs well in the described benchmark setting achieving an area under the ROC curve between 91% and 98%. However, these numbers apply only for diseases with associated linkage intervals. Furthermore, disease families used in this evaluation are, with about seven genes per family, much larger than the average disease in OMIM with about 1.25 genes (see Section 7.1). Thus, it remains unclear how the performance is influenced by a smaller but more realistic number of genes per disease.

The work of Köhler *et al.* (2008) differs in three aspects from ours. First, the method depends on associated genomic regions and disregards diseases without known linkage intervals. Second, it is only applicable for proteins with protein interaction data. Albeit predicted interactions are included in the global interaction network to overcome the incompleteness of interaction data, other sources of functional relationships are not considered. Third, the presence and impact of hub proteins on the distance measure are not taken into account in this work. Thus, hub proteins might receive higher similarity scores as they will be connected through multiple paths although they are not necessarily relevant for the disease.

6.4.2.2 PRINCE

Vanunu *et al.* (2010) developed a propagation-based approach which integrates protein interaction data and disease information in terms of phenotype similarity to score the strength of a potential association of proteins with a disease of interest. A scoring function is defined based on a network propagation algorithm that simulates an iterative process where proteins with prior information pump flow to their network neighbors. Each protein propagates the flow it received in the previous iteration to its neighboring proteins. The scoring is designed to be smooth over the network, i.e., adjacent nodes are assigned with similar values to exploit prior information on the involvement of proteins in the same or similar diseases. The propagated flow converges after several iterations and the final score for each proteins is determined by the amount of flow a protein received during the iterations.

Cross-validation has been performed to evaluate PRINCE on 1,369 OMIM diseases with at least one associated gene product in the interaction data using artificial linkage intervals. 34% of the top-scoring candidates have been correctly re-discovered. PRINCE outperformed two other state-of-the art global approaches: RWR and Cipher (Wu *et al.*, 2008) achieved inferior results with 28.8% and 24.7%, respectively, when considering the top-ranked candidate as prediction.

PRINCE is one of the few prioritization methods that has been applied in a genome-wide setting although its performance has so far only be assessed on shorter linkage intervals. In contrast to our framework, the algorithm does not reduce the number of relevant candidates by considering only functionally related proteins for the ranking. This complicates the prioritization process as ranking in such large networks becomes much more difficult (Wu *et al.*, 2008). Albeit PRINCE exploits protein interaction data and phenotype similarity, other relevant data, such as function or predicted information, are not taken into account. Lastly, the effect of hub proteins on the flow simulation is neither examined nor taken into account. Consequently, hub proteins will receive a larger amount of flow due to their higher connectivity within the network.

6.4.3 Disease module-based methods

The modularity of the cellular interactome indicates that many genes perform their function as components of protein complexes or functional modules (see Section 2.3.2.4) which might also have implications for diseases. For instance, mutations in single genes might disrupt the complete module while mutations in multiple proteins constituting such a module might induce the same phenotype. Thus, several approaches follow a modular approach toward identifying novel disease-causing. Most approaches proceed from a number of disease associated genes and grow a network around them by adding physical or functional interaction data. Once disease-related modules have been identified, scoring strategies known from direct methods can be applied to prioritize candidates from the network (Chen *et al.*, 2006; Gonzalez *et al.*, 2007; Özgür *et al.*, 2008).

6.4.3.1 Method of Chen *et al.* (2006)

Chen *et al.* (2006) proposed a method to identify proteins associated with Alzheimer's Disease (AD). They first collect an initial seed list of genes known to be involved in AD which is expanded by integrated protein interaction data to generate an Alzheimer-specific protein interaction (sub-)network. The AD-specific network is then analyzed to prioritize proteins according to their relevance for AD using a heuristic scoring function that ranks proteins based on their connectedness in the network, reflecting their overall role and contribution to the AD-specific interaction (sub-)network.

A major drawback of this approach is its bias toward known AD-related proteins. When building the network, only interactions among seeds and between seeds and their direct neighboring proteins are considered while interactions among the neighbors are disregarded. In consequence, only one out of 20 proteins presents a novel finding when assessing the top-20 candidates. The remaining proteins are initial seeds. See below for

a comparison to our method.

6.4.3.2 Method of Özgür *et al.* (2008)

Özgür *et al.* (2008) proposed a similar approach using literature mining and network analysis. They start with an initial list of seed genes known to be associated with a disease of interest and generate a disease-specific interaction network by extracting interactions among the seed genes and their neighbors from the biomedical literature. Next, network centrality analysis is applied to rank the genes in the network according to their relevance to the disease, assuming that genes which are central in the disease-specific network are likely to be related with the disease. This assumption was verified on prostate cancer using four centrality measures: degree, eigenvector, closeness and betweenness centrality metrics. Degree and eigenvector centrality have been shown to achieve highly accurate results, for instance, 95% of the top-20 genes are actually related to the disease. Closeness and betweenness centrality yield genes that are currently not associated with the disease of interest.

Although both seed-based methods are very similar to ours, they differ in several points. First, only direct interaction partners are considered when constructing the network while we also include indirect interactions. In addition, Chen *et al.* (2006) disregard interactions between non-seed proteins which favors known disease proteins. Özgür *et al.* (2008) derived interactions only from the literature. However, methods relying on text mining data are inherent to a knowledge bias and thus might perform better on known historical data than in a prospective setting in which new disease-related genes are explored (Lage *et al.*, 2007; Köhler *et al.*, 2008). Third, no additional data sources have been exploited which limits the coverage of each method. The true predictive power of both approaches remains unclear as cross-validation has not been performed.

6.4.3.3 Phenome-interactome protein complexes implicated in genetic disorders

A more advanced approach has been proposed by Lage *et al.* (2007) based on the observation that mutations in different members of a protein complex lead to comparable phenotypes. Lage *et al.* (2007) apply a Bayesian predictor to prioritize candidate genes from linkage intervals by assigning candidates to protein complexes based on the phenotypes associated with its members. Given a particular phenotype and a linkage interval, candidate genes are ranked as follows:

- First, direct interaction partners are extracted for each positional candidate which are used to constitute a so-called candidate complex.
- Second, proteins within each complex are annotated with disease information. The similarity of each protein to the candidate is determined by measuring the phenotypic overlap among the proteins and the candidate using text mining.
- In the final step, a Bayesian predictor is used to score each candidate by assigning posterior probabilities based on the phenotypes associated with the proteins in the candidate complex.

Positional candidates are ranked according to this score and a prediction is made if the score exceeds a certain threshold. Validation was performed on 1,404 linkage intervals and only candidates scoring above 0.1 were considered as predictions which accounted for 25% of the candidate complexes. 45% of the candidates ranked as top-scoring proteins have been correctly identified as gene products relevant to the disease of interest. This very good performance can be attributed to a combination of different factors. First, protein interaction data are integrated with a phenotype similarity scheme which allows to take advantage of the entire clinical spectrum of related human diseases. In addition, only high confidence protein interaction data (either supported by network topology, different publications, reliable small-scale experiments, reproducibility or a combination of these) are considered. Further, cross-species interaction data are incorporated to increase the limited coverage of human interaction data and to provide a comprehensive data basis.

The major drawback of this approach, compared to the method developed in this thesis, is its dependence on linkage intervals as source for positional candidates which renders it inapplicable for diseases without associated linkage intervals. Another limiting aspect is the usage of only direct interaction partners when generating candidate complexes. Relationships through indirect interactions cannot be captured which hinders the ability to incorporate interaction partners associated with a similar phenotype as the relevant protein into the candidate complexes.

6.4.4 Integrative approaches

As most of the interaction-based methods are limited by the quality and sparseness of the experimental data, several techniques follow an integrative approach, leveraging, for instance, functional annotations, gene expression data, protein sequences and their features to complement protein interaction data (Chen *et al.*, 2007b). The current state-of-the-art method among the more integrative systems is Endeveaour (Aerts *et al.*, 2006).

6.4.4.1 Gene prioritization through genomic data fusion

Endeavour is based on the integration and comparison of various gene characteristics to prioritize candidate genes according to their similarity to a set of known disease genes. The prioritization is carried out as a three-step analysis:

- First, information is gathered from a set of (training) genes known to be associated with the process of interest by considering various data sources, e.g., functional annotation, literature, EST and microarray expression, and protein domains.
- In the second step, a set of candidate genes is defined as, e.g., list of differentially expressed genes, chromosomal region, linkage interval or full genome. Candidate genes are then ranked according to their similarity to the functional properties reflected in the training set. This yields one prioritized list for each data source.
- In the last step, the rankings of each data source are fused into one global ranking using order statistics. Order statistics is able to handle missing values, thereby

avoiding penalizing incomplete genomic data sources while minimizing the bias for well-characterized genes.

One of the key strength of Endeavour is the usage of multiple data sources. Currently 26 distinct data sources can be selectively incorporated (Tranchevent *et al.*, 2008). Endeavour is also one of the few methods, including ours, which is capable of prioritizing genes involved in particular biological pathways. However, albeit Endeveaour is able to perform genome-wide prioritization this has not been evaluated yet for human diseases but only for receptor-signaling pathways in Drosophila (Aerts *et al.*, 2009).

7 Evaluation of Disease Gene Identification

In this chapter, we present the evaluation of the algorithm for genome-wide identification of disease-related genes described in Chapter 6. We apply our strategy to diseases documented in OMIM. We compile disease-specific networks for each disease with at least one associated gene and study each network with respect to the disease. Throughout this evaluation, we focus in particular on the effects of utilizing indirect interactions and predicted functions as well as on the impact of hub correction.

Chapter 7 is organized as follows. First, we describe in Section 7.1 the disease data providing the basis for this evaluation. We also investigate functional relationships between proteins associated with the same disease. In Section 7.2 we verify whether disease proteins are central in disease networks. We show that predicted functions enhance the ranking of disease-relevant proteins. Furthermore, we study the impact of global hub proteins on the ranking and discuss the effect of the proposed normalization strategy. We demonstrate that our hub correction decreases the fraction of highly ranked hub proteins while increasing the fraction of disease proteins. Section 7.3 proceeds with an extensive leave-one-out cross-validation of our proposed method. We show that indirect interactions significantly improve cross-validation recovery rates. In Section 7.3.1 we mimic a more constrained search by filtering for chromosomal regions which increases our recovery rates significantly. We also investigate whether the number of disease-associated proteins or the disease type influences the performance of our method (see Sections 7.3.2 and 7.3.3).

To test the ability of our algorithm to handle complex phenotypes not associated with any particular genomic region, we further assess its performance on classical Hodgkin Lymphoma (cHL) and colorectal cancer (CRC). In Section 7.3.4 we apply our method to epigenetic and gene expression data from cHL to (i) re-identify genes related to cHL pathogenesis and to (ii) discover new candidates that are not yet associated with this phenotype. Genes highly ranked by our method (i) overlap significantly with transcripts identified by *in vitro* cHL studies and (ii) are known to be involved in Hodgkin-related pathways. Novel candidates, such as *MYC*, show a number of interesting features making them important targets for further investigations. In a similar setting we compile a CRC-specific network from genes associated with this type of cancer in OMIM (see Section 7.3.5). Based on our method we infer novel CRC-related proteins from this network. We analyze the potential association of the most promising candidates by considering knowledge from literature, KEGG pathways and expression profiles.

Section 7.4 reports on the performance comparison with two network-based state-of-the-art approaches for associating diseases with genes, namely RWR (Köhler *et al.*, 2008) and PRINCE (Vanunu *et al.*, 2010). We apply our disease-specific approach as well as the two related methods to different disease sets and compare their performance. We

show that our approach performs significantly better than PRINCE across all disease settings we studied. A comparable performance can be achieved when comparing to RWR.

In another application we modify our framework to infer surface membrane factors that contribute to HIV-1 infection, a phenotype which cannot be reduced to a genomic region (see Section 7.5). We identify ten surface proteins that are involved in a cascade of events in HIV-1 infection. Their involvement ranges from serving as co-receptors for cell entry (*CCR1* and *CCBP2*), mediating trans-infection (*DARC*), activating immune cells (*CD97*) to inducing viral production from latently infected cells (*CSF3R*, *TNFRSF3* and *CD2*).

7.1 Disease Data

The most comprehensive source for human disease-gene association data is the *Online Mendelian Inheritance in Man* (OMIM) database, curated by the NCBI and Johns Hopkins University (McKusick, 2007). OMIM catalogs all human diseases with a genetic component, and links them – when possible – to the relevant genes in the human genome. Additionally, further references are provided as well as tools for genomic analysis of the documented genes. OMIM initially focused on classic monogenic disorders but has been extended to include complex traits and their associated genetic mutations that confer susceptibility to these common disorders. Although this focus introduces some bias, and the disease gene record is still far from being complete, OMIM represents currently the most complete and up-to-date repository of known disease genes and the disorders they contribute to (Goh *et al.*, 2007).

For this reason, we utilize OMIM as the source for disease-gene association data to evaluate our method. We used OMIMs Morbid Map[10] to extract diseases including their corresponding disease names and cytogenetic location(s). Genes associated with a disease have been retrieved from the OMIM Gene Map. As of May 2011, 7,061 mendelian diseases are documented in OMIM (see Table 7.1). 4,061 are associated with a defined phenotypic locus and 3,077 are associated with at least one gene contributing to the disease outcome. In turn, the underlying molecular basis of app. 4,000 diseases remains to be characterized.

Throughout this evaluation, we consider all disease-gene associations that encompass gene products in our data set. This comprises 3,077 diseases with on average 1.25 disease-related genes per disease (std $\pm$ 1.28, max = 27). Before we proceed with the different evaluation scenarios, we first investigate the functional relationships utilized in our framework. To this end, we study direct and indirect interaction relationships as well as functional similarity between proteins associated with the same disease and compare our findings against the same number of randomly selected protein pairs.

Figure 7.1 shows the fraction of disease proteins as well as randomly selected proteins that interact either directly or indirectly with each other. Figure 7.1(a) indicates that

[10]The OMIM Morbid Map presents a list of diseases documented in OMIM and their associated cytogenetic locations.

Table 7.1: OMIM statistics (May 2011). Number of disease entries in OMIM by entry type and genetic origin. (+) describes genes associated with a sequence and a disease phenotype. (#) indicates phenotypes with multiple loci (with and without associated genes). (%) denotes a confirmed mendelian phenotype or phenotypic locus for which the underlying molecular basis is not known. Phenotypes marked with (+) and (#) contribute to the 3077 diseases considered in this work. Note that the total number of both entry types is larger than 3077 as not all phenotypes are associated with a gene.

Entry type	Autosomal	X-linked	Y-linked	Mitochondrial	Total
+ Gene with known sequence and phenotype	314	18	0	2	334
# Phenotype description with multiple loci	2725	236	4	28	2993
% Mendelian phenotype or locus, molecular basis unknown	1632	134	5	0	1771
Other, mainly phenotypes with suspected mendelian basis	1831	130	2	0	1963
Total	6502	518	11	30	7061

a significantly larger number of disease proteins interact indirectly with each other than directly (p-value = $4.1 \cdot 10^{-10}$). On average 15% of the proteins associated with the same disease interact directly with each other while 28% of them interact through a common interaction partner. There are three possible explanations for this difference:

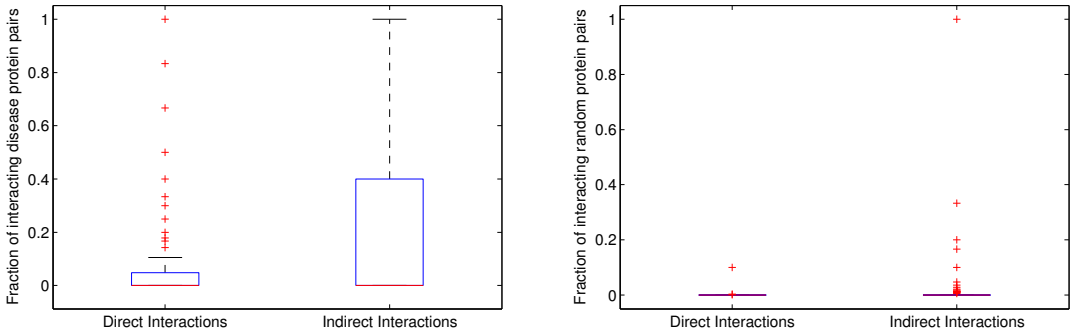
- First, given the incompleteness of human interaction data (see Section 2.2.2.2), not all relationships between disease proteins are represented in the data yet.
- Second, the smaller fraction of direct interactions between disease proteins might also indicate a stronger indirect relationship between disease proteins.
- Third, disease proteins do not necessarily interact with each other.

These findings underline the potential of including indirect interaction partners when attempting to identify disease-associated genes. Figure 7.1(b) shows the fraction of interactions among randomly selected protein pairs. Contrary to disease proteins, only 0.3% and 1% of the proteins interact directly or indirectly with each other, respectively. Although the fraction of directly interacting disease proteins is fairly low, it is still significantly larger (p-value = $4.3 \cdot 10^{-15}$) than the fraction of interactions among random protein pairs.

Figure 7.2 presents functional similarity for disease proteins and random protein pairs in terms of molecular function, biological process and cellular component. In general, the functional similarity among disease proteins is significantly higher than between random protein pairs across all subontologies (p-value $\leq 6.3 \cdot 10^{-62}$). The highest correlation between disease relatedness and functional similarity is detected for biological process followed by cellular component and molecular function. The difference between the subontologies is highly significant for molecular function (p-value = $1.5 \cdot 10^{-13}$) and still significant for cellular component (p-value = $1.1 \cdot 10^{-4}$), respectively. As disease

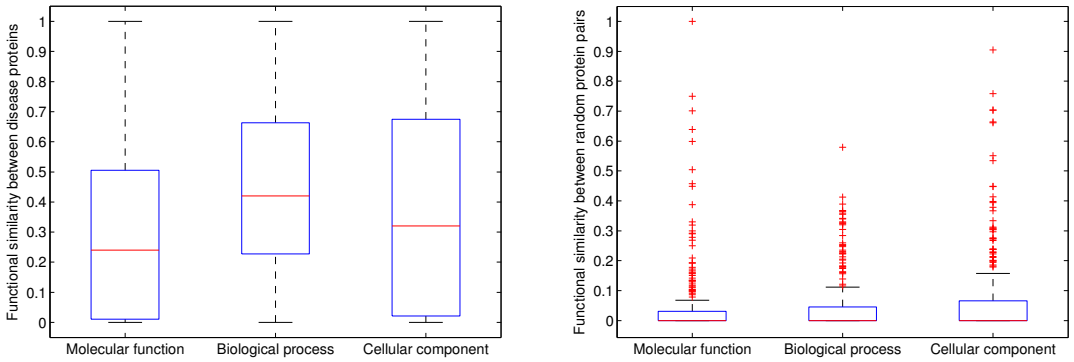
relationships are reflected best in biological process, we consider only this subontology when exploiting functional similarity for disease gene association.

In summary, our findings indicate that proteins associated with the same disease are more likely to interact with each other. Furthermore, they also tend to share common functions to a higher extent than non-disease proteins.



(a) Disease protein pairs (b) Same number of random protein pairs

Figure 7.1: Fraction of (a) disease proteins (involved in the same disease) and (b) randomly selected protein pairs that interact either directly or indirectly with each other. Note that the fraction of directly interacting random protein pairs approximates $\frac{|PPI|}{|P|^2}$.



(a) Disease protein pairs (b) Same number of random protein pairs

Figure 7.2: Average functional similarity between (a) disease proteins and (b) randomly selected protein pairs per subontology.

7.2 Centrality of Disease Proteins

As proof-of-concept, we first verify whether disease proteins are central in disease networks. To this end, we determine the number of seed proteins among the top $k\%$ ranked proteins (see Section 6.3.3) across four different disease network configurations:

Table 7.2: Disease network characteristics. For each network configuration the average (and median) number of proteins and edges is specified as well as the size of the largest disease network.

	Average (median) network size		Largest network	
Disease network type	# Proteins	# Edges	Proteins	Edges
$DN - GO\ d_1$	100 (58)	4,750 (440)	1,688	163,689
$DN - GO\ d_1$ *enrich*	130 (73)	8,295 (867)	2,089	172,469
$DN - GO\ d_2$	674 (359)	11,132 (3,397)	8,017	179,138
$DN - GO\ d_2$ *enrich*	700 (384)	15,797 (4,639)	8,125	236,206

- *DN – GO* d_1: direct interaction data as well as manually curated functional annotation
- *DN – GO* d_2: direct and indirect interaction data as well as manually curated functional annotation
- *DN – GO* d_1 *enrich*: direct interaction data as well as manually curated and predicted functional annotation
- *DN – GO* d_2 *enrich*: direct and indirect interaction data as well as manually curated and predicted functional annotation

The network-specific characteristics, i.e., average number of nodes and edges of the different network types, are summarized in Table 7.2.

Before we assess if predicted functional annotations and indirect interactions enhance the ranking of disease proteins, we investigate the performance of the four centrality measures described in Section 2.3.2.3: degree centrality, closeness centrality, betweenness centrality, and PageRank centrality. The difference in the ranking among these measures is illustrated in Figure 7.3. Betweenness centrality clearly outperforms the other centrality measures. However, it does not only perform best on our data but also shows favorable characteristics for generating new hypothesis on disease-gene associations (Özgür *et al.*, 2008). Accordingly, we use this centrality measure for all further experiments.

Figure 7.4 shows the fraction of highly ranked seed proteins within the four different disease network configurations. In general, we find more than 88% of the seed proteins among the 5% most central proteins which supports our hypothesis that disease proteins are central in their specific disease networks. Clearly, we expect the majority of seed proteins to be highly ranked in the prioritized list since we build the disease networks around them which puts them naturally in a central position. Yet, high centrality in disease networks is a characteristical feature of disease proteins (see Figure B.4). Comparing the ranking within disease and random networks, i.e., built from random proteins sets, shows that disease proteins exhibit a significantly higher centrality (p-value = $4.5{\cdot}10^{-7}$) in the generated networks than random proteins.

The best ranking with respect to disease proteins is obtained when considering disease networks built from direct links including predicted functions, followed by using only original annotations (see Figure 7.4). The inclusion of predicted functions yields superior

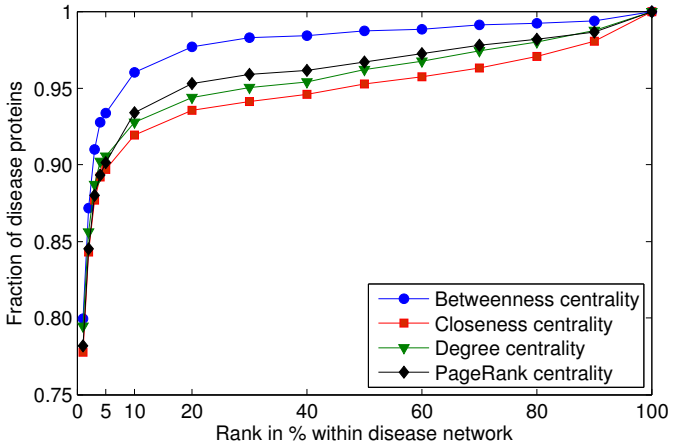


Figure 7.3: Comparison of the performance of the four different centrality measures: betweenness, degree, closeness and PageRank, in non-enriched d_1 disease networks.

ranking in both d_1 and d_2 networks compared to the non-enriched networks, in particular for $k \leq 10\%$, which emphasizes the power of integrating predicted functional information into the disease gene ranking framework.

Using indirect links leads in the first place to a lower fraction of seed proteins among the top ranked proteins due to a higher amount of (disease unrelated) hub proteins (see Section 6.3.2.1). Such hubs do not strongly impact the ranking in d_1 disease networks, since in those they are mainly only connected to the seeds. However, the inclusion of indirect neighbors also integrates all their interaction partners which puts the hubs into a central position.

7.2.1 Normalization for hub proteins

To verify whether unspecific hubs compromise the ranking, we consider the number of highly ranked hub proteins within the networks as well as the relative frequency of each protein across all networks. Hub proteins are determined by examining the degree distribution across the human interaction network. The node degree at the 90th percentile of this distribution is set to be the cut-off for defining a hub protein. Thus, we consider any protein as hub if it has more than 23 interactions in the human interaction network (consistent with Aragues *et al.* (2007)) which applies to 1,485 proteins (10% of proteins).

Figure 7.5 shows the distribution of proteins across disease networks generated from direct as well as indirect interaction data. The comparison indicates that proteins from d_2 networks occur more frequently in the different disease networks. For instance, the most frequent protein in d_1 disease networks, namely GRB2, is included in 8.5% of the disease networks. In contrast, UBC, UBB, RPS27A and UBA52 occur in almost half of the generated d_2 networks (47.3%). Such highly frequent proteins are unlikely to be specific for particular disease networks. In turn, studying highly ranked proteins reveals that the amount of hub proteins within the most central network regions is significantly

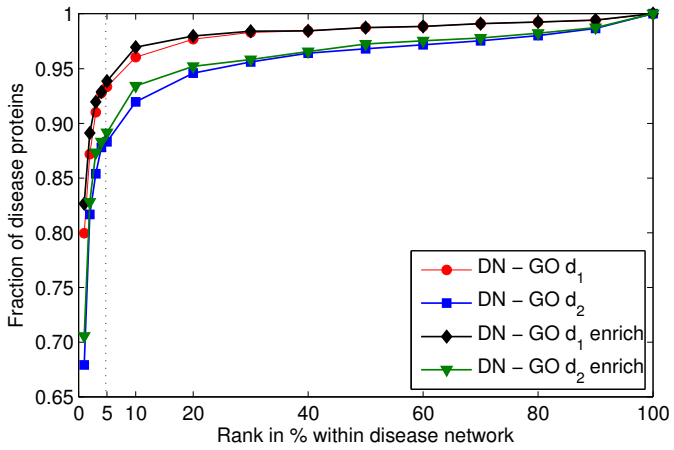


Figure 7.4: Centrality of disease proteins within disease networks (DN). Disease networks are compiled from direct and indirect interaction data (d_1 and d_2) as well as manually curated and predicted functional annotations (*enrich*).

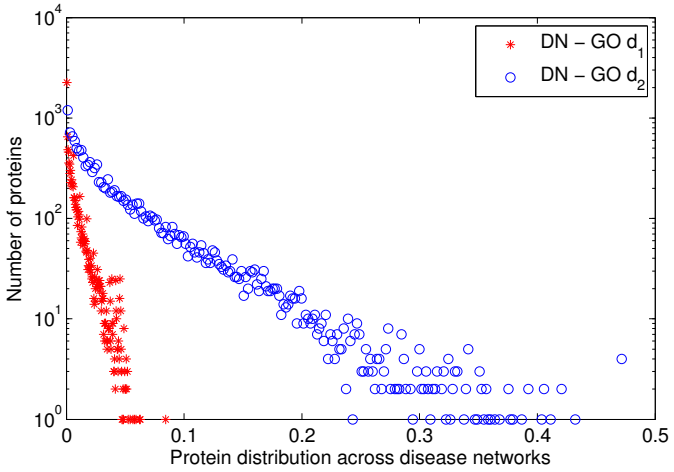


Figure 7.5: Comparison of the protein distribution across disease networks compiled from manually curated functional annotations and direct as well as indirect interaction data (y-axis in log-scale).

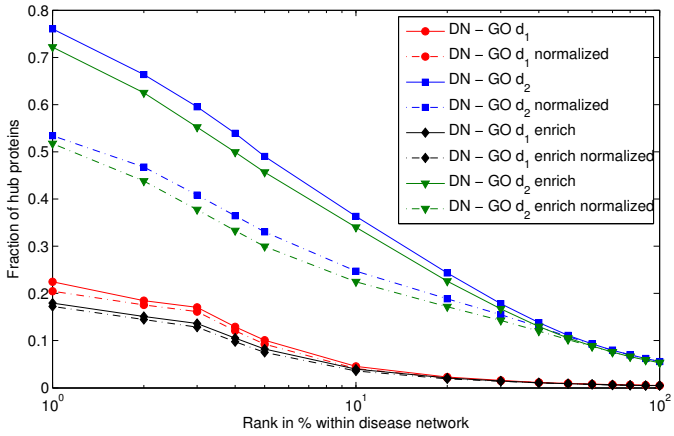


Figure 7.6: Fraction of highly ranked hub proteins in the four disease network types with (dashed lines) and without hub correction (solid lines).

higher for d_2 disease networks, e.g., 72% and 76% compared to 18% and 22% in d_1 networks for $k = 1\%$ (see Figure 7.6).

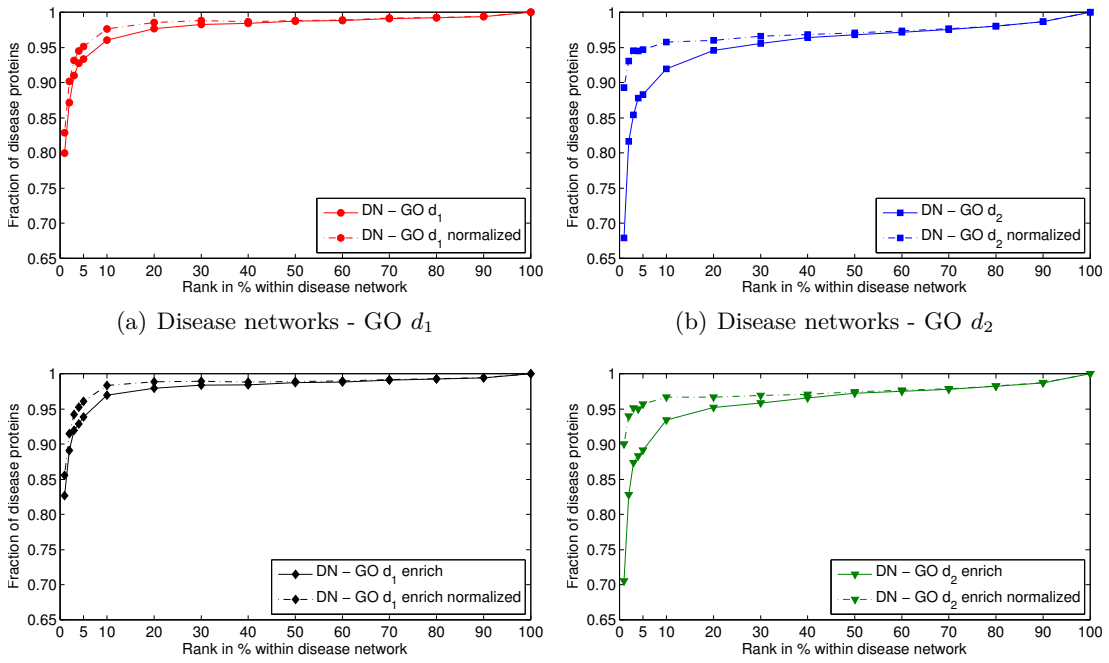
To account for unspecific hub proteins we normalize the betweenness centrality score of each protein by the number of its occurrence across all networks (see Section 6.3.2.1, Eq. 6.1). Figure 7.7(a) – 7.7(d) shows that the proposed hub correction leads to a considerable increase in the fraction of disease proteins (up to 22%) among the most central proteins in the d_2 disease networks. Overall, the impact of the normalization is most striking for $k < 10\%$ and decreases the more proteins of a disease network are considered. The hub normalization also improves the ranking in d_1 networks but its impact is less significant than in d_2 networks. The improvement in ranking with respect to disease proteins is also reflected in a decreasing number of highly ranked hub proteins (see Figure 7.6). The fraction of highly ranked hub proteins decreases, for instance, by 23% for enriched d_2 networks. Both observations demonstrate that our normalization is an effective approach to filter proteins unrelated to the given disease.

In the following evaluations we only consider rankings with hub normalization.

7.3 Cross-validation

For assessing the ability of our method to identify novel disease genes in a genome-wide setting we performed leave-one-out cross-validation. In contrast to the centrality analysis in the previous section, this is the 'true' setting for evaluating disease gene identification methods. As described in Section 6.3.3.1, we only consider diseases with at least two seed proteins during cross-validation, since our method needs at least one disease protein as seed. This applies for 284 out of 3,077 diseases with at least one protein. For each left-out disease protein we determine whether it can be re-discovered using our method. Figure 7.8 shows the cross-validation recovery rates with hub normalization across the ranked networks using disease networks generated from different functional associations.

In total, we re-discover 41%, 46%, 55% and 59% of the blinded disease-protein as-



(a) Disease networks - GO d_1

(b) Disease networks - GO d_2

(c) Disease networks - GO d_1 enrich

(d) Disease networks - GO d_2 enrich

Figure 7.7: Impact of the hub correction on the ranking of disease relevant proteins within the four different disease network types: (a) direct interactions and original annotations, (b) direct and indirect interactions and original annotations, (c) direct interactions and original and predicted annotations and (d) direct and indirect interactions and original and predicted annotations.

sociations when considering direct interaction and functional annotation alone, and in combination with predicted functions and indirect neighbors, respectively. Combining indirectly linked proteins and functionally enriched networks significantly increases the amount of re-discovered proteins, up to 20%.

The inclusion of additional evidence leads to larger networks and thus to a higher number of potential candidates. In consequence, the ratio between promising and false positive candidates decreases with the increasing number of proteins in a top-k% list. To assess whether we truly gain additional information from our extended networks, we also study the absolute recovery rates. To this end, we performed an analysis using only the top 100 proteins within each network. The results confirm that using the enriched networks clearly leads to better results (see Figure 7.9) which in turn underlines that the improved recovery rates result from a more comprehensive representation of the proteins within the networks rather than from the larger network sizes.

7.3.1 Filtering chromosomal regions

Comparing the performance of methods for predicting disease-related genes is a difficult undertaking. Related methods usually constrain the set of genes under study, either

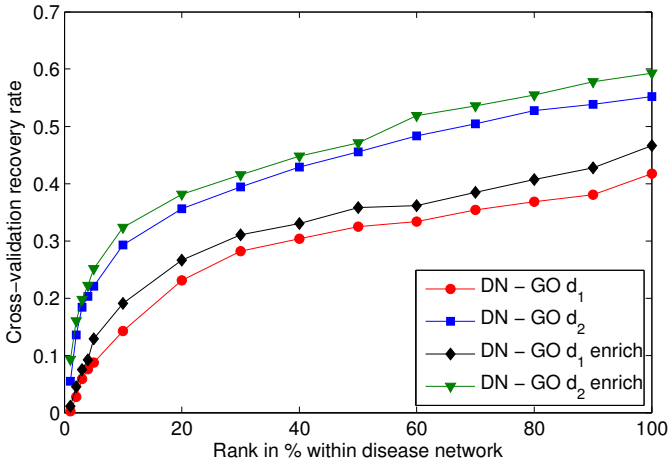


Figure 7.8: Cross-validation recovery rates (with hub correction) from disease networks with direct and indirect interaction as well as original and predicted functional annotations.

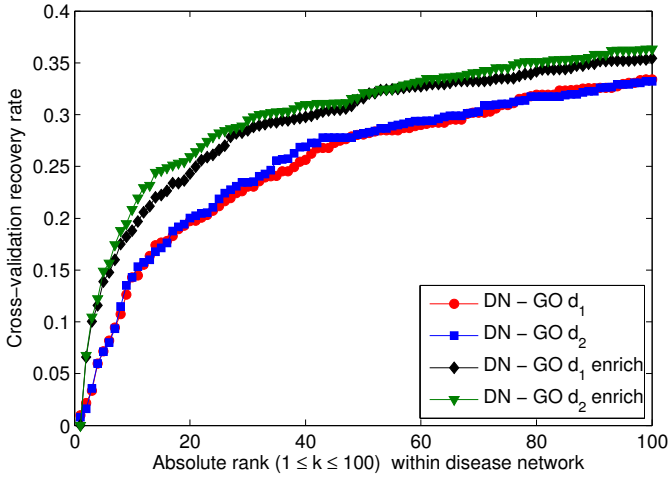
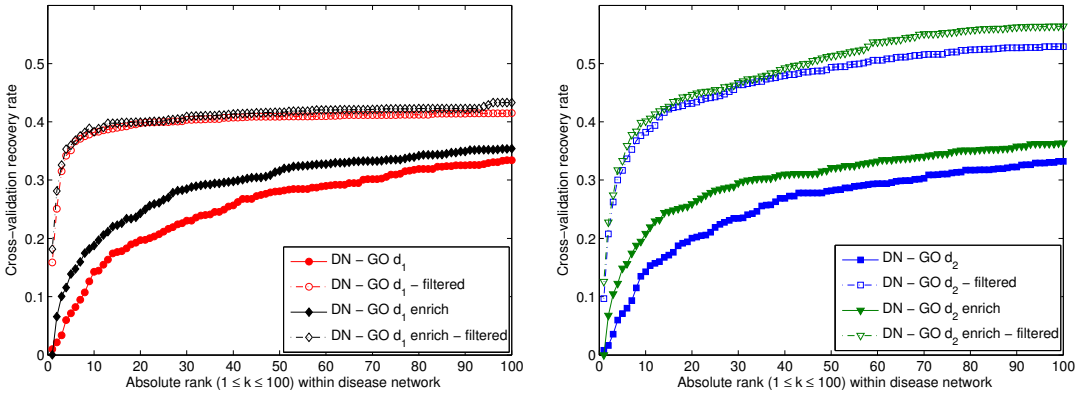


Figure 7.9: Absolute cross-validation recovery rates (with hub correction). Considering only the top $1 \leq k \leq 100$ proteins in disease networks with direct and indirect interaction as well as original and predicted functional annotations.



(a) Filtering in d_1 disease networks

(b) Filtering in d_2 disease networks

Figure 7.10: Effect of chromosomal filtering on the absolute recovery rates, $1 \leq k \leq 100$, in (a) d_1 and (b) d_2 disease networks.

by focusing on particular disease-gene families or, most often, by focusing on defined chromosomal regions. Such constraints act as stringent filters, making the resulting disease networks much smaller and thus strongly restricting the set of gene candidates. Most approaches are commonly evaluated on artificial linkage intervals with around 100 to 108 genes including the target gene (Perez-Iratxeta *et al.*, 2002; Franke *et al.*, 2006; Lage *et al.*, 2007; Wu *et al.*, 2008). In contrast, our unconstrained networks usually contain several hundred proteins. Enriched d_2 networks, for instance, involve on average about 700 proteins (std $\pm$ 405, see Table 7.2). Ranking in such large networks becomes much more difficult (Wu *et al.*, 2008), but only region-independent methods are applicable to diseases where no regions are associated yet. This applies to about 43% of the OMIM diseases.

To test how much our method would benefit from utilizing information on chromosomal regions, we performed another leave-one-out evaluation in which we filtered all proteins from the ranked lists which are not located on the same chromosome as the left-out protein, which mimics a search constrained to a chromosome. Note that human chromosomes contain on average 1,341 genes, ranging from 379 genes on chromosome 11 to 4,220 on chromosome 1. Figure 7.10 shows that such filtering improves the recovery rates significantly. For instance, the recovery rate almost doubled when considering the top $k = 20$ proteins in the enriched d_2 networks (see Figure 7.10(b)). Note that this task is still considerably more difficult than the one solved by most other methods as we still need to first reach the target protein while growing the disease network – in contrast to artificial linkage intervals where the target gene is initially included.

7.3.2 Impact of the number of initial seeds on the performance

OMIM documents Mendelian disorders and a number of more complex multifactorial diseases that comprise several genes and disease loci. Currently, the number of known

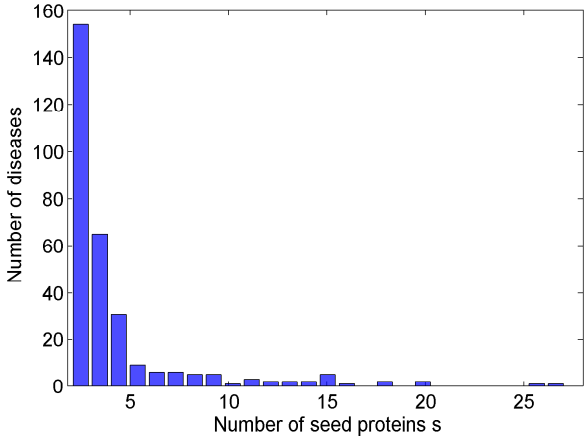


Figure 7.11: Distribution of the number of seed proteins per OMIM disease. Note that diseases with only one seed protein (2,771) are disregarded in the figure as they are not considered in our validation.

disease genes in OMIM ranges from one to 27, but on average only 1.28 gene is associated with each disease (see Section 7.1). Most studies validate their methods on diseases with a large number of known causative genes or specifically defined disease families (Aerts *et al.*, 2006; Köhler *et al.*, 2008; Chen *et al.*, 2007a, 2009b). Yet, a good predictive performance is also necessary for diseases with only few known genes. To this end, we assess the impact of the number of known causative genes on the performance of our method by analyzing the recovery rates according to the number of seeds s available for a disease. Figure 7.11 gives an overview on the number of seed proteins per OMIM disease. As expected the number of diseases decreases with the increase in the number of seed proteins. For instance, 154 diseases are associated with two disease proteins while 23 are associated with eleven disease genes or more. To obtain statistically sound conclusions for seed numbers with only a small number of diseases, we group diseases according to their number of seed proteins for larger s.

Figure 7.12 shows the seed-number-dependent recovery rates for OMIM diseases with $s = \{2, 3, 4, 5, 6-10, 11-15, 16-27\}$ seed proteins computed from enriched d_2 disease networks (see Figure B.5 for individual recovery rates per s). The overall recovery rates correlate clearly with the number of disease genes known *a priori*. The recovery rate increases, for instance, from 51% for diseases associated with two disease proteins up to 79% for diseases with 16 seeds or more which emphasizes that finding relevant genes for a disease is even more challenging when only little is yet known on that disease.

Comparing seed-size-specific results for d_1 and d_2 disease networks highlights again the benefit of using indirect interaction data. Recovery rates for $s = \{2, 3, 4\}$ increase significantly, e.g., from 35% to 51% and from 46% to 61% for $s = 2$ and $s = 4$, respectively, when considering indirect interaction data (see Figure B.6). In general, a successful recovery of a known disease protein correlates with the number of available seed genes as most methods perform better on diseases with more seed genes. However, using

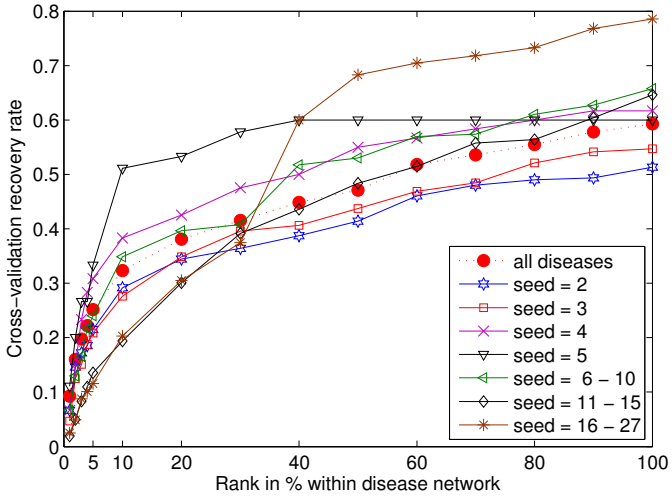


Figure 7.12: Seed-number-specific recovery rates for OMIM diseases with $s = \{2, 3, 4, 5, 6-10, 11-15, 16-27\}$ seed proteins.

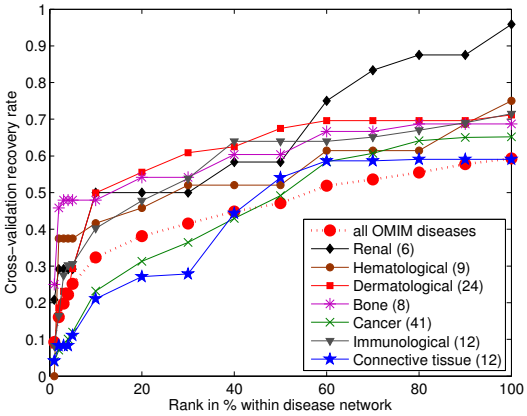
indirect neighbors increases the recovery rates for diseases with only few known genes significantly which underlines the value of indirect functional links.

7.3.3 Results for different disease types

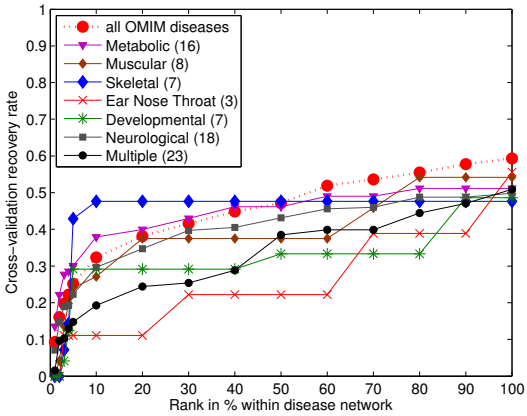
In addition to the influence of the number of seeds, we study whether the disease type impacts the performance of our method. To distinguish between different disease types we used a classification scheme proposed by Goh *et al.* (2007). Goh *et al.* manually classified OMIM diseases into 22 distinct types of disorders according to the physiological system perturbed by the disease, e.g., immunological, metabolic or neurological system (see SI Table S1 in Goh *et al.* (2007) for details). Disorders with multiple clinical features are assigned to a "multiple" class while disorders without sufficient information for clear assignment were associated with an "unclassified" class. Using this classification scheme we assign 1,757 diseases to one of 22 disorder classes (see Table B.3).

Again, we perform cross-validations across all diseases with two or more seed proteins and determine the recovery rates with respect to their associated disorder class. In total, 256 of the 284 diseases have been considered in this particular cross-validation. Figure 7.13 shows the disease-specific recovery rates for the different types of diseases. Overall the performance of our method varies widely when performing cross-validation on a per-disease type basis. For more clarity we grouped disease types according to their performance compared with the average cross-validation recovery rate obtained for functionally enriched disease networks with direct and indirect interaction data (DN – $GO\ d_2\ enrich$).

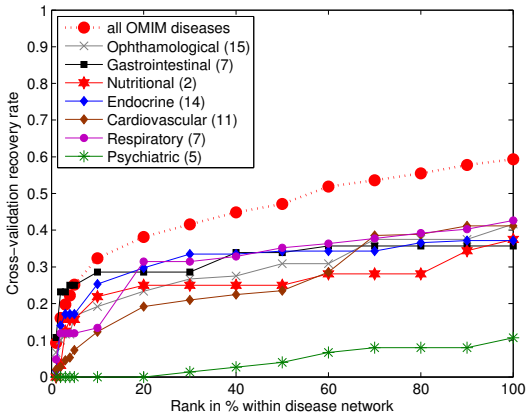
Figure 7.13(a) shows disease classes with superior performance. Diseases affecting, for instance, the dermatological, hematological or the renal system yield strikingly high recovery rates of 71%, 75% and 96%, respectively. However, also for cancer, which is



(a) Disease types with superior performance



(b) Disease types with similar performance

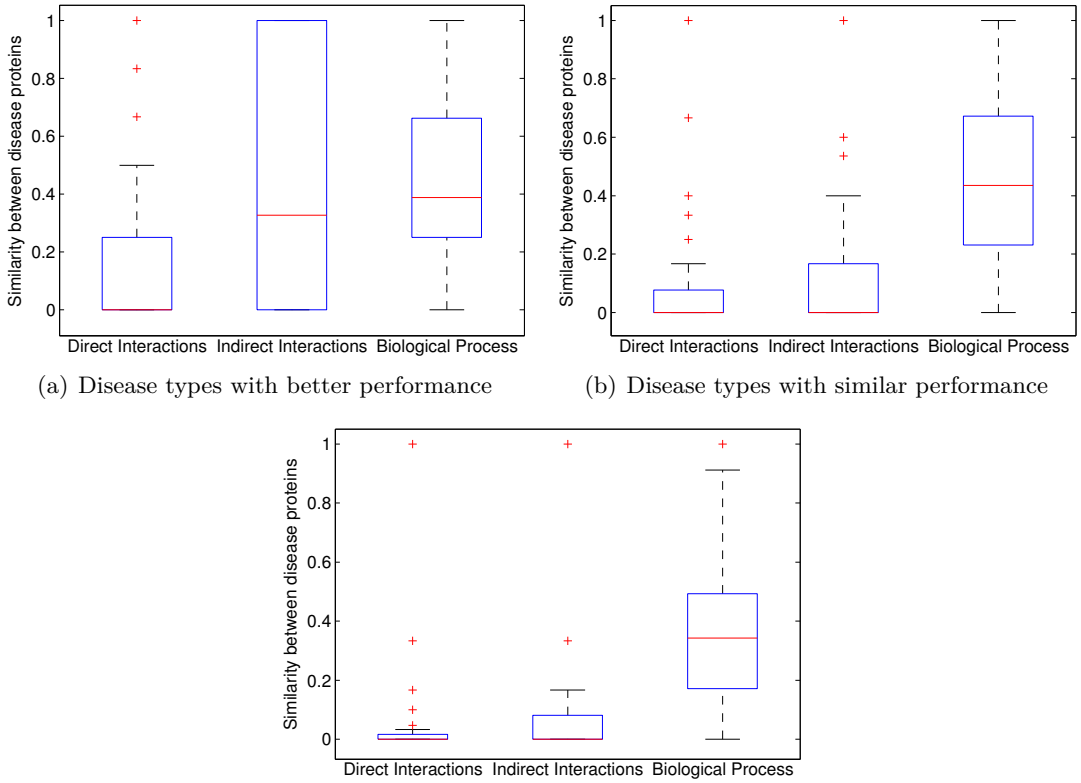


(c) Disease types with inferior performance

Figure 7.13: Disease-specific cross-validation recovery rates. Disease classes are grouped according to their performance compared to the average cross-validation recovery rate obtained for functionally enriched d_2 disease networks: (a) disease classes with superior performance, (b) disease classes with similar performance and (c) disease classes with inferior performance.

known to be a particular complex disease caused by several genomic alterations, we achieve a recovery rate of 65%. Figure 7.13(b) illustrates disease classes on which our methods performs similar or slightly worse compared to the average recovery rate of 59%. The overall recovery rates range between 47% and 55%. In turn, for disease types shown in Figure 7.13(c) our method performs poorly. Only 11% to 33% of the blinded disease proteins can be recovered during cross-validation. The compiled disease networks seem to provide only little information for diseases assigned to, e.g., the endocrine or respiratory class, indicating that additional data is required to study such diseases successfully.

Based on this assumption, we compared the functional relationships between disease proteins associated with disease types performing either better, similar or worse than the average recovery rate across all OMIM diseases. Figure 7.14 shows that disease proteins



(a) Disease types with better performance

(b) Disease types with similar performance

(c) Disease types with inferior performance

Figure 7.14: Disease-class-specific protein characteristics. Functional relationships between disease proteins associated with disease types with (a) superior performance, (b) similar performance or (c) inferior performance.

assigned to disease types with the best recovery rates interact to a significantly higher extent with each other than, for instance, proteins from disease types with low recovery rates. There are two possible explanations for this difference. On the one hand, proteins of such disease types are more likely to share other protein characteristics rather than interaction partners. On the other hand, proteins involved in endocrine, respiratory or cardiovascular diseases might be less studied. Therefore, less functional information is available for them yet, which in turn hinders the discovery of novel proteins related to such diseases.

7.3.4 Classical Hodgkin Lymphoma[†]

To show the ability of our method to handle highly complex diseases involving complex genomic alterations, we apply our approach to unravel molecular mechanisms involved in the pathogenesis of classical Hodgkin Lymphoma (cHL). cHL is a peculiar type of

[†]Joint work with Karin Zimmermann (Humboldt Universtität zu Berlin), Volkhard Seitz and Michael Hummel (Charite - Universitätsmedizin Berlin).

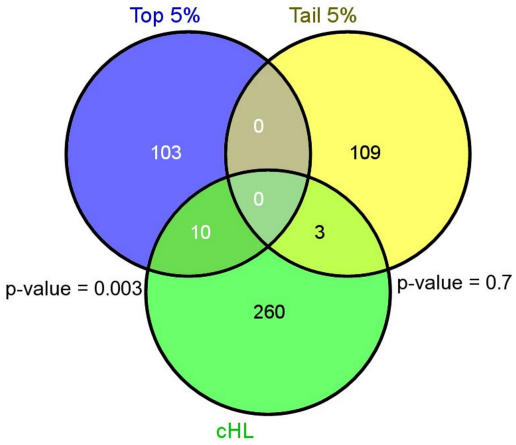


Figure 7.15: Venn diagram (created with Venny (Oliveros, 2007)) displaying the intersect of the top 5% and the tail 5% of the predicted gene list with the upregulated genes in classical Hodgkin Lymphoma (cHL).

lymphoma genotypically derived from B-cells (Küppers *et al.*, 1994) which reside in an extensive cellular background of various types of non-malignant bystander cells (WHO Classification 2008).

For identifying novel disease-related candidates for cHL, we first analyze epigenetic data to define an initial set of genes involved in the pathogenesis of cHL. To this end, we compared two independent data sets: (i) specifically acetylated genes in Hodgkin cell lines ($n = 172$) and (ii) genes being up-regulated upon epigenetic treatment of B-cell lines ($n = 435$) with demethylating and acetylating agents (5-Aza-dC/Trichostatin A) (Seitz *et al.*, 2011) inducing a Hodgkin-like phenotype (Ehlers *et al.*, 2008).

Based on 22 experimentally linked cHL-seed genes (see Table B.4) we compiled an enriched d_1 lymphoma-specific network with 2,258 proteins and rank its proteins according to their normalized betweenness centrality. We first find that many cHL-related proteins are highly ranked in the corresponding network. 12 out of 22 seeds are found among the top 33 proteins. To associate novel proteins with cHL we selected the top 5% proteins from the network, 119 proteins not including the 22 seed proteins, for further evaluation (see Table 7.3 for the top 20 candidates and Table B.5 for the full list). We compare this list to a set of Hodgkin-characteristic transcripts that are differentially expressed in Hodgkin cell lines versus B cell lines. From the initial set of 396 genes described by Seitz *et al.* (2011), 273 transcripts could be mapped to gene products in our data. The overlap between these two sets (10 genes) is highly significant (p-value 0.003, see Figure 7.15) and contains many genes known to be cHL-related, such as *STAT3*, *FAS*, *NFKB2* and *CFLAR* (Seitz *et al.*, 2011). In contrast, no significant overlap (p-value 0.70) is found when conducting the same comparison for the lowest ranked 5% proteins.

The remaining 109 proteins have not been previously discussed as Hodgkin-related and may represent an important and independent expansion of the present knowledge. We studied those using expert knowledge and by searching the literature. 10 proteins are related to elevated proteasome activity. Proteasome inhibition is know to block the

Table 7.3: Top 20 candidate proteins inferred from the lymphoma-specific network that are not associated with cHL (sorted by rank). Candidates are specified by gene symbol, gene id, name and Uniprot id.

Gene				
Symbol	ID	UniProt	Name	Mentioned in cHL context
HIST1H1C	3006	P16403	Histone H1.2	Hodgkin-related
ACTL6B	51412	O94805	Actin-like protein 6B	
HIST1H2AM	8329	P0C0S8	Histone H2A type 1	
HIST1H3J	8350	P68431	Histone H3.1	
SMCHD1	23347	O75141	Structural maintenance of chromosomes flexible hinge domain containing 1	
HIST2H3A	126961	Q71DI3	Histone H3.2	
NUF2	83540	Q5SXK4	NDC80 kinetochore complex component, homolog	
EPS15	2060	P42566	Epidermal growth factor receptor substrate 15	
VPS25	84313	Q9BRG1	Vacuolar protein-sorting-associated protein 25	
PREB	10113	Q9HCU5	Prolactin regulatory element-binding protein	
HIST2H2BF	440689	Q5QNW6	Histone H2B type 2-F	
VPS36	51028	Q86VN1	Vacuolar protein-sorting-associated protein 36	
HIST3H3	8290	Q16695	Histone H3	
CHD3	1107	Q12873	Chromodomain-helicase-DNA-binding protein 3	
PSMA1	5682	P25786	Proteasome subunit alpha type-1	Proteasome complex
LCP1	3936	P13796	Plastin-2	
HIST1H2BB	3018	P33778	Histone H2B type 1-B	
TRA@	6955	Q6PJ56	TRA@ T cell receptor alpha locus	
TSC22D3	1831	Q99576	TSC22 domain family protein 3	
HNRNPD	3184	Q12771	Heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa)	

pro-apoptotic NF-κB activity which in turn induces apoptosis of Hodgkin and Reed-Sternberg (HRS) cells. This molecular mechanism is currently discussed as a therapy option for patients with cHL (Zhao *et al.*, 2008). Furthermore, various signaling pathways are constitutively active in cHL, most importantly the nuclear factor-κB (NF-κB) and Janus Kinase (Jak)-Stat pathway (Küppers, 2009). Interestingly, several proteins related to those known Hodgkin-related pathways, such as TNF-Receptors, Jak/STAT and NF-κB, are also found within the 103 proteins but were not identified by other approaches (Salghetti *et al.*, 1999). Genetic lesions in these pathways are thought to be involved in the activation in HRS cells (Küppers, 2009). The complete list of candidates including supporting evidence for an association with cHL is given in Table B.5.

MYC, an important oncogene, which was so far not in focus of cHL research is also present within the top 5% proteins. *MYC* is a central transcription factor known to be involved in many cellular activities including cell proliferation and apoptosis (Li *et al.*, 2003a). *MYC* also regulates genes involved in ubiquitin-mediated proteolysis which is thought to be responsible for *MYC* degradation (Salghetti *et al.*, 1999; Li *et al.*, 2003a). Although *MYC* is not specific for cHL, the identification of *MYC* by our approach highlights its potential role in cHL on protein level and hints to hitherto unknown functions of *MYC*.

7.3.5 Colorectal cancer

We applied our strategy to identify genes that are directly or indirectly involved in the pathogenesis of colorectal cancer (CRC), the third most common cause of cancer deaths for both men and women in the United States and Europe (Grothey *et al.*, 2004). CRC arises from the colorectal epithelium in consequence to the accumulation of genetic aberrations in defined oncogenes and tumor suppressor genes as well as epigenetic alterations including aberrant DNA methylation and chromatin modifications (Grady and Carethers, 2008). Most CRC-causing mutations are somatic, i.e., occurring in the perturbed tissue during carcinogenesis. Yet, similar to most cancer types, CRC also has a hereditary component caused by mutations which affect the germline and account for the initiation of carcinogenesis (de la Chapelle, 2004). So-called high-penetrance[11] mutations confer susceptibility to CRC, for instance, in Lynch syndrome caused by mutations in mismatch repair genes (Lynch and Smyrk, 1996), and familial adenomatous polyposis involving alterations in the tumor suppressor *APC* (Half *et al.*, 2009). Low-penetrance mutations accounting for the remaining familial cases as well as the large proportion of sporadic CRC are less understood.

For associating proteins with CRC, we extracted all phenotype entries from OMIM that are associated with CRC. Albeit several phenotypes describe different variants of this cancer type, only few are already associated with causal genes (see Table B.6). To identify gene products related to CRC in general we combine the different subtypes to one set of 27 genes (see Table B.7) and grow a CRC-specific d_2 network around these seeds with 8,137 proteins. Before inferring CRC-related genes we first perform a cross-validation over this set to study the trade-off between potential candidates and false positives. For cross-validation we remove one seed protein from the initial list and generate a CRC network from the remaining seeds in which we rank the proteins according to their network centrality. Subsequently, we determine whether the left-out protein can be re-discovered and at which position of the ranked list. We repeat this procedure for each seed and determine the average recovery rate across all seeds which is then normalized by the number of proteins considered at each rank k.

Both the original and the normalized recovery rate are shown in Figure 7.16. In total, we re-discover 24 out of 27 colon cancer seed genes across the respective networks. When considering the top 1% proteins of the networks (81 proteins), we find two of the blinded

[11] Penetrance indicates the frequency with which individuals exhibit the phenotype linked to a particular mutation.

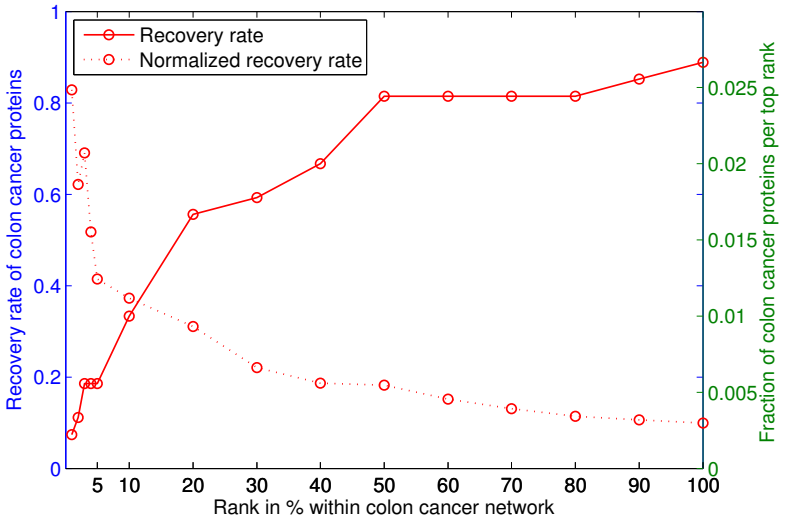


Figure 7.16: Original and normalized cross-validation recovery rate for the colon cancer specific seed gene set.

colon cancer proteins within the 81 most central proteins. Thus, the minimum likelihood of identifying unknown genes relevant for colon cancer equals $2/81 = 2.5\%$. Note, this probability is estimated from the cross-validation on known data and therefore provides a lower bound since all novel findings are counted as false positives during cross-validation. Naturally, this likelihood decreases significantly the more proteins of the network are considered. Hence, we choose a cut-off of 1% when assessing the final predictions.

Applying this cut-off to the CRC-specific network results in 81 non-seed proteins which are investigated with respect to CRC. Table 7.4 presents the top 20 candidates (see Table B.8 for full list). Furthermore, we gathered evidence from literature, KEGG pathways and expression profiles to assess the potential association of these proteins with colorectal cancer. The large majority of the candidates is highly overexpressed in cancerous colon tissue compared to healthy colon tissue (Yanai *et al.*, 2005).

For at least 17 candidates we find strong evidence in the literature for an involvement in the development and progression of colon cancer (see underlined entries in Table 7.4 and Table B.8). Some of them, e.g., *SMAD2*, *SMAD3* and *SMAD4*, have already strong support from the experimental field (Malek *et al.*, 2002; Xie *et al.*, 2003). Smad proteins, for instance, are key components of the TGF-β signaling pathway which regulates a wide range of cellular processes including cell proliferation, differentiation and apoptosis (Massagué and Chen, 2000). TGF-β stimulation induces the phosphorylation and activation of Smad2/3 which in turn initiates the assembly of heteromeric complexes with Smad4. These complexes accumulate in the nucleus where they regulate the transcription of target genes, i.e., genes crucial for cell cycle control (Sameer *et al.*, 2010). Mutations in Smad proteins impair the natural function of the TGF-β pathway (Woodford-Richens *et al.*, 2001) providing cellular resistance to TGF-β-induced growth inhibition which is often observed in tumor cells (see Figure 7.17). Smad4 inactivation is particularly linked

Table 7.4: Top 20 candidates predicted to be involved in colorectal cancer (sorted by rank). Candidates are specified by gene symbol, gene id, name and Uniprot id. Pathway information are derived from KEGG. Literature evidence supporting potential associations with colon cancer is provided in Table B.8. Highly relevant predictions for CRC are underlined.

Gene				
Symbol	Id	UniProt	Name	Pathways
PLA2R1	22925	Q13018	Secretory phospholipase A2 receptor	–
CLN3	1201	Q13286	Battenin	–
ATXN1	6310	P54253	Ataxin-1	–
MYC	4609	P01106	Myc proto-oncogene protein	Colorectal cancer
YWHAG	7532	P61981	14-3-3 protein gamma	–
EGFR	1956	P00533	Epidermal growth factor receptor	Colorectal cancer
YWHAZ	7534	P63104	14-3-3 protein zeta–delta	–
SRC	6714	P12931	Proto-oncogene tyrosine-protein kinase Src	–
SH3GLB2	56904	Q9NR46	Endophilin-B2	–
SFN	2810	P31947	14-3-3 protein sigma	–
SLX4	84464	Q8IY92	Structure-specific endonuclease subunit SLX4	–
COPS6	10980	Q7L5N1	COP9 signalosome complex subunit 6	–
SMAD2	4087	Q15796	Mothers against decapentaplegic homolog 2	Colorectal cancer
PIK3R1	5295	P27986	Phosphatidylinositol 3-kinase regulatory subunit alpha	Colorectal cancer
UBE2I	7329	P63279	SUMO-conjugating enzyme UBC9	–
CTNNB1	1499	P35222	Catenin beta-1	Colorectal cancer
MUC2	4583	Q02817	Mucin-2	–
RELA	5970	Q04206	Transcription factor p65	Pathways in cancer
PLK1	5347	P53350	Serine–threonine-protein kinase PLK1	–
SMAD4	4089	Q13485	Mothers against decapentaplegic homolog 4	Colorectal cancer

with late stage or metastatic colorectal cancer (Miyaki *et al.*, 1999; Maitra *et al.*, 2000). Smad2 may act as a tumor suppressor in colorectal cancer while mutations in Smad3 have been associated with colorectal adenocarcinoma in mice (Zhu *et al.*, 1998).

Other candidates, such as *SRC* and *MYC*, are known oncogenes encoding for proteins that control cell proliferation, apoptosis, or both (Croce, 2008). Both genes are found to be over-expressed and highly activated in a variety of human cancers (Irby and Yeatman, 2000; Nilsson and Cleveland, 2003) including colon cancer. The frequent dysregulation of *SRC* in human colon cancer cells indicates its potential role in the development of this cancer type (Malek *et al.*, 2002). Furthermore, the increased activity of *SRC* has been shown to enhance metastasis and the malignant progression of colon cancer (Kline *et al.*, 2009). The contribution of *MYC* is less conclusive yet. However, the depletion of *MYC* in colon cancer cells inhibits cell growth and induces apoptosis (Hongxing *et al.*, 2008). Despite the strong evidence in the literature, the discussed candidates have not yet been established in the particular databases.

A number of predictions, e.g., *PLK1*, are thought to be potential prognostic markers for the disease. Polo-like kinase 1 (over)expression, for instance, is associated with advanced tumor stages in colon cancer (Weichert *et al.*, 2005). Further studies confirmed

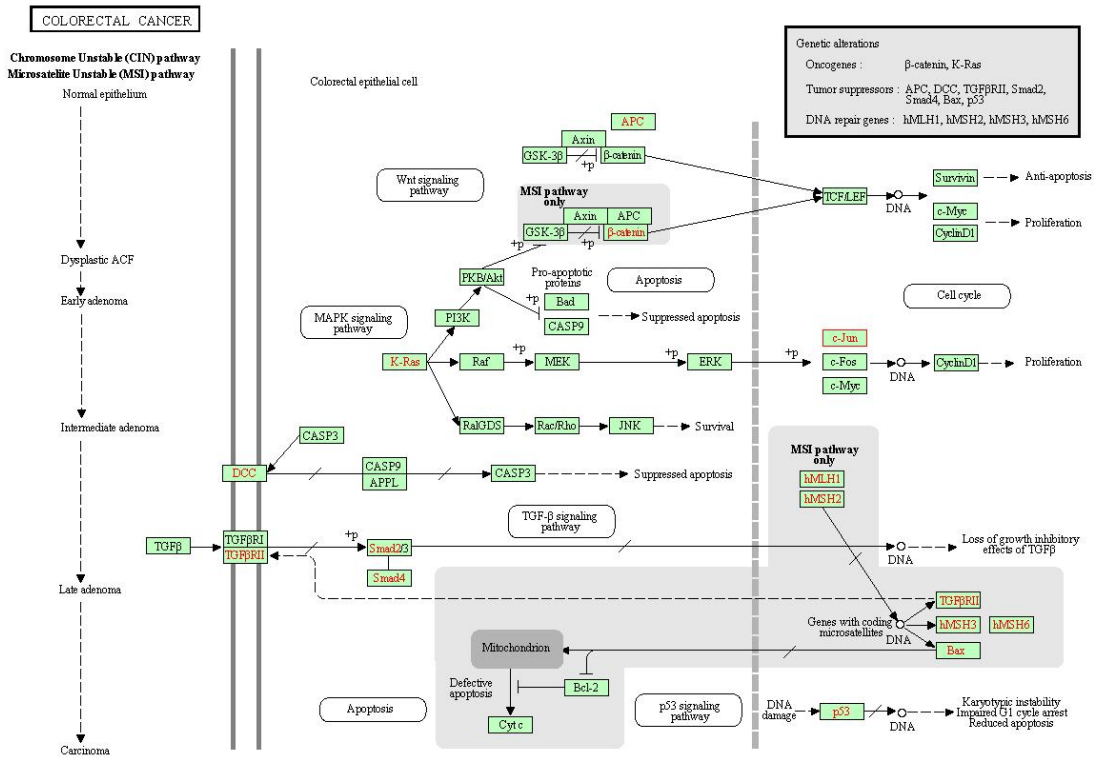


Figure 7.17: Colon cancer pathway (hsa05210) from KEGG.

the correlation of *PLK1* expression with patient prognosis indicating that this kinase is a prognostic marker for colon carcinoma patients (Takahashi *et al.*, 2003). For the remaining candidates we did not find literature evidence. Yet, several of them are involved in cell cycle control (*CDC20*), in colon cancer (*RAF1*, see Figure 7.17) or other cancer-related pathways. These findings emphasize that our method generates novel hypotheses that are relevant for colorectal cancer. Yet, their true relevancy needs to be elucidated in systematic follow-up experiments.

7.4 Comparison to related methods

For evaluating the performance of our developed algorithm we compared it with two state-of-the-art methods for disease gene prioritization, namely PRINCE and RWR (see Section 6.4 for details). Both algorithms have been shown to outperform existing local approaches significantly (Köhler *et al.*, 2008; Vanunu *et al.*, 2010; Navlakha and Kingsford, 2010). For this reason, we focus on the two methods in this performance comparison[12]. PRINCE has been obtained from Vanunu *et al.* (2010) in June 2010. An

[12]Note that we also considered Endeavour as one of the state-of-the-art approaches for our performance evaluation. However, for technical reasons we were not able to perform cross-validation in a genome-wide setting. Therefore, we could not include Endeavour into this analysis.

implementation of RWR is included in the DADA suite (Degree-aware algorithms for network-based disease gene prioritization, Erten *et al.* (2011)) which is freely available on the project website[13] (downloaded June 2011).

The benchmarking has been performed on a subset of diseases classified by Goh *et al.* (2007) (see Section 7.3.3). From the distinct disease types presented in Table B.3 we selected six disease classes according to the performance of our method:

- *Superior performance*: dermatological diseases and cancer diseases
- *Average performance*: metabolic diseases and neurological diseases
- *Inferior performance*: respiratory diseases and endocrine diseases

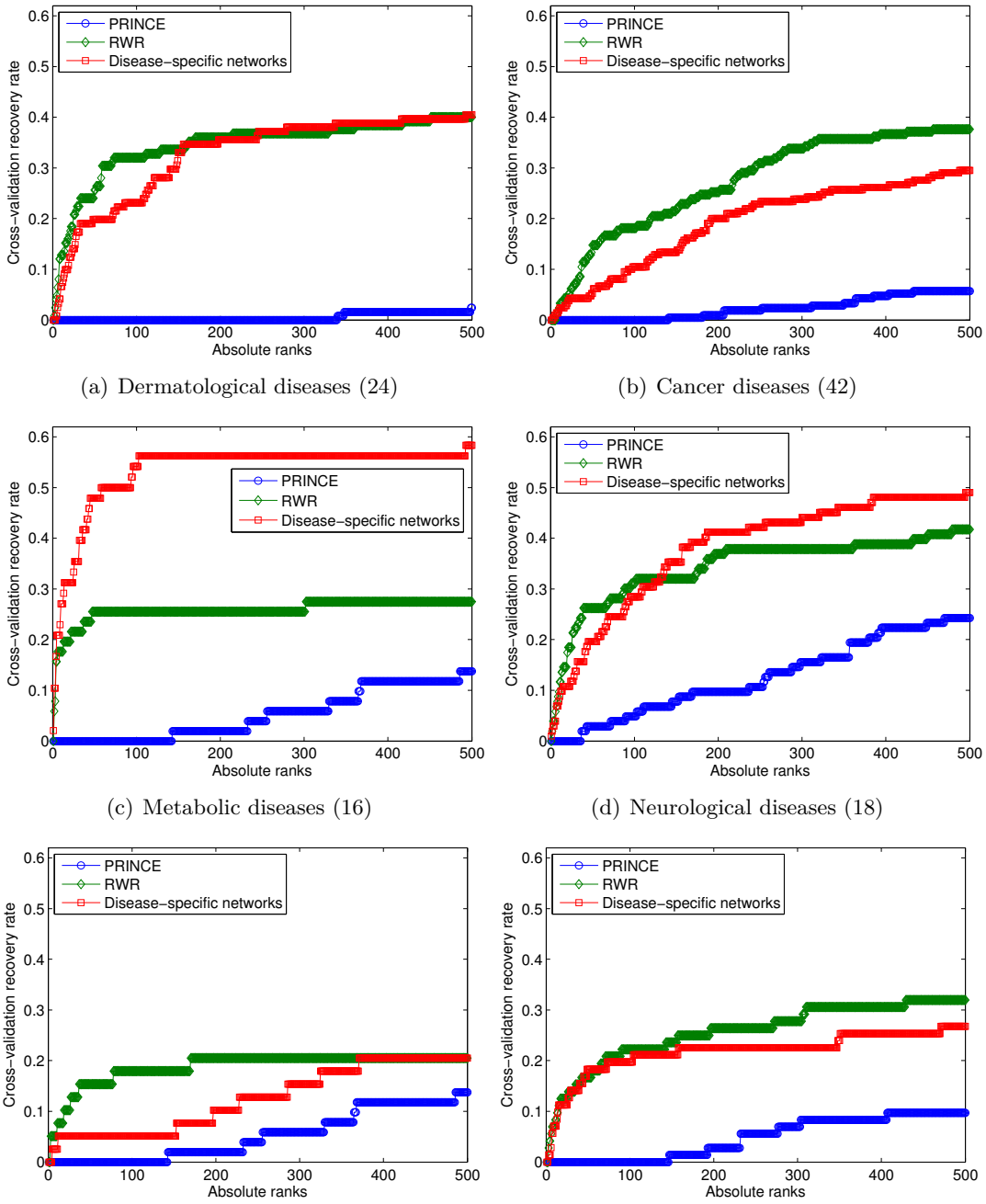
RWR and PRINCE have been applied to the human interaction network generated in this work. To assess and compare the performance of these methods we conducted leave-one-out cross-validation as described in Section 6.3.3.2. Note that for a fair comparison we considered the absolute ranks when determining the average recovery rates (using micro average) as disease-specific networks and the human interaction network differ largely in their size. Figure 7.18 shows the performance of the disease-specific approach and of the two related approaches on the six different disease sets. Note that we focus on the top 500 proteins in the prioritized disease and human interaction network as biologists are only interested in the most promising candidates, i.e., the top ranked candidates, rather than in several thousands of genes.

The direct comparison of the different cross-validation recovery rates shows that our approach clearly outperforms PRINCE. Most importantly, our method recovers the major fraction of blinded disease proteins at an earlier stage in the networks. Thus, a smaller number of genes has to be analyzed to find true disease-related proteins. This is an important feature over PRINCE as disease gene identification methods aim for reducing the number of potential candidates while delivering novel biological hypotheses. This observation holds for disease classes with superior performance (see Figure 7.18(a) and 7.18(b)) but also for diseases with average or inferior performance (see Figure 7.18(c) – 7.18(f)).

When considering the recovery rates of RWR our disease-specific approach performs comparably well. For instance, for dermatological, respiratory and neurological diseases we achieve fairly similar or slightly better results. For cancer and endocrine diseases, on the other hand, our overall recovery rate is lower than for RWR. Yet, the difference is only minor. In contrast to these disease types, the recovery rate obtained for metabolic diseases outperforms RWR significantly. Strikingly, about 50% of the blinded disease proteins are found among the top 50 proteins within the disease-specific networks; twice as much as for the other two methods.

Overall, we show that our disease-specific approach performs comparably well or even better than state-of-the-art methods. Our analysis also indicates that, in comparison to evaluations on linkage intervals, the performance of global network-based approaches decreases significantly when no information on genomic regions is available. In particular,

[13] `http://compbio.case.edu/dada/`



(a) Dermatological diseases (24)

(b) Cancer diseases (42)

(c) Metabolic diseases (16)

(d) Neurological diseases (18)

(e) Respiratory diseases (7)

(f) Endocrine diseases (14)

Figure 7.18: Performance comparison with PRINCE and RWR for six disease classes. For each disease type and method we determine the average recovery rate among the top 500 proteins in the prioritized disease-specific networks and the prioritized human interaction networks.

PRINCE benefits largely from the filter that linkage intervals provide. These observations emphasize the need for efficient prioritization strategies that identify disease-gene associations accurately even if no disease loci have been associate with the disease of interest. An advantage of our module-based approach over global methods is the inherent division of the original genomic data into smaller proportions. Using disease-specific networks, for instance, reduces the complexity for prioritization by yielding shorter lists of disease associated genes, in particular when following a genome-wide approach.

7.5 Case study: Inference of Surface Membrane Factors contributing to HIV-1 Infection[‡]

One of the important characteristics of Human Immunodeficiency Virus (HIV) is its ability to interact with many cell types and its capacity to alter the function of chemokines that otherwise work in harmony with the immune system. This interaction depends on the phenotype of the virus, the receptor type residing on the cell as well as the chemokines present in the environment. Structurally, its genome has evolved to interact with many human proteins from various cellular pathways through viral proteins, such as Tat, Gp120 or Nef (Cook *et al.*, 2002; Piguet and Trono, 1999; Yang *et al.*, 2009).

Typically, a HIV infection originates from the binding of HIV envelope proteins gp120 and gp41 to cell surface receptors CD4 and CCR5/CXCR4 which affects populations of T helper cells, dendritic cells and macrophages. Cell types which are targeted in the course of HIV infection often have different receptor expression profiles and do not necessarily harbor main co-receptors CCR5 or CXCR4, which suggests the involvement of other surface membrane factors (Gorry *et al.*, 2007). Binding of HIV to cell surface factors other than CD4 and chemokine receptors does not always permit viral entry but leads to endocytosis of the viral particles. This promotes relocation of the infectious virions, future trans-infection of adjacent cells (Dong *et al.*, 2007) and leads to the activation of the immune system. Therefore, it is imperative to bear in mind that there are surface membrane factors interacting with HIV proteins, hence affecting the course of infection indirectly. These observations lead to the following questions: What is the extent of surface membrane factors contributing to HIV-1 infection and how do they influence the outcome of the treatment?

HIV exploits the existing signaling and regulatory pathways in its host. The different receptors or surface membrane proteins that are targeted in different cell types are likely to be involved in the same (or closely related) functional pathways, because the range of processes and pathways available to the virus is limited. The complexity in finding the right factors arises from the several hundreds of surface membrane proteins expressed on a wide variety of cells. However, experimental testing of hundreds of targets from numerous pathways is not feasible. Therefore, we adapt the strategy from disease gene discovery described in Chapter 6 to generate high quality hypotheses for wet-lab experiments with the aim to identify surface membrane host factors contributing to

[‡]Joint work with Gökhan Ertaylan and David van Dijk (University of Amsterdam).

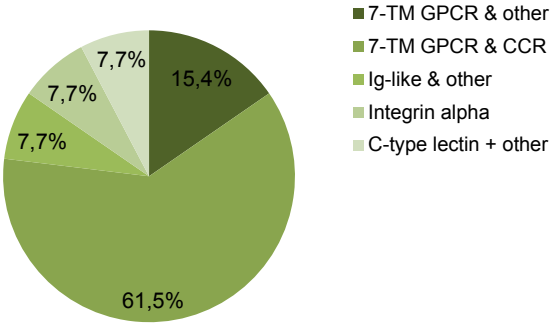


Figure 7.19: Distribution of the functional protein domains across the HIV seed receptors documented in Table B.9.

HIV-1 disease outcome based on receptors that are known to interact with HIV.

Translating the developed framework into the context of identifying surface membrane factors interacting with HIV-1 implies that proteins, which are related to known HIV receptors through functional similarity or interaction with the same ligand(s), tend to be part of the same pathway and often share the same biological function. Therefore, if a network is built based on documented surface membrane factors (see Table B.9) that is extended with related genes, yet undiscovered surface proteins should also be central in the resulting network.

In the following, we briefly describe the setting of this case study and its outcomes. First, we introduce the underlying data basis. Second, we translate the developed framework into the context of identifying surface membrane. Finally, we present a set of novel surface membrane factors and assess their relevancy for HIV infection by exploring the literature, functional domains and their protein interfaces.

7.5.1 Human immunodeficiency virus type 1

In this case study, we use a set of known HIV receptors, their functional annotations and the human protein interaction data as scaffold for building an HIV receptor network. The initial list is compiled by mining the literature and the 'HIV-1, Human Protein Interaction Database' (Fu *et al.*, 2009). A receptor is included if it is reported by at least two independent studies. This applies to 16 HIV receptors. However, three of them, namely Rdc1, Gpr15 and ChemR23, are not documented in the data gathered from the protein interaction databases and thus have not been used in this study. Table B.9 shows the final list of 13 HIV receptors including protein domain information (InterPro, see also Figure 7.19) and their role in HIV infection. The list covers established primary receptors such as CD4 and DC-SIGN, HIV co-receptors CCR5 and CXCR4 as well as alternative co-receptors CCR2 and CCR3. Only recently reported co-receptors, such as XCR1 (Shimizu *et al.*, 2009), have not yet been included since they were not documented by the time the study was conducted. However, cell surface proteins in Table B.9 are reported to interact with HIV in a broad sense. Therefore, we do not limit our prediction method to receptors that only permit the entry of HIV into the primary cells.

7.5.2 Predicting novel HIV surface membrane factors

To determine yet uncharacterized surface membrane proteins based on their functional similarity and topological closeness to receptors that are known to interact with HIV, we integrated protein interaction, protein function and network centrality analysis as elaborated in Chapter 6. The details of this process are summarized below:

- According to the framework proposed in Section 6.3.1, we first compiled an enriched HIV receptor network from the 13 documented surface membrane factors (see Table B.9) by populating it with functionally related proteins that either interact directly with or show significant functional similarity to any known factor. The resulting network comprises 739 proteins (726 candidates) and ∼80,000 functional relationships.
- Subsequently, we used the PageRank centrality measure to discover novel surface membrane factors that are involved in HIV-1 infection. Accordingly, we ranked all proteins with respect to their PageRank centrality within the network. The underlying principle of the centrality analysis presumes that the most central proteins in the HIV-specific network are likely to be of high functional relevance (van Dijk *et al.*, 2010).
- The list of centrality-ranked proteins is further analyzed to identify potentially novel surface membrane factors. To this end, we investigated the trade-off between discovering potential candidates vs. false positives by means of leave-one-out cross-validation. The receptor-per-protein ratios are then used to define a cut-off to select candidates from the prioritized list. We chose 3% as threshold, since it presents a sensible trade-off between potential candidates and false positives while yielding a reasonable number of novel candidates. Thus, the top 21 proteins in the ranked list are considered as surface membrane factor candidates; seed receptors are removed from this list since they are (by definition) highly ranked. Table 7.5 presents the top-ranked candidates including their InterPro domains and cell types.

Table 7.5: List of inferred surface membrane factors. Potential surface membrane proteins resulting from our method, including functional domains and cell types. Predictions associated with HIV in earlier studies are marked with '+'. '−' indicates predictions with negative evidence while for predictions without literature on interaction the association remains unclear (shown by '?').

Receptor	Receptor-specific domains	Cell types	Association with HIV
7-TM GPCR and Other			
HTR6	Not applicable	Uniform expression[14]	+
HTR1B	5HT1B_rcpt	Uniform expression[14]	?
HTR1E	5HT1F_rcpt	Uniform expression[14]	?

Continued on next page

[14]Uniform expression in CD34, endothelial, B lymphoblasts, dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, and whole blood.

Table 7.5 – (continued)

Receptor	Receptor-specific domains	Cell types	Association with HIV
RXFP2	LDL_rcpt_classA_cys-rich_rpt, Leu-rich_rpt, LRR-contain_N, Leu-rich_rpt_typical-subtyp, Relaxin_rcpt	Low expression	?
RXFP1	LDL_rcpt_classA_cys-rich, Leu-rich_rpt, LRR-contain_N, Leu-rich_rpt_typical-subtyp, Relaxin_rcpt	No expression profiles available	?
GPR17	P2_purnocptor	Uniform expression[14]	?
GPR182	G10D_rcpt	Uniform expression[14]	?
NPBWR2	Neuropept_W_rcpt	Uniform expression[14]	−
7-TM GPCR and CCR_rcpt			
CCR1	CC_1_rcpt	High expression: whole blood, monocytes, myeloid, dendritic cell	+
CCBP2	CXC_4_rcpt	Uniform expression[14]	+/−
7-TM GPCR			
DARC	Duffy_cmk_rcpt	High expression: (early) erythroid, endothelial cells	+
Ig-like and Other			
CD2	Ag_CD2, Ig-like_fold, Ig_C2-set, Ig_V-set, T-cell_adhesion_molc_CD2	High expression: dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, whole blood	+
CSF3R	FN_III, Hematopoietin_rcpt_gp130_CS, IgC2-like_lig-bd	High expression: myeloid cells, monocytes and whole blood	+
IL1R1	Ig, Ig-like_fold, Ig_sub, IL1_rcpt_1, IL1R_rcpt	No expression profile available	–
CD79B	Ig-like_fold, Ig_sub, Ig_V-set, Phos_immunorcpt_sig_ITAM	High expression: CD34, endothelial and dendritic cells	+
IL6ST	FN_III, Hematopoietin_rcpt_gp130_CS, Ig-like_fold, IgC2-like_lig-bd	Uniform expression[14]	+
TNFR_Cys_rich_reg and Other			
TNFRSF5	Fas_rcpt	High expression: B lymphoblasts	+
TNFRSF3	TNFR_3_LTBR	High expression: myeloid, monocytes and whole blood	+
Other			
CD97	EGF-type_Asp/Asn_hydroxyl_site, EGF_Ca_bd_2, GPCR_2_CD97, GPCR_2_secretin-like, GPS_dom	High expression: CD34, B lymphoblast, dendritic cells, CD8 and CD4 T-cells, NK, myeloid, monocytes	+
GP1BB	LRR-contain_N, Cys-rich_flank_reg_C	High expression: CD34, monocytes and whole blood	?
GYPB	Glycophorin	High expression: (early) erythroid and endothelial cells	?

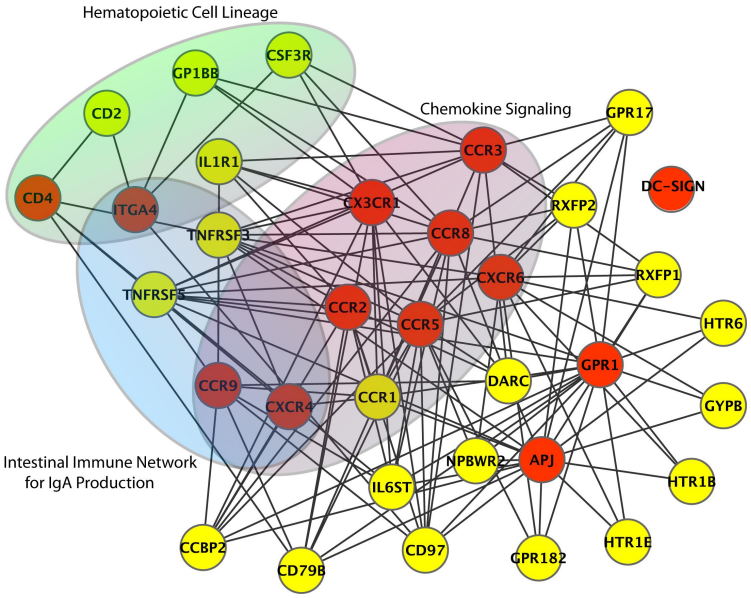


Figure 7.20: Subnetwork of the generated HIV receptor network. The subnetwork focuses on the functional relationships between the seed receptors (red) and the predicted surface membrane proteins (yellow) within the HIV receptor network. Non-seed and non-candidate proteins are not shown to avoid confusion. Significantly enriched pathways within this subnetwork are additionally highlighted.

Figure 7.20 illustrates a subnetwork from the full HIV receptor network that exhibits only the direct functional relationships between seed receptors and predicted surface membrane factors. The analysis of the known and predicted surface membrane factors regarding their annotated KEGG pathways (Kanehisa *et al.*, 2010) revealed the involvement of three pathways, namely the chemokine signaling pathway (hsa04062), the hematopoietic cell lineage (hsa04640) and the intestinal immune network for IgA production (hsa04672).

7.5.3 Support for predictions

In total, we predicted 21 surface membrane HIV factors that are potentially involved in the different stages of infection influencing the progression of the disease. The relevancy of these candidates is assessed by using evidence that supports an association with HIV. We investigate the predictions with respect to functional domains, cell types, chromosomal locations and matching protein interfaces. Furthermore, we explore the literature on expression levels, associated SNPs and reported clinical evidence.

7.5.3.1 Receptor domains

We analyze the most promising predictions by comparing their functional protein domains to the domains of the known seed receptors assuming that overlapping functional

domains indicate similar protein properties, e.g., binding the same ligand, and functional similarity (Zhang, 2009). The most common protein domains of the seed receptors are:

1. G-protein-coupled receptors (GPCR) rhodopsin-like superfamily and 7 transmembrane (7-TM) GPCR rhodopsin-like domains (7-TM GPCR)
2. Chemokine receptor domains (CCR_rcpt)
3. Immunoglobulin and related domains (Ig-like)
4. C-type lectin and related domains (C-type lectin like)
5. Integrin alpha and related domains (Integrin alpha)

The distribution of the domains among the seed receptors is shown in Figure 7.19.

Predicted surface membrane factors are grouped according to their functional domains (see Table 7.5) which results in GPCR with chemokine domains, GPCR without chemokine domains, Ig-like receptors and receptors without any overlapping domains. The respective domain distribution is displayed in Figure 7.21. The largest domain overlap is found for 7-TM GPCR rhodopsin-like domains. Half of the predictions have this particular domain, which is also overrepresented in the set of seed receptors (10 of 13, see Figure 7.19). In addition, CCR1 and CCBP2 share a chemokine domain, which is very frequent in the set of initial receptors (8 of 13). Moreover, five predicted surface membrane factors have Ig-like domains that match the primary HIV receptor CD4.

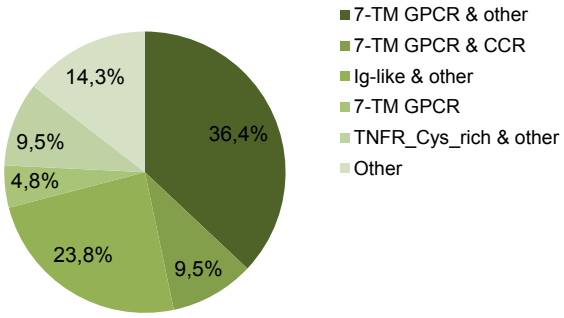


Figure 7.21: Distribution of the protein domains for the predicted surface membrane factors documented in Table 7.5.

The amount of overlapping functional domains indicates that the functional characteristics of the initial HIV binding receptors are reflected in predicted surface factors. In particular, GPCRs have a broad usage spectrum as co-receptors by primary isolates of HIV (Shimizu *et al.*, 2009) and specifically chemokine receptors are known as co-receptors for HIV (Broder and Collman, 1997). Strikingly, CCR1 and CCBP2 share both 7-TM GPCR rhodopsin-like and chemokine domains and are reported as co-receptors of HIV. However, receptors without any overlapping domains might present unprecedented characteristics that are not documented in the initial set but are reflected in their complementary domain diversity (see Figure 7.21).

7.5.3.2 Chromosomal locations

Genes with similar properties are sometimes located in the same regions of the human genome. Thus, the genomic location of a gene is often taken into account when new candidate genes are associated with a disease. The reason is that mapping those candidates to a region containing other genes related to the same disease further supports the association. For example, HIV binding human CC chemokine receptor genes are known to cluster within the 3p21.3 region of the genome (Maho *et al.*, 1999).

We determine the chromosomal location of the predicted surface proteins and study whether they cluster together with other candidates or known seed factors. The chromosomal location for each seed and prediction retrieved from EntrezGene is shown in Table B.10. When considering the known receptors there is a group of six chemokine receptors that map to the CCR cluster within 3p21.3, and also two receptors, CCR1 and CCBP2, from the predicted set are associated with this region. However, the remaining ones are located on different chromosomes. Only CD97 and DC-SIGN, and GPR17 and CXCR4 are mapped together to 19p13 and 2q21, respectively.

7.5.3.3 Literature support

We explored the literature to gather further evidence to support the relevancy of the predicted surface membrane factors. Overall, the involvement of co-receptors and surface membrane proteins assisting HIV-1 infection and contributing to viral pathogenesis has always been underestimated (Shimizu *et al.*, 2009). Only a limited number of studies aim to elucidate the role of surface membrane factors interacting with viral proteins, even though they are potential amenable drug targets for HIV therapeutics (Zhou and He, 2008; Dunn *et al.*, 2004).

Remarkably, we inferred ten surface proteins that are involved in a cascade of events in HIV infection. Among these cell surface proteins, three have confirmed functions in HIV infection while seven have been reported by at least two other studies. Their involvement ranges from serving as co-receptors for cell entry (CCR1 and CCBP2) (Shimizu *et al.*, 2009; Neil *et al.*, 2005), mediating trans-infection (DARC) (He *et al.*, 2008), activating immune cells (CD97) (Zhou and He, 2008) to inducing viral production from latently infected cells (CSF3R, TNFRSF3 and CD2) (Dunn *et al.*, 2004; Coleman and Wu, 2009; Shen *et al.*, 2007). Our findings on experimentally confirmed predictions and predictions with indirect experimental support are discussed in detail in (Jaeger *et al.*, 2010b).

We also present eleven original predictions that deserve experimental investigation (see Table 7.5). In particular, the platelet glycoprotein Ib (GPIb) is a surface membrane protein of platelets. Mutations in the GPIb beta subunit are associated with Bernard-Soulier syndrome which is characterized by thrombocytopenia, circulating giant platelets, and prolonged bleeding time (Hadjkacem *et al.*, 2009). We speculate that the prolonged interaction of blood platelet expressed GP1BB with HIV might be responsible for thrombocytopenia observed in HIV infection. Furthermore, the relaxin receptors RXFP1 and RXFP2 are known to be expressed on the acrosome of elongated spermatids (Filonzi *et al.*, 2007; Gianesello *et al.*, 2009). Their intron-rich gene organi-

zation indicates alternatively spliced variants. This suggests the existence of different protein isoforms that contribute to their diverse expression *in-vivo*. Their association with HIV might explain the different rates of evolution observed in seminal versus blood plasma of infected patients (Ghosn *et al.*, 2004). Moreover, either one or both receptors might be involved in viral hijacking of the spermatozoa in viral transmission (Kern and Bryant-Greenwood, 2009).

Several seed receptors, such as CCR5, CCR2 and CX3CR1 (Passam *et al.*, 2007; Singh *et al.*, 2008), have been associated with SNPs that contribute to different disease outcome. Among the 21 predicted factors, except for the controversial –46C/C in DARC, SNPs in CCR1, CCBP2, HTR6, HTR1B, HTR1E, CSF3R, IL1R1, TNFRSF5 are associated with one or more clinical phenotypes but their relation to HIV infection has not been investigated. Thus, we encourage investigating the SNPs from the predicted surface membrane factors for association with HIV to study their potential effect on HIV infection.

7.5.3.4 Structural Matching

As indicated above, a large number of the predicted surface membrane factors are likely to be involved in the different stages of HIV infection influencing the progression of the disease. Albeit literature curation largely confirms the relevancy of our findings, we do not predict with which particular HIV protein these factors might interact. To study this, we used PRISM (Protein Interactions by Structural Matching, Keskin *et al.* (2008)) to predict putative interactions between the predicted surface membrane factors and HIV proteins. PRISM identifies potential interactions among proteins by comparing their interfaces and structures against a subset of structurally and evolutionary representative interactions from PDB. The rationale of this approach is that if two protein structures exhibit particular surface regions that complement known interfaces, they are likely to interact through these regions.

PRISM uses a template set of known interaction interfaces to infer potential interactions between a set of target proteins. The template set is constructed from mapping binary interactions between human and HIV proteins to known protein complexes. This set characterizes virus-host interactions with respect to their physical and chemical properties. Note, only little information on the structural characteristics of interactions between human and HIV proteins is known yet which limits the structural coverage of our template set. Given the template set we run PRISM on the target set comprised of known structures of the predicted surface membrane factors and HIV proteins.

Figure 7.22 presents the predicted binary interactions between the inferred surface factors and HIV proteins. Six surface membrane factors are predicted to interact with six HIV proteins according to their complementary protein interfaces. As proteins interact through their interfaces, this structural analysis adds another level of confidence strongly supporting our predictions. Predicted binding sites for CSF3R with gp120 and gp41 are illustrated in Figure 7.23.

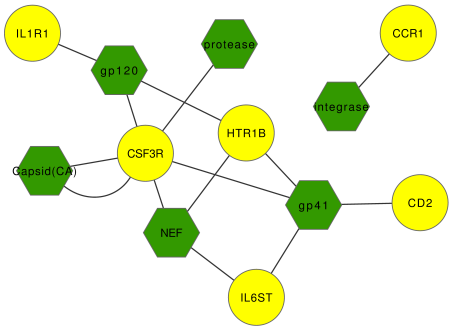
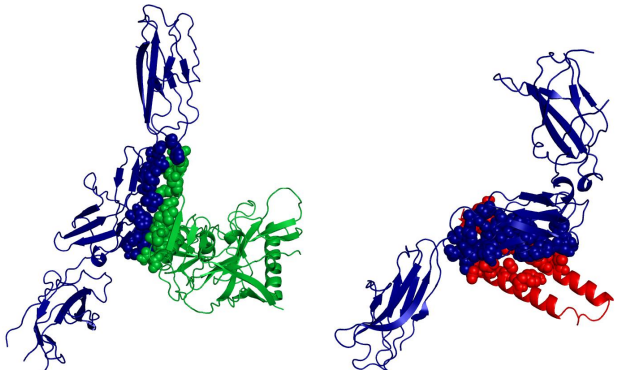


Figure 7.22: Predicted interactions between the inferred surface membrane factors (yellow circles) and HIV proteins (green hexagon).



(a) CSF3R with gp120 (b) CSF3R with gp41

Figure 7.23: Predicted protein interactions of CSF3R with (a) gp120, and (b) p41. Physical binding between these proteins is inferred based on the structural matching of their interfaces (Keskin *et al.*, 2008). CSF3R is shown in blue while gp120 and gp41 are colored in green and red, respectively. Predicted binding sites are highlighted with spheres.

8 Summary and Outlook

This thesis focused on the computational analysis of one of the most commonly studied types of biological networks – protein interaction networks – which have become particularly important for functional analysis in several organisms, particularly in human. Protein interaction networks are crucial to many aspects of cellular function (Piehler, 2005). On the one hand, they present direct and robust manifestations of functional relationships (Sharan *et al.*, 2007). On the other hand, alterations in protein interactions perturb natural cellular processes and contribute to many diseases (Ideker and Sharan, 2008). Both correlations, the functional and the pathological one, have been considered in this work to infer novel protein function for uncharacterized proteins as well as to associate yet uncharacterized proteins with disease phenotypes, respectively.

As first main contribution we presented a novel approach to predict protein function from protein interaction networks of multiple species. The key to our method is to study proteins within modules defined by evolutionary conserved processes, combining comparative cross-species genomics and functional linkage within interaction networks. To this end, interologs are assembled to highly conserved protein sub-networks, so-called connected and conserved subgraphs (CCS). Within each conserved subgraph we infer novel protein functions from orthology relationships across species and along conserved interactions of neighboring proteins within a species.

- Altogether, we integrate three different sources of evidence, namely evolutionary conservation of functional modules, orthology relationships, and direct and indirect protein interactions into a single, comprehensive prediction method which yields high-quality predictions with very good coverage.
- We show that results can be further improved by processing large CCS in an adequate manner. Failing to do so either restricts coverage of the method or leads to higher false positive rates.
- In comparison to three related methods, CCS-based function prediction clearly outperforms Neighbor Counting and χ^2. A comparable or even better performance is achieved when comparing against FS-Weighted Averaging.
- Overall, we infer thousands of protein functions for every species included in the analysis at varying, yet always high levels of precision. A large amount of novel functions can not be validated directly which shows that our method also generates novel functional knowledge rather than only reproducing known functions for well-characterized proteins.

As second main contribution we developed a region-independent, network-based framework in which we integrate protein interaction, protein function, and network centrality analysis to identify yet uncharacterized disease-related gene products. Given a

disease, we first extract all genes known to be involved in this disease. We compile a disease-specific network by integrating directly and indirectly linked gene products using protein interaction and functional information. Proteins in this network are ranked based on their network centrality.

The general approach of our method follows the lines of others but in contrast to previous methods, our approach does not depend on the availability of associated chromosomal regions. This makes it applicable to a much wider range of diseases, such as disorders with very few or even only a single known disease protein, diseases with multiple, very large, or no associated loci, and even diseases without genetic origin. As discovering disease-related genes is particularly challenging if no chromosomal regions are associated yet, we employed distinctive features to address this complexity and to enhance the disease gene discovery process:

- As disease genes are often not directly linked, we also include indirectly linked proteins during network construction which increases cross-validation re-discovery rates significantly, up to 20%.
- This extension lowers the precision since larger networks naturally integrate many global "hub" proteins which are highly central but mostly unspecific for a disease. We adjust the ranking for a bias towards hub proteins in disease networks which decreases the fraction of highly ranking hub proteins (by 23%) while increasing the fraction disease proteins up to 22%.
- Further, we integrate predicted functional information to overcome the incomplete functional coverage of the human genome which is still one of the main limitations in finding disease-related genes. Predicted functions increase the outreach of our networks and assist the proper ranking of proteins without functional annotations.
- In a benchmark comparison with related approaches our disease-specific framework outperforms PRINCE significantly and performs comparably well against RWR.
- In a case study, we identify 21 novel surface membrane factors that contribute to HIV-1 infection; three have confirmed functions in HIV infection, seven have been identified by at least two other studies, and eleven are novel predictions and thus excellent targets for experimental investigation (Jaeger *et al.*, 2010b).

Future directions

Protein function prediction and disease gene identification remain important challenges in the post-genomic era (Friedberg, 2006; Botstein and Risch, 2003). In the following, we will discuss several aspects to further improve our proposed approaches.

CCS-based function prediction

Our function prediction method is primarily based on functional modules defined by evolutionarily conserved processes. Thus, the accurate detection of CCS is an important aspect for precise function prediction. As indicated before, high coverage of our method

is partially achieved by using a relaxed definition of interaction conservation when studying multiple species. A logical extension of such approximate CCS is the inclusion of orthologous groups which do not have direct counterparts in the species under consideration. This can be implemented by considering gaps and mismatches during the network comparison procedure (Ogata *et al.*, 2000; Kelley *et al.*, 2003). Analogously to sequence alignments, gaps indicate that a protein interaction in one network omits a protein in the other network. Mismatches occur if aligned proteins do not share sequence similarity. Both concepts allow to account for evolutionary variations and experimental errors on the protein level which in turn will improve module detection and function prediction.

As shown in Section 5.3.4.1, processing large CCS generates significantly more predictions with mostly better precision. So far we splitted CCS with more than 25 proteins into smaller subgraphs (see Section 4.2.5) since biological processes typically involve only between 5 and 25 proteins (Spirin and Mirny, 2003). We initially chose the size of 25 without exploring other CCS sizes. Thus, it would be interesting to study whether the definition of large CCS used throughout this thesis is optimal with respect to precision and recall.

Apart from the promising results of our prediction approach, our method currently only provides lists of yes/no predictions. This binary behavior is implicit in the way we compute CCS and how we determine predicted terms and targets of prediction. For further improvement and applicability we extended our approach in a diploma thesis by deriving confidence scores for each prediction based on multiple biological evidence (Pollex, 2011). Predictions ranked by reliability allow to focus experimental resources on hypotheses (predictions) that are more likely to be true. This is essential for biologists to decide which proteins and predictions should be investigated further, e.g., in follow-up experiments. Pollex (2011) introduced a method that represents annotations as vectors in a feature space, in which every dimension presents specific evidence or feature of the annotation. Confidence scores have been derived by using the weighted sum of all elements in the feature vector obtained for a given annotation, so-called *Sum of Scores*. Evaluating the *Sum of Scores* approach against our binary methods indicated that combining all evidence into one score, rather than discarding weak evidence, improves the overall coverage without decreasing precision. This promising approach can be further improved, in particular in terms of precision, by assigning weights to the individual features based on the idea that distinct features are more important for discriminating between correct and incorrect predictions. Determining such weights brings up two further challenges: (i) defining a target function to maximize/minimize and (ii) determining a set negative annotation for optimizing the target function. In addition, more features, such as the conservation of an interaction, could be incorporated into the score in order to model function prediction even more accurately.

Disease gene identification

One of the key aspects in finding novel disease genes is the underlying data representing relationships between gene products (Tranchevent *et al.*, 2010). Both high quality and high coverage data sources are essential to derive precise predictions. We have shown

that utilizing indirect interaction data partially addresses the current incompleteness of the human interactome. However, there is still a large number uncharacterized genes for which only little or no functional data exist in the public databases. To further improve the inclusion of such genes we plan to incorporate (i) functional relationships extracted from the literature as wells as (ii) interactions indirectly inferred from CCS by using the relaxed interolog definition. Furthermore, less common data, e.g., quantitative protein expression, describing unique features not captured by the most widely used data sources should be investigated. Such complementary data will yield more comprehensive networks reflecting the processes related to a particular disease more accurately.

In contrast to our function prediction approach in which we use interologs to filter for spurious interactions, we do not account for the varying quality of protein interaction data yet. However, false interactions compromise disease gene identification as novel hypotheses might be derived from relationships without biological relevance (Navlakha and Kingsford, 2010). To avoid such cases it will important to assign confidence scores to protein interactions (Braun *et al.*, 2009). Several concepts have been discussed recently for increasing the quality within interaction data sets (Lage *et al.*, 2007; Chua *et al.*, 2007). Confidence scores might be based on the experimental setup as large-scale experiments generally contain more false positives than small-scale experiments (von Mering *et al.*, 2002). Additionally, the number of distinct publications documenting an interaction might be used as a score since interactions are often more reliable if they have been reproduced in more than one individual experiment.

Another crucial point is the ranking of proteins with respect to their relevance for a particular disease. For now we used the normalized betweenness centrality as score for ranking proteins within disease networks. Yet, the more comprehensive the disease networks become the more difficult becomes the ranking. To further improve the scoring and consequently the ranking we plan to use a more probabilistic approach which models the probability of a protein to be disease-related given its centrality score within the respective network. The underlying idea of this model is based on the assumption that true disease proteins receive high scores in their disease networks while unrelated proteins obtain much lower centrality scores. Such a probability can be determined by considering the ratio between the probability that a protein with a particular score is disease-related and the probability that this protein is not disease-related. Both likelihoods are based on the relative frequency of disease and non-disease genes, respectively, to receive a particular score estimated across all disease networks. Thus, if a candidate protein receives a centrality score which is more characteristic for disease genes than for non-disease genes, its probability to be also involved in the disease will be higher than for proteins with scores resembling non-disease genes. A further refinement of the proposed disease probability could be achieved by also integrating the protein distribution across the distinct disease networks into the probability model.

Using such a probability as prediction score also allows to determine a cut-off at which we make predictions for a disease. Currently, we derive predictions for any disease under consideration. Yet, as shown in Section 7.3.3 the performance of our approach varies largely depending on the disease type; for some we perform very well while we are less successful for others. By applying a probabilistic threshold we can further adjust disease

gene identification toward precision and recall. The higher the threshold the higher will be the likelihood that high-scoring candidates contribute to the disease in question. This means, considering predictions for a disease only if the top-scoring candidate exceeds a particular threshold will implicitly filter for diseases with weaker hypotheses and focuses on diseases with more reliable predictions. In turn, a lower threshold will yield a larger number of potential candidates. This might be useful if a broader picture about the gene products involved in disease-related processes needs to be obtained.

Vision for Network Systems Biology

Network biology and its various applications for understanding cellular function and organization contributed greatly to biomedical science in the last decade (Barabási and Oltvai, 2004; Pujol *et al.*, 2010; Barabási *et al.*, 2011). However, the dynamic principles within biological networks have been often neglected although dynamic interactions are crucial for regulating the function of cells and organisms. While understanding individual genes and proteins remains to be important, the current focus in research shifts toward network systems biology which integrates the quantitative component missing in network biology.

To move from static to dynamic interaction networks, physical and functional relationships between the cellular components have to be studied according to their spatial, contextual or temporal context (Przytycka *et al.*, 2010). Yet, most established large-scale technologies for identifying protein interactions, such as Y2H and TAP-MS, do not provide any spatial, temporal or contextual information. In absence of such data, genome-wide experimental data can be utilized to associate static interactions with dynamic information. Gene expression data, CHIP-chip data or phenotypic responses to perturbations obtained from knock-out studies or expression-quantitative trait loci (e-QTL) are often integrated to elucidate protein interaction and network dynamics. Han *et al.* (2004) assessed, for instance, temporal characteristics of hub proteins in yeast using gene expression data. According to their analysis, hub proteins exhibit condition- or location-specific features which indicate dynamic modularity in interaction networks. Dittrich *et al.* (2008) integrated lymphoma-specific expression data with static interaction data to detect functional modules beyond classical pathways by means of differentially expressed regions in human interaction networks.

Incorporating other data sources reflecting different types of dynamic information, such as quantitative protein expression levels, protein localization and modification data, will further advance network systems biology and its applications. However, also novel experimental techniques are needed for directly capturing interaction dynamics in large-scale. A promising approach with respect to protein interaction is FRET which has been applied to detect protein interactions in the *E. coli* chemotaxis pathway for identifying stimulation induced changes (Kentner and Sourjik, 2009). Yet, this is a small-scale approach not applicable for large-scale analysis.

Appendix A – Databases and terminologies

In the following, we provide information on databases and terminologies used in this thesis.

DIP – Database of Interacting Proteins

- DIP documents experimentally verified protein interactions of several species. Interaction data are derived from literature, PDB, and high-throughput methods, including Y2H, DNA and protein microarrays, and TAP-MS. The various sources are combined to create a single, consistent set of protein-protein interactions. In addition, core subsets are provided for each species comprising only the most reliable interactions.
- DIP is a member of the IMEX Consortium (International Molecular Exchange) and adheres to standards developed for exchanging data with other interaction databases (IntAct, MINT, etc.) to improve data quality and curation.
- *Reference:* Salwinski *et al.* (2004), `http://dip.doe-mbi.ucla.edu`

MIPS–MPPI – MIPS Mammalian Protein-Protein Interaction Database

- MIPS–MPPI collects manually curated mammalian interaction data. High-quality data sets are compiled from the scientific literature by expert curators. MPPI includes only data from individually performed experiments as they usually provide the most reliable evidence for physical interactions.
- *Reference:* Pagel *et al.* (2005), `http://mips.helmholtz-muenchen.de/proj/ppi`

IntAct – Molecular Interaction Database

- IntAct provides information on physical interactions for various species derived from small-scale and large-scale studies.
- IntAct is also a member of the IMEX Consortium and follows its standards to cooperate with similar interaction databases, such as MINT and DIP.
- *Reference:* Hermjakob *et al.* (2004a), `http://www.ebi.ac.uk/intact`

BioGRID – Database of Protein and Genetic Interactions

- BioGRID is a general repository for interaction data sets, protein and genetic interactions, compiled from comprehensive curation efforts. BioGRID documents molecular interactions derived from both high-throughput and small-scale experiments.
- BioGRID is part of the IMEX Consortium.
- *Reference:* Stark *et al.* (2006), `http://thebiogrid.org/`

MINT – Molecular INTeraction database

- MINT is a public repository for molecular interactions reported in peer-reviewed journals. It focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators.
- MINT adopted the PSI-MI standards for the annotation and for the representation of molecular interactions and is a member of the IMEx consortium.
- *Reference:* Chatr-aryamontri *et al.* (2007), `http://mint.bio.uniroma2.it/mint`

HPRD – Human Protein Reference Database

- HPRD represents a centralized platform to describe and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome. Information in HPRD are manually extracted from the literature by expert biologists who read, interpret and analyze the published data.
- *Reference:* Peri *et al.* (2003), `http://www.hprd.org/`

HIV-1, Human Protein Interaction Database

- The HIV-1, Human Protein Interaction Database catalogs the numerous interactions between human immunodeficiency virus type 1 (HIV-1) proteins and human proteins reported in the literature.
- *Reference:* Fu *et al.* (2009), `http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/`

UniProt – Protein sequence and functional information

- UniProt is a central repository of protein sequences and functional information. Annotations in UniProt include, amongst many others, sequence and related features, isoforms, protein domains, Gene Ontology annotations, interactions, SNPs, homology, associations with diseases, tissue specificity, enzymatic reactions, and encoding genes.
- *Reference:* UniProt Consortium (2010), `http://www.uniprot.org`

Entrez Gene – Database of genes from RefSeq genomes

- Entrez Gene curates gene-center information, including sequence, taxa, names and synonyms, isoforms, chromosomal location, functional annotations and description found in PubMed abstracts (GeneRIFs), gene products, associations with diseases, interactions, and pathways.
- *Reference:* Wheeler *et al.* (2008), `http://www.ncbi.nlm.nih.gov/sites/entrez/?db=gene`

Gene Ontology – Standardization of gene and gene product annotations

- GO is a widely accepted vocabulary for representing gene and protein function in a systematic manner. GO is organized as an ontology, covering three major aspects of function, each structured as an independent subontology: *molecular function*, *biological process*, and *cellular component*.
- *Reference:* Ashburner *et al.* (2000), `http://www.geneontology.org`

OMIM – Online Mendelian Inheritance in Man

- OMIM is a comprehensive compendium of genetic disorders and human genes. It catalogs all diseases with a genetic component, and links them – if possible – to the relevant genes in the human genome. In addition, it provides references for further research and tools for genomic analysis of each cataloged gene.
- *Reference:* McKusick (2007), `http://www.ncbi.nlm.nih.gov/omim/`

Appendix B – Additional Results

Table B.1: Complete results of the strict and relaxed network comparisons for pairs of species and three, four, five and six species combinations. We compute CCS for 15 combinations of two species, 20 comparisons with three, 11 with four species, six with five species, and one considering all species. For each combination the number of orthologous groups, interologs and CCS from strict and relaxed definition are presented as well as the size of the largest CCS. Species combinations discussed throughout Chapter 5 are highlighted in gray.

Species	Criteria	# OrthoMCL groups	# Interologs	# CCS (≥3)	Largest CCS Proteins (Edges)
rno sce	strict	403	42	18 (3)	14 (14)
	relaxed		3586	2 (1)	400 (3585)
hsa cel	strict	1470	233	146 (34)	10 (11)
	relaxed		7698	13 (1)	1412 (7686)
mmu dme	strict	1182	94	57 (13)	10 (13)
	relaxed		2688	13 (3)	1079 (2672)
dme cel	strict	1236	97	75 (14)	4 (3)
	relaxed		2568	17 (3)	1085 (2550)
rno cel	strict	508	12	10 (2)	3 (2)
	relaxed		905	8 (2)	475 (897)
rno dme	strict	717	17	16 (1)	3 (2)
	relaxed		1251	16 (1)	673 (1235)
rno hsa	strict	1229	459	101 (28)	236 (287)
	relaxed		6562	4 (1)	1221 (6557)
mmu cel	strict	723	49	46 (2)	3 (2)
	relaxed		1328	9 (1)	644 (1318)
rno mmu	strict	791	95	47 (8)	29 (33)
	relaxed		2026	11 (1)	762 (2016)
mmu hsa	strict	2801	1550	188 (53)	914 (1249)
	relaxed		22472	14 (1)	2777 (22456)
mmu sce	strict	551	126	48 (17)	13 (26)
	relaxed		4762	2 (1)	539 (4761)
hsa dme	strict	2782	488	167 (50)	85 (102)
	relaxed		17030	13 (2)	2684 (17017)
hsa sce	strict	1484	1192	142 (22)	464 (969)
	relaxed		19428	7 (1)	1468 (19422)

Continued on next page

Table B.1 – (continued)

Species	Criteria	# OrthoMCL groups	# Interologs	# CCS (≥3)	Largest CCS Proteins (Edges)
dme sce	strict	1219	210	94 (18)	25 (50)
	relaxed		12732	6 (1)	1188 (12727)
cel sce	strict	693	118	81 (15)	7 (7)
	relaxed		6494	6 (1)	667 (6489)
rno hsa cel	strict	405	8	7 (1)	3 (2)
	relaxed		116	46 (11)	41 (49)
mmu hsa cel	strict	618	28	27 (1)	3 (2)
	relaxed		244	66 (14)	74 (101)
mmu dme cel	strict	625	20	20 (0)	2 (1)
	relaxed		86	60 (13)	5 (4)
rno dme cel	strict	414	3	2 (1)	3 (2)
	relaxed		33	25 (5)	4 (3)
rno hsa dme	strict	586	7	6 (1)	3 (2)
	relaxed		160	58 (11)	55 (66)
rno mmu sce	strict	235	5	4 (0)	2 (1)
	relaxed		29	15 (3)	9 (11)
rno mmu cel	strict	324	6	6 (0)	2 (1)
	relaxed		36	19 (3)	10 (13)
rno dme sce	strict	346	2	2 (0)	2 (1)
	relaxed		78	23 (5)	16 (17)
hsa dme sce	strict	1114	119	65 (12)	13 (12)
	relaxed		959	127 (23)	344 (727)
hsa dme cel	strict	1116	67	57 (7)	4 (3)
	relaxed		322	135 (34)	27 (30)
rno mmu dme	strict	432	6	6 (0)	2 (1)
	relaxed		43	21 (4)	12 (16)
rno hsa sce	strict	348	12	10 (1)	3 (2)
	relaxed		237	47 (12)	55 (90)
rno mmu hsa	strict	707	71	39 (5)	20 (24)
	relaxed		539	34 (5)	337 (501)
rno cel sce	strict	267	3	3 (0)	2 (1)
	relaxed		56	25 (6)	16 (18)
mmu hsa dme	strict	1009	60	41 (11)	5 (5)
	relaxed		497	88 (20)	230 (309)
dme cel sce	strict	600	35	28 (6)	4 (3)
	relaxed		176	78 (15)	28 (48)
mmu hsa sce	strict	485	69	31 (6)	13 (25)
	relaxed		474	68 (20)	144 (321)

Continued on next page

Table B.1 – (continued)

Species	Criteria	# OrthoMCL groups	# Interologs	# CCS (≥3)	Largest CCS Proteins (Edges)
hsa cel sce	strict	657	74	53 (7)	7 (6)
	relaxed		541	87 (13)	215 (409)
mmu dme sce	strict	460	20	16 (2)	4 (3)
	relaxed		159	48 (11)	36 (43)
mmu cel sce	strict	333	9	9 (0)	2 (1)
	relaxed		85	44 (13)	8 (8)
rno mmu dme sce	strict	211	0	–	–
	relaxed		44	18 (3)	11 (13)
mmu hsa dme cel	strict	510	9	9 (0)	2 (1)
	relaxed		263	58 (14)	113 (166)
rno hsa dme cel	strict	330	2	2 (0)	2 (1)
	relaxed		121	41 (6)	36 (50)
mmu hsa cel sce	strict	286	7	7 (0)	2 (1)
	relaxed		243	45 (13)	90 (168)
rno mmu dme cel	strict	269	2	2 (0)	2 (1)
	relaxed		38	17 (3)	11 (15)
rno mmu hsa sce	strict	202	5	4 (0)	2 (1)
	relaxed		145	39 (14)	36 (73)
rno mmu hsa cel	strict	269	3	3 (0)	2 (1)
	relaxed		130	32 (7)	50 (79)
rno hsa dme sce	strict	291	0	0 (0)	0 (0)
	relaxed		186	46 (10)	31 (58)
rno mmu hsa dme	strict	368	3	3 (0)	2 (1)
	relaxed		175	35 (6)	89 (120)
rno mmu cel sce	strict	167	1	1 (0)	1 (1)
	relaxed		31	17 (3)	7 (8)
rno hsa cel sce	strict	222	3	3 (0)	2 (1)
	relaxed		139	36 (9)	33 (62)
rno dme cel sce	strict	234	0	–	–
	relaxed		73	21 (4)	17 (20)
mmu hsa dme sce	strict	395	16	11 (3)	4 (3)
	relaxed		433	53 (14)	146 (324)
mmu dme cel sce	strict	308	6	6 (0)	2 (1)
	relaxed		119	46 (12)	18 (24)
hsa dme cel sce	strict	552	22	19 (3)	3 (2)
	relaxed		477	67 (12)	200 (372)
rno mmu dme cel sce	strict	166	0	–	–
	relaxed		7	6 (0)	2 (1)

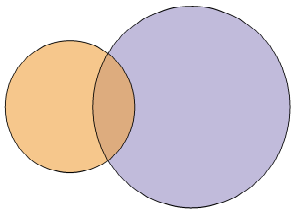
Continued on next page

Table B.1 – (continued)

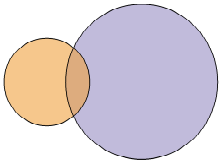
Species	Criteria	# OrthoMCL groups	# Interologs	# CCS (≥3)	Largest CCS Proteins (Edges)
mmu hsa dme cel sce	strict	271	3	3 (0)	2 (1)
	relaxed		64	35 (9)	8 (11)
rno mmu hsa cel sce	strict	147	1	1 (0)	1 (1)
	relaxed		16	11 (1)	3 (3)
rno mmu hsa dme sce	strict	181	0	–	–
	relaxed		23	12 (3)	5 (6)
rno mmu hsa dme cel	strict	234	1	1 (0)	2 (1)
	relaxed		30	17 (3)	8 (12)
rno hsa dme cel sce	strict	206	0	–	–
	relaxed		35	19 (2)	11 (12)
rno mmu hsa dme cel sce	strict	141	0	–	–
	relaxed		24	15 (2)	5 (6)

Table B.2: Impact of processing large CCS on function prediction in multiple species. CCS with more than 25 proteins are splitted into smaller, overlapping sub-subgraphs.

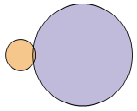
		0.3			0.5			0.7	
Species	# Terms	P	R_{pp}	# Terms	P	R_{pp}	# Terms	P	R_{pp}
Non-split									
mmu	15945	0.59	0.37	4262	0.73	0.48	449	0.82	0.49
hsa	22742	0.47	0.42	831	0.75	0.62	594	0.89	0.67
dme	13590	0.54	0.38	1746	0.72	0.31	873	0.86	0.63
sce	11314	0.75	0.39	8876	0.74	0.31	947	0.85	0.54
Split									
mmu	22505	0.60	0.14	12508	0.70	0.17	4941	0.83	0.27
hsa	21552	0.61	0.14	10762	0.75	0.22	3757	0.87	0.32
dme	17023	0.59	0.14	9863	0.72	0.21	3411	0.85	0.35
sce	17112	0.75	0.23	14224	0.77	0.15	8625	0.85	0.23
Non-split									
hsa	21046	0.51	0.36	352	0.49	0.17	0	–	–
dme	9115	0.52	0.23	3981	0.67	0.34	230	0.83	0.34
cel	4903	0.53	0.25	4125	0.56	0.24	72	0.65	0.13
sce	8173	0.82	0.31	7134	0.83	0.31	689	0.91	0.34
Split									
hsa	23497	0.63	0.17	14722	0.69	0.20	6380	0.88	0.23
dme	12813	0.58	0.13	7484	0.74	0.18	4593	0.81	0.29
cel	7801	0.51	0.11	5043	0.57	0.09	1747	0.68	0.10
sce	13006	0.85	0.16	12136	0.86	0.16	9420	0.91	0.23



Threshold 0.3

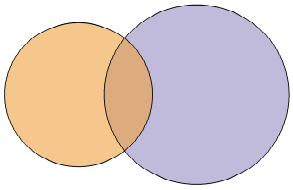


Threshold 0.5

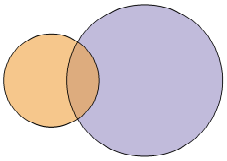


Threshold 0.7

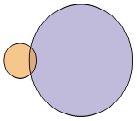
(a) Human



Threshold 0.3

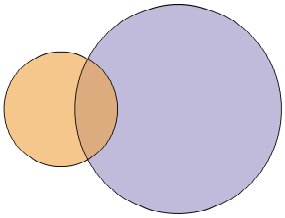


Threshold 0.5

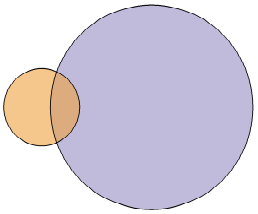


Threshold 0.7

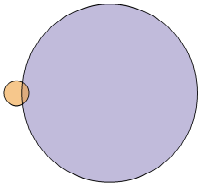
(b) Fly



Threshold 0.3



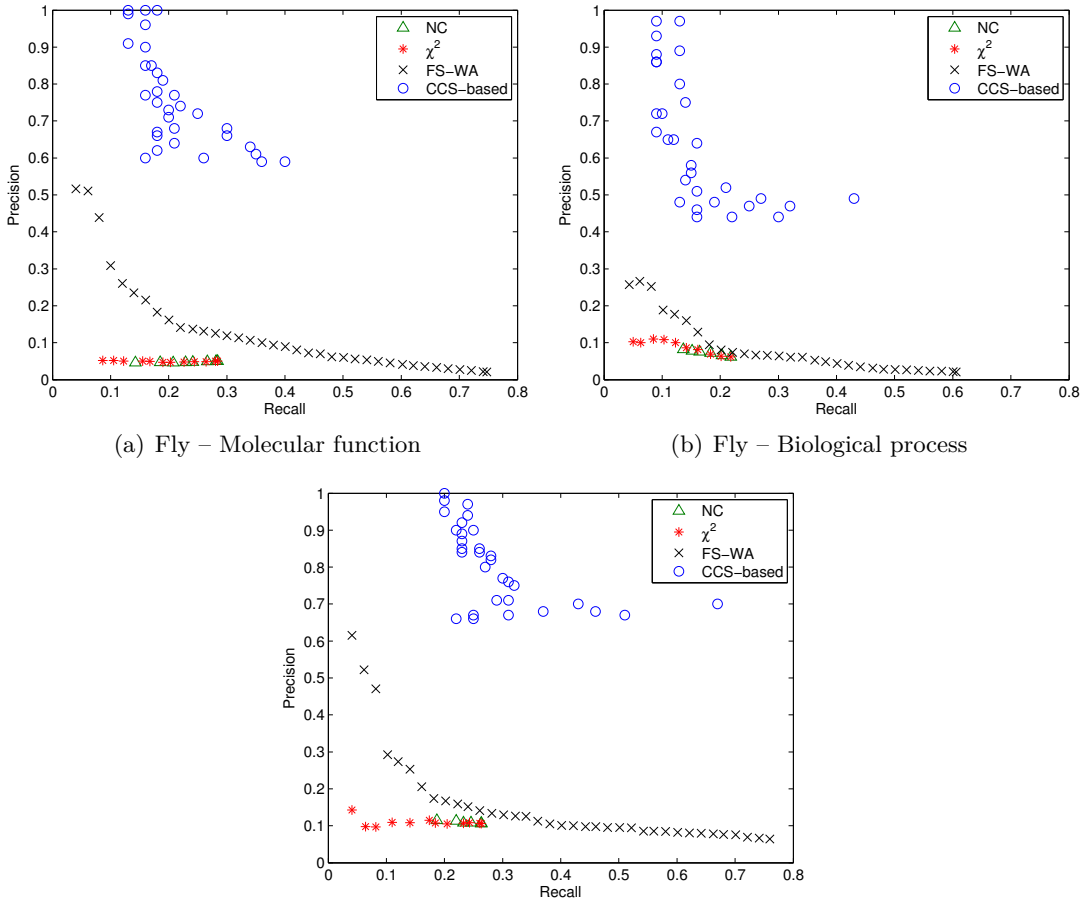
Threshold 0.5



Threshold 0.7

(c) Yeast

Figure B.1: Overlap within predictions derived from the orthology- (orange) and link-based (olive) strategy for human, fly and yeast proteins from *hsa-dme-sce*.

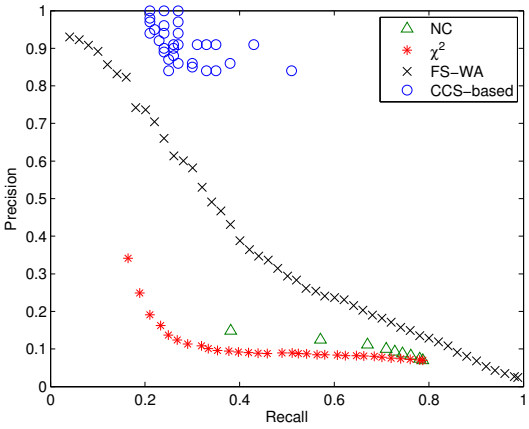


(a) Fly – Molecular function

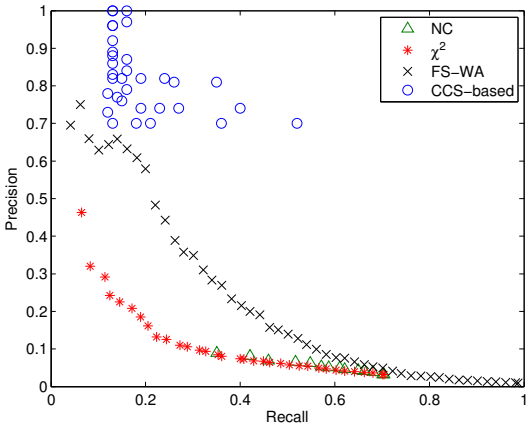
(b) Fly – Biological process

(c) Fly – Cellular component

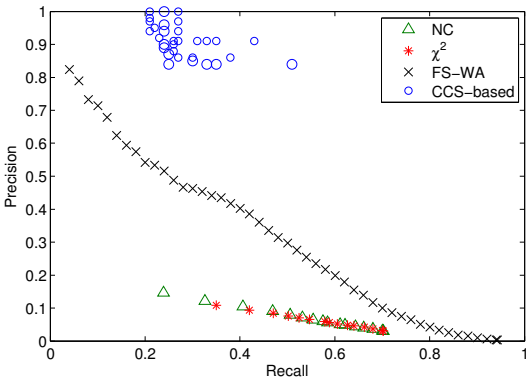
Figure B.2: Performance comparison for fly. CCS-based precision and recall are compared against *Neighbor Counting* (NC), χ^2 *statistics* and *FS-Weighted Averaging* (FS-WA) for molecular function, biological process and cellular component.



(a) Yeast – Molecular function



(b) Yeast – Biological process



(c) Yeast – Cellular component

Figure B.3: Performance comparison for yeast. CCS-based precision and recall are compared against *Neighbor Counting* (NC), χ^2 *statistics* and *FS-Weighted Averaging* (FS-WA) for molecular function, biological process and cellular component.

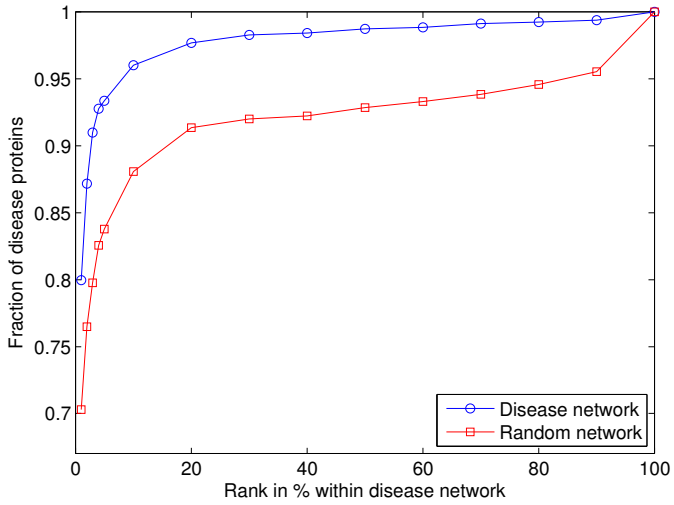
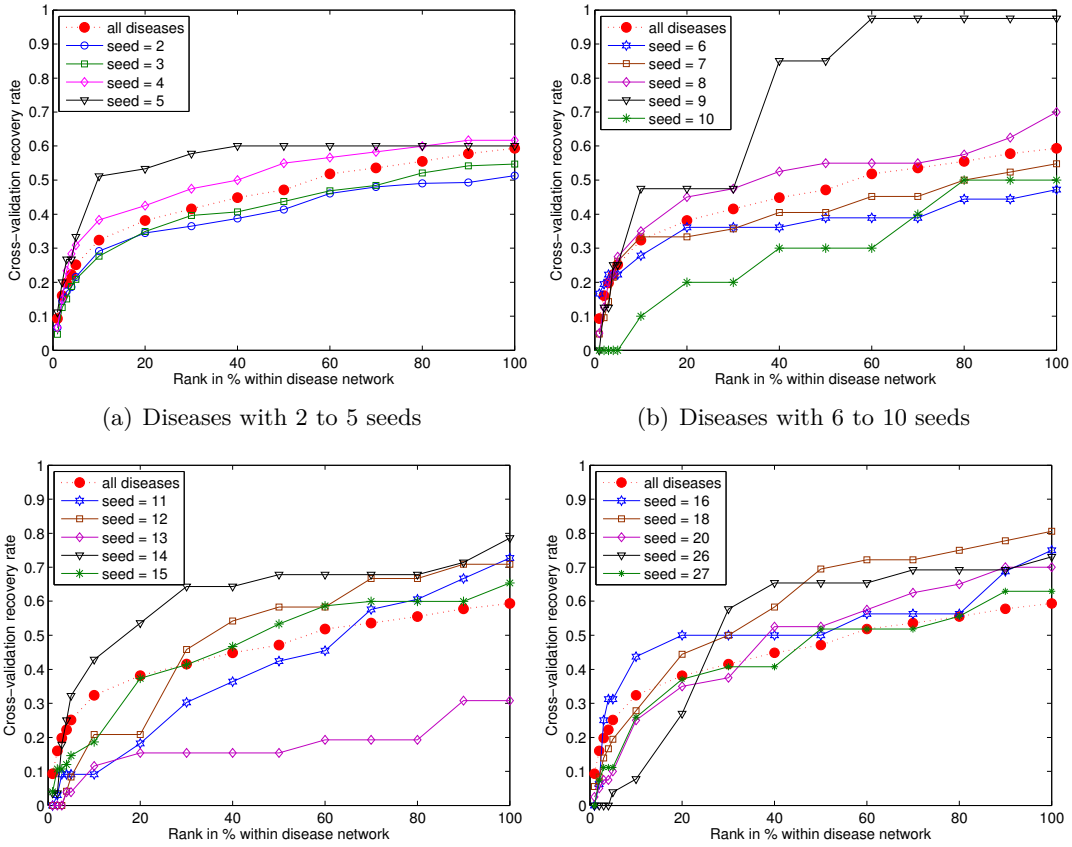


Figure B.4: Comparison of the ranking in disease and random d_1 networks.

Table B.3: List of disease classes as defined by Goh *et al.* (2007). In addition, the number of OMIM diseases associated with each class is given as well as the number of OMIM diseases per class considered in the disease-specific crossvalidation (CV) in Section 7.3.3.

Disease class	# OMIM diseases	# OMIM diseases studied in CV
Ophthamological	123	15
Immunological	70	12
Dermatological	84	24
Metabolic	248	16
Gastrointestinal	24	7
Muscular	70	8
Skeletal	72	7
Ear,Nose,Throat	46	3
Nutritional	7	2
Connective tissue disorder	42	12
Endocrine	74	14
Cardiovascular	84	11
Cancer	101	41
Hematological	113	9
Renal	49	6
Bone	46	8
Developmental	39	7
Respiratory	13	7
Neurological	234	18
Psychiatric	19	5
Unclassified	18	1
Multiple	181	23
Total	1757	284



(a) Diseases with 2 to 5 seeds

(b) Diseases with 6 to 10 seeds

(c) Diseases with 11 to 15 seeds

(d) Diseases with 16 to 27 seeds

Figure B.5: Seed-number-specific cross-validation recovery rates for $s = \{2, 3, ..., 27\}$ (complete results).

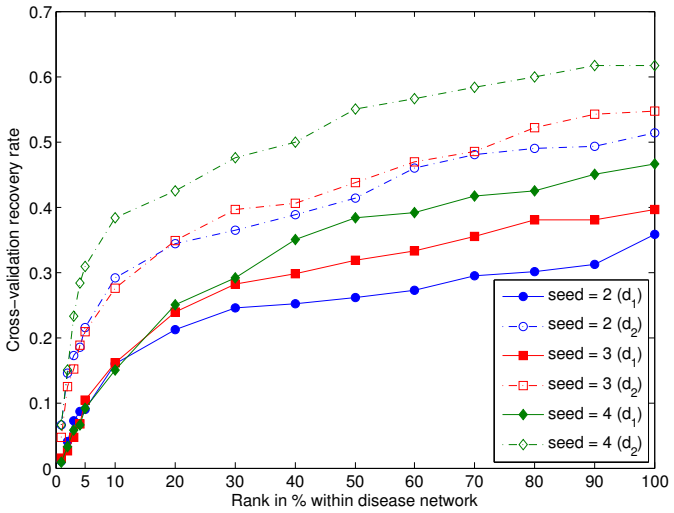


Figure B.6: Comparison of the seed-number specific recovery rates between enriched d_1 and d_2 disease networks for $s = \{2, 3, 4\}$.

Table B.4: List of cHL-associated seed proteins experimentally linked to cHL by epigenetic studies.

Gene Symbol	Gene ID	UniProt	Gene Name
TNFAIP3	7128	P21580	Tumor necrosis factor, alpha-induced protein 3
NCF2	4688	P19878	Neutrophil cytosol factor 2
VIM	7431	P08670	Vimentin
HIST2H4B	121504	P62805	Histone H4
RNF11	26994	Q9Y3C5	RING finger protein 11
HSPA1A	3303	P08107	Heat shock 70 kDa protein 1
FSCN1	6624	Q16658	Fascin
STAT1	6772	P42224	Signal transducer and activator of transcription 1, 91kDa
HIST1H2AC	8334	Q93077	Histone H2A type 1-C
IER3	8870	P46695	Radiation-inducible immediate-early gene IEX-1
MGST3	4259	O14880	Microsomal glutathione S-transferase 3
HIST1H2BI	-	P62807	Histone H2B type 1-C/E/F/G/I
OPTN	10133	Q96CV9	Optineurin
SLC2A3	6515	P11169	Solute carrier family 2, facilitated glucose transporter member 3
BCL2A1	597	Q16548	Bcl-2-related protein A1
JUN	3725	P05412	Transcription factor AP-1
ATF3	467	P18847	Cyclic AMP-dependent transcription factor ATF-3
NOTCH2	4853	Q04721	Neurogenic locus notch homolog protein 2 precursor
RYBP	23429	Q8N488	RING1 and YY1-binding protein
ZMIZ2	83637	Q8NF64	Zinc finger MIZ domain-containing protein 2
ID2	3398	Q02363	DNA-binding protein inhibitor ID-2
CCR7	1236	P32248	C-C chemokine receptor type 7

Table B.5: Novel candidates inferred from the lymphoma-specific network that were not yet associated to cHL (sorted by rank).

Gene Symbol	Gene ID	UniProt	Gene Name	Mentioned in cHL context
HIST1H1C	3006	P16403	Histone H1.2	Hodgkin-related
ACTL6B	51412	O94805	Actin-like protein 6B	
HIST1H2AM	8329	P0C0S8	Histone H2A type 1	
HIST1H3J	8350	P68431	Histone H3.1	
SMCHD1	23347	O75141	Structural maintenance of chromosomes flexible hinge domain containing 1	
HIST2H3A	126961	Q71DI3	Histone H3.2	
NUF2	83540	Q5SXK4	NDC80 kinetochore complex component, homolog	
EPS15	2060	P42566	Epidermal growth factor receptor substrate 15	
VPS25	84313	Q9BRG1	Vacuolar protein-sorting-associated protein 25	
PREB	10113	Q9HCU5	Prolactin regulatory element-binding protein	
HIST2H2BF	440689	Q5QNW6	Histone H2B type 2-F	

Continued on next page

Table B.5 – (continued)

Gene Symbol	Gene ID	UniProt	Gene Name	Mentioned in cHL context
VPS36	51028	Q86VN1	Vacuolar protein-sorting-associated protein 36	
HIST3H3	8290	Q16695	Histone H3	
CHD3	1107	Q12873	Chromodomain-helicase-DNA-binding protein 3	
PSMA1	5682	P25786	Proteasome subunit alpha type-1	Proteasome complex
LCP1	3936	P13796	Plastin-2	
HIST1H2BB	3018	P33778	Histone H2B type 1-B	
TRA@	6955	Q6PJ56	TRA@ T cell receptor alpha locus	
TSC22D3	1831	Q99576	TSC22 domain family protein 3	
HNRNPD	3184	Q12771	Heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa)	
DIAPH3	81624	Q9NSV4	Protein diaphanous homolog 3	
ZFYVE16	9765	Q7Z3T8	ENDOFIN, zinc finger, FYVE domain containing 16	
DCP1A	55802	Q9NPI6	mRNA-decapping enzyme 1A	
PSMA3	5684	P25788	Proteasome subunit alpha type-3	Proteasome complex
TNFRSF10D	8793	Q9UBN6	Tumor necrosis factor receptor superfamily member 10D	TNF
DPF2	5977	Q92785	Zinc finger protein ubi-d4	Apoptosis
NAP1L5	266812	Q96NT1	Nucleosome assembly protein 1-like 5	
KIF15	56992	Q9NS87	Kinesin-like protein KIF15	
PDE3A	5139	Q14432	cGMP-inhibited 3',5'-cyclic phosphodiesterase A	
DCP2	167227	Q8IU60	mRNA-decapping enzyme 2	
HIST1H2AE	3012	P28001	Histone H2A type 1-E	
OCRL	4952	Q01968	Inositol polyphosphate 5-phosphatase OCRL-1	
DEDD	9191	O75618	Death effector domain-containing protein	Apoptosis
YWHAZ	7534	P63104	14-3-3 protein zeta/delta	
PLS3	5358	P13797	Plastin-3	Hodgkin-related
ATG12	9140	O94817	Autophagy-related protein 12	
CASP8AP2	9994	Q9UKL3	CASP8-associated protein 2	Apoptosis, TNF, NF-κ-B pathway
STUB1	10273	Q9UNE7	STIP1 homology and U box-containing protein 1	U-protein ligase activity
PDE4DIP	9659	Q5VU43	Myomegalin	
SAP30BP	29115	Q9UHR5	SAP30-binding protein	Apoptosis
MAPK14	1432	Q16539	Mitogen-activated protein kinase 14	
PSMB1	5689	P20618	Proteasome (prosome, macropain) subunit, beta type, 1	Proteasome complex
KLK3	354	Q8NCW4	Prostate specific antigen precursor	
CAPN1	823	P07384	Calpain-1 catalytic subunit	

Continued on next page

Table B.5 – (continued)

Gene Symbol	Gene ID	UniProt	Gene Name	Mentioned in cHL context
PRKAB1	5564	Q9Y478	5'-AMP-activated protein kinase subunit beta-1	
NME2	4831	P22392	Nucleoside diphosphate kinase B	
GH1	2688	P01241	Somatotropin precursor	
PAN2	9924	Q504Q3	PAB-dependent poly	
TRAF4	9618	Q9BUZ4	TNF receptor-associated factor 4	TNF
MCL1	4170	Q07820	Induced myeloid leukemia cell differentiation protein Mcl-1	Apoptosis
TRAF6	7189	Q9Y4K3	TNF receptor-associated factor 6	TNF
VDAC1	7416	P21796	Voltage-dependent anion-selective channel protein 1	
TTR	7276	P02766	Transthyretin precursor	
USP9X	8239	Q59EZ5	Ubiquitin specific peptidase 9, X-linked	
CORO2B	10391	Q9UQ03	Coronin-2B	
APIP	51074	Q96GX9	APAF1-interacting protein	
DCP1B	196513	Q8IZD4	mRNA-decapping enzyme 1B	
TIAF1	9220	O95411	TGFB1-induced anti-apoptotic factor 1	
SUMO2	6613	P61956	Small ubiquitin-related modifier 2	
CASP4	837	P49662	Caspase-4	Apoptosis
PSMA6	5687	P60900	Proteasome subunit alpha type-6	Proteasome complex
TRADD	8717	Q15628	Tumor necrosis factor receptor type 1-associated DEATH domain protein	TNF
H1F0	3005	P07305	Histone H1.0	
NFKBIB	4793	Q15653	NF-kappa-B inhibitor beta	NF-κ-B pathway
TNIK	23043	Q9UKE5	TRAF2 and NCK-interacting protein kinase	NF-κ-B pathway
HIST1H2BD	3017	P58876	Histone H2B type 1-D	
DEDD2	162989	Q8WXF8	DNA-binding death effector domain-containing protein 2	Apoptosis
SLC2A2	6514	P11168	Solute carrier family 2, facilitated glucose transporter member 2	
TRAF3	7187	Q13114	TNF receptor-associated factor 3	TNF
LMO7	4008	Q8WWI1	LIM domain only protein 7	
PSMA5	5686	P28066	Proteasome subunit alpha type-5	Proteasome complex
ATXN1	6310	P54253	Ataxin-1	Hodgkin-related
HIRIP3	8479	Q9BW71	HIRA-interacting protein 3	
STAT3	6774	P40763	Signal transducer and activator of transcription 3	Hodgkin-related
RIPK2	8767	O43353	Receptor-interacting serine/threonine-protein kinase 2	NF-κ-B pathway
CASP3	836	P42574	Caspase-3	Apoptosis
TP53BP2	7159	Q05BL1	TP53BP2 protein	Hodgkin-related, Apoptosis

Continued on next page

Table B.5 – (continued)

Gene Symbol	Gene ID	UniProt	Gene Name	Mentioned in cHL context
CFLAR	8837	O15519	CASP8 and FADD-like apoptosis regulator	Hodgkin-related, Apoptosis
PSMA7	5688	O14818	Proteasome subunit alpha type-7	Proteasome complex
HIPK3	10114	Q9H422	Homeodomain-interacting protein kinase 3	Apoptosis
KRT16	3868	P08779	Keratin, type I cytoskeletal 16	
PAK2	5062	Q13177	Serine/threonine-protein kinase PAK 2	
PSMC2	5701	P35998	26S protease regulatory subunit 7	Proteasome complex
PSMF1	9491	Q92530	Proteasome inhibitor PI31 subunit	Proteasome complex
TNFRSF18	8784	Q5T7K5	Tumor necrosis factor receptor superfamily, member 18	TNF
BIRC3	330	Q13489	Baculoviral IAP repeat-containing protein 3	Apoptosis, TNF
RPS6KA5	9252	O75582	Ribosomal protein S6 kinase alpha-5	
NCF1C	653361	P14598	Neutrophil cytosol factor 1	
TNFRSF1A	7132	P19438	Tumor necrosis factor receptor superfamily member 1A	TNF
NFKB2	4791	Q00653	Nuclear factor NF-kappa-B p100 subunit	Hodgkin-related, NF-κ-B pathway
FASLG	356	P48023	Tumor necrosis factor ligand superfamily member 6	TNF
CDC42	998	P60953	Cell division control protein 42 homolog	TNF
TNFRSF25	8718	Q93038	Tumor necrosis factor receptor superfamily member 25	TNF
CTSL1	1514	P07711	Cathepsin L1 precursor	
ROCK1	6093	Q13464	Rho-associated protein kinase 1	
MYC	4609	P01106	Myc proto-oncogene protein	
PASK	23178	Q96RG2	PAS domain-containing serine/threonine-protein kinase	Hodgkin-related
FAS	355	P25445	Tumor necrosis factor receptor superfamily member 6 precursor	Hodgkin-related, TNF
HIST2H2AC	8338	Q16777	Histone H2A type 2-C	
PSMA2	5683	P25787	Proteasome subunit alpha type-2	Proteasome complex
NFKB1	4790	P19838	Nuclear factor NF-kappa-B p105 subunit	NF-κ-B pathway
STAT2	6773	P52630	Signal transducer and activator of transcription 2	JAK-Stat-Signaling
VHL	7428	P40337	Von Hippel-Lindau disease tumor suppressor	Proteasome complex
TSC22D1	8848	Q6IBU1	TSC22 domain family, member 1	
BCAR3	8412	O75815	Breast cancer anti-estrogen resistance protein 3	

Continued on next page

Table B.5 – (continued)

Gene Symbol	Gene ID	UniProt	Gene Name	Mentioned in cHL context
BAG2	9532	O95816	BAG family molecular chaperone regulator 2	
CDV3	55573	Q9UKY7	Protein CDV3 homolog	
F10	2159	P00742	Coagulation factor X precursor	
PSMD10	5716	O75832	26S proteasome non-ATPase regulatory subunit 10	Proteasome complex
CAPN2	824	P17655	Calpain-2 catalytic subunit precursor	Hodgkin-related
RARA	5914	P10276	Retinoic acid receptor alpha	
DFFA	1676	O00273	DNA fragmentation factor subunit alpha	Apoptosis
HIST3H2BB	128312	Q8N257	Histone H2B type 3-B	
E2F4	1874	Q16254	Transcription factor E2F4	

Table B.6: Colorectal cancer types extracted from OMIM.

OMIM ID	Phenotype
#114500	COLORECTAL CANCER; CRC
#120435	LYNCH SYNDROME I
#175100	ADENOMATOUS POLYPOSIS OF THE COLON; APC
+190182	COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 6, INCLUDED; HNPCC6, INCLUDED
246470	LEUKEMIA, ACUTE MYELOCYTIC, WITH POLYPOSIS COLI AND COLON CANCER
+600258	COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 3, INCLUDED; HNPCC3, INCLUDED
+600259	COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 4, INCLUDED; HNPCC4, INCLUDED
+600678	COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 5, INCLUDED; HNPCC5, INCLUDED
+604395	COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 7, INCLUDED; HNPCC7, INCLUDED
#608456	COLORECTAL ADENOMATOUS POLYPOSIS, AUTOSOMAL RECESSIVE
#608615	OLIGODONTIA-COLORECTAL CANCER SYNDROME
%608812	COLORECTAL CANCER, SUSCEPTIBILITY TO, 1; CRCS1
#609310	COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2, INCLUDED; HNPCC2, INCLUDED
%611469	COLORECTAL CANCER, SUSCEPTIBILITY TO, 2; CRCS2
#612229	COLORECTAL CANCER, SUSCEPTIBILITY TO, 3; CRCS3
%612230	COLORECTAL CANCER, SUSCEPTIBILITY TO, 5; CRCS5
%612231	COLORECTAL CANCER, SUSCEPTIBILITY TO, 6; CRCS6
%612232	COLORECTAL CANCER, SUSCEPTIBILITY TO, 7; CRCS7
%612589	COLORECTAL CANCER, SUSCEPTIBILITY TO, 8; CRCS8
%612590	COLORECTAL CANCER, SUSCEPTIBILITY TO, 9; CRCS9
%612591	COLORECTAL CANCER, SUSCEPTIBILITY TO, 10; CRCS10
%612592	COLORECTAL CANCER, SUSCEPTIBILITY TO, 11; CRCS11

Table B.7: List of proteins associated with colorectal cancer in OMIM.

Gene Symbol	Gene ID	UniProt	Gene Name
PLA2G2A	P14555	5320	Phospholipase A2, membrane associated
CCND1	P24385	595	G1/S-specific cyclin-D1
PMS2	P54278	5395	Mismatch repair endonuclease PMS2
MSH6	P52701	2956	DNA mismatch repair protein Msh6
PMS1	P54277	5378	PMS1 protein homolog 1
MLH1	P40692	4292	DNA mismatch repair protein Mlh1
TLR4	O00206	7099	Toll-like receptor 4
ODC1	P11926	4953	Ornithine decarboxylase
AURKA	O14965	6790	Serine/threonine-protein kinase 6
BUB1B	O60566	701	Mitotic checkpoint serine/threonine-protein kinase BUB1 beta
MUTYH	Q9UIF7	4595	A/G-specific adenine DNA glycosylase
SMAD7	O15105	4092	Mothers against decapentaplegic homolog 7
APC	P25054	324	Adenomatous polyposis coli protein
GALNT12	Q8IXK2	79695	Polypeptide N-acetylgalactosaminyltransferase 12
PDGFRL	Q15198	5157	Platelet-derived growth factor receptor-like protein
EP300	Q09472	2033	Histone acetyltransferase p300
MLH3	Q9UHC1	27030	DNA mismatch repair protein Mlh3
PTPRJ	Q12913	5795	Receptor-type tyrosine-protein phosphatase eta
AXIN2	Q9Y2T1	8313	Axin-2
FLCN	Q8NFG4	201163	Folliculin
TP53	P04637	7157	Cellular tumor antigen p53
TGFBR2	P37173	7048	TGF-beta receptor type-2
PIK3CA	P42336	5290	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha isoform
TLR2	O60603	7097	Toll-like receptor 2
MSH2	P43246	4436	DNA mismatch repair protein Msh2
NRAS	P01111	4893	GTPase NRas
AKT1	P31749	207	RAC-alpha serine/threonine-protein kinase

Table B.8: Proteins predicted to be involved in colorectal cancer. Each gene product is specified by gene symbol, gene id and uniprot id. Relevant literature supporting a potential association with colorectal cancer is referenced by the respective PubMed Ids (PMIDs) and information on colorectal or cancer pathways are derived from KEGG. Highly relevant predictions for CRC are underlined.

Gene symbol	Gene Id	UniProt	Gene name	PMIDs	Pathways
PLA2R1	22925	Q13018	Secretory phospholipase A2 receptor	17970059	–
CLN3	1201	Q13286	Battenin	–	–
ATXN1	6310	P54253	Ataxin-1	–	–
MYC	4609	P01106	Myc proto-oncogene protein	20065031, 18454692	Colorectal cancer
YWHAG	7532	P61981	14-3-3 protein gamma	–	–
EGFR	1956	P00533	Epidermal growth factor receptor	20072938, 20070321, 19451802, 17562274, 19014499, 19033715, 18497962, 18413774	Colorectal cancer
YWHAZ	7534	P63104	14-3-3 protein zeta–delta	21385632	–
SRC	6714	P12931	Proto-oncogene tyrosine-protein kinase Src	12370817, 12420216, 19620276, 18839319	–
SH3GLB2	56904	Q9NR46	Endophilin-B2	–	–
SFN	2810	P31947	14-3-3 protein sigma	–	–
SLX4	84464	Q8IY92	Structure-specific endonuclease subunit SLX4	–	–
COPS6	10980	Q7L5N1	COP9 signalosome complex subunit 6	–	–
SMAD2	4087	Q15796	Mothers against decapentaplegic homolog 2	12967141	Colorectal cancer
PIK3R1	5295	P27986	Phosphatidylinositol 3-kinase regulatory subunit alpha	18245521, 19962665	Colorectal cancer
UBE2I	7329	P63279	SUMO-conjugating enzyme UBC9	–	–
CTNNB1	1499	P35222	Catenin beta-1	20514474, 19190323, 12810642	Colorectal cancer
MUC2	4583	Q02817	Mucin-2	11850585, 16816167	–
RELA	5970	Q04206	Transcription factor p65	15112579, 15484295	Pathways in cancer

Continued on next page

Table B.8 – (continued)

Gene symbol	Gene Id	UniProt	Gene name	PMIDs	Pathways
PLK1	5347	P53350	Serine–threonine-protein kinase PLK1	16237758, 16052696	–
SMAD4	4089	Q13485	Mothers against decapentaplegic homolog 4	15814640, 15711891, 10340381, 12967141	Colorectal cancer
UBB	7314	P0CG47	Polyubiquitin-B	–	–
UBC	7316	P0CG48	Polyubiquitin-C	–	–
RPS27A	6233	P62979	Ubiquitin-40S ribosomal protein S27a	–	–
UBA52	7311	P62987	Ubiquitin-60S ribosomal protein L40	–	–
SETDB1	9869	Q15047	Histone-lysine N-methyltransferase SETDB1	–	–
RB1	5925	P06400	Retinoblastoma-associated protein	–	Pathways in cancer
ATG4C	84938	Q96DT6	Cysteine protease ATG4C	–	–
CRMP1	1400	Q14194	Dihydropyrimidinase-related protein 1	–	–
CCNDBP1	23582	O95273	Cyclin-D1-binding protein 1	–	–
ATG4B	23192	Q9Y4P1	Cysteine protease ATG4B	–	–
MUC7	4589	Q8TAX7	Mucin-7	–	–
PRKACA	5566	P17612	cAMP-dependent protein kinase catalytic subunit alpha	–	–
PRKAB2	5565	O43741	5ˊAMP-activated protein kinase subunit beta-2	–	–
AKTIP	64400	Q9H8T0	AKT-interacting protein	–	–
RAF1	5894	P04049	RAF proto-oncogene serine–threonine-protein kinase	–	Colorectal cancer
PTP4A3	11156	O75365	Protein tyrosine phosphatase type IVA 3	11598267, 17440740	–
NDC80	10403	O14777	Kinetochore protein NDC80 homolog	–	–
FOXO3	2309	O43524	Forkhead box protein O3	17615082	–

Continued on next page

Table B.8 – (continued)

Gene symbol	Gene Id	UniProt	Gene name	PMIDs	Pathways
ZBTB16	7704	Q05516	Zinc finger and BTB domain-containing protein 16	–	Pathways in cancer
IGSF1	3547	Q8N6C5	Immunoglobulin superfamily member 1	–	–
UBE2N	7334	P61088	Ubiquitin-conjugating enzyme E2 N	–	–
JUN	3725	P05412	Transcription factor AP-1	17510524 , 15139522	Colorectal cancer
PCNA	5111	P12004	Proliferating cell nuclear antigen	–	–
MED31	51003	Q9Y3C7	Mediator of RNA polymerase II transcription subunit 31	–	–
MAPK14	1432	Q16539	Mitogen-activated protein kinase 14	18444174, 19845689	Colorectal cancer
MAPK3	5595	P27361	Mitogen-activated protein kinase 3	15735687, 18533112	Colorectal cancer
ESR1	2099	P03372	Estrogen receptor	14500559, 20064828, 18727987, 18706253, 16788818	–
CREBBP	1387	Q92793	CREB-binding protein	–	Pathways in cancer
SMAD1	4086	Q15797	Mothers against decapentaplegic homolog 1	–	–
CDC20	991	Q12834	Cell division cycle protein 20 homolog	–	–
DDB1	100290337	Q16531	DNA damage-binding protein 1	–	–
APOA1	335	P02647	Apolipoprotein A-I	–	–
MYD88	4615	Q99836	Myeloid differentiation primary response protein MyD88	–	–
BAG6	7917	P46379	Large proline-rich protein BAT3	–	–
PPP1CA	5499	P62136	Serine–threonine-protein phosphatase PP1-alpha catalytic subunit	–	–
AR	367	P10275	Androgen receptor	–	Pathways in cancer

Continued on next page

Table B.8 – (continued)

Gene symbol	Gene Id	UniProt	Gene name	PMIDs	Pathways
MAP1LC3B	81631	Q9GZQ8	Microtubule-associated proteins 1A–1B light chain 3B	–	–
SERPINA1	5265	P01009	Alpha-1-antitrypsin	–	–
SMAD3	4088	P84022	Mothers against decapentaplegic homolog 3	9753318, 16528675	Colorectal cancer
PRKDC	5591	P78527	DNA-dependent protein kinase catalytic subunit	–	–
SUMO2	6613	P61956	Small ubiquitin-related modifier 2	–	–
MDM2	4193	Q00987	E3 ubiquitin-protein ligase Mdm2	–	Pathways in cancer
RAC1	5879	P63000	Ras-related C3 botulinum toxin substrate 1	16551621, 18165265	Colorectal cancer
SMURF2	64750	Q9HAU4	E3 ubiquitin-protein ligase SMURF2	–	–
RND2	8153	P52198	Rho-related GTP-binding protein RhoN	–	–
XRCC6	2547	P12956	X-ray repair cross-complementing protein 6	11731412	–
TSC22D1	8848	Q15714	TSC22 domain family protein 1	–	–
LRSAM1	90678	Q6UWE0	E3 ubiquitin-protein ligase LRSAM1	–	–
SP1	6667	P08047	Transcription factor Sp1	15883203, 19593667	–
PPM1A	5494	P35813	Protein phosphatase 1A	–	–
HDAC2	3066	Q92769	Histone deacetylase 2	–	Pathways in cancer
PPP2CA	5515	B3KUN1	Serine–threonine-protein phosphatase	–	–
CASP3	836	P42574	Caspase-3	17805550, 11894121	Colorectal cancer
TOLLIP	54472	Q9H0E2	Toll-interacting protein	–	–
CTDSP1	58190	Q9GZU7	Carboxy-terminal domain RNA polymerase II polypeptide A small phosphatase 1	–	–

Continued on next page

Table B.8 – (continued)

Gene symbol	Gene Id	UniProt	Gene name	PMIDs	Pathways
DDX24	57062	Q9GZR7	ATP-dependent RNA helicase DDX24	–	–
OAZ1	4946	P54368	Ornithine decarboxylase antizyme 1	–	–
TSG101	7251	Q99816	Tumor susceptibility gene 101 protein	–	–
STX5	6811	Q13190	Syntaxin-5	–	–
RUVBL2	10856	Q9Y230	RuvB-like 2	–	–

Table B.9: Initial set of HIV seed receptors, including their role in HIV infection indicated by the receptor type and their functional domains. Receptors are grouped according to their functional domains (see Figure 7.19 for the distribution of those domains).

Receptor	Receptor type	InterPro domains
Ig-like and Other		
CD4	Primary receptor for HIV	Ag_CD4, CD4-extracel, Ig-like, Ig-like_fold, Ig_C2-set, Ig_sub, Ig_V-set_sub
7-TM GPCR and CCR_rcpt		
CCR5	Co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_5_rcpt
CCR3	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_3_rcpt
CCR2	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_2_rcpt, CC_5_rcpt
CCR8	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_8_rcpt
CCR9	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_9_rcpt
CXCR4	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CXC_4_rcpt
CXCR6	Co-receptor	7TM_GPCR_Rhodpsn, CXC_6_rcpt
CX3CR1	Co-receptor with CD4	7TM_GPCR_Rhodpsn, CX3C_fract_rcpt
7-TM GPCR and Other		
APJ	Alternative co-receptor	7TM_GPCR_Rhodpsn, APJ_rcpt
GPR1	Alternative co-receptor	7TM_GPCR_Rhodpsn, GPR1_rcpt
Integrin-α		
ITGA4	Co-receptor with CD4	Int_alpha_beta-p, Integrin_alpha, Integrin_alpha-2, Integrin_alpha_C
C-type lectin and Other		
DC-SIGN	Receptor for HIV	AntifreezeII, C-type_lectin

Table B.10: Chromosomal locations of known and predicted surface membrane factors.
Similar genomic regions are colored similarly.

Known factors	Locus	Predicted factor	Locus
CXCR4	2q21	CD2	1q13.1
GPR1	2q33.3	DARC	1q21-q22
ITGA4	2q31.3	HTR6	1p36-p35
CCR9	3p21.3	CSFR3	1p35-p34.3
CCR3	3p21.3	IL1R1	2q12
CCR2	3p21.3	GPR17	2q21
CCR5	3p21.31	CCR1	3p21
CX3CR1	3p21|3p21.3	CCBP2	3p21.3
CXCR6	3p21	RXFP1	4q32.1
CCR8	3p22	GYPB	4q28-q31
APJ	11q12	IL6ST	5q11
CD4	12pter-p12	HTR1B	6q13
DC-SIGN	19p13	HTR1E	6q14-q15
		TNFRSF3	12p13
		GPR182	12q13.3
		RXFP2	13q13.1
		CD79B	17q23
		CD97	19p13
		TNFRSF5	20q12-q13.2
		NPBWR2	20q13.3
		GP1BB	22q11.21-q11.23|22q11.21

Bibliography

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**(8), 1021–1023.

Aerts, S., Lambrechts, D., Maity, S., Loo, P. V., Coessens, B., Smet, F. D., Tranchevent, L.-C., Moor, B. D., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nat Biotechnol*, **24**(5), 537–544.

Aerts, S., Vilain, S., Hu, S., Tranchevent, L.-C., Barriot, R., Yan, J., Moreau, Y., Hassan, B. A., and Quan, X.-J. (2009). Integrating computational biology and forward genetics in Drosophila. *PLoS Genet*, **5**(1), e1000351.

Aigelsreiter, A., Janig, E., Stumptner, C., Fuchsbichler, A., Zatloukal, K., and Denk, H. (2007). How a cell deals with abnormal proteins. Pathogenetic mechanisms in protein aggregation diseases. *Pathobiology*, **74**(3), 145–158.

Albert, R. (2005). Scale-free networks in cell biology. *J Cell Sci*, **118**(Pt 21), 4947–4957.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**(3), 291–294.

Alexeyenko, A., Lindberg, J., Perez-Bercoff, A., and Sonnhammer, E. L. L. (2006). Overview and comparison of ortholog databases. *Drug Discovery Today: Technologies*, **3**(2), 137–143.

Almaas, E. (2007). Biological impacts and context of network theory. *J Exp Biol*, **210**(Pt 9), 1548–1558.

Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., and Kanaya, S. (2006). Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, **7**, 207.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402.

Ambrogelly, A., Palioura, S., and Söll, D. (2007). Natural expansion of the genetic code. *Nat Chem Biol*, **3**(1), 29–35.

Aragues, R., Jaeggi, D., and Oliva, B. (2006). PIANA: protein interactions and network analysis. *Bioinformatics*, **22**(8), 1015–1017.

Aragues, R., Sali, A., Bonet, J., Marti-Renom, M. A., and Oliva, B. (2007). Characterization of protein hubs by inferring interacting motifs from protein interactions. *PLoS Comput Biol*, **3**(9), 1761–1771.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, **38**(Database issue), D525–D531.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**(1), 25–29.

Asyali, M. H., Colak, D., Demirkaya, O., and Inan, M. S. (2006). Gene Expression Profile Classification: A Review. *Current Bioinformatics*, **1**, 55–73.

Azuaje, F., Al-Shahrour, F., and Dopazo, J. (2006). Ontology-driven approaches to analyzing data in functional genomics. *Methods Mol Biol*, **316**, 67–86.

Bader, G. D. and Hogue, C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, **20**(10), 991–997.

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

Bader, G. D., Betel, D., and Hogue, C. W. V. (2003). BIND: The Biomolecular Interaction Network Database. *Nucleic Acids Res*, **31**(1), 248–250.

Baenziger, J. U. (2003). A major step on the road to understanding a unique posttranslational modification and its role in a genetic disease. *Cell*, **113**(4), 421–422.

Baker, S. M., Plug, A. W., Prolla, T. A., Bronner, C. E., Harris, A. C., Yao, X., Christie, D. M., Monell, C., Arnheim, N., Bradley, A., Ashley, T., and Liskay, R. M. (1996). Involvement of mouse Mlh1 in DNA mismatch repair and meiotic crossing over. *Nat Genet*, **13**(3), 336–342.

Ban, C. and Yang, W. (1998). Crystal structure and ATPase activity of MutL: implications for DNA repair and mutagenesis. *Cell*, **95**(4), 541–552.

Barabasi and Albert (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5**(2), 101–113.

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.

Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, **23**(13), i41–i48.

Berg, J., Lässig, M., and Wagner, A. (2004). Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol Biol*, **4**, 51.

Berger, S. I. and Iyengar, R. (2009). Network analyses in systems pharmacology. *Bioinformatics*, **25**(19), 2466–2472.

Berman, H. M. (2008). The Protein Data Bank: a historical perspective. *Acta Crystallogr A*, **64**(Pt 1), 88–95.

Bieri, T., Blasiar, D., Ozersky, P., Antoshechkin, I., Bastiani, C., Canaran, P., Chan, J., Chen, N., Chen, W. J., Davis, P., Fiedler, T. J., Girard, L., Han, M., Harris, T. W., Kishore, R., Lee, R., McKay, S., Müller, H.-M., Nakamura, C., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E. M., Spooner, W., Tuli, M. A., Auken, K. V., Wang, D., Wang, X., Williams, G., Durbin, R., Stein, L. D., Sternberg, P. W., and Spieth, J. (2007). WormBase: new content and better access. *Nucleic Acids Res*, **35**(Database issue), D506–D510.

Birgbauer, E., Oster, S. F., Severin, C. G., and Sretavan, D. W. (2001). Retinal axon growth cones respond to ephb extracellular domains as inhibitory axon guidance cues. *Development*, **128**(15), 3041–3048.

Bittar, G. J. and Sonderegger, B. P. (2009). An Introduction to Phylogenetics and Its Molecular Aspects. In *Bioinformatics: A Swiss Perspective*, chapter 11, pages 285–328. World Scientific Pub Co.

Blow, N. (2009). Systems biology: Untangling the protein web. *Nature*, **460**(7253), 415–418.

Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, **14**(3), 292–299.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, **33 Suppl**, 228–237.

Brambilla, R. and Klein, R. (1995). Telling axons where to grow: a role for eph receptor tyrosine kinases in guidance. *Mol Cell Neurosci*, **6**(6), 487–495.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Roncari, L., de Smet, A.-S., Venkatesan, K., Rual, J.-F., Vandenhaute, J., Cusick, M. E., Pawson, T., Hill, D. E., Tavernier, J., Wrana, J. L., Roth, F. P., and Vidal, M. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*, **6**(1), 91–97.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, **30**(1-7), 107–117.

Brocchieri, L. and Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res*, **33**(10), 3390–3400.

Broder, C. C. and Collman, R. G. (1997). Chemokine receptors and HIV. *J Leukoc Biol*, **62**(1), 20–29.

Brown, D. and Sjölander, K. (2006). Functional classification using phylogenomic inference. *PLoS Comput Biol*, **2**(6), e77.

Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, **21**(9), 2076–2082.

Brueckner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci*, **10**(6), 2763–2788.

Brunner, H. G. and van Driel, M. A. (2004). From syndrome families to functional genomics. *Nat Rev Genet*, **5**(7), 545–551.

Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., Blake, J. A., and Group, M. G. D. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res*, **36**(Database issue), D724–D728.

Butland, G., Peregrín-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005). Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, **433**(7025), 531–537.

Camon, E. B., Barrell, D. G., Dimmer, E. C., Lee, V., Magrane, M., Maslen, J., Binns, D., and Apweiler, R. (2005). An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, **6 Suppl 1**, S17.

Cardon, L. R. and Bell, J. I. (2001). Association study designs for complex diseases. *Nat Rev Genet*, **2**(2), 91–99.

Ceol, A., Aryamontri, A. C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res*, **38**(Database issue), D532–D539.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Res*, **35**(Database issue), D572–D574.

Chatr-Aryamontri, A., Ceol, A., Licata, L., and Cesareni, G. (2008). Protein interactions: integration leads to belief. *Trends Biochem Sci*, **33**(6), 241–2; author reply 242–3.

Chaurasia, G., Iqbal, Y., Hänig, C., Herzel, H., Wanker, E. E., and Futschik, M. E. (2007). UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res*, **35**(Database issue), D590–D594.

Chen, F., Mackey, A. J., Vermunt, J. K., and Roos, D. S. (2007a). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, **2**(4), e383.

Chen, J., Xu, H., Aronow, B. J., and Jegga, A. G. (2007b). Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, **8**, 392.

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

Chen, J. Y., Shen, C., and Sivachenko, A. Y. (2006). Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac Symp Biocomput*, pages 367–378.

Chen, J. Y., Mamidipalli, S., and Huan, T. (2009b). HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics*, **10 Suppl 1**, S16.

Chen, X.-W. and Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**(24), 4394–4400.

Chiti, F. and Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem*, **75**, 333–366.

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, **22**(13), 1623–1630.

Chua, H. N., Sung, W.-K., and Wong, L. (2007). Using indirect protein interactions for the prediction of Gene Ontology functions. *BMC Bioinformatics*, **8 Suppl 4**, S8.

Coleman, C. M. and Wu, L. (2009). HIV interactions with monocytes and dendritic cells: viral latency and reservoirs. *Retrovirology*, **6**, 51.

Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J. (2004). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, **20**(1), 45–50.

Cook, J. A., August, A., and Henderson, A. J. (2002). Recruitment of phosphatidylinositol 3-kinase to CD28 inhibits HIV transcription by a Tat-dependent mechanism. *J Immunol*, **169**(1), 254–260.

Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Kranz, J. E., Olsen, P., Robertson, L. S., Skrzypek, M. S., Braun, B. R., Hopkins, K. L., Kondu, P., Lengieza, C., Lew-Smith, J. E., Tillberg, M., and Garrels, J. I. (2001). YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res*, **29**(1), 75–79.

Couto, F. M., Silva, M. J., and Pedro Coutinho, P. M. (2007). Measuring semantic similarity between Gene Ontology terms. *Data Knowl. Eng*, **61**(1), 137–152.

Craig, A., Sidaway, J., Holmes, E., Orton, T., Jackson, D., Rowlinson, R., Nickson, J., Tonge, R., Wilson, I., and Nicholson, J. (2006). Systems toxicology: integrated genomic, proteomic and metabonomic analysis of methapyrilene induced hepatotoxicity in the rat. *J Proteome Res*, **5**(7), 1586–1601.

Croce, C. M. (2008). Oncogenes and cancer. *N Engl J Med*, **358**(5), 502–511.

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

Cusick, M. E., Klitgord, N., Vidal, M., and Hill, D. E. (2005). Interactome: gateway into systems biology. *Hum Mol Genet*, **14 Spec No. 2**, R171–R181.

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P., and Vidal, M. (2009). Literature-curated protein interaction datasets. *Nat Methods*, **6**(1), 39–46.

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

Date, S. V. and Marcotte, E. M. (2005). Protein function prediction using the Protein Link EXplorer (PLEX). *Bioinformatics*, **21**(10), 2558–2559.

Davy, A., Bello, P., Thierry-Mieg, N., Vaglio, P., Hitti, J., Doucette-Stamm, L., Thierry-Mieg, D., Reboul, J., Boulton, S., Walhout, A. J., Coux, O., and Vidal, M. (2001). A protein-protein interaction map of the Caenorhabditis elegans 26S proteasome. *EMBO Rep*, **2**(9), 821–828.

de la Chapelle, A. (2004). Genetic predisposition to colorectal cancer. *Nat Rev Cancer*, **4**(10), 769–780.

Delgado, M. D. and León, J. (2006). Gene expression regulation and cancer. *Clin Transl Oncol*, **8**(11), 780–787.

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

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.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nat Biotechnol*, **23**(12), 1499–1501.

Dickson, B. J. (2002). Molecular mechanisms of axon guidance. *Science*, **298**(5600), 1959–1964.

Dimmer, E. C., Huntley, R. P., Barrell, D. G., Binns, D., Draghici, S., Camon, E. B., Hubank, M., Talmud, P. J., Apweiler, R., and Lovering, R. C. (2008). The Gene Ontology - Providing a Functional Role in Proteomic Studies. *Proteomics*.

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–i231.

Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, **426**(6968), 884–890.

Dolinski, K. and Botstein, D. (2007). Orthology and functional conservation in eukaryotes. *Annu Rev Genet*, **41**, 465–507.

Dong, C., Janas, A. M., Wang, J.-H., Olson, W. J., and Wu, L. (2007). Characterization of human immunodeficiency virus type 1 replication in immature and mature dendritic cells reveals dissociable cis- and trans-infection. *J Virol*, **81**(20), 11352–11362.

Dunn, S. J., Khan, I. H., Chan, U. A., Scearce, R. L., Melara, C. L., Paul, A. M., Sharma, V., Bih, F.-Y., Holzmayer, T. A., Luciw, P. A., and Abo, A. (2004). Identification of cell surface targets for HIV-1 therapeutics using genetic screens. *Virology*, **321**(2), 260–273.

Ehlers, A., Oker, E., Bentink, S., Lenze, D., Stein, H., and Hummel, M. (2008). Histone acetylation and DNA demethylation of B cells result in a Hodgkin-like phenotype. *Leukemia*, **22**(4), 835–841.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, **95**(25), 14863–14868.

Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, **405**(6788), 823–826.

Eisenberg, E. and Levanon, E. Y. (2003). Human housekeeping genes are compact. *Trends Genet*, **19**(7), 362–365.

Engelhardt, B. E., Jordan, M. I., Repo, S. T., and Brenner, S. E. (2009). Phylogenetic molecular function annotation. *J Phys*, **180**(1), 12024.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**(6757), 86–90.

Enright, A. J., Dongen, S. V., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, **30**(7), 1575–1584.

Erdeniz, N., Nguyen, M., Deschenes, S. M., and Liskay, R. M. (2007). Mutations affecting a putative MutLα endonuclease motif impact multiple mismatch repair functions. *DNA Repair (Amst)*, **6**(10), 1463–1470.

Erten, S., Bebek, G., Ewing, R. M., and Koyuturk, M. (2011). DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. *BioData Min*, **4**(1), 19.

Estrada, E. (2006). Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, **6**(1), 35–40.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*, **3**, 89.

Farber, C. R. and Lusis, A. J. (2008). Integrating global gene expression analysis and genetics. *Adv Genet*, **60**, 571–601.

Fell, D. A. and Wagner, A. (2000). The small world of metabolism. *Nat Biotechnol*, **18**(11), 1121–1122.

Ferrari, L. D. and Aitken, S. (2006). Mining housekeeping genes with a Naive Bayes classifier. *BMC Genomics*, **7**, 277.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, **340**(6230), 245–246.

Filonzi, M., Cardoso, L. C., Pimenta, M. T., Queiróz, D. B. C., Avellar, M. C. W., Porto, C. S., and Lazari, M. F. M. (2007). Relaxin family peptide receptors Rxfp1 and Rxfp2: mapping of the mRNA and protein distribution in the reproductive tract of the male rat. *Reprod Biol Endocrinol*, **5**, 29.

Finley, R. L. and Brent, R. (1994). Interaction mating reveals binary and ternary connections between Drosophila cell cycle regulators. *Proc Natl Acad Sci U S A*, **91**(26), 12980–12984.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Res*, **38**(Database issue), D211–D222.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**(6669), 806–811.

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

FlyBase Consortium (2003). The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res*, **31**(1), 172–175.

Forslund, K. and Sonnhammer, E. L. L. (2008). Predicting protein function from domain content. *Bioinformatics*, **24**(15), 1681–1687.

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), 1011–1025.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, **296**(5568), 750–752.

Freedman, D., Pisani, R., and Purves, R. (1998). *Statistics*. New York: W.W. Norton and Company, 3 edition.

Friedberg, I. (2006). Automated protein function prediction–the genomic challenge. *Brief Bioinform*, **7**(3), 225–242.

Frishman, D. (2007). Protein annotation at genomic scale: the current status. *Chem Rev*, **107**(8), 3448–3466.

Fu, W., Sanders-Beer, B. E., Katz, K. S., Maglott, D. R., Pruitt, K. D., and Ptak, R. G. (2009). Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res*, **37**(Database issue), D417–D422.

Gabaldón, T. and Huynen, M. A. (2004). Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci*, **61**(7-8), 930–944.

Galperin, M. Y. and Koonin, E. V. (2000). Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol*, **18**(6), 609–613.

Galperin, M. Y., Tatusov, R. L., and Koonin, E. V. (1999). Comparing microbila genomes: how the gene set determines the lifestyle. In *Organization of the prokaryotic genome (Robert L. Charlebois)*, chapter 6, pages 91–108. ASM Press, Washington, DC.

Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, **38**(3), 285–293.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Duempelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**(7084), 631–636.

Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet*, **29**(4), 482–486.

Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S. M., and Aburatani, H. (2005). Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, **86**(2), 127–141.

Gene Ontology Consortium (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, **38**(Database issue), D331–D335.

George, R. A., Liu, J. Y., Feng, L. L., Bryson-Richardson, R. J., Fatkin, D., and Wouters, M. A. (2006). Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res*, **34**(19), e130.

Ghosn, J., Viard, J.-P., Katlama, C., de Almeida, M., Tubiana, R., Letourneur, F., Aaron, L., Goujard, C., Salmon, D., Leruez-Ville, M., Rouzioux, C., and Chaix, M.-L. (2004). Evidence of genotypic resistance diversity of archived and circulating viral strains in blood and semen of pre-treated HIV-infected men. *AIDS*, **18**(3), 447–457.

Giallourakis, C., Henson, C., Reich, M., Xie, X., and Mootha, V. K. (2005). Disease gene discovery through integrative genomics. *Annu Rev Genomics Hum Genet*, **6**, 381–406.

Gianesello, L., Ferlin, A., Menegazzo, M., Pepe, A., and Foresta, C. (2009). RXFP1 is expressed on the sperm acrosome, and relaxin stimulates the acrosomal reaction of human spermatozoa. *Ann N Y Acad Sci*, **1160**, 192–193.

Gibson, S. L., Narayanan, L., Hegan, D. C., Buermeyer, A. B., Liskay, R. M., and Glazer, P. M. (2006). Overexpression of the DNA mismatch repair factor, PMS2, confers hypermutability and DNA damage tolerance. *Cancer Lett*, **244**(2), 195–202.

Gillberg, C. (1998). Chromosomal disorders and autism. *J Autism Dev Disord*, **28**(5), 415–425.

Glazier, A. M., Nadeau, J. H., and Aitman, T. J. (2002). Finding genes that underlie complex traits. *Science*, **298**(5602), 2345–2349.

Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., Droege, A., Lindenberg, K. S., Knoblich, M., Haenig, C., Herbst, M., Suopanki, J., Scherzinger, E., Abraham, C., Bauer, B., Hasenbank, R., Fritzsche, A., Ludewig, A. H., Büssow, K., Buessow, K., Coleman, S. H., Gutekunst, C.-A., Landwehrmeyer, B. G., Lehrach, H., and Wanker, E. E. (2004). A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell*, **15**(6), 853–865.

Goh, C.-S. and Cohen, F. E. (2002). Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol*, **324**(1), 177–192.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proc Natl Acad Sci U S A*, **104**(21), 8685–8690.

Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H., Lamb, B. T., and Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics*, **24**(15), 1743–1744.

Gong, C.-X., Liu, F., Grundke-Iqbal, I., and Iqbal, K. (2005). Post-translational modifications of tau protein in Alzheimer's disease. *J Neural Transm*, **112**(6), 813–838.

Gonzalez, G., Uribe, J. C., Tari, L., Brophy, C., and Baral, C. (2007). Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. *Pac Symp Biocomput*, pages 28–39.

Gorry, P. R., Dunfee, R. L., Mefford, M. E., Kunstman, K., Morgan, T., Moore, J. P., Mascola, J. R., Agopian, K., Holm, G. H., Mehle, A., Taylor, J., Farzan, M., Wang, H., Ellery, P., Willey, S. J., Clapham, P. R., Wolinsky, S. M., Crowe, S. M., and Gabuzda, D. (2007). Changes in the V3 region of gp120 contribute to unusually broad coreceptor usage of an HIV-1 isolate from a CCR5 Delta32 heterozygote. *Virology*, **362**(1), 163–178.

Grady, W. M. and Carethers, J. M. (2008). Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology*, **135**(4), 1079–1099.

Gregersen, N. (2006). Protein misfolding disorders: pathogenesis and intervention. *J Inherit Metab Dis*, **29**(2-3), 456–470.

Grigoriev, A. (2003). On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res*, **31**(14), 4157–4161.

Grothey, A., Sargent, D., Goldberg, R. M., and Schmoll, H.-J. (2004). Survival of patients with advanced colorectal cancer improves with the availability of fluorouracil-leucovorin, irinotecan, and oxaliplatin in the course of treatment. *J Clin Oncol*, **22**(7), 1209–1214.

Guarné, A., Junop, M. S., and Yang, W. (2001). Structure and function of the N-terminal 40 kDa fragment of human PMS2: a monomeric GHL ATPase. *EMBO J*, **20**(19), 5521–5531.

Habraken, Y., Sung, P., Prakash, L., and Prakash, S. (1997). Enhancement of MSH2-MSH3-mediated mismatch recognition by the yeast MLH1-PMS1 complex. *Curr Biol*, **7**(10), 790–793.

Hadjkacem, B., Elleuch, H., Gargouri, J., and Gargouri, A. (2009). Bernard-Soulier syndrome: novel nonsense mutation in GPIbbeta gene affecting GPIb-IX complex expression. *Ann Hematol*, **88**(5), 465–472.

Hakes, L., Robertson, D. L., Oliver, S. G., and Lovell, S. C. (2007). Protein interactions from complexes: a structural perspective. *Comp Funct Genomics*, page 49356.

Half, E., Bercovich, D., and Rozen, P. (2009). Familial adenomatous polyposis. *Orphanet J Rare Dis*, **4**, 22.

Hall, M. C., Shcherbakova, P. V., and Kunkel, T. A. (2002). Differential ATP binding and intrinsic ATP hydrolysis by amino-terminal domains of the yeast Mlh1 and Pms1 proteins. *J Biol Chem*, **277**(5), 3673–3679.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P., and Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**(6995), 88–93.

Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. (2002). Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18 Suppl 1**, S145–S154.

Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol*, **7**(11), 120.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, **402**(6761 Suppl), C47–C52.

Hayete, B. and Bienkowska, J. R. (2005). Gotrees: predicting go associations from protein domain composition using decision trees. *Pac Symp Biocomput*, pages 127–138.

He, W., Neil, S., Kulkarni, H., Wright, E., Agan, B. K., Marconi, V. C., Dolan, M. J., Weiss, R. A., and Ahuja, S. K. (2008). Duffy antigen receptor for chemokines mediates trans-infection of HIV-1 from red blood cells to target cells and affects HIV-AIDS susceptibility. *Cell Host Microbe*, **4**(1), 52–62.

He, X. and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genet*, **2**(6), e88.

Hene, L., Sreenu, V. B., Vuong, M. T., Abidi, S. H. I., Sutton, J. K., Rowland-Jones, S. L., Davis, S. J., and Evans, E. J. (2007). Deep analysis of cellular transcriptomes - LongSAGE versus classic MPSS. *BMC Genomics*, **8**, 333.

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., and Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**(5338), 609–614.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004a). IntAct: an open source molecular interaction database. *Nucleic Acids Res*, **32**(Database issue), D452–D455.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004b). The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, **22**(2), 177–183.

Himanen, J.-P. and Nikolov, D. B. (2003). Eph receptors and ephrins. *Int J Biochem Cell Biol*, **35**(2), 130–134.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, **18**(6), 523–531.

Hollstein, M., Sidransky, D., Vogelstein, B., and Harris, C. C. (1991). p53 mutations in human cancers. *Science*, **253**(5015), 49–53.

Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R. S., Oughtred, R., Skrzypek, M. S., Weng, S., Wong, E. D., Zhu, K. K., Dolinski, K., Botstein, D., and Cherry, J. M. (2008). Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res*, **36**(Database issue), D577–D581.

Hongxing, Z., Nancai, Y., Wen, S., Guofu, H., Yanxia, W., Hanju, H., Qian, L., Wei, M., Yandong, Y., and Hao, H. (2008). Depletion of c-Myc inhibits human colon cancer colo 320 cells' growth. *Cancer Biother Radiopharm*, **23**(2), 229–237.

Hsiao, L. L., Dangond, F., Yoshida, T., Hong, R., Jensen, R. V., Misra, J., Dillon, W., Lee, K. F., Clark, K. E., Haverty, P., Weng, Z., Mutter, G. L., Frosch, M. P., MacDonald, M. E., Milford, E. L., Crum, C. P., Bueno, R., Pratt, R. E., Mahadevappa, M., Warrington, J. A., Stephanopoulos, G., Stephanopoulos, G., and Gullans, S. R. (2001). A compendium of gene expression in normal human tissues. *Physiol Genomics*, **7**(2), 97–104.

Hsieh, P. and Yamane, K. (2008). DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mech Ageing Dev*, **129**(7-8), 391–407.

Huang, T.-W., Tien, A.-C., Huang, W.-S., Lee, Y.-C. G., Peng, C.-L., Tseng, H.-H., Kao, C.-Y., and Huang, C.-Y. F. (2004). POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, **20**(17), 3273–3276.

Huberts, D. H. E. W. and van der Klei, I. J. (2010). Moonlighting proteins: an intriguing mode of multitasking. *Biochim Biophys Acta*, **1803**(4), 520–525.

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nat Rev Genet*, **6**(4), 287–298.

Hunter, L. and Cohen, K. B. (2006). Biomedical language processing: what's beyond PubMed? *Mol Cell*, **21**(5), 589–594.

Huot, J. (2004). Ephrin signaling in axon guidance. *Prog Neuropsychopharmacol Biol Psychiatry*, **28**(5), 813–818.

Huynen, M. A., Snel, B., von Mering, C., and Bork, P. (2003). Function prediction and protein networks. *Curr Opin Cell Biol*, **15**(2), 191–198.

ichi Yoshioka, K., Yoshioka, Y., and Hsieh, P. (2006). ATR kinase activation mediated by MutSalpha and MutLalpha in response to cytotoxic O6-methylguanine adducts. *Mol Cell*, **22**(4), 501–510.

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

Ikegaki, N., Tang, X. X., Liu, X. G., Biegel, J. A., Allen, C., Yoshioka, A., Sulman, E. P., Brodeur, G. M., and Pleasure, D. E. (1995). Molecular characterization and chromosomal localization of DRT (EPHT3): a developmentally regulated human protein-tyrosine kinase gene of the EPH family. *Hum Mol Genet*, **4**(11), 2033–2045.

Irby, R. B. and Yeatman, T. J. (2000). Role of Src expression and activation in human cancer. *Oncogene*, **19**(49), 5636–5642.

Irvine, G. B., El-Agnaf, O. M., Shankar, G. M., and Walsh, D. M. (2008). Protein aggregation in the brain: the molecular basis for Alzheimer's and Parkinson's diseases. *Mol Med*, **14**(7-8), 451–464.

Isserlin, R., El-Badrawi, R. A., and Bader, G. D. (2011). The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database (Oxford)*, **2011**, baq037.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, **98**(8), 4569–4574.

Jaeger, S. and Leser, U. (2007). High-Precision Function Prediction using Conserved Interactions. In C. Falter, A. Schliep, J. Selbig, M. Vingron, and D. Walther, editors, *Proceedings of the German Conference on Bioinformatics, GCB 2007, September 26-28, 2007, Potsdam, Germany*, volume 115 of *LNI*, pages 146–162. GI.

Jaeger, S., Sers, C. T., and Leser, U. (2010a). Combining modularity, conservation, and interactions of proteins significantly increases precision and coverage of protein function prediction. *BMC Genomics*, **11**(1), 717.

Jaeger, S., Ertaylan, G., van Dijk, D., Leser, U., and Sloot, P. (2010b). Inference of surface membrane factors of HIV-1 infection through functional interaction networks. *PLoS One*, **5**(10), e13139.

Jaikaran, E. T. and Clark, A. (2001). Islet amyloid and type 2 diabetes: from molecular misfolding to islet pathophysiology. *Biochim Biophys Acta*, **1537**(3), 179–203.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res*, **12**(1), 37–46.

Jensen, L. J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H. H., Rapacki, K., Workman, C., Andersen, C. A. F., Knudsen, S., Krogh, A., Valencia, A., and Brunak, S. (2002). Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol*, **319**(5), 1257–1265.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, **37**(Database issue), D412–D416.

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**(6833), 41–42.

Jiricny, J. (2000). Mediating mismatch repair. *Nat Genet*, **24**(1), 6–8.

Jiricny, J. (2006). MutLalpha: at the cutting edge of mismatch repair. *Cell*, **126**(2), 239–241.

Jones, S. and Thornton, J. M. (2004). Searching for functional sites in protein structures. *Curr Opin Chem Biol*, **8**(1), 3–7.

Jonsson, P. F. and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, **22**(18), 2291–2297.

Jorde, L. B. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res*, **10**(10), 1435–1444.

Jothi, R., Cherukuri, P. F., Tasneem, A., and Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J Mol Biol*, **362**(4), 861–875.

Junker, B. H., Koschützki, D., and Schreiber, F. (2006). Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, **7**, 219.

Kalaev, M., Smoot, M., Ideker, T., and Sharan, R. (2008). NetworkBLAST: comparative analysis of protein networks. *Bioinformatics*, **24**(4), 594–596.

Kamath, R. S. and Ahringer, J. (2003). Genome-wide RNAi screening in Caenorhabditis elegans. *Methods*, **30**(4), 313–321.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, **38**(Database issue), D355–D360.

Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R., and Kasif, S. (2004). Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A*, **101**(9), 2888–2893.

Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R., and Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A*, **100**(20), 11394–11399.

Kemmeren, P., van Berkum, N. L., Vilo, J., Bijma, T., Donders, R., Brazma, A., and Holstege, F. C. P. (2002). Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell*, **9**(5), 1133–1143.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**(4610), 662–666.

Kennedy, R. D. and D'Andrea, A. D. (2005). The Fanconi Anemia/BRCA pathway: new faces in the crowd. *Genes Dev*, **19**(24), 2925–2940.

Kentner, D. and Sourjik, V. (2009). Dynamic map of protein interactions in the Escherichia coli chemotaxis pathway. *Mol Syst Biol*, **5**, 238.

Kern, A. and Bryant-Greenwood, G. D. (2009). Mechanisms of relaxin receptor (LGR7/RXFP1) expression and function. *Ann N Y Acad Sci*, **1160**, 60–66.

Keseler, I. M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A. G., and Karp, P. D. (2009). EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res*, **37**(Database issue), D464–D470.

Keskin, O., Nussinov, R., and Gursoy, A. (2008). PRISM: protein-protein interaction prediction by structural matching. *Methods Mol Biol*, **484**, 505–521.

Khanin, R. and Wit, E. (2006). How scale-free are biological networks. *J Comput Biol*, **13**(3), 810–818.

King, A. D., Przulj, N., and Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics*, **20**(17), 3013–3020.

Kline, C. L. B., Olson, T. L., and Irby, R. B. (2009). Src activity alters alpha3 integrin expression in colon tumor cells. *Clin Exp Metastasis*, **26**(2), 77–87.

Klingström, T. and Plewczynski, D. (2010). Protein-protein interaction and pathway databases, a graphical review. *Brief Bioinform*.

Koegl, M. and Uetz, P. (2007). Improving yeast two-hybrid screening systems. *Brief Funct Genomic Proteomic*, **6**(4), 302–312.

Kolodny, R., Koehl, P., and Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol*, **346**(4), 1173–1188.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, **39**, 309–338.

Koschützki, D. and Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Bio*, **2**, 193–201.

Koyutürk, M., Grama, A., and Szpankowski, W. (2004). An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, **20 Suppl 1**, i200–i207.

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

Krueger, K. E. and Srivastava, S. (2006). Posttranslational protein modifications: current implications for cancer detection, prevention, and therapeutics. *Mol Cell Proteomics*, **5**(10), 1799–1810.

Kumar, S. R., Scehnet, J. S., Ley, E. J., Singh, J., Krasnoperov, V., Liu, R., Manchanda, P. K., Ladner, R. D., Hawes, D., Weaver, F. A., Beart, R. W., Singh, G., Nguyen, C., Kahn, M., and Gill, P. S. (2009). Preferential induction of EphB4 over EphB2 and its implication in colorectal cancer progression. *Cancer Res*, **69**(9), 3736–3745.

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), 949–958.

Kühn, R., Schwenk, F., Aguet, M., and Rajewsky, K. (1995). Inducible gene targeting in mice. *Science*, **269**(5229), 1427–1429.

Küppers, R. (2009). The biology of Hodgkin's lymphoma. *Nat Rev Cancer*, **9**(1), 15–27.

Küppers, R., Rajewsky, K., Zhao, M., Simons, G., Laumann, R., Fischer, R., and Hansmann, M. L. (1994). Hodgkin disease: Hodgkin and Reed-Sternberg cells picked from histological sections show clonal immunoglobulin gene rearrangements and appear to be derived from B cells at various stages of development. *Proc Natl Acad Sci U S A*, **91**(23), 10962–10966.

Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, **25**(3), 309–316.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*, **8**(12), 995–1005.

Lehne, B. and Schlitt, T. (2009). Protein-protein interaction databases: keeping up with growing interactomes. *Hum Genomics*, **3**(3), 291–297.

Li, C., Norris, P. S., Ni, C.-Z., Havert, M. L., Chiong, E. M., Tran, B. R., Cabezas, E., Reed, J. C., Satterthwait, A. C., Ware, C. F., and Ely, K. R. (2003a). Structurally distinct recognition motifs in lymphotoxin-beta receptor and CD40 for tumor necrosis factor receptor-associated factor (TRAF)-mediated signaling. *J Biol Chem*, **278**(50), 50523–50529.

Li, G.-M. (2008). Mechanisms and functions of DNA mismatch repair. *Cell Res*, **18**(1), 85–98.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003b). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, **13**(9), 2178–2189.

Li, S., Iakoucheva, L. M., Mooney, S. D., and Radivojac, P. (2010). Loss of post-translational modification sites in disease. *Pac Symp Biocomput*, pages 337–347.

Li, X.-L., Tan, Y.-C., and Ng, S.-K. (2006). Systematic gene function prediction from gene expression data by using a fuzzy nearest-cluster method. *BMC Bioinformatics*, **7 Suppl 4**, S23.

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), 1482–1493.

Lin, C.-C., Hsiang, J.-T., Wu, C.-Y., Oyang, Y.-J., Juan, H.-F., and Huang, H.-C. (2010). Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy. *BMC Syst Biol*, **4**, 138.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th ICML*, pages 296–304, Madison WI.

Lin, Y. and Wilson, J. H. (2009). Diverse effects of individual mismatch repair components on transcription-induced CAG repeat instability in human cells. *DNA Repair (Amst)*, **8**(8), 878–885.

Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R., and Tatusova, T. A. (2002). The relationship of protein conservation and sequence length. *BMC Evol Biol*, **2**, 20.

Liu, Y., Kim, I., and Zhao, H. (2008). Protein interaction predictions from diverse sources. *Drug Discov Today*, **13**(9-10), 409–416.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, **14**(13), 1675–1680.

Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A., Ploegh, H., and Matsudaira, P. (2007). *Molecular Cell Biology*. W. H. Freeman, 6 edition.

Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J., and Tramontano, A. (2009). Protein function annotation by homology-based inference. *Genome Biol*, **10**(2), 207.

Lubin, D. J., Butler, J. S., and Loh, S. N. (2010). Folding of tetrameric p53: oligomerization and tumorigenic mutations induce misfolding and loss of function. *J Mol Biol*, **395**(4), 705–716.

Lynch, H. T. and Smyrk, T. (1996). Hereditary nonpolyposis colorectal cancer (Lynch syndrome). An updated review. *Cancer*, **78**(6), 1149–1167.

Lynch, M. (2007). The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet*, **8**(10), 803–813.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**(5494), 1151–1155.

Mackay, J. P., Sunde, M., Lowry, J. A., Crossley, M., and Matthews, J. M. (2007). Protein interactions: is seeing believing? *Trends Biochem Sci*, **32**(12), 530–531.

Mackenbach, J. P. (2006). The origins of human disease: a short story on "where diseases come from". *J Epidemiol Community Health*, **60**(1), 81–86.

Maho, A., Bensimon, A., Vassart, G., and Parmentier, M. (1999). Mapping of the CCXCR1, CX3CR1, CCBP2 and CCR9 genes to the CCR cluster within the 3p21.3 region of the human genome. *Cytogenet Cell Genet*, **87**(3-4), 265–268.

Maitra, A., Molberg, K., Albores-Saavedra, J., and Lindberg, G. (2000). Loss of Dpc4 expression in colonic adenocarcinomas correlates with the presence of metastatic disease. *Am J Pathol*, **157**(4), 1105–1111.

Makino, T. and Gojobori, T. (2007). Evolution of protein-protein interaction network. *Genome Dyn*, **3**, 13–29.

Malek, R. L., Irby, R. B., Guo, Q. M., Lee, K., Wong, S., He, M., Tsai, J., Frank, B., Liu, E. T., Quackenbush, J., Jove, R., Yeatman, T. J., and Lee, N. H. (2002). Identification of Src transformation fingerprint in human colon cancer. *Oncogene*, **21**(47), 7256–7265.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, **39**(7), 906–913.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999a). A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**(6757), 83–86.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999b). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**(5428), 751–753.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet*, **24**(3), 133–141.

Massagué, J. and Chen, Y. G. (2000). Controlling TGF-beta signaling. *Genes Dev*, **14**(6), 627–644.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*, **11**(12), 2120–2126.

Mayes, A. E., Verdone, L., Legrain, P., and Beggs, J. D. (1999). Characterization of Sm-like proteins in yeast and their association with U6 snRNA. *EMBO J*, **18**(15), 4321–4331.

McCarroll, S. A. and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nat Genet*, **39**(7 Suppl), S37–S42.

McCarthy, M. I. (2011). The importance of global studies of the genetics of type 2 diabetes. *Diabetes Metab J*, **35**(2), 91–100.

McKusick, V. A. (2007). Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet*, **80**(4), 588–604.

Mendel, G. (1866). *Versuche über Pflanzenhybriden*. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, Abhandlungen, 3–47.

Merlini, G. and Bellotti, V. (2003). Molecular mechanisms of amyloidosis. *N Engl J Med*, **349**(6), 583–596.

Merskey, H. (1986). Variable meanings for the definition of disease. *J Med Philos*, **11**(3), 215–232.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, **11**(1), 31–46.

Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., and Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, **30**(1), 31–34.

Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A*, **102**(31), 10930–10935.

Miozzi, L., Piro, R. M., Rosa, F., Ala, U., Silengo, L., Cunto, F. D., and Provero, P. (2008). Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data. *PLoS One*, **3**(6), e2439.

Miyaki, M., Iijima, T., Konishi, M., Sakai, K., Ishii, A., Yasuno, M., Hishima, T., Koike, M., Shitara, N., Iwama, T., Utsunomiya, J., Kuroki, T., and Mori, T. (1999). Higher frequency of Smad4 gene mutation in human colorectal cancer with distant metastasis. *Oncogene*, **18**(20), 3098–3103.

Morgan, T. H. (1910). SEX LIMITED INHERITANCE IN DROSOPHILA. *Science*, **32**(812), 120–122.

Murali, T. M., Wu, C.-J., and Kasif, S. (2006). The art of gene function prediction. *Nat Biotechnol*, **24**(12), 1474–5; author reply 1475–6.

Myers, C. L., Barrett, D. R., Hibbs, M. A., Huttenhower, C., and Troyanskaya, O. G. (2006). Finding function: evaluation methods for functional genomic data. *BMC Genomics*, **7**, 187.

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–i310.

Navlakha, S. and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, **26**(8), 1057–1063.

Neil, S. J. D., Aasa-Chapman, M. M. I., Clapham, P. R., Nibbs, R. J., McKnight, A., and Weiss, R. A. (2005). The promiscuous CC chemokine receptor D6 is a functional coreceptor for primary isolates of human immunodeficiency virus type 1 (HIV-1) and HIV-2 on astrocytes. *J Virol*, **79**(15), 9618–9624.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, **103**(23), 8577–8582.

Nigro, J. M., Baker, S. J., Preisinger, A. C., Jessup, J. M., Hostetter, R., Cleary, K., Bigner, S. H., Davidson, N., Baylin, S., and Devilee, P. (1989). Mutations in the p53 gene occur in diverse human tumour types. *Nature*, **342**(6250), 705–708.

Nilsson, J. A. and Cleveland, J. L. (2003). Myc pathways provoking cell suicide and cancer. *Oncogene*, **22**(56), 9007–9021.

Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (1992). *Enzyme Nomenclature*. Academic Press. ISBN: 0–122–27164–5.

Nussinov, R. and Tsai, C.-J. (2005). Protein–protein interactions: principles and predictions. *Physical Biology*, **2**(2).

Ofran, Y. and Rost, B. (2007). Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol*, **3**(7), e119.

Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M. (2000). A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*, **28**(20), 4021–4028.

Oliveros, J. (2007). VENNY. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html.

Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stümpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Woollard, P., Salama, J. J., Moore, S., Wojcik, J., Bader, G. D., Vidal, M., Cusick, M. E., Gerstein, M., Gavin, A.-C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., Rivas, J. D. L., Prieto, C., Perreau, V. M., Hogue, C., Mewes, H.-W., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, G., and Hermjakob, H. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat Biotechnol*, **25**(8), 894–898.

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), 691–698.

Ouzounis, C. A., Coulson, R. M. R., Enright, A. J., Kunin, V., and Pereira-Leal, J. B. (2003). Classification schemes for protein structure and function. *Nat Rev Genet*, **4**(7), 508–519.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, **96**(6), 2896–2901.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.-W., Ruepp, A., and Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**(6), 832–834.

Pandey, G., Kumar, V., and Steinbach, M. (2006). Computational Approaches for Protein Function Prediction: A Survey. Technical Report 06-028, Department of Computer Science and Engineering, University of Minnesota, Twin Cities.

Pandey, J., Koyutürk, M., and Grama, A. (2010). Functional characterization and topological modularity of molecular interaction networks. *BMC Bioinformatics*, **11 Suppl 1**, S35.

Pasquale, E. B. (2005). Eph receptor signalling casts a wide net on cell behaviour. *Nat Rev Mol Cell Biol*, **6**(6), 462–475.

Passam, A. M., Sourvinos, G., Krambovitis, E., Miyakis, S., Stavrianeas, N., Zagoreos, I., and Spandidos, D. A. (2007). Polymorphisms of Cx(3)CR1 and CXCR6 receptors in relation to HAART therapy of HIV type 1 patients. *AIDS Res Hum Retroviruses*, **23**(8), 1026–1032.

Pastor-Satorras, R., Smith, E., and Solé, R. V. (2003). Evolving protein interaction networks through gene duplication. *J Theor Biol*, **222**(2), 199–210.

Patel, K. J. and Joenje, H. (2007). Fanconi anemia and DNA replication repair. *DNA Repair (Amst)*, **6**(7), 885–890.

Patil, A., Nakai, K., and Nakamura, H. (2011). HitPredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res*, **39**(Database issue), D744–D749.

Pawson, T. (2002). Regulation and targets of receptor tyrosine kinases. *Eur J Cancer*, **38 Suppl 5**, S3–10.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**(8), 4285–4288.

Peltonen, L. and McKusick, V. A. (2001). Genomics and medicine. dissecting human disease in the postgenomic era. *Science*, **291**(5507), 1224–1229.

Pennacchio, L. A. (2003). Insights from human/mouse genome comparisons. *Mamm Genome*, **14**(7), 429–436.

Perdew, G. H., Heuvel, J. P. V., and Peters, J. M. (2006). *Regulation of Gene Expression: Molecular Mechanisms*. Humana Pr, 1 edition.

Pereira-Leal, J. B., Enright, A. J., and Ouzounis, C. A. (2004). Detection of functional modules from protein interaction networks. *Proteins*, **54**(1), 49–57.

Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nat Genet*, **31**(3), 316–319.

Perez-Iratxeta, C., Wjst, M., Bork, P., and Andrade, M. A. (2005). G2D: a tool for mining genes associated with disease. *BMC Genet*, **6**, 45.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G. N., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, **13**(10), 2363–2371.

Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). Transient protein-protein interactions: structural, functional, and network properties. *Structure*, **18**(10), 1233–1243.

Perrett, D. (2007). From 'protein' to the beginnings of clincial proteomics. *PROTEOMICS CLINICAL APPLICATIONS*, **1**, 720–783.

Perutz, M. F. (1960). Structure of hemoglobin. *Brookhaven Symp Biol*, **13**, 165–183.

Phizicky, E. M. and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, **59**(1), 94–123.

Piehler, J. (2005). New methodologies for measuring protein interactions in vivo and in vitro. *Curr Opin Struct Biol*, **15**(1), 4–14.

Piguet, V. and Trono, D. (1999). The Nef protein of primate lentiviruses. *Rev Med Virol*, **9**(2), 111–120.

Polacco, B. J. and Babbitt, P. C. (2006). Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*, **22**(6), 723–730.

Pollex, B. (2011). *Scoring Protein Function Prediction*. Diploma thesis, Humboldt Universität zu Berlin, Germany.

Porollo, A. A., Adamczak, R., and Meller, J. (2004). POLYVIEW: a flexible visualization tool for structural and functional annotations of proteins. *Bioinformatics*, **20**(15), 2460–2462.

Porter, I. H. (1982). Control of hereditary disorders. *Ann Rev. Public Health*, **3**, 277–319.

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

Prieto, C. and Rivas, J. D. L. (2006). APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res*, **34**(Web Server issue), W298–W302.

Przytycka, T. M., Singh, M., and Slonim, D. K. (2010). Toward the dynamic interactome: it's about time. *Brief Bioinform*, **11**(1), 15–29.

Pujol, A., Mosca, R., Farrés, J., and Aloy, P. (2010). Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol Sci*, **31**(3), 115–123.

Punta, M. and Ofran, Y. (2008). The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput Biol*, **4**(10), e1000160.

Quackenbush, J. (2006). Computational approaches to analysis of DNA microarray data. *Yearb Med Inform*, pages 91–103.

Que, Q. Q. and Winzeler, E. A. (2002). Large-scale mutagenesis and functional genomics in yeast. *Funct Integr Genomics*, **2**(4-5), 193–198.

Radivojac, P., Baenziger, P. H., Kann, M. G., Mort, M. E., Hahn, M. W., and Mooney, S. D. (2008). Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*, **24**(16), i241–i247.

Ramirez, 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. and Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, **67**(2 Pt 2), 026112.

Read, A. P. and Strachan, T. (2003). *Human Molecular Genetics 3*. Taylor & Francis, 3 edition.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskya, O. G., Ideker, T., Dolinski, K., Batada, N. N., and Tyers, M. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol*, **5**(4), 11.

Reis, C. A., Osorio, H., Silva, L., Gomes, C., and David, L. (2010). Alterations in glycosylation as biomarkers for cancer detection. *J Clin Pathol*, **63**(4), 322–329.

Rhee, S. Y., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nat Rev Genet*, **9**(7), 509–515.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, **17**(10), 1030–1032.

Rison, S. C., Hodgman, T. C., and Thornton, J. M. (2000). Comparison of functional annotation schemes for genomes. *Funct Integr Genomics*, **1**(1), 56–69.

Rivas, J. D. L. and Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, **6**(6), e1000807.

Roizen, N. J. and Patterson, D. (2003). Down's syndrome. *Lancet*, **361**(9365), 1281–1289.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O., and Ofran, Y. (2003). Automatic prediction of protein function. *Cell Mol Life Sci*, **60**(12), 2637–2650.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**(7062), 1173–1178.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M., and Mewes, H. W. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*, **32**(18), 5539–5545.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res*, **38**(Database issue), D497–D501.

Saeed, R. and Deane, C. (2008). An assessment of the uses of homologous interactions. *Bioinformatics*, **24**(5), 689–695.

Salghetti, S. E., Kim, S. Y., and Tansey, W. P. (1999). Destruction of Myc by ubiquitin-mediated proteolysis: cancer-associated and transforming mutations stabilize Myc. *EMBO J*, **18**(3), 717–726.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, **32**(Database issue), D449–D451.

Sameer, A. S., Abdullah, S., Banday, M. Z., Syeed, N., and Siddiqi, M. A. (2010). Colorectal cancer, TGF-β signaling and SMADs. *Int J Genet Mol Biol*, **2**(6), 101–11.

Sanger, F. and Tuppy, H. (1951a). The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem J*, **49**(4), 481–490.

Sanger, F. and Tuppy, H. (1951b). The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J*, **49**(4), 463–481.

Schaefer, A. M., Taylor, R. W., Turnbull, D. M., and Chinnery, P. F. (2004). The epidemiology of mitochondrial disorders–past, present and future. *Biochim Biophys Acta*, **1659**(2-3), 115–120.

Schapira, A. H. V. (2006). Mitochondrial disease. *Lancet*, **368**(9529), 70–82.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**(5235), 467–470.

Schlessinger, J. (2000). Cell signaling by receptor tyrosine kinases. *Cell*, **103**(2), 211–225.

Schlessinger, J. and Lemmon, M. A. (2003). SH2 and PTB domains in tyrosine kinase signaling. *Sci STKE*, **2003**(191), RE12.

Schloss, P. D. and Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol*, **6**(8), 229.

Schug, J., Diskin, S., Mazzarelli, J., Brunk, B. P., and Stoeckert, C. J. (2002). Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res*, **12**(4), 648–655.

Schuster-Böckler, B. and Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biol*, **9**(1), R9.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol*, **18**(12), 1257–1261.

Seitz, V., Thomas, P. E., Zimmermann, K., Paul, U., Ehlers, A., Joosten, M., Dimitrova, L., Lenze, D., Sommerfeld, A., Oker, E., Leser, U., Stein, H., and Hummel, M. (2011). Classical Hodgkin's lymphoma shows epigenetic features of abortive plasma cell differentiation. *Haematologica*, **96**(6), 863–870.

Selkoe, D. J. (2003). Folding proteins in fatal ways. *Nature*, **426**(6968), 900–904.

Seo, J. and Lee, K.-J. (2004). Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. *J Biochem Mol Biol*, **37**(1), 35–44.

Shah, K. B., Inoue, Y., and Mehra, M. R. (2006). Amyloidosis and the heart: a comprehensive review. *Arch Intern Med*, **166**(17), 1805–1813.

Shapiro, L. and Harris, T. (2000). Finding function through structural genomics. *Curr Opin Biotechnol*, **11**(1), 31–35.

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M., and Ideker, T. (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.

Shcherbakova, P. V., Hall, M. C., Lewis, M. S., Bennett, S. E., Martin, K. J., Bushel, P. R., Afshari, C. A., and Kunkel, T. A. (2001). Inactivation of DNA mismatch repair by increased expression of yeast MLH1. *Mol Cell Biol*, **21**(3), 940–951.

Shen, A., Yang, H.-C., Zhou, Y., Chase, A. J., Boyer, J. D., Zhang, H., Margolick, J. B., Zink, M. C., Clements, J. E., and Siliciano, R. F. (2007). Novel pathway for induction of latent virus from resting CD4(+) T cells in the simian immunodeficiency virus/macaque model of human immunodeficiency virus type 1 latency. *J Virol*, **81**(4), 1660–1670.

Shimizu, N., Tanaka, A., Oue, A., Mori, T., Ohtsuki, T., Apichartpiyakul, C., Uchiumi, H., Nojima, Y., and Hoshino, H. (2009). Broad usage spectrum of G protein-coupled receptors as coreceptors by primary isolates of HIV. *AIDS*, **23**(7), 761–769.

Shoemaker, B. A. and Panchenko, A. R. (2007a). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol*, **3**(3), e42.

Shoemaker, B. A. and Panchenko, A. R. (2007b). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, **3**(4), e43.

Sikić, M., Tomić, S., and Vlahovicek, K. (2009). Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. *PLoS Comput Biol*, **5**(1), e1000278.

Singh, P., Kaur, G., Sharma, G., and Mehra, N. K. (2008). Immunogenetic basis of HIV-1 infection, transmission and disease progression. *Vaccine*, **26**(24), 2966–2980.

Sjölander, K. (2004). Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**(2), 170–179.

Skolnick, J., Fetrow, J. S., and Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. *Nat Biotechnol*, **18**(3), 283–287.

Sleator, R. D. and Walsh, P. (2010). An overview of in silico protein function prediction. *Arch Microbiol*, **192**(3), 151–155.

Song, J. and Singh, M. (2009). How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics*, **25**(23), 3143–3150.

Sonnhammer, E. L. L. and Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*, **18**(12), 619–620.

Spirin, V. and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, **100**(21), 12123–12128.

Sprinzak, E., Sattath, S., and Margalit, H. (2003). How reliable are experimental protein-protein interaction data? *J Mol Biol*, **327**(5), 919–923.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, **34**(Database issue), D535–D539.

Stein, L. (2001). Genome annotation: from sequence to biology. *Nat Rev Genet*, **2**(7), 493–503.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoez, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**(6), 957–968.

Stone, J. E. and Petes, T. D. (2006). Analysis of the proteins involved in the in vivo repair of base-base mismatches and four-base loops formed during meiotic recombination in the yeast saccharomyces cerevisiae. *Genetics*, **173**(3), 1223–1239.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E., and Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**(5813), 848–853.

Stumpf, M. P. H., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proc Natl Acad Sci U S A*, **105**(19), 6959–6964.

Stumvoll, M., Goldstein, B. J., and van Haeften, T. W. (2005). Type 2 diabetes: principles of pathogenesis and therapy. *Lancet*, **365**(9467), 1333–1346.

Sturz, A., Bader, B., Thierauch, K. H., and Glienke, J. (2004). EphB4 signaling is capable of mediating ephrinB2-induced inhibition of cell migration. *Biochem Biophys Res Commun*, **313**(1), 80–88.

Sumner, J. B. (1926). The isolation and crystallization of the enzyme urease. *The Journal of Biological Chemistry*, **69**, 435–441.

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.

Takahashi, T., Sano, B., Nagata, T., Kato, H., Sugiyama, Y., Kunieda, K., Kimura, M., Okano, Y., and Saji, S. (2003). Polo-like kinase 1 (PLK1) is overexpressed in primary colorectal cancers. *Cancer Sci*, **94**(2), 148–152.

Taketo, M. M. (2006). Mouse models of gastrointestinal tumors. *Cancer Sci*, **97**(5), 355–361.

Tanaka, M., Ohashi, R., Nakamura, R., Shinmura, K., Kamo, T., Sakai, R., and Sugimura, H. (2004). Tiam1 mediates neurite outgrowth induced by ephrin-B1 and EphA2. *EMBO J*, **23**(5), 1075–1088.

Tanford, C. and Reynolds, J. (2001). *Nature's Robots: A History of Proteins*. Oxford University Press, USA, 1 edition.

Tao, Y., Sam, L., Li, J., Friedman, C., and Lussier, Y. A. (2007). Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, **23**(13), i529–i538.

Tiffin, N., Andrade-Navarro, M. A., and Perez-Iratxeta, C. (2009). Linking genes to diseases: it's all in the data. *Genome Med*, **1**(8), 77.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson,

R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M., and Venter, J. C. (1997). The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature*, **388**(6642), 539–547.

Tranchevent, L. A.-C., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2010). A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*.

Tranchevent, L.-C., Barriot, R., Yu, S., Vooren, S. V., Loo, P. V., Coessens, B., Moor, B. D., Aerts, S., and Moreau, Y. (2008). ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res*, **36**(Web Server issue), W377–W384.

Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T., and Sun, F. (2006). Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, **7**, 31.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., and Wodak, S. J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database (Oxford)*, **2010**, baq026.

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.

Twigger, S. N., Shimoyama, M., Bromberg, S., Kwitek, A. E., Jacob, H. J., and Team, R. G. D. (2007). The Rat Genome Database, update 2007–easing the path from disease to data and back again. *Nucleic Acids Res*, **35**(Database issue), D658–D662.

Tyson, J. J., Chen, K., and Novak, B. (2001). Network dynamics and cell physiology. *Nat Rev Mol Cell Biol*, **2**(12), 908–916.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**(6770), 623–627.

UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, **38**(Database issue), D142–D148.

Valencia, A. (2005). Automatic annotation of protein function. *Curr Opin Struct Biol*, **15**(3), 267–274.

Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, **12**(3), 368–373.

van Dijk, D., Ertaylan, G., Boucher, C. A., and Sloot, P. M. (2010). Identifying potential survival strategies of HIV-1 through virus-host protein interaction networks. *BMC Syst Biol*, **4**, 96.

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.

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.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A.-S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabasi, A.-L., and Vidal, M. (2009). An empirical framework for binary interactome mapping. *Nat Methods*, **6**(1), 83–90.

Vidal, M. and Furlong, E. E. M. (2004). From OMICS to systems biology. *Nature Reviews Genetics*, **5**(10), poster.

Vogelstein, B., Lane, D., and Levine, A. J. (2000). Surfing the p53 network. *Nature*, **408**(6810), 307–310.

Voight, B. F., Scott, L. J., Steinthorsdottir, V., Morris, A. P., Dina, C., Welch, R. P., Zeggini, E., Huth, C., Aulchenko, Y. S., Thorleifsson, G., McCulloch, L. J., Ferreira, T., Grallert, H., Amin, N., Wu, G., Willer, C. J., Raychaudhuri, S., McCarroll, S. A., Langenberg, C., Hofmann, O. M., Dupuis, J., Qi, L., Segrè, A. V., van Hoek, M., Navarro, P., Ardlie, K., Balkau, B., Benediktsson, R., Bennett, A. J., Blagieva, R., Boerwinkle, E., Bonnycastle, L. L., Bengtsson Boström, K., Bravenboer, B., Bumpstead, S., Burtt, N. P., Charpentier, G., Chines, P. S., Cornelis, M., Couper, D. J., Crawford, G., Doney, A. S. F., Elliott, K. S., Elliott, A. L., Erdos, M. R., Fox, C. S., Franklin, C. S., Ganser, M., Gieger, C., Grarup, N., Green, T., Griffin, S., Groves, C. J., Guiducci, C., Hadjadj, S., Hassanali, N., Herder, C., Isomaa, B., Jackson, A. U., Johnson, P. R. V., Jørgensen, T., Kao, W. H. L., Klopp, N., Kong, A., Kraft, P., Kuusisto, J., Lauritzen, T., Li, M., Lieverse, A., Lindgren, C. M., Lyssenko, V., Marre, M., Meitinger, T., Midthjell, K., Morken, M. A., Narisu, N., Nilsson, P., Owen, K. R., Payne, F., Perry, J. R. B., Petersen, A.-K., Platou, C., Proença, C., Prokopenko, I., Rathmann, W., Rayner, N. W., Robertson, N. R., Rocheleau, G., Roden, M., Sampson, M. J., Saxena, R., Shields, B. M., Shrader, P., Sigurdsson, G., Sparsø, T., Strassburger, K., Stringham, H. M., Sun, Q., Swift, A. J., Thorand, B., Tichet, J., Tuomi, T., van Dam, R. M., van Haeften, T. W., van Herpt, T., van Vliet-Ostaptchouk, J. V., Walters, G. B., Weedon, M. N., Wijmenga, C., Witteman,

J., Bergman, R. N., Cauchi, S., Collins, F. S., Gloyn, A. L., Gyllensten, U., Hansen, T., Hide, W. A., Hitman, G. A., Hofman, A., Hunter, D. J., Hveem, K., Laakso, M., Mohlke, K. L., Morris, A. D., Palmer, C. N. A., Pramstaller, P. P., Rudan, I., Sijbrands, E., Stein, L. D., Tuomilehto, J., Uitterlinden, A., Walker, M., Wareham, N. J., Watanabe, R. M., Abecasis, G. R., Boehm, B. O., Campbell, H., Daly, M. J., Hattersley, A. T., Hu, F. B., Meigs, J. B., Pankow, J. S., Pedersen, O., Wichmann, H.-E., Barroso, I., Florez, J. C., Frayling, T. M., Groop, L., Sladek, R., Thorsteinsdottir, U., Wilson, J. F., Illig, T., Froguel, P., van Duijn, C. M., Stefansson, K., Altshuler, D., Boehnke, M., McCarthy, M. I., , M. A. G. I. C. i., and , G. I. A. N. T. C. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet*, **42**(7), 579–589.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**(6887), 399–403.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, **18**(7), 1283–1292.

Wagner, A. (2003). How the global structure of protein interaction networks evolves. *Proc Biol Sci*, **270**(1514), 457–466.

Walker, M. G., Volkmuth, W., Sprinzak, E., Hodgson, D., and Klingler, T. (1999). Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res*, **9**(12), 1198–1203.

Wallace, D. C. (1999). Mitochondrial diseases in man and mouse. *Science*, **283**(5407), 1482–1488.

Walsh, C. T. (2006). *Posttranslational modification of proteins: expanding nature's inventory*. Roberts and Company Publishers.

Wang, Y., Lin, F., and Qin, Z.-H. (2010). The role of post-translational modifications of huntingtin in the pathogenesis of Huntington's disease. *Neurosci Bull*, **26**(2), 153–162.

Warrington, J. A., Nair, A., Mahadevappa, M., and Tsyganskaya, M. (2000). Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics*, **2**(3), 143–147.

Watson, J. D., Bartlett, G. J., and Thornton, J. M. (2009). Inferring protein function from structure. In *Structural Bioinformatics.*, chapter 21, pages 515–537. John Wiley & Sons, 2 edition.

Weichert, W., Kristiansen, G., Schmidt, M., Gekeler, V., Noske, A., Niesporek, S., Dietel, M., and Denkert, C. (2005). Polo-like kinase 1 expression is a prognostic factor in human colon cancer. *World J Gastroenterol*, **11**(36), 5644–5650.

Weinberg, R. A. (1996). How cancer arises. *Sci Am*, **275**(3), 62–70.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **36**(Database issue), D13–D21.

Whisstock, J. C. and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Q Rev Biophys*, **36**(3), 307–340.

Wiles, A. M., Doderer, M., Ruan, J., Gu, T.-T., Ravi, D., Blackman, B., and Bishop, A. J. R. (2010). Building and analyzing protein interactome networks by cross-species comparisons. *BMC Syst Biol*, **4**, 36.

Winklhofer, K. F., Tatzelt, J., and Haass, C. (2008). The two faces of protein misfolding: gain- and loss-of-function in neurodegenerative diseases. *EMBO J*, **27**(2), 336–349.

Wong, J. M. S., Ionescu, D., and Ingles, C. J. (2003). Interaction between brca2 and replication protein a is compromised by a cancer-predisposing mutation in brca2. *Oncogene*, **22**(1), 28–33.

Woodford-Richens, K. L., Rowan, A. J., Gorman, P., Halford, S., Bicknell, D. C., Wasan, H. S., Roylance, R. R., Bodmer, W. F., and Tomlinson, I. P. (2001). SMAD4 mutations in colorectal cancer probably occur before chromosomal instability, but after divergence of the microsatellite instability pathway. *Proc Natl Acad Sci U S A*, **98**(17), 9719–9723.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T. P., and Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nat Methods*, **6**(1), 75–77.

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Mol Syst Biol*, **4**, 189.

Xie, W., Rimm, D. L., Lin, Y., Shih, W. J., and Reiss, M. (2003). Loss of smad signaling in human colorectal cancer is associated with advanced disease and poor prognosis. *Cancer J*, **9**(4), 302–312.

Xu, J. and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, **22**(22), 2800–2805.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., and Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**(5), 650–659.

Yang, B., Akhter, S., Chaudhuri, A., and Kanmogne, G. D. (2009). HIV-1 gp120 induces cytokine expression, leukocyte adhesion, and transmigration across the blood-brain barrier: modulatory effects of STAT1 signaling. *Microvasc Res*, **77**(2), 212–219.

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, H.-G., Chan, D. W., Huang, Z.-Q., Li, J., Fondell, J. D., Qin, J., and Wong, J. (2003). Purification and functional characterization of the human N-CoR complex: the roles of HDAC3, TBL1 and TBLR1. *EMBO J*, **22**(6), 1336–1346.

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

Zhang, A. (2009). *Protein Interaction Networks: Computational Analysis*. Cambridge University Press, 1 edition.

Zhang, L. and Li, W.-H. (2004). Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol*, **21**(2), 236–239.

Zhao, X., Qiu, W., Kung, J., Zhao, X., Peng, X., Yegappan, M., Yen-Lieberman, B., and Hsi, E. D. (2008). Bortezomib induces caspase-dependent apoptosis in Hodgkin lymphoma cell lines and is associated with reduced c-FLIP expression: a gene expression profiling study with implications for potential combination therapies. *Leuk Res*, **32**(2), 275–285.

Zhou, D. and He, Y. (2008). Extracting interactions between proteins from the literature. *J Biomed Inform*, **41**(2), 393–407.

Zhou, X., Kao, M.-C. J., and Wong, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U S A*, **99**(20), 12783–12788.

Zhu, J., He, F., Hu, S., and Yu, J. (2008). On the nature of human housekeeping genes. *Trends Genet*, **24**(10), 481–484.

Zhu, Y., Richardson, J. A., Parada, L. F., and Graff, J. M. (1998). Smad3 mutant mice develop metastatic colorectal cancer. *Cell*, **94**(6), 703–714.

Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*, **4**(8), e1000140.

Özgür, A., Vu, T., Erkan, G., and Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, **24**(13), i277–i285.

List of Figures

List of Tables

Selbstständigkeitserklärung

Hiermit erkläre ich,

- dass ich die vorliegende Arbeit mit dem Titel "*Network-based inference of protein function and disease-gene association*" selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe,
- dass ich keinen Doktorgrad im Fach Informatik besitze,
- und dass mir die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät II der Humboldt-Universität zu Berlin vom 17.01.2005, zuletzt geändert am 13.02.2006, veröffentlicht im Amtlichen Mitteilungsblatt Nr. 34/2006, bekannt ist.

Berlin, Mäerz 2012

Samira Jaeger