

ANALYSIS, INTEGRATION AND APPLICATIONS OF THE HUMAN INTERACTOME

DISSERTATION

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer.nat.)

im Fach Biophysik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät I

Humboldt Universität zu Berlin

von

M.Sc. Bioinf. Gautam Kumar Chaurasia

Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I

Prof. Dr. Andreas Herrmann

Gutachter:

1. Prof. Dr. Hanspeter Herzel
2. Prof. Dr. Joachim Selbig, PhD
3. Prof. Dr. Erich E. Wanker

Tag der mündlichen Prüfung: 06.01.2012

Abstract

Protein interaction networks aim to provide the scaffold maps for systematic studies of the complex molecular machinery in the cell. The complexity of protein interactions poses, however, large experimental and computational challenges regarding their identification, validation and annotation. Additionally, storage and linking is demanding since new data are rapidly accumulating. In this research work, I addressed these issues and provided solutions to overcome the limitations of current human protein-protein interaction (PPI) maps. In particular, my thesis can be partitioned into two parts:

- In the first part, I conducted a comparative assessment of eight recently constructed human protein-protein interaction networks to identify experimental biases. To this end, I scrutinized PPI networks with respect to their overlap, functional composition and topological properties. Results showed strong selection and detection biases which are necessary to take into consideration in future applications of these maps. One of the important conclusions of this study was that the current human interaction networks contain complementary information; hence, their integration could be beneficial. To this end, a database was developed, termed as Unified Human Interactome (UniHI), integrating human PPI data from twelve major sources. Several new tools were included for querying, analyzing and visualizing human PPI networks, enabling researchers to target their analysis and prioritize candidates for follow-up studies.
- In the second part of this research work, I applied the data from UniHI to study the two aspects. First, I aimed to study the modular organization of human interactome. Results from this study showed a larger number of modules including many known protein complexes, linked via many overlapping key proteins. Further functional and expression analysis of detected modules enabled a direct comparison between stable and dynamic modules. Second, UniHI dataset was applied to characterize the genetic modifiers involved in a specific disease: Chorea Huntington (HD), an autosomal dominant neurodegenerative disease. To find the modifiers, a network-based modeling approach was implemented by integrating huntingtin-specific protein interaction network with gene expression data from HD patients in multiple steps. Using this approach, a Caudate Nucleus-specific HD protein interaction (PPI) network was predicted, connecting 14 potentially dysregulated proteins directly or indirectly to the

disease protein. Follow-up analysis showed the highly significant overrepresentation of network proteins participating in pro-apoptotic pathways, cell survival, anti-apoptotic, growth, and neuronal diseases, demonstrating the essentiality of this prediction approach.

Zusammenfassung

Protein-Protein Interaktions (PPI) Netzwerke liefern ein Grundgerüst für systematische Untersuchungen der komplexen molekularen Maschinerie in der Zelle. Die Komplexität von Protein-Wechselwirkungen stellt jedoch in Bezug auf ihre Identifizierung, Validierung und Annotation eine große experimentelle und rechnerische Herausforderung dar. Darüber hinaus ist die Speicherung und Verknüpfung anspruchsvoll, da die Menge der relevanten Daten rasch anwächst. In dieser Arbeit analysierte ich diese Probleme und lieferte Lösungen, um die Limitierungen aktueller humanen PPI Netzwerke zu überwinden. Meine Arbeit kann in zwei Teile aufgeteilt werden:

- Im ersten Teil führte ich einen kritischen Vergleich von acht unabhängig konstruierten humanen PPI Netzwerken durch, um mögliche experimentellen Verzerrungen zu erkennen. Zu diesem Zweck habe ich PPI Netzwerke hinsichtlich ihrer Überlappung, funktionalen Zusammensetzung und topologischen Eigenschaften geprüft. Die Ergebnisse zeigten starke Tendenzen bezüglich der Selektion und Detektion von Interaktionen, die in zukünftigen Anwendungen dieser Netzwerke berücksichtigt werden sollten. Einer der wichtigsten Schlussfolgerungen dieser Studie war, dass die derzeitigen humanen Interaktions Netzwerke komplementär sind und deshalb kann eine Integration von diesen Karten von großen Nutzen sein. Zu diesem Zweck wurde eine Datenbank mit der Bezeichnung *Unified Human Interaktome* (UniHI) entwickelt, die menschliche PPI Daten aus zwölf wichtigsten Quellen integriert. Mehrere neue Tools wurden für die Abfrage, Analyse und Visualisierung von Protein Interaktionen entwickelt. Diese Tools ermöglichen Forschern die Analyse von Interaktions-Netzwerken und Auswahl der interessanter Kandidaten für weiterführende Studien.
- Im zweiten Teil dieser Forschungsarbeit benutzte ich die Daten aus der UniHI Datenbank, um zwei Aspekte zu untersuchen. Erstens wurde die modulare Struktur der menschlichen Interaktoms von mir analysiert. Als Ergebnis dieser Studie ergab sich die Detektion einer größeren Zahl von Modulen, darunter viele bekannte Protein-Komplexe, die über einzelne Proteinen verknüpft waren. Weitere funktionelle Studien und Expressions-Analysen hinsichtlich der gefunden Module ermöglichten eine Unterscheidung zwischen stabilen und dynamischen Modulen. Zweitens wurde der UniHI Datensatz von mir angewandt, um die genetischen Modifikatoren in einer bestimmten Krankheit, Chorea Huntington (HD) eine autosomal dominante

neurodegenerative Erkrankung, zu charakterisieren. Um die Proteine zu identifizieren, die den Krankheitsverlauf modifizieren können, wurde eine netzwerk-basierte Methode implementiert. Diese basiert auf der Integration von Interaktionsdaten für das Huntingtin-Protein und Genexpressionsdaten von HD-Patienten in Kombination mit einem Mehrschritt-Filterungsverfahren. Mit dem neuartigen Ansatz wurde ein Nucleus caudatus-spezifische Protein-Interaktion HD (PPI)-Netzwerk vorhergesagt, das 14 potentiell dysregulierten Proteine direkt oder indirekt mit dem Huntingtin-Protein verlinkt. Funktionelle Analysen zeigten, dass die Proteine dieses Netzwerk auf hoch-signifikante Weise zu wichtigen molekularen Prozessen wie z.B. Apoptose, Metabolismus, neuronale Entwicklung assoziiert sind.

Table of Contents

ABSTRACT	III
ZUSAMMENFASSUNG	V
TABLE OF CONTENTS.....	VII
GRATITUDE	XI
1 INTRODUCTION	1
1.1 CELL: NETWORK OF NETWORKS	1
1.2 SYSTEMS BIOLOGY: A FRAMEWORK TO STUDY THE BEHAVIOUR OF THE CELL.....	1
1.3 NETWORK BIOLOGY: A KEY COMPONENT OF SYSTEMS BIOLOGY	2
1.4 PPI NETWORKS: CORE OF THE NETWORK BIOLOGY	3
1.5 CURRENT CHALLENGES	4
1.6 AIMS AND OVERVIEW OF THE CURRENT RESEARCH WORK	6
1.6.1 <i>Comparative analysis of human PPI networks</i>	6
1.6.2 <i>Implementation of an integrated platform for PPI networks</i>	7
1.6.3 <i>Identification of modules in PPI networks</i>	7
1.6.4 <i>Characterization of brain-specific dys-regulated processes and modifiers for Huntington's disease</i> 7	
1.7 OUTLINE OF THE THESIS	8
2 REVIEW OF LITERATURE.....	9
2.1 CELLS AS BUILDING BLOCKS OF LIFE.....	10
2.1.1 <i>Central dogma of molecular biology</i>	10
2.1.2 <i>Genes</i>	10
2.1.3 <i>Gene expression profiling</i>	11
2.1.4 <i>Proteins as workhorse</i>	12
2.1.5 <i>Protein misfolding and diseases</i>	13
2.2 PROTEIN-PROTEIN INTERACTION NETWORKS	13
2.2.1 <i>Methods for generating large-scale protein-protein interaction networks</i>	13
2.2.2 <i>Topological properties of PPI network</i>	18
2.2.3 <i>Databases for human protein interactions</i>	20
2.2.4 <i>Tools for analysis and visualization of interaction networks</i>	21
2.2.5 <i>Application of interactomics</i>	23
2.2.6 <i>Role of the PPI networks in disease research</i>	23
3 COMPARISON OF LARGE-SCALE MAPS OF THE HUMAN PROTEIN INTERACTOME	29
3.1 INTRODUCTION	30
3.2 MATERIALS AND METHODS	31

3.2.1	<i>Assembly of protein-protein interaction maps</i>	31
3.2.2	<i>Overlap of interaction maps</i>	32
3.2.3	<i>Gene Ontology analysis</i>	34
3.2.4	<i>Graph analysis</i>	36
3.3	RESULTS	36
3.3.1	<i>Common proteins and interactions</i>	37
3.3.2	<i>Overlap and intersection</i>	38
3.3.3	<i>Functional assessment</i>	42
3.3.4	<i>Graph-theoretical comparison</i>	43
3.3.5	<i>Analysis of network hubs</i>	46
3.4	DISCUSSION AND CONCLUSIONS	48
4	UNIHI: INTEGRATION OF HUMAN INTERACTOME	52
4.1	INTRODUCTION	54
4.1.1	<i>Highly divergent and distributed PPI networks</i>	54
4.1.2	<i>Quality of human PPI networks</i>	55
4.1.3	<i>Regular updates</i>	55
4.1.4	<i>Functional interpretation of PPI networks</i>	55
4.2	MATERIALS AND METHODS	56
4.2.1	<i>PPI data sources</i>	56
4.2.2	<i>Gene and protein identifiers</i>	56
4.2.3	<i>Gene annotation</i>	56
4.2.4	<i>Gene expression data</i>	57
4.2.5	<i>Pathway information</i>	57
4.3	RESULTS	57
4.3.1	<i>Architecture of the UniHI</i>	58
4.3.2	<i>Mapping of proteins</i>	60
4.3.3	<i>Data quality assessment</i>	60
4.3.4	<i>Data query, analysis and visualization</i>	61
4.3.5	<i>Integration of PPI data with Gene Expression and Pathway Data</i>	63
4.4	DISCUSSION AND CONCLUSIONS	67
5	FUNCTIONAL AND TRANSCRIPTIONAL COHERENCY OF MODULES IN THE HUMAN PROTEIN INTERACTION NETWORK	69
5.1	INTRODUCTION	69
5.2	MATERIALS AND METHODS	70
5.2.1	<i>Human protein-protein interaction data</i>	70
5.2.2	<i>Identification of modules in the protein interaction network</i>	71
5.2.3	<i>Generation of random graphs</i>	71
5.2.4	<i>Protein annotation</i>	72
5.2.5	<i>Expression data</i>	72
5.3	RESULTS	73

5.3.1	<i>Identification of modular structures in the human interaction network</i>	73
5.3.2	<i>Community size distribution</i>	73
5.3.3	<i>Distribution of proteins</i>	74
5.3.4	<i>Functional annotation of the detected modular structures</i>	74
5.3.5	<i>Localization of modules</i>	77
5.3.6	<i>Co-expression of modules</i>	77
5.3.7	<i>Overlap between modules and identification of linking proteins</i>	79
5.4	DISCUSSION AND CONCLUSIONS	80
6	NETWORK-BASED CHARACTERIZATION OF BRAIN SPECIFIC HUNTINGTON'S DISEASE MODIFIERS	83
6.1	INTRODUCTION	84
6.2	MATERIALS AND METHODS	87
6.2.1	<i>PPI data source</i>	87
6.2.2	<i>Microarray data analysis</i>	87
6.2.3	<i>Functional enrichment analysis using Gene Ontology database</i>	88
6.2.4	<i>Functional analysis by manual curation</i>	88
6.3	RESULTS	88
6.3.1	<i>In silico construction and analysis of a Huntingtin focused protein interaction network</i>	90
6.3.2	<i>Prioritization by multi-level filtering using gene expression data</i>	91
6.3.3	<i>Functional analysis of dysregulated HD network</i>	93
6.3.4	<i>Enrichment analysis using annotated targets for HD therapy development</i>	94
6.3.5	<i>Precision of predicted HD network</i>	94
6.3.6	<i>Specificity of predicted HD network</i>	95
6.3.7	<i>Grade-associated analysis of predicted HD modifiers</i>	96
6.4	DISCUSSION AND CONCLUSIONS	98
7	SUMMARY AND OUTLOOK	102
7.1	REVIEW OF FINDINGS	102
7.1.1	<i>Analysis and integration of human Protein-Protein interaction networks</i>	102
7.1.2	<i>Analysis of modular structure of human PPI networks</i>	104
7.1.3	<i>Prediction of Huntington disease modifier</i>	104
7.2	FUTURE DIRECTIONS	105
7.2.1	<i>Scope and extension of UniHI</i>	105
7.2.2	<i>Quality of PPI maps</i>	106
7.2.3	<i>Implementation of the network-based strategy for the prediction of disease genes in UniHI</i>	107
7.3	CONCLUSIONS	108
	APPENDIX A	110
	APPENDIX B	130
	<i>B.1 Design and Implementation</i>	130

<i>B.2 Data Integration</i>	130
APPENDIX C	135
<i>C.1 Details of the datasets used for the Precision analysis</i>	135
BIBLIOGRAPHY	137
LIST OF PUBLICATIONS	151
SELBSTÄNDIGKEITSERKLÄRUNG	153

Gratitude

First and foremost I would like to thank Dr. Matthias Futschik, Head of Bioinformatics and Systems Biology Group at CBME, University of Algarve, for giving me this wonderful opportunity to carry out this research work. Indeed, it has been a great pleasure to work under him. I acknowledge and highly appreciate his scientific support and guidance, without him this work would not have been possible. His calm, understated, yet meticulous manner of dealing with all matters have been instrumental in helping me to deal, with more equanimity. The times spent with him will remain some of my most cherished memories.

I attribute my sincere thanks to Prof. Dr. Hanspeter Herzel, Head of Theoretical Biology group at Institute for Theoretical Biology, for being my supervisor and ensuring the necessary infrastructure so that work never hit a snag for want of anything. I would like to acknowledge him for numerous discussions and lectures on related topics that helped me improve my knowledge in the area. His easy candour will always be remembered.

I express my humble reverence to Prof. Dr. Erich Wanker, Head of Proteomics lab at Max Delbrück Center for Molecular Medicine of the Max-Dellbrück Centrum Berlin, for his untiring support. I feel extremely fortunate in having him as a mentor; his boisterous camaraderie as well as his supportive attitude at work always ensured an exemplary environment. I am really grateful to him for many discussions related to my work, and especially for his valuable suggestions.

I thank Prof. Dr. Joachim Selbig, PhD, Professor for Bioinformatics at Max-Planck-Institute for Molecular Plant Physiology, Postdam University, Potsdam, for agreeing to be the third reviewer of this thesis.

My sincere appreciation goes to Dr. Martin Strödicke for his valuable support. I heartily acknowledge Dr. Ilka Axmann for her advices and suggestions. Jenny, Christian, Uli, Vinu, Yacine, Maciej, Frau Pisch and all other fellows of Proteomics lab deserve special thanks and acknowledgement that were always helpful when approached.

I am highly indebted and thankful to my wife Soniya, for her tireless support, unconditional love and affection. She supported me all the way and I am especially grateful to her that she always encouraged me and boost my morale during the tough time.

And last, but definitely, by far not the least, I would like to thank my Parents for their constant support and care. Behind all this there was one big support from God! And I really thank God for the blessings and every thing given to me.

1 INTRODUCTION

1.1 Cell: Network of Networks

A cell is the fundamental unit of all known living organisms, and is often referred as the building block of life. Its basic functions result from complex networks of interacting constituents such as DNA, RNA, proteins, and small molecules. Advances in high-throughput experimental methods have enabled the study of complex interactions on a global genome-wide level and have generated different types of biological networks. In particular they comprise protein-protein, protein-DNA, protein-metabolite and genetic interactions networks. However, these networks do not work in isolation, but carefully constitute a network of networks at different space and time that is responsible for functions and the structure of the cell (Barabasi and Oltvai, 2004). Therefore, a key challenge is to understand the role of these diverse networks and the interactions between them that solely define the behaviour of the cell. This especially requires the development of a framework for the study of the cellular systems as a whole and thus helps us to reveal the complex nature of the biological systems.

1.2 Systems Biology: a framework to study the behaviour of the Cell

Recently, a new discipline has emerged with the advent of large-scale biological data sets, termed *Systems Biology*. It can be viewed as a complementary - but not opposing – approach to the classical reductionist strategy for the study of the biological processes. In contrast to reductionist approaches based on the dissection of processes into their most elementary levels, systems biology is more holistically orientated. The guiding principle of systems biology is that the total system can be more than the sum of its parts and can acquire properties that are not implicated in the single components.

Following this principle, we seek to study a biological system as a whole. The aim is to determine the rules governing its behaviour and eventually to generate qualitative and quantitative predictions concerning its response to potential perturbations and modifications. To achieve this, two requirements have to be fulfilled: i) a sufficient

amount of data and information describing the system has to be available and ii) a computational model of the system has to be designed. Whereas the first requirement is increasingly met with the development of new high-throughput techniques, the second necessity still demands considerable efforts. For instance, when we aim to represent the whole system, we need to choose an adequate level of resolution. Finding this level is challenging, since there is usually a trade-off between computational feasibility and detailed representation of the molecular systems due to their mere size and complexity. The inclusion of too many components can lead to ill-determined models of the system with many parameters unknown, whereas a too severe restriction can result in an incomplete model with a lack of coherence. In fact, the choice of a suitable model depends not only on the research objective, but also, more practically, on the quality and quantity of data and information present.

1.3 Network Biology: a key component of Systems Biology

In response to the challenges posed by systems biology, various methodologies for different levels of resolution have been brought forward to date. A nowadays very popular approach is based on the representation of biological systems as mathematical graphs and has laid the ground for the blooming field called *network biology*. In the context of molecular systems, for instance, the molecules are typically represented as nodes and their interactions as edges (figure 1.1). Although this type of representation is clearly a strong simplification of the underlying physical system, a major advantage of this approach is that the analysis of large networks becomes feasible. Also, the underlying graph-theory has been well developed and offers researchers a variety of tools. In fact, with its beginning dating back to Leonard Euler in 1736, graph theory has made profound impact in social, physical and computer sciences (Euler, 1736). The application of graph-theory to biology seems to be well suited where large networks are involved in the process of interest. Thus, it is not surprising that the concepts of network biology have been especially applied to elucidate the complex processes during several diseases and to consolidate the hitherto divergent observations (Wachi *et al.*, 2005; Jonsson and Bates, 2006; Hernandez *et al.*, 2007; Platzer *et al.*, 2007).

Before we can study any complex disease within a network-based framework, we need to assess the availability of data and information necessary for such endeavour.

Currently, most disease-related data are produced in large sequencing and transcriptional profiling projects. Although their output has given us unprecedented details of the molecular changes during pathogenesis, they cannot give us *per se* causal relationships leading to the observed changes. To gain new insights, the knowledge of the relationships (i.e. interactions) between the involved biomolecules is crucial. Ideally, we would like to have the complete set of molecular interactions (i.e. *interactome*) that take place within the human body. Such human interactome would include a variety of different types of interactions such as transient or constitutive protein-DNA, RNA-RNA, protein-protein and protein-ligand interactions. Similarly to the sequencing of the human genome which supply us with a catalogue of the molecular parts of a cell, the charting of the human interactome would give us the blueprints how they are put together to function. In reality, however, we are still far away from a complete map of molecular interactions within the human body. Thus, at a practical level, most network-based approaches to decipher most of the diseases are restricted to certain types of interactions between a limited set of molecular entities. For example, in the context of cancer, most research efforts to date have been focused on the analysis of physical protein-protein and regulatory DNA-protein interaction networks. In particular, protein-protein interaction (PPI) networks have been extensively scrutinized with tools of network biology to advance our understanding of the complex molecular processes involved in diseases (Chuang *et al.*, 2007; Ergun *et al.*, 2007; Pujana *et al.*, 2007).

1.4 PPI Networks: core of the Network Biology

In a living organism, proteins interact with other proteins to carry out vital cellular functions, such as signal transduction, DNA replication, transcription, protein transport, or metabolic catalysis. Also, many major diseases such as neurological disorders or cancer are characterized by complex interactions of multiple proteins (Goehler *et al.*, 2004; Calvano *et al.*, 2005; Lim *et al.*, 2006; Oti *et al.*, 2006; Ideker and Sharan, 2008). The study of human protein interactions may therefore help (i) to improve our general understanding of biological processes (Figure 1.1) and (ii) to decipher the molecular basis of complex diseases and to provide new potential therapeutic targets.

For many years, interactions between proteins have been studied in small-scale experiments. This situation has however dramatically changed in the last decade. The availability of fully sequenced genomes (Ruder and Winstead) and advances in high-throughput approaches (Fields and Song, 1989; Figeys *et al.*, 2001; Puig *et al.*, 2001; Koegl and Uetz, 2007) have led to large-scale studies of protein–protein interactions on a genome-wide scale and to efforts to map the complete PPI network for an organism, termed also as interactome. Indeed, we have recently witnessed many large-scale protein interaction mapping projects in several model organisms such as *S. cerevisiae* (Schwikowski *et al.*, 2000; Uetz *et al.*, 2000; Ito *et al.*, 2001; Ho *et al.*, 2002), *D. melanogaster* (Giot *et al.*, 2003) and *C. elegans* (Li *et al.*, 2004). Now, the focus has moved towards a systematic mapping of human PPI maps (Aranda *et al.*, ; Bader *et al.*, 2003; Lehner and Fraser, 2004; Salwinski *et al.*, 2004; Brown and Jurisica, 2005; O'Brien *et al.*, 2005; Pagel *et al.*, 2005; Persico *et al.*, 2005; Ramani *et al.*, 2005; Rual *et al.*, 2005; Stelzl *et al.*, 2005; Ewing *et al.*, 2007; Breitkreutz *et al.*, 2008; Matthews *et al.*, 2009; Prasad *et al.*, 2009). The constructed human PPI maps have been derived from both experimental and computational approaches (Fields and Song, 1989; Figeys *et al.*, 2001; Matthews *et al.*, 2001; Puig *et al.*, 2001), and offer not only a wealth of information but are also expected to be of great assist for the biomedical research community (Goehler *et al.*, 2004; Goh *et al.*, 2007; Braun *et al.*, 2008; Ideker and Sharan, 2008). However, utilization of these interaction maps is impeded and the current limitations are manifold.

1.5 Current Challenges

In the postgenomic era, one of the daunting tasks of proteomics is to chart complete protein-protein interaction networks that occur within cells. Although, the availability of many large genome sequences and advances in high-throughput methods provided us a platform to construct PPI maps, interactomes of many organisms are far from complete. A major problem of currently available approaches is therefore that they are unable to capture interactions in a comprehensive manner (Hart *et al.*, 2006). Even, in a recent study, in which the authors claimed to developed an improved version of currently available high-throughput methods to identify the yeast PPI network, termed as generation-2 methods, the coverage of all possible interactions in *S. cerevisiae* was estimated to reach only ~20% (Yu *et al.*, 2008).

The Human Protein-Protein Interactome

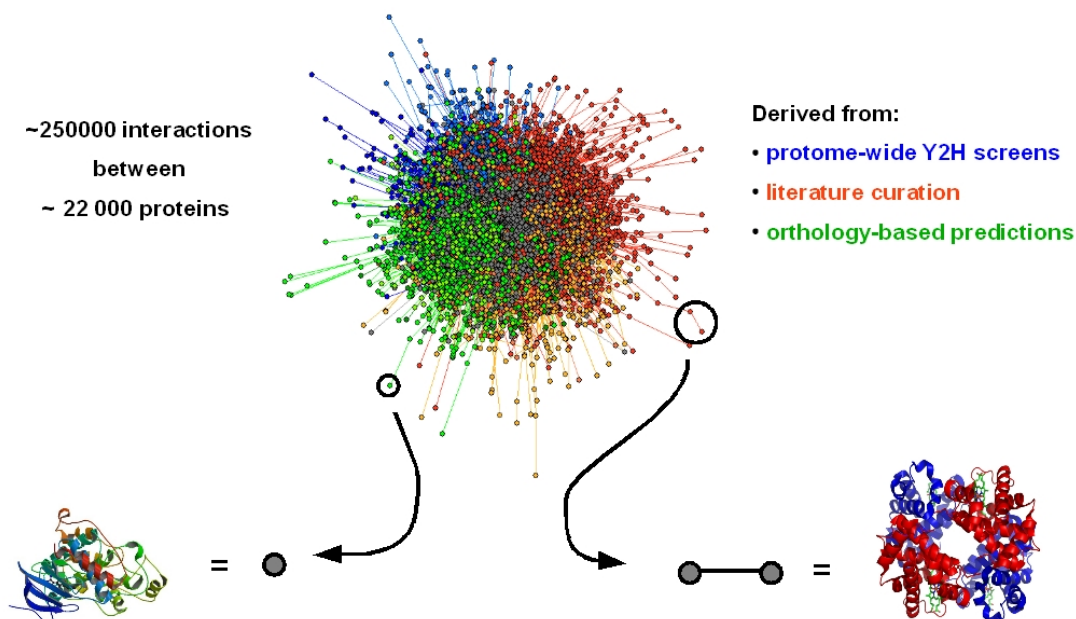


Figure 1.1: Overview of the human interactome. Graphical representation of the current human protein-protein interactome as stored in the UniHI database (<http://www.unihi.org>). Altogether, it comprises over a quarter of a million of interactions derived from experimental resources and by computational prediction. The figure also illustrates the grade of simplification achieved by the graph-theoretical approach. The highlighted nodes representing to protein structures (left: mitogen activated protein kinase; right: haemoglobin complex consisting of alpha and beta chains) are depicted for illustration only; they do not represent the actual location of these proteins in the interactome. Displayed structures were taken from the Protein Data Bank (Berman *et al.*, 2007).

Similar problems also exist in human PPI. Recent studies have shown that the current human PPI maps are incomplete and highly unsaturated. For example, given a total size of human interactome of ~650000 interactions as a recent study estimated (Stumpf *et al.*, 2008), even HPRD, a manually-curated literature database for human protein interaction maps (Prasad *et al.*, 2009), as one of the largest sources, covers not more than 5% of the total interactome (Stumpf *et al.*, 2008).

The quality of PPI data also remains a critical issue. It has been observed in many studies, that the data, produced by high-throughput experimental methods, contains high rate of false positive or negative interactions. Additionally, these methods may

also have experimental biases toward certain protein types and cellular localizations, which demands for the improvement in high throughput methods. For example, Y2H-based mapping approaches offer rapid screens between thousands of proteins, but might be compromised by large false positive rates. The extent, however, how much the resulting interaction maps are influenced by the choice of mapping strategy, is less clear. Thus, it is important to critically assess and compare quality and reliability of produced maps. Comparative analysis of interaction maps in lower eukaryotes showed a surprising divergence between different interaction maps (Mrowka *et al.*, 2001; Bader and Hogue, 2002; von Mering *et al.*, 2002). Human PPI maps are likely to be no exceptions, but a comparison was still lacking for human protein despite their expected importance for biomedical research. Thus, critical evaluation of the available human interaction maps has been necessary regarding the method chosen for network generation.

1.6 Aims and overview of the current research work

This thesis primarily focuses on the analysis, integration and applications of the large-scale human protein interaction networks. In particular, my work aims to contribute to the usability of protein interactions in biomedical research as well as to apply protein interaction networks for the study of physiological and pathological processes. The objectives of my doctoral research can be summarized as follows:

1.6.1 Comparative analysis of human PPI networks

Current PPI networks are often error-prone and unsaturated and (Mrowka *et al.*, 2001; Bader and Hogue, 2002; von Mering *et al.*, 2002). Moreover, these networks might contain biases, i.e. over- or under-representation of proteins from some certain categories, due to the sensitiveness of experimental methods towards specific type of proteins. For example, membrane proteins have been found under-represented in interaction networks generated by Y2H-methods (Mrowka *et al.*, 2001; von Mering *et al.*, 2002). Whether, the same problem also exists in human PPI network remained to be answered. To address these issues, a first systematic analysis of eight human PPI networks, generated by either yeast-two-hybrid methods, or manual curation, or computational prediction approaches, was conducted. These PPI networks were analyzed regarding their overlaps, functional constitution and topological organization. Results of these analyses are discussed in more detail in chapter 3.

1.6.2 Implementation of an integrated platform for PPI networks

Comparative evaluation of PPI maps, described in chapter 3, provided me with important findings. Especially, the overlaps of interactions between the networks were found to be rather small, suggesting the highly complementary nature of the current human PPI networks. Therefore, an integration of these networks could be a milestone towards achieving a comprehensive human interactome. But, integrating the data from diverse sources is not an easy task and poses many challenges. In chapter 4, I addressed these problems and described a strategy to overcome them for a successful integration of PPI networks.

1.6.3 Identification of modules in PPI networks

PPI networks are scale-free and organized in a hierarchical manner (Barabasi and Oltvai, 2004). Several studies have been performed to study the modular structure of PPI networks in lower eukaryotes (Rives and Galitski, 2003; Spirin and Mirny, 2003). However, such analysis was still missing for human interactome. In chapter 5, I examined the modular structure of human interactome, and described the important findings

1.6.4 Characterization of brain-specific dys-regulated processes and modifiers for Huntington's disease

Huntington's disease is an autosomal dominant late-onset neurodegenerative disorder, caused by an expansion of polyglutamine tract. The pathogenic outcome of HD leads to disturbance in muscle coordination and some cognitive functions. Network-based approaches are powerful predictive tools, and have been successfully applied in several studies to characterize modifiers in complex diseases such as cancer, ataxia, multiple sclerosis (Goehler *et al.*, 2004; Lim *et al.*, 2006; Chuang *et al.*, 2007; Pujana *et al.*, 2007; Baranzini *et al.*, 2009). In this study, I developed a network-based method by systematically integrating huntingtin-specific human protein interaction network with the gene activity data from Huntington disease patients in multiple steps to identify the genes which are altered during HD pathogenesis and may provide a basis for new treatments. More details on the bioinformatic analysis, their findings and the statistical validations are provided in

chapter 6.

1.7 Outline of the thesis

Chapter 1 (this chapter) introduces the fundamentals of systems biology. It further provides the basics of PPI networks, their current problems, and aims of this research work. Chapter 2 gives brief details about biological background required for reading this work. Next, it summarizes the methods for generating large-scale PPI network, limitations of current approaches, databases and the tools for the storing and analyzing the human PPI networks. Especially, it reviews published research works where PPI networks have been applied to characterize disease genes and the associated biological processes. Chapter 3, 4, 5, and 6 present the results from different research articles, which are either published or submitted. All these articles have a similar structure, and contain four sections: section one provides introduction to related work, section two describes the material and methods applied, section three discusses the results, followed by last the section discussion and conclusions. In particular, Chapter 3 presents the results from three different published research articles, discusses the findings of the systematic comparative evaluation of current human PPI networks. Chapter 4 discusses the several challenges and necessary steps for the integration of human PPI networks. Chapter 5 summarizes the findings from modularity analysis performed using human PPI network integrated within UniHI database. Chapter 6 presents newly implemented network-based method, to characterize the brain-specific modifiers for Huntington's disease. Chapter 7 summarizes the contributions of the research described in this thesis and discusses its impact on future investigations.

2 REVIEW OF LITERATURE

A brief introduction to the aim of this research work has been provided in Chapter 1. In this chapter, I will review few topics related to the current work. Section one introduces the basic concepts of molecular biology. Additionally, this section also elaborates basic mechanism of protein misfolding and the related diseases. Section two described several aspects of PPI networks. First, different methods for generating large-scale protein interaction networks and their limitations are discussed. Further it reviews few studies, in which these approaches have been employed to generate large-scale PPI networks. Additionally, it also provides details on fundamentals of network theory. Finally, in section three, several databases for housing data on human protein interaction, tools for networks analysis, and their applications in biomedical sciences are discussed.

2.1 Cells as building blocks of life

A cell is one of the most basic units of life. All living creatures are made of cells. They can be classified into two major categories: prokaryotes and eukaryotes. Most prokaryotes are single cells organism. In contrast, eukaryotes are highly evolved multi-cellular organisms. Examples of eukaryotes are all plants and higher animals. Prokaryotic cells do not contain nucleus and their DNA lies in the same compartment as the cytoplasm, while eukaryotic cells contain membrane-bound compartments in which specific metabolic activities take place. Most important among these compartments is the nucleus, which houses the eukaryotic cell's DNA

2.1.1 Central dogma of molecular biology

One of the fundamental mechanisms of molecular biology is the flow of information from DNA to RNA and subsequently from RNA to proteins (figure 2.1). This process is also known as central dogma of molecular biology. Deoxyribonucleic acid (DNA) is considered to be a cellular library that contains the genetic instructions (i.e. genes) used in the development and functioning of all known living organisms. The transfer of information from DNA to RNA is known as transcription, creating the copies of messenger RNAs (mRNA). After transcription, mRNAs are converted into proteins by a process called translation. For this, mRNAs contain required information encoded in nucleotide triplets called codons, which are translated into proteins by some rules known as genetic code.

2.1.2 Genes

Genes are the basic units of heredity in DNA, and are associated with regulatory, transcribed and other functional sequence regions. Basically, they contain coding sequences that are required for its expression. The molecules resulting from gene expression, whether proteins or RNA, are known as gene products, and are responsible for the development and functioning of all living things. The process by which production of a gene product is controlled is called gene regulation, which is usually carried out through interactions among DNA, RNA and proteins. It increases the versatility and adaptability of an organism by allowing the cell to express a protein for specific function when it is needed. Abnormal changes in the expression level of a gene may damage a healthy cell and lead to a dangerous disease. On the other

hand, expression changes can also counter the effects of a malignant cell, and thereby help to protect the internal cellular environment. A classical example is TP53 gene. Mutation or deletion of TP53 can lead to cancer disease, whereas, increased amount of TP53, prevents tumour cells from spreading, but can also cause premature aging (Tyner et al., 2002).

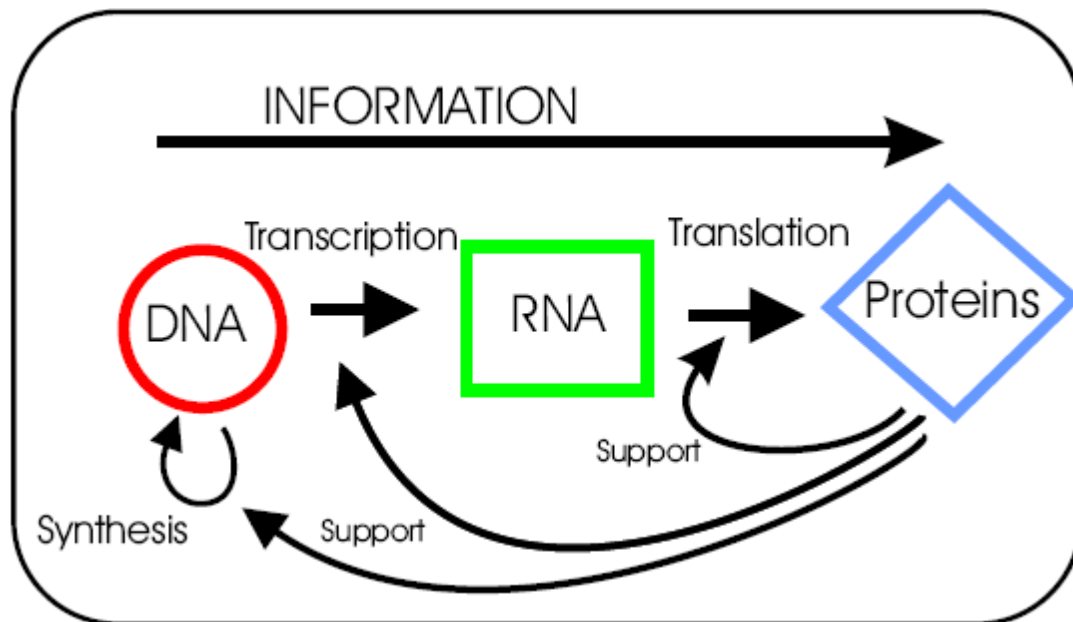


Figure 2.1: From DNA to mRNA to protein. Image taken from (Futschik, 2003).

2.1.3 Gene expression profiling

DNA Microarrays are an essential platform that enables scientist to monitor the expression levels of thousands of genes simultaneously on a global, genome-wide level. A typical cDNA microarray consists of a glass slide containing complementary sequences of many genes arranged in a regular pattern. These arrays can be applied to study the effects of certain treatments, diseases, and developmental stages on gene expression. In particular, they can be applied to study the behaviour of a cell by comparing the expression levels of a set of genes under in diseased and normal conditions. There are several databases and repositories which manages the massive amounts of data produced by microarray experiments. Popular examples are the Gene Expression Omnibus (Barrett *et al.*, 2009) and the Array Express (Brazma *et al.*, 2003).

2.1.4 Proteins as workhorse

Proteins are the most active elements of cells. They control and mediate in many of the biological activities that make the cell work. The chief characteristic of proteins that enable them to be multi-functional player is their ability to bind with other molecules specifically and tightly. The range of the protein functions includes formation of structural complexes, intracellular signaling, synthesis, repair and replication of DNA, membrane transport, and many others. In order to perform these tasks, they exhibit outstanding richness in their structure, which can be described using four hierarchical orders. The primary structure of a protein is its linear sequence of amino acids. Secondary structure is a regularly repeating local structures stabilized by hydrogen bonds. The most common examples are the alpha helix, beta sheet and turns. Tertiary structure refers to the three-dimensional structure of the entire polypeptide chain, which is responsible for the functions of a protein. Quaternary structure is structure formed by several protein molecules, also termed as single protein complex. Proteins complexes are the foundation of many cellular processes and together they constitute different types of molecular machinery that execute plenty of biological functions.

Several databases have been developed for storing and managing the information on protein. These databases can be classified into two main categories based on their content: structural databases (SDs) and functional databases (FDs). Structural databases provide information on secondary and tertiary structures of proteins, protein families, domains and functional sites, and protein fingerprints (motifs). Examples of such databases are PDB (Berman *et al.*, 2007), PFAM (Bateman *et al.*, 2004), SCOP (Murzin *et al.*, 1995), CATH (Pearl *et al.*, 2005), and INTERPRO (Mulder *et al.*, 2005). In contrast to the structural databases, functional databases store information on the functions of proteins on a hierarchical level. Popular examples for housing functional classification on proteins are Enzyme Commission (EC) hierarchical classification (Bairoch, 2000) the Gene Ontology (GO) (Ashburner *et al.*, 2000). Especially, GO provides structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in many organism.

2.1.5 Protein misfolding and diseases

Protein folding is an essential mechanism by which polypeptide chain of a protein folds into a functional three-dimensional structure (Berg *et al.*, 2002). Under normal condition, folding of a protein depends on many environmental factors, involving solvent (water or lipid bilayer) (van den Berg *et al.*, 2000), concentration of salts, temperature, and more importantly the presence of molecular chaperones. However, changes in these factors result in a denatured protein, leading to the misfolding and aggregation of the protein. Defects in protein folding usually produce inactive proteins with different properties including toxic prions. Various diseases have been reported to protein misfolding which may result in the formation of insoluble protein plaques in the brain or other organs. These diseases include prion diseases such as bovine spongiform encephalopathy (BSE), Creutzfeld-Jakob disease (CJD), amyloid -related illnesses such as Alzheimer's disease and familial amyloid cardiomyopathy or polyneuropathy, as well as intracytoplasmic aggregation diseases such as Huntington's and Parkinson's disease (Glenner, 1980; Selkoe, 2003; Ross and Poirier, 2004; Chiti and Dobson, 2006).

2.2 Protein-protein interaction networks

Protein-protein interactions underlie most of the molecular mechanisms essential for any living organisms. Intensive research in last decades has revealed many details of the fascinating multifaceted capacity of proteins to gain divers functionality by interaction. Although these efforts have supplied us with a tremendous amount of information for single proteins, they also indicated that most proteins function in a highly cooperative manner. A comprehensive protein interaction network is, therefore, an important framework to study the complex cellular processes and a prerequisite for accurate models in systems biology. In the following sections, I will provide details on the methods for network generation, topological properties, and their use in biomedical sciences.

2.2.1 Methods for generating large-scale protein-protein interaction networks

In last decade, various high-throughput experimental approaches have been developed to construct large-scale protein-protein interaction maps. These approaches comprise yeast-two-hybrid (Y2H) (Fields and Song, 1989; Koegl and

Uetz, 2007), affinity purifications or immunoprecipitation (Puig *et al.*, 2001) followed by mass-spectrometry (Figeys *et al.*, 2001), the coordinated efforts in systematic charting of interactions by human experts (Prasad *et al.*, 2009) as well as the progress in computational text-mining (Ramani *et al.*, 2005) and prediction methods (Matthews *et al.*, 2001). As all these methods can lead to considerably divergent protein interaction maps (Bader and Hogue, 2002; von Mering *et al.*, 2002), it is important to have a basic understanding of the applied methodologies. In the following sections, I therefore introduce several current methods and discuss their pros and cons.

Yeast-two hybrid system

The Yeast-two hybrid (Y2H) method is based on a screening approach using a set of modified proteins (Fields and Song, 1989). The physical basis of Y2H is the reconstitution of multi-domain transcription factor (such as GAL4 or ADE2). In particular, a protein-encoding cDNA of interest is cloned into a bait vector, and fused with the DNA binding domain of the multi-domain transcription factor. A second cDNA encoding a potentially interacting protein is cloned into a prey vector and fused to the transcription factor's activation domain. Subsequently, the two yeast strains carrying the bait and prey hybrid proteins in plasmids are mated. If the bait and prey proteins interact, a functional transcription factor is reconstituted leading to the transcription of a reporter gene such as *lacZ* encoding for β -galactosidase. In a high-throughput mode, whole libraries of bait and prey vectors can be screened for interactions. Thus, a main advantage of this approach is that it provides a platform for the generation of large-scale protein-protein interaction networks that need not to be biased toward known interactions. Another attractive feature of Y2H is that weak transient interactions can be detected. A disadvantage, however, is that interacting proteins have to be located to the nucleus for detection which can cause difficulties for the examination of membrane proteins. Another crucial problem in Y2H systems is that interactions are measured outside the native surrounding (except for yeast proteins) and thus potential protein modification specific to e.g. humans may not take place. Moreover, the false positive rate for Y2H screens might be considerable and frequently exceed the estimated true positive rates (Hart *et al.*, 2006).

Nevertheless, the Y2H system has been an important method in the field of interactomics and been extensively used in the generating the protein interaction

maps for various model organism (Schwikowski *et al.*, 2000; Uetz *et al.*, 2000; Ito *et al.*, 2001; Giot *et al.*, 2003; Li *et al.*, 2004). Recently, the Y2H system was applied in two large-scale studies to screen for protein interaction in human (Rual *et al.*, 2005; Stelzl *et al.*, 2005). Stelzl and co-workers used a combination of fetal brain cDNA library and a set of full length open reading frames (ORFs) to create over 11 000 Y2H clones. Applying a pooling approach, more than 25 million protein pairs were tested resulting in an identification of over 3000 interactions. Independently, Rual and collaborators performed an Y2H screen based on more than 8000 ORFs and detected ~2800 interactions. Together, both projects identified over ~5,500 new protein interactions, of which a selected sub-set was experimentally validated.

Affinity purification and mass spectroscopy approaches

Affinity purification using tagged bait proteins in combination with mass spectrometry have been performed to generate large-scale protein interaction maps in different organisms (Gavin *et al.*, 2002; Ho *et al.*, 2002). For this approach, proteins of interest are fused with a specific tag and expressed in cells where they form native complexes with other proteins. Using the tags, these bait proteins are precipitated after cell lyses and the composition of the obtained complexes subsequently determined by mass spectrometry. Note that identified interaction partners do not necessarily directly interact with the bait protein, but might be linked to the bait by indirect interactions. This has important consequences in cases when we like to represent the obtained complex by a set of binary interactions in order to facilitate the analysis of the global network. Since the internal structure of the complex is generally not known, two generic models are frequently used for this task: 1) the *matrix* model assumes that all proteins in a complex interact with each other. Especially for large complexes, this assumption leads to a large number of direct interactions and thus potentially implies also a large number of false positives. 2) The *spike* model entails that the only direct interactions are those between the bait and the co-precipitated proteins neglecting any other internal structures.

Using this methodology, Gavin & colleagues created first protein interaction maps for yeast (Gavin *et al.*, 2002; Ho *et al.*, 2002). A general conclusion of these studies was that a considerable part of the proteome can be organized in protein complexes. For instance, Gavin and colleagues could identify over 200 mostly novel complexes for an initial set of over 1700 tagged yeast proteins. Detailed analysis of the obtained

dataset showed resulted most protein complexes are linked constituting a higher order network beyond the level of binary interactions

More recently, Ewing and co-workers utilized a similar approach to capture the interactions between human proteins involved in important biological processes or associated with diseases (Ewing *et al.*, 2007). For the study, they selected an initial set of 338 bait proteins, which were functionally enriched in biological processes such as protein modification, cell cycle, transcription and signal transduction, or were associated with diseases such as breast cancer, colon cancer, diabetes or obesity. To identify interaction partners for this set of proteins, a large-scale immunoprecipitation experiment with subsequent mass spectrometry was performed. Despite starting with a rather small set of bait proteins, they detected ~24500 protein interactions using a human cell line (HEK293) for expression. To exclude redundant or potential false positive interactions, the detected interactions were computationally processed using statistical analysis and confidence measures. The final set was comprised of approximately 6500 novel interactions between over 2200 proteins including many novel and pathway-related associations.

Literature curation and text-mining

Besides high-throughput experimental approaches, the numerous small-scale experiments described in the literature can be exploited to create large-scale protein interaction maps. Tapping into the wealth of published experiments, information on protein interactions is systematically extracted from the literature either by human experts or text-mining algorithms. The advantages of such procedure are that it is not biased towards a particular experimental technique and that the chartered interactions are determined under a broad range of conditions and protocols. Characteristic disadvantages are the inherent difficulty to estimate the false positive rate and the biases towards highly studied proteins.

For human, several research groups have followed this strategy to create large-scale protein interaction maps (Aranda *et al.*, ; Bader *et al.*, 2003; Salwinski *et al.*, 2004; Breitkreutz *et al.*, 2008; Prasad *et al.*, 2009) (table 2.1). As a prime example, the Human Protein Reference Database (HPRD) has been established for manual curation of interactions described in the published literature. Currently, it includes almost 38000 interactions constituting the largest set of literature-curated human

protein interactions. Obviously, the selection of the literature for curation influences directly the resulting interaction map. In the context of HPRD, for instance, the curation efforts are focused on proteins which are disease-associated. Recently, several other major databases have joined forces to capture all interactions described in the literature. The established International Molecular Exchange (IMEx) consortium is expected to contribute to coordination and thus minimizing the efforts of single databases (Orchard *et al.*, 2007). However, it should also be noted that such procedure might eventually reduce independent curation which is often an asset in judging the quality of interactions.

Since the manual curation of scientific literature is highly labour-intensive, computational text-mining approaches have become a cost-effective alternative. One of the simplest text-mining strategies is to count how often a pair of protein names occurs in the same scientific text. If this count is higher than the one expected by chance, we might infer that both proteins are functionally associated and potentially interacting. Although this approach has recently applied to construct from Medline abstracts a network including ~3,700 human proteins, one should keep in mind that the deduced interactions need not be physical (Ramani *et al.*, 2005) (table 2.1). More elaborated computational search algorithms capturing the semantics and syntax might give us more precise interaction networks (Hoffmann and Valencia, 2004).

Computational prediction

Alternative to the large-scale experimental and manually-curated approaches, *in silico* prediction method have been used to build large-scale protein-protein interaction maps (Lehner and Fraser, 2004; Brown and Jurisica, 2005; Persico *et al.*, 2005). This strategy is based on the assumption that protein interactions are likely to be evolutionarily conserved between orthologous proteins from different species, and therefore interaction between proteins in lower organisms can be extrapolated to their human orthologs (O'Brien *et al.*, 2005). A main advantage of this method is that it is entirely computational and thus enables rapid and cost-effective construction of protein-protein interaction maps. Disadvantages are that it is purely predictive in nature and false positive can arise through erroneous mapping to human orthologs or that the interactions are simply lost during evolution.

Two different groups applied InParanoid algorithm (Berglund *et al.*, 2008) to find human orthologs from various model organism (Lehner and Fraser, 2004; Persico *et al.*, 2005). Lehner & colleagues used seven experimental and four computationally predicted protein-interaction maps from three model organisms *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Caenorhabditis elegans*. An interaction was predicted if both interaction partners from a model organism have one or more human orthologs. Using this strategy, authors were readily able to generate a human interaction network comprising ~71,000 interactions between ~6,000 human proteins. The generated map was further scored using co-expression measures and Gene Ontology annotation to identify a core network of 9641 interactions between 3503 unique proteins.

A similar study was undertaken by Perisco and co-workers. Besides using interactions from lower organisms, they analysed the domain composition of human proteins to refine the predictions of interactions. In contrast, Brown and Jurisica (Brown and Jurisica, 2005) applied a BLASP and reciprocal best-hit approach to extrapolate interactions between organisms. They created first an integrated interaction dataset from various model organisms and mapped it to human orthologs by BLASTing proteins from each model organism against SWISS-Prot database filtered for human proteins. As a next step, each top BLAST hit (with an E-value $<10^{-5}$) was BLASTed back against the set of all model organism protein sequences. A protein (with an E-value $< 10^{-5}$) was then considered as a potential ortholog, if it matched the original query protein in reverse direction. Following this method, they generated a human PPI map containing ~25000 interactions between ~4000 proteins (table 2.1).

2.2.2 Topological properties of PPI network

Cellular functions are attributed to interactions among many molecules. In a cell, these molecules are organized in a complex manner and together they form a network, whose complexity reflects to a large degree those in other types of networks, such as the social networks, internet, or organizational networks. This astonishing semblance indicates that most complex networks in nature might generally be controlled by the similar universal laws, and therefore the learning from these well-studied non-biological systems may be extrapolated to cellular networks to study the complex association that regulate the molecular functions (Barabasi and

Oltvai, 2004). The graph-based theoretical approaches offer possibilities to study these relationships. For example, modelling molecular networks with graph can thus help us to visualize how molecules in a biological system work together in concerted manners. Graph-theoretical measurement such as connectivity or centrality of a node within network might indicate their functional importance as a hub protein (He and Zhang, 2006). To this end, molecular network has to be first converted into a mathematical graph. For protein interaction networks, for instance, proteins are commonly represented as nodes and their physical interactions as undirected edges. The resulting graphs can be analyzed using various graph-theoretical measures described in the following sub-sections.

Scale-free networks and hubs

A fundamental characteristic of a node in a mathematical graph is its degree i.e. the number of connections or edges that a node has to other nodes. The degree distribution $P(k)$ of a network is then defined as fraction of proteins in the network with k interactions. It is an important feature to distinguish different network classes. Of special importance here is the power-law distribution ($P(k) \sim k^{-\gamma}$) which is characteristic for the class of scale-free networks. It has been shown that such network architecture is more robust against random failure of single components (Barabasi and Albert, 1999; Albert et al., 2000; Han et al., 2005). A consequence of the scale-free topology is the emergence of so-called network hubs i.e. highly connected nodes. Hubs are of particular importance for the network integrity and were associated with essential proteins (Jeong et al., 2001; Wuchty, 2004; He and Zhang, 2006). However, recent studies have demonstrated that the essentiality of protein hubs is better explained by the number of shortest paths going through them (Yu et al., 2007).

Small-world effect

A common feature of many networks is that most of the nodes within network can be reached from every other by a small number of edges (Watts and Strogatz, 1998). The path length between two proteins is called as shortest path length. This feature was first observed in social networks, but other networks such as World Wide Web and the metabolic network also demonstrate this property

Cluster coefficient

Cluster Coefficient is a fundamental measurement that assesses the degree to which nodes tend to cluster together (Watts and Strogatz, 1998). It is defined as $C = 2n/m(m-1)$ where n is the number of links between m neighbors. A large clustering coefficient indicates that neighbors of a node are likely to interact to each other. The clustering coefficient of a network, C , is the average of C_i over all vertices. A function $C_{(k)}$ defines the average clustering coefficient over the vertices with degree k . When a network is modular and hierarchical, the clustering function follows a power law $C_{(k)} \sim k - 1$ (Barabasi and Oltvai, 2004).

Centrality and betweenness

Centrality seeks to describe the relative importance of a protein within the graph, by evaluating the location within a network. Frequently, centrality of a node is defined by the number of shortest paths passing through this node. An alternative centrality measure is betweenness., nodes that occur on many shortest paths between other nodes tend to have higher betweenness than those that do not (Jeong *et al.*, 2001).

Community

A community structure describes a group of nodes that have a high number of edges within them, but low number of edges to nodes of other groups. This is also a common feature which exists in many real world networks including biological networks.

2.2.3 Databases for human protein interactions

Several human protein interaction databases have been established to help researchers to find and analyze interaction partners of their interest of protein (table 2.1). These databases can generally be divided into two different categories: The first one is based on the manual curation of published literature or datasets and includes e.g. the Human Protein Reference Database (HPRD) (Prasad *et al.*, 2009), Biological General Repository for Interaction Datasets (BioGRID) (Breitkreutz *et al.*, 2008), IntAct molecular interaction database (Aranda *et al.*), Database of Interacting Proteins (DIP) (Salwinski *et al.*, 2004), Biomolecular Interaction Network Database (BIND) (Bader *et al.*, 2003), Molecular INTeraction database (MINT) (Chatr-

aryamontri *et al.*, 2007) and MIPS Mammalian Protein-Protein Interaction Database (MPPI) (Pagel *et al.*, 2005). The other category of databases is based primarily on computationally predicted interactions; examples of such databases are the Online Predicted Human Interaction *Database* (OPHID) (Brown and Jurisica, 2005) and HomoMINT (Persico *et al.*, 2005). Currently, HPRD is one of the major sources for human interaction data and - as the name implies - dedicated to store data a variety of information for human proteins. Besides interactions, it also provides information on domain architecture, post- translational modifications, and disease association and biological pathways. Other databases e.g. BioGRID, IntAct, DIP, BIND, and MINT are the repositories for a more diverse set of organisms and provide access to interaction data for other model organisms such as yeast, worm and fly.

2.2.4 Tools for analysis and visualization of interaction networks

Besides the generation of interaction data, one major focus of interactomics is the development of tools for the analysis of complex networks. As outlined before, the representation of biological networks as mathematical graphs facilitates the computational analysis tremendously. At present, several types of tools supporting the computational examination of interaction networks are available. They can be broadly classified based on their implementation in standalone software and web-based analysis tools.

Popular examples of standalone solutions are the statistical programming environment R-Bioconductor (Gentleman *et al.*, 2004) and the Cytoscape bioinformatic platform (Shannon *et al.*, 2003). Whereas the first is generally widely applied in analysis of high-throughput data, the latter is dedicated to the visualization and interrogation of network structures. Both software tools provide many open source add-on packages for graph-theoretical analyses making them highly versatile in their application. However, users might be required to pre-compile data collection for investigation. An alternative standalone software is the Osprey package which provides links to the BioGRID database (Breitkreutz *et al.*, 2008).

Table 2.1: Overview of the currently available human protein-protein interaction maps

Resource	Proteins	Interactions	Method	References
MDC-Y2H	1703	3186	Y2H SCREEN	(Stelzl <i>et al.</i> , 2005)
CCSB-Y2H	1549	2754	Y2H SCREEN	(Rual <i>et al.</i> , 2005)
HPRD-BIN	8788	38800	LITERATURE	(Prasad <i>et al.</i> , 2009)
DIP	1085	1397	LITERATURE	(Salwinski <i>et al.</i> , 2004)
BIOGRID	7953	24624	LITERATURE	(Breitkreutz <i>et al.</i> , 2008)
INTACT	7273	19404	LITERATURE	(Aranda <i>et al.</i>)
BIND	5286	7394	LITERATURE	(Bader <i>et al.</i> , 2003)
COCIT	3737	6580	TEXT MINING	(Ramani <i>et al.</i> , 2005)
REACTOME	1554	37332	LITERATURE	(Matthews <i>et al.</i> , 2009)
ORTHO	6225	71466	ORTHOLOGY	(Lehner and Fraser, 2004)
HOMOMINT	4127	10174	ORTHOLOGY	(Chatr-aryamontri <i>et al.</i> , 2007)
OPHID	4785	24991	ORTHOLOGY	(Brown and Jurisica, 2005)

To offer more convenient interfaces for the network analysis, especially for researchers less acquainted with bioinformatical analyses, various web servers have been implemented, some of which I exemplarily introduce in the following. A fairly generic collection of algorithms for graph analysis is provided by the Network analysis tools (NeAT) server (Brohee *et al.*, 2008). Users can perform basic operations on graphs as well as detect cluster structures. More focused is the hub objects analyser (Hubba), which comprises several algorithms to identify highly connected proteins in interaction networks (Lin *et al.*, 2008). Notably, the interaction data has to be supplied by the user for both tools. To relieve users from the burden of collecting and pre-processing interactions, several web-servers additionally provide precompiled data sets. Popular examples of this kind of integrative platforms are VisANT (Hu *et al.*, 2009) and BiologicalNetworks (Baitaluk *et al.*, 2006).

2.2.5 Application of interactomics

Large-scale protein interaction network appears to be attractive resources in many research fields. Traditionally, they have been applied for the functional characterization of an unknown protein, based on the assumption that the two interacting partners are likely to be involved in a same biological process, where they perform similar functions. Several algorithms, which use this principle, have been developed, and are classified as direct methods (Sharan *et al.*, 2007). Examples of such direct methods are neighbourhood counting, graph-based methods that are based on topological properties of a network, integrative systems biology that integrates the information from multiple sources in the combination with machine-learning algorithms which use features of a known protein to characterize the functions of its partner (Sharan *et al.*, 2007). In contrast to direct methods, module-assisted approaches uses the topological properties of the network to identify the group of proteins or modules whose interactions can be attributed to certain biological function (Hartwell *et al.*, 1999). The assumption is here that the identified module might contain both known and unknown proteins, and therefore help us in the prediction of the function of unknown proteins by using the function of its known partners. Popular examples of module-detection methods include hierarchical clustering (Rives and Galitski, 2003), graph-based clustering (Spirin and Mirny, 2003; Przulj *et al.*, 2004), network topology (Bader and Hogue, 2003), or data integration in the combination with machine learning algorithms (Tanay *et al.*, 2004; Kelley and Ideker, 2005). Modularity aspect will be elaborated in more detail in chapter 5.

Other areas of interactomics are the identification of domain-domain interactions (Guimaraes *et al.*, 2006), network motifs (Milo *et al.*, 2004), comparison between model organism and human (Gandhi *et al.*, 2006), and the several applications in disease research detailed in the following section.

2.2.6 Role of the PPI networks in disease research

PPI networks have been recently extensively applied in disease research, due to fact that several diseases, such as cancer, neurodegenerative etc., do not result from one protein, instead, caused by misregulation of interactions between many proteins. Therefore, characterization of those proteins that modulate these interactions may

provide new insights in the pathogenesis of the disease. Moreover, functional and structural analysis of these modulators may give us possible solution and basis for the new treatments. Several studies have been performed using PPI networks to identify the genetic modifiers in disease such as Chorea Huntington (Goehler et al., 2004), Ataxis (Lim et al., 2006), inflammation (Calvano et al., 2005) and various types of cancers (Wachi et al., 2005; Jonsson and Bates, 2006; Chuang et al., 2007; Pujana et al., 2007). In the next following sections, I will describe the role of PPI networks in disease research using several aspects.

Topological analysis of disease genes

One of the first questions addressed by network-based approaches in disease research is also one of the most intriguing: What makes a gene to a disease gene? Although such naïve question may be somewhat puzzling at first, it makes naturally sense in network biology to ask whether disease-associated genes have characteristic properties within interaction networks. To address this question, graph-based methods can be applied to study network properties of disease genes. Several research groups have applied such concepts to reveal the graph-theoretical properties and the role of cancer genes in human protein interaction networks (Wachi *et al.*, 2005; Jonsson and Bates, 2006; Hernandez *et al.*, 2007; Platzer *et al.*, 2007). For the analysis, the set of disease-associated genes has first to be determined, for which commonly databases or microarray studies are used. As a second step, a disease network is created by integrating the disease genes products (i.e. proteins encoded by disease-associated genes) with available large-scale protein interaction networks. Finally, the topological properties (e.g. degree distribution, centrality) of disease genes within this network are computed and compared to those of genes that have not been associated with disease.

Wachi and co-worker applied the above outlined strategy to study the centrality of genes that are differentially expressed in cancer (Wachi *et al.*, 2005). Their analysis showed that upregulated genes tend to be highly connected and more centrally located in the network compared to randomly selected genes. Downregulated genes tended to be also more highly connected but not significantly. Furthermore, they did not show an increased centrality. Based on their findings, the authors suggested that a core set of central genes has to be activated during the course of carcinogenesis.

Similar results were reported in a separate topological analysis performed by Jonsson and Bates (Jonsson and Bates, 2006). Results indicated that the cancer proteins show higher degrees than non-cancer proteins. Cancer proteins also tend to function as central hubs, reflecting their role as a key player in protein-protein interaction network. Clustering analysis additionally showed that cancer proteins, on average, are more frequently located in the interfaces between clusters indicating an enhanced role in the coordination of different cellular processes.

A similar approach was undertaken by Goni & colleges for the analysis of the topological properties of genes associated with neurodegenerative disorders, such as Multiple Sclerosis (MS) brain and blood and Alzheimer Disease (AD) brain and blood (Goni *et al.*, 2008). The aim of this study was the comparative assessment of the centrality related features such as degree and betweenness between disease genes products (seed proteins) and its neighbours. Comparative topological analysis of seed proteins in all four networks displayed a lower average degree with respect to the degree of their PPI neighbors in both diseases and in both tissues. Remarkably, seed proteins showed a higher betweenness in AD-brain and MS-blood networks, and a low correspondence between their degree and betweenness in all four networks. These findings suggested that critical proteins in disease pathogenesis are not highly connected, but tend to be located in bottleneck regions, and therefore less extensively connected proteins might be more appropriate therapeutic targets than hyper-connected ones, at least in neurodegenerative diseases.

Network-based prediction of new disease genes

A second area in which protein interaction networks have been utilized in disease research is the identification of new disease-associated genes. The rationale behind these investigations is that interacting proteins are likely linked to the same or similar phenotype. A leading example is Fanconi anemia, a genetic disease, for which seven of the nine associated proteins form a physical complex involved in DNA repair. Although interaction data can provide a suitable first basis for *de novo* identification of disease-causing genes, additional information has commonly been utilized to improve specificity.

For many years, genetic linkage studies were the most potent approach to find new disease-causing genes. A major difficulty, however, is to pick the right gene within

rather extended chromosomal regions that have been linked to a disease. Oti *et al.* showed that this task can be considerably facilitated using protein interaction data (Oti *et al.*, 2006). For genetically homogenous diseases, they predicted new disease associations when genes fell within an identified susceptibility locus and have protein interactions with a gene known to cause this disease. This simple method of data integration led to a 10-fold increased specificity compared to randomly selected candidate genes at the same locus. Notably, Oti *et al.* also deduced that protein interactions added as much information as localization to the prediction accuracy. In a similar study, Franke *et al.* extended the protein interaction network by including microarray and gene annotation to generate a functional interaction network (Franke *et al.*, 2006). Also, new candidate genes were identified in the larger network neighbourhood of known disease genes, avoiding the restriction to direct interactors only.

One requirement of these studies is that we have to know already a set of genes associated with a certain disease. This set can be then used to ‘anchor’ a disease in the human interactome. If however no such genes are known, this approach cannot be used. To overcome this limitation, Lage *et al.* catalogued human phenotypes in a computational tractable manner (Lage *et al.*, 2007). Their motivation was that similar diseases might share the same molecular basis. Having defined a score for the similarity of phenotypes, information for a specific disease can then be deduced from similar diseases. Thus, candidate genes can be predicted even if no other gene associated with the specific disease is known yet. For prediction, Lage *et al.* integrated human protein interaction with linkage data in a similar manner as Oti *et al.* and Franke *et al.* Using an *in silico* pull-down approach and the similarity of phenotypes, they extracted known and new complexes and predicted several novel candidate disease genes involved in disorders such as cancer, Alzheimer’s, diabetes and coronary heart diseases. Detailed analysis for epithelial ovarian cancer lead to the identification of a new candidate gene, Fanconi anemia group D2 protein (*FANCD2*) placed in a complex with breast cancer type 1 susceptibility protein (*BRCA1*) and breast cancer type 2 susceptibility protein (*BRCA2*). This protein has been associated with different types of cancer, but not with epithelial ovarian cancer so far.

A conceptually similar network-based modelling approach was applied by Pujana *et al.* to predict new candidate genes involved in breast cancer (Pujana *et al.*, 2007). They assumed that genes, which are functionally related or showed conserved co-expression across species, might cause a similar phenotype. To test their hypothesis, they created a cancer-specific network with four known breast cancer-associated genes: *BRCA1*, *BRCA2*, *ATM*, and *CHEK2*. Neighbours of each reference gene set were further ranked using scoring system based on coexpression, phenotypic similarity, and genetic or physical interactions among orthologs of the proteins in other species. They identified a new gene (*HMMR*) that was found to be associated with an increased risk of breast cancer.

Network-based prediction of cellular processes

In addition to prediction of novel disease-associated genes, interaction networks were also employed to unravel disease-related molecular processes. As one example, Chuang *et al.* applied a network-based classifier to identify sub-networks as markers for breast cancer prognosis (Chuang *et al.*, 2007). To find the sub-networks, they mapped the gene-expression profiles of metastatic and non-metastatic patients on a human protein–protein interaction network. Subsequently, they computed activity scores of all associated members to rank the sub-network as a whole. Their finding showed that high scoring sub-networks were enriched in many cancer-related biological processes such as apoptosis, proliferation, tissue remodelling, signalling and survival. Their analysis also indicated that identified modules were more reproducible than individual genes selected without network information, and that they achieve a higher accuracy in the classification of metastatic versus non-metastatic tumours. Another advantage of this approach is that it also captures those genes which may have not been detected based on gene expression data alone. Such non-differentially expressed genes could be integral part of a complex and be required for connecting high scoring proteins in a sub-network. In fact, Chuang *et al.* found that a large number of the identified network structures contained at least one protein that was not significantly expressed in metastasis while most of them served as a bridge between high scoring proteins in a sub-network. This integration provides the opportunity to analyze the relationships between members of the complexes, and also increases the accuracy of the overall prediction.

In another study, Baranzini and colleagues applied the similar approach for identifying sub-networks involved in multiple sclerosis (MS), a neurodegenerative disorder, (Baranzini *et al.*, 2009). For the identification of sub-networks, disease genes were collected from two genome-wide SNP markers association studies in MS with nominal evidence of association ($P < 0.05$), and further superimposed on a human protein interaction network. Their findings suggested that several identified sub-networks were found to be over-represented in immunological pathways including cell adhesion, communication and signalling, neural pathways, and synaptic potentiation. Especially, authors claimed to report for the first time the potential involvement of neural pathways in MS susceptibility. Such findings are crucial, since mechanisms of MS is still very much under investigation, and network-based approaches may help identify different, and even unrelated, biological processes as responsible for disease pathogenesis (Baranzini *et al.*, 2009).

Common genes to many disorders

PPI networks have also been applied for identification of common genes to many disorders (Goh *et al.*, 2007). Notably, for this application, Barabasi and colleagues integrated human PPI network with disease phenotype data, obtained from OMIM, to create first ever human diseasome. Using a bipartite graph-approach, they created a human disease network showing many-to-many relationship between genes and disorders. Such genome-wide network-based approaches are essential for biomedical research and may help to enhance our understanding of the genetic links between disorders and disease genes.

In summary, the described network studies give us a first overview about the role of PPI networks in biomedical research. Nevertheless, care has to be taken in interpretation as current interaction networks often show divergence in structure due to different methods used for their assembly.

3 COMPARISON OF LARGE-SCALE MAPS OF THE HUMAN PROTEIN INTERACTOME

This chapter is an extended version of following three papers and some unpublished data:

References:

Gautam Chaurasia, Hanspeter Herzel, Erich E. Wanker and Matthias E. Futschik, Systematic functional assessment of human proteins-protein interaction maps, *Genome Informatics*, 17(1), 36-45, 2006.

Matthias E. Futschik, **Gautam Chaurasia** and Hanspeter Herzel, Comparison of Human Protein-Protein Interaction Maps, *Bioinformatics*, 23(5):605-611, 2007.

Matthias E. Futschik, Anna Tschaut, **Gautam Chaurasia**, and Hanspeter Herzel (2007) Graph-Theoretical Comparison Reveals Structural Divergence of Human Protein Interaction Networks, *Genome Informatics*, 18, 141-151, 2007.

3.1 Introduction

Protein-protein interactions are essential for the vast majority of cellular processes. This pivotal role has led to intensive studies of large-scale mappings of protein-protein interaction networks. In last decade, several strategies to catalog the human interactome have been proposed and pursued (Lehner and Fraser, 2004; Brown and Jurisica, 2005; Persico *et al.*, 2005; Rual *et al.*, 2005; Stelzl *et al.*, 2005; Ewing *et al.*, 2007; Prasad *et al.*, 2009). To date, most of these approaches can be assigned to one of the following type: i) literature-based, ii) orthology-based and iii) high-throughput yeast-two-hybrid-based (Y2H) or mass spectrometry-based interaction maps. Each of these approaches has its own known strengths and weaknesses. However, how the resulting interaction maps are affected is less clear.

At the same time, first attempts in biological and medical research to systematically utilize interaction data sets have been undertaken (Goh *et al.*, 2007; Yildirim *et al.*, 2007; Ideker and Sharan, 2008). Although the results have been promising, it may not be denied that the number of successful efforts to exploit protein interaction maps is still limited. An important reason for this situation might be the missing integration of separated maps. Clearly, it has been tempting to immediately start the unification of hitherto disconnected interaction maps. However, it has been also evident that quality and reliability of diverse interaction maps has to be stringently assessed first, especially if their methods of generation are distinct. A comparison was therefore timely because efforts towards reciprocal adjustment and updating of currently separated interaction databases can be expected or are already in process. While such integration facilitates the data access for researchers, it may cloud possible biases of the distinct mapping approaches in single databases.

Comparative assessments of interaction maps have already been performed for *S. cerevisiae* regarding the overlap, coverage and reliability (Bader and Hogue, 2002; von Mering *et al.*, 2002). von Mering *et al.* performed a comparative analysis to measure accuracy and as well as to identify biases, strengths and weaknesses of each method used for generating yeast interaction data. Their analysis indicated that currently available interaction data are highly divergent, mainly due to the presence of high false-positive rate, and some methods may have selection and detection biases, resulting in complementarities between the methods. Bader & colleagues (Bader and Hogue, 2002) pointed out that the low overlap could arise from either a high false-

discovery rate or high false-negative rate. Their analysis showed that 25% to 45% of the reported interactions in yeast, worm and fly are likely false positives, in which membrane proteins have higher false-discovery rates on average, and signal transduction proteins have lower rates.

Whereas most of the previous studies were focused on the comparative evaluation of yeast PPI network, such extensive comparisons were still lacking for human protein interaction maps. A simple extrapolation of the results from yeast to human maps might be misleading regarding the different underlying biology and mapping approaches. Therefore, a systematic evaluation of currently available human protein-protein interaction maps is warranted to gain a better understanding and insight into their functional composition and topological structure. To this end, a comparative analysis of currently available eight human interaction maps was performed in this work. For the analysis, first, overall number of common proteins and interactions were examined in the analyzed maps. To investigate the composition of interaction maps with regard to protein function, biological processes and cellular location, Gene Ontology (GO) annotation database was utilized. Next, two different approaches were introduced to subsequently assess the functional coherency of interaction maps. Finally, the topological properties of each interaction map and properties of hubs proteins in different networks were analyzed and compared.

This chapter is organized as follows: section 3.2 gives brief details about materials and methods. Section 3.3 presents several results. Finally, in section 3.4, I conclude this chapter by discussing the importance of this study.

3.2 Materials and Methods

3.2.1 Assembly of protein-protein interaction maps

For comparison, eight interaction maps were chosen that were considered as representative for different approaches listed above: three literature-based, three orthology-based and two Y2H-based maps (table 3.1). This study was aimed to restrict the analysis of binary interactions to facilitate a direct comparison between Y2H-based and alternative approaches. As sources for the assembly of the first two literature-based interaction maps, the Human Protein Reference Database (HPRD) (Prasad *et al.*, 2009) and Biomolecular Interaction Network Database (BIND) (Bader

et al., 2003) were chosen. The third literature-based interaction map comprises protein interactions found by Ramani and co-workers using a text-mining approach (Ramani *et al.*, 2005). It is derived by computational identification of co-cited proteins in Medline abstracts. Thus, this map is distinguished from HPRD and BIND as it is computationally generated, but it is similarly based on literature search. This map was referred to as the COCIT map. Notably, it does not distinguish between physical and functional interactions and lacks self-interactions due to the computational approach taken.

A first orthology-based map was assembled from interactions predicted for human proteins by Lehner and Fraser. These so-called interlogs are based on interactions in yeast, worm and fly. Here only the set of core interactions were used, since they were shown to be more reliable (Lehner and Fraser, 2004). The map constructed is referred to as the ORTHO map in this study. Two further orthology-based maps were constructed based on predicted interactions in Online Predicted Human Interaction Database (OPHID) (Brown and Jurisica, 2005) and HOMOMINT database (Persico *et al.*, 2005).

Finally, the two large-scale Y2H screens for human protein interactions by Stelzl *et al.* and Rual *et al.* (Rual *et al.*, 2005; Stelzl *et al.*, 2005) were included in the comparisons. They are referred to as MDC-Y2H and CCSB-H1, respectively. For comparative analysis, all proteins were mapped to their corresponding EntrezGene ID either using the original ID provided with the data or utilizing Ensmart (Kasprzyk *et al.*, 2004) and HGNC (Eyre *et al.*, 2006).

3.2.2 Overlap of interaction maps

As protein interaction maps are constituted by proteins and interactions, comparisons can either be performed with regard to proteins or to interactions. For pair-wise comparison of interaction maps (A , B) in regard to proteins included, common proteins between the two maps were identified. This defines the intersection $P_{AB} = P_A \cap P_B$ where $P_{A,B}$ are the sets of proteins in map A or B respectively. Subsequently, the intersection was normalized with respect to the number of proteins in A and B ($P^A_{AB} = P_{AB} / P_A$; $P^B_{AB} = P_{AB} / P_B$). The average of P^A_{AB} and P^B_{AB} was referred to as (relative) protein overlap between A and B .

Table 3.1: Overview of protein-protein interactions maps compared. Numbers were based on the proteins which could be mapped to their corresponding EntrezGene ID. The number of proteins and interactions before mapping and further information regarding the interaction maps can be found in the supplementary material (Appendix table A.1).

Network	Proteins	Interactions	Self-interactions	Percentage of self-interactions	Average Degree	Method	Reference
MDC-Y2H	1703	3186	36	1.1	1.9	Y2H-ASSAY	(Stelzl <i>et al.</i> , 2005)
CCSB-H1	1549	2754	143	5.1	1.8	Y2H-ASSAY	(Rual <i>et al.</i> , 2005)
HPRD	5908	15658	679	4.2	2.7	LITERATURE	(Prasad <i>et al.</i> , 2009)
BIND	2677	4233	614	13.5	1.7	LITERATURE	(Bader <i>et al.</i> , 2003)
COCIT	3737	6580	0	0	1.8	LITERATURE	(Ramani <i>et al.</i> , 2005)
OPHID	2284	8962	0	0	3.9	ORTHOLOGY	(Brown and Jurisica, 2005)
ORTHO	3503	9641	199	2.0	2.8	ORTHOLOGY	(Lehner and Fraser, 2004)
HOMOMINT	2556	5582	471	8.1	2.3	ORTHOLOGY	(Persico <i>et al.</i> , 2005)

Comparison of common interactions was also done in similar manner. For normalization of the intersections, however, only the number of interactions between common proteins was used. This procedure avoids confounding the interaction overlap with the protein overlap, as otherwise a small protein overlap would always lead to a small interaction overlap. Thus, the interaction overlap is defined as the average percentage of shared interactions between common proteins.

Although intuitive, the described measures of interaction overlap have the drawback, that they only assess concurrence of the observed interactions, but not of missing interactions. Therefore, I additionally used a log-likelihood ratio (*LLR*) score introduced previously to compare of interaction sets (I_1, I_2) (Lee *et al.*, 2004). In contrast to simple overlap measures, the *LLR* evaluate both concurrences in observed as well as in missing interactions. It is defined as:

$$LLR(I_1, I_2) = \ln\left(\frac{P(I_1 | I_2)}{P(I_1 | \sim I_2)}\right)$$

Where $P(I_1|I_2)$ is the probability of observing an interaction in I_1 conditioned on observing the same interaction in I_2 . Respectively, $P(I_1|\sim I_2)$ is the probability of observing an interaction in I_1 conditioned on not observing the same interaction in I_2 . A large LLR indicates high similarity of the set of interactions whereas the LLR is expected to be zero for comparison of sets of random interactions.

To assess the statistical significance of interaction overlaps observed between maps, two types of permutation tests were used (Balasubramanian *et al.*, 2004). The first test is based on repeated random re-labelling of nodes (proteins) before intersection (node label permutation), while the second test is based on the permutation of edges before intersection (edge permutation). The probability of obtaining at least the same number of interactions as in the observed intersection provides the significance. Although differing in their algorithm, the two permutation tests usually give similar results.

3.2.3 Gene Ontology analysis

To investigate the functional composition and coherency of interaction maps, I utilized the Gene Ontology (GO) annotation database as it presently provides the most comprehensive functional annotation for human genes (Ashburner *et al.*, 2000). GO includes gene annotations regarding molecular function (MF), biological process (BP) and cellular component (CC) using a defined hierarchical ontology.

Functional composition of interaction maps

To determine the statistical significance that proteins of specific GO category are overrepresented in a map, Fisher's exact test was used. It is based on the hypergeometric distribution and was employed in our study as follows: The significance of observing k proteins of a chosen category can be derived from the probability P to observe k or more proteins of the category if I proteins would be randomly drawn:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{l-i}}{\binom{N}{l}}$$

Where M is the total number of proteins attributed to the category, N is the total number of proteins annotated and l is the number of proteins in the corresponding map. Likewise, the significance can be calculated assessing the underrepresentation of GO categories in maps. Since multiple testing was performed, p -values were converted to false discovery rate (FDR) using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Functional coherency of interaction networks

For the analysis of functional coherency of protein interactions, a similar approach was applied as proposed by von Mering and co-workers (von Mering *et al.*, 2002). First, the frequencies k_{mn} were calculated i.e. a protein assigned to GO category m interacts with a protein assigned to GO category n . If only proteins of the same category interact, a diagonal matrix should emerge. Hence, the functional coherency of a network can be assessed by inspection of the interaction matrix. However, this procedure underlies the assumption that proteins are assigned to exactly one category. Notably, this does not hold for GO annotations where proteins are frequently assigned to multiple categories. To overcome this short-coming, the current approach was modified, and frequencies k_{mn} were compared between the observed interaction matrix and the interaction matrix expected for corresponding random scale-free networks. Log odds were used to measure the deviation of the observed frequency distribution k for an expected base line distribution f^0 :

$$LOD(f_i, f_i^0) = \log_2 \frac{f_i}{f_i^0}$$

Alternatively, association between the GO annotation and interaction maps can be examined based on the similarity of GO terms assigned to interacting proteins (Jansen *et al.*, 2003). Here, the similarity of GO terms was quantified by calculating their shared path lengths (from the root term) within the GO tree. Similar GO terms are expected to have large shared paths. To test the significance, random graphs of conserved degree distribution were generated. The distribution obtained for the original network was subsequently compared to those obtained for the randomized

networks and log odds computed. The analysis was performed in the R language using the Bioconductor environment including the packages *graph*, *GO*, *GOStats* and *GraphAT* (Balasubramanian *et al.*, 2004; Gentleman *et al.*, 2004; Carey *et al.*, 2005).

3.2.4 Graph analysis

Protein interaction maps can be converted to graphs with proteins as nodes and interactions as links or edges. Graphs can be characterized using a variety of graph-theoretical measures. To scrutinize the network properties of each map, several topological measurements such as connectivity, small-world property, degree-distribution, cluster coefficient and hubs proteins, were computed and compared. More detailed definition of these measurements can be found in Chapter 2 section 2.2.2.

To avoid artefacts, self-interactions were excluded in the graph-theoretical analysis and all calculations were performed based on the largest connected graph for each map. The significance of the results was assessed by comparison to two background network models: i) Random graphs with the same number of nodes and interactions, but without conservation of the degree distribution. ii) Random graphs with conservation of number of nodes and interactions as well as of the degree distribution. Such graphs were constructed using the original networks and repeated random exchange of interactions: Edge between node A and B (A-B) and between C and D (C-D) will be changed to A-D and B-D, if such edges are not present yet (Maslov and Sneppen, 2002).

3.3 Results

In total, 57095 interactions could be mapped between 10769 proteins uniquely identified by their EntrezGene ID. For most interaction maps, the majority of proteins could be assigned to their corresponding EntrezGene IDs (Appendix table A.1). The interaction maps differ considerably in size by a factor up to five (table 3.1). The largest map with over 15000 binary interactions was derived from HPRD. The ORTHO and OPHID maps were similarly sized (both including around 9000 interactions) followed by COCIT, HOMOMINT and BIND each incorporating around 5000 interactions. The smallest interactions maps resulted from the Y2H assays (MDC-Y2H: 3186, CCSB-H1: 2754). The average number of interactions per protein

are similar for different networks ranging between 1.7 (BIND) and 3.9 (OPHID). Given previous estimates for the average number of interactions of 3-10 (von Mering *et al.*, 2002; Bork *et al.*, 2004), all the interaction maps compared are likely to be highly unsaturated. A more distinguishing feature between the maps is the percentage of self-interactions included differing more than ten-fold and ranging from 1.1% (MDC) to 13.5% (BIND). Notably, no homodimers were recorded in COCIT and OPHID.

3.3.1 Common proteins and interactions

For direct pair-wise comparison of maps, I first calculated the number of common proteins and interactions between the interaction maps produced (figure 3.1A). This showed that an ample number of proteins can be found in multiple maps. Whereas the majority of proteins are assigned to at least two maps (60%), only a very limited number is included in more than half of the maps (less than 6%). The number of proteins found in all maps is vanishingly small: Only a total of 10 proteins (0.1%) fulfilled this criterion. All in all, however, a substantial part of their protein space is shared between single maps.

A smaller overlap is observed if interactions of different maps are compared (figure 3.1B). The vast majority of interactions (over 90%) can be found in only one map while missing in the others. None common interaction exists in more than six networks and only a minute number of eight interactions are shared by five maps. Interestingly, three of these eight interactions are homodimers. This observation may be related to the finding that homodimers generally tend to be overrepresented in protein-protein interaction networks (Ispolatov *et al.*, 2005). The small number of common interactions is somewhat surprising considering the large number of shared proteins, but resembles similar observations in previous comparisons for yeast (Bader and Hogue, 2002; von Mering *et al.*, 2002). Possible explanations could be that saturation has not been reached; that maps include a large number of false positives and that certain approaches can only detect subsets of all possible interactions due to technical reasons (Mrowka *et al.*, 2001; von Mering *et al.*, 2002). Figure 3.2 shows the combined protein interaction network. Notably, single maps are tightly connected to each other by the large number of common proteins. The combined network does not split into separate networks despite the small interaction overlap. However, the figure also shows that the sets of interactions in different maps are quite distinct.

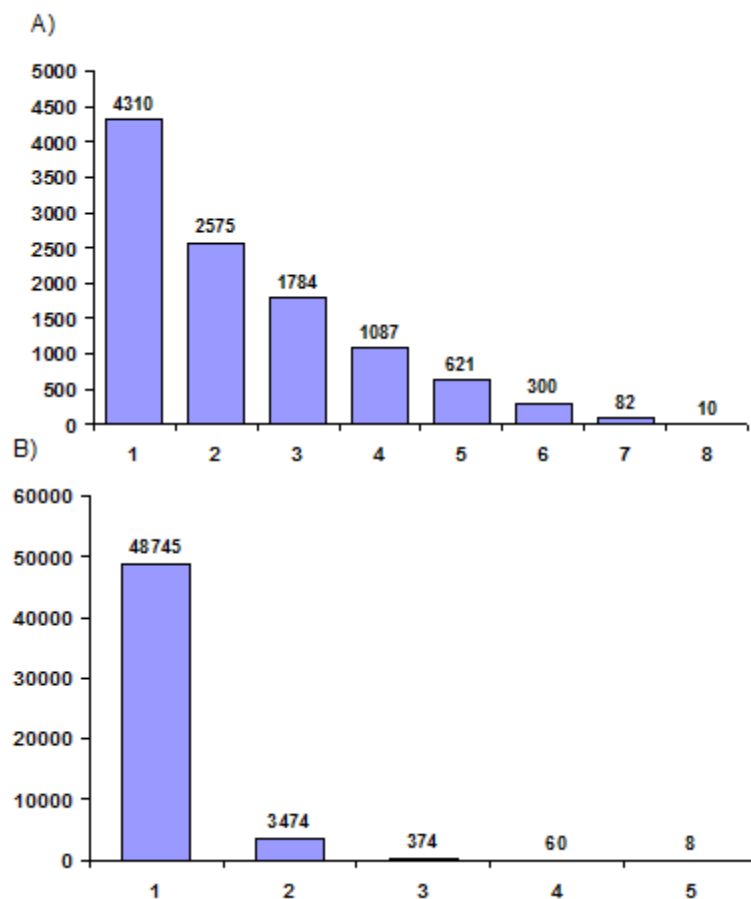


Figure 3.1 Frequency of common proteins and interactions. The plots display the number of proteins (A) and interactions (B) with respect to the number of maps in which they are included.

3.3.2 Overlap and intersection

After the general assessment, I examined in more detail the overlap between single maps (Appendix table A.2). The protein overlap varies from 16% (between CCSB-H1 and COCIT) to 58% (between OPHID and HOMOMINT) with a mean value of 31%. To detect tendencies in the sets of proteins covered by different maps, I clustered maps based on the protein overlaps. Any strong clustering structure would indicate selection bias. Indeed, distinct clusters were observed for literature-based as well as orthology-based maps. Thus, both approaches tend to sample interactions from divergent sets of proteins (figure 3.3A). For literature-based maps, such selection bias can be explained by a preference in small-scale studies to examine ‘popular’ proteins. Similarly, the orthology-based maps might be expected to be restricted to well-conserved proteins. In contrast, both Y2H-based maps appear to be more

separated from the other maps. Thus, Y2H-based maps tend to sample interactions between sets of proteins which are not covered by other mapping approaches.

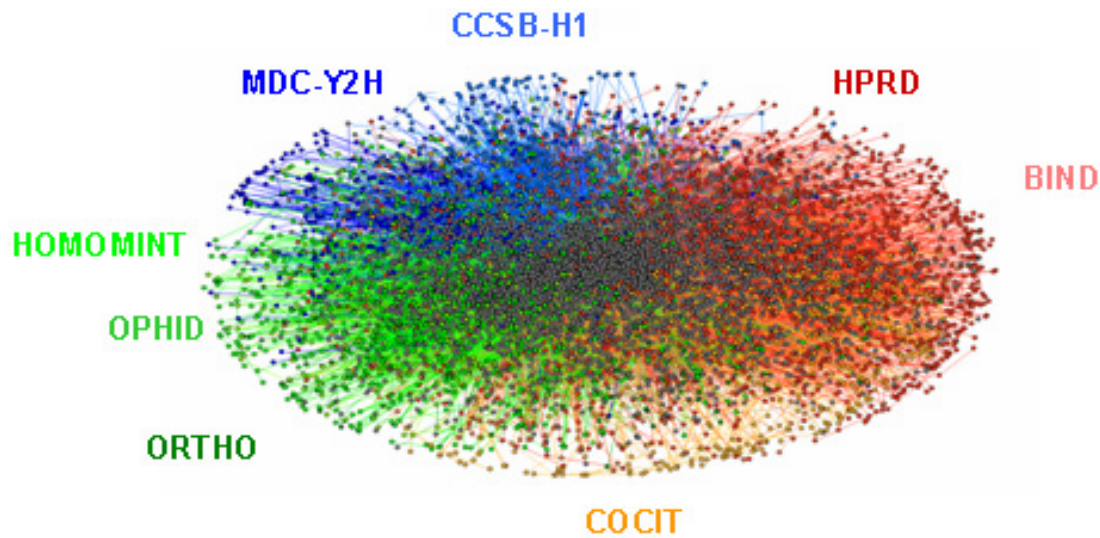


Figure 3.2: Graphical representation of the combined network of interaction maps. Nodes and edges symbolize proteins and interactions, respectively. For clarity, only the connected network was visualized. The color encoding is following: MDC-Y2H - dark blue, CCSB-H1 - light blue, HPRD - dark red, BIND - light red; COCIT - orange; OPHID - dark green; ORTHO - green; HOMOMINT - light green. Gray color indicates that proteins or interactions are found in multiple maps. Although the number of shared interactions between maps is small, the combined network does not disintegrate into separate networks as a large number of proteins are common to multiple networks. For visualization, the software package Pajek was used (Batagelj and Mrvar, 2003).

Next, the degree of shared interactions between in different maps was analyzed. The overlap ranges from 1.5% (MDC-Y2H - COCIT) to 38% (HOMOMINT – ORTHO) with an average of 15% (Appendix table A.3). To identify potential detection bias of mapping approaches, I subsequently clustered the maps based on the interaction overlap. As before, characteristic clusters were obtained (figure 3.3B). Literature-, orthology- and Y2H-based maps form three distinct clusters structure pointing to a pronounced detection bias in the maps compared. The clustering results demonstrate are more convergent if they are derived by similar approach.

The applied measure of interaction overlap is purely based on the occurrence of observed interactions. To assess the concurrence of missing interactions as well, I subsequently utilized the log likelihood ratio *LLR* (as defined in Materials and

Methods) between all pairs of maps. The *LLR* ranges from 1.5 (OPHID-MDC) to 7.1 (BIND-HPRD) with an average 4.6 (Appendix table A.4). Notably, no pair-wise comparison resulted in a *LLR* close to zero which would be the expected value for comparison of random sets of interactions. Thus, the observed interaction overlap has not simply resulted by chance despite of being rather small. This conclusion is also supported by the outcome of the permutation tests assessing the significance of shared interactions (Materials and Methods sec 3.2.2). For all but two comparisons, the detected concurrency of interactions was highly significant with $p < 0.001$.

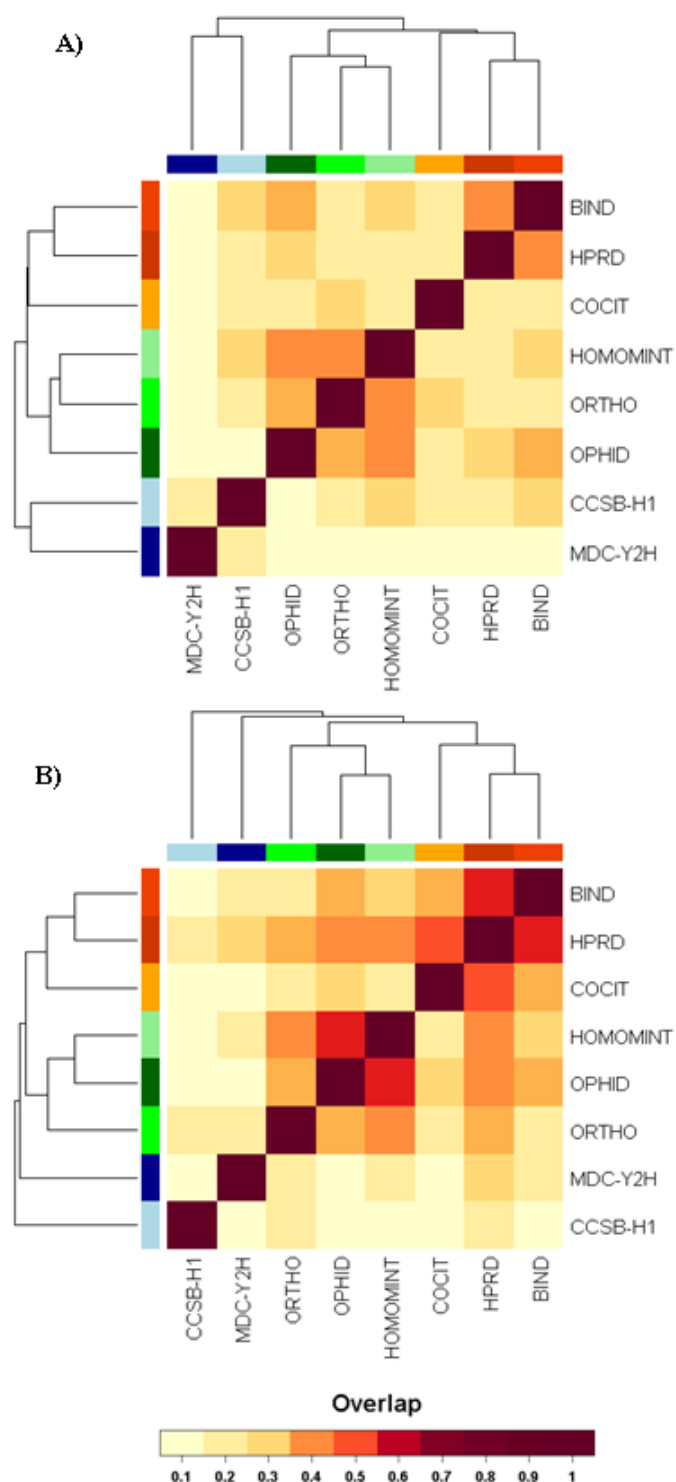


Figure 3.3: Hierarchical clustering of maps based on protein (A) or interaction (B) overlap. The matrices display the overlap between all possible pairs of maps. Large overlap signifies a large number of common proteins and interactions, respectively. On top and right side of each matrix, the resulting clustering trees are shown. Clustering was based on average linkage with the distance Δ_{ij} between map i and j defined as $\Delta_{ij} = 1 - O_{ij}$ (O_{ij} : average protein/interaction overlap between map i and j).

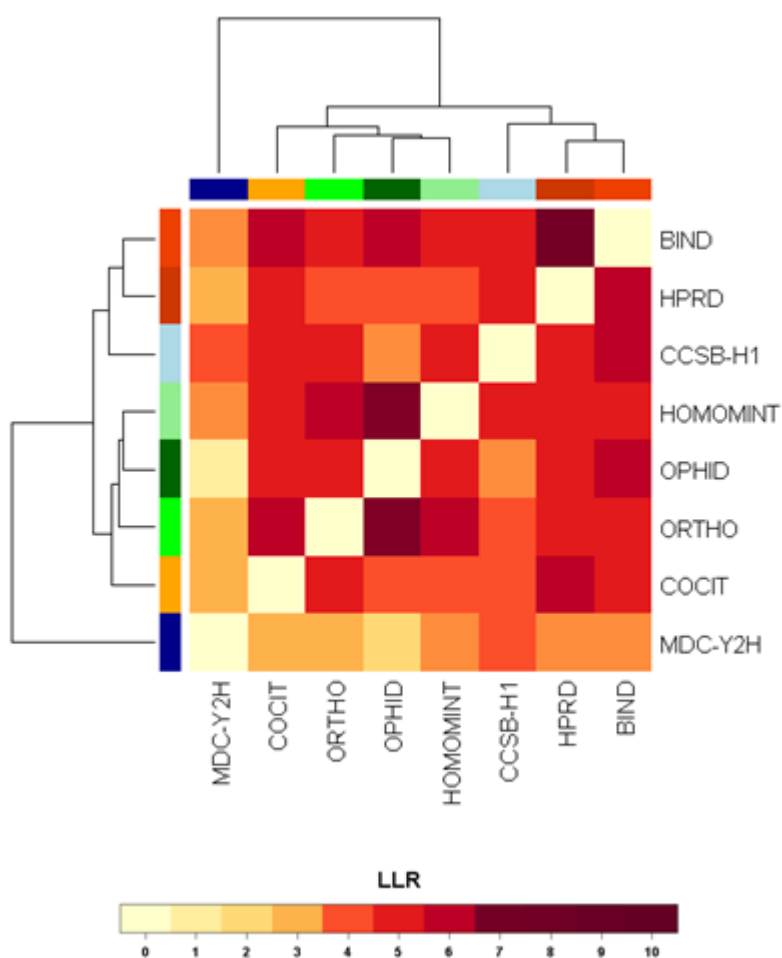


Figure 3.4: Hierarchical clustering of interaction maps based on the log likelihood ratio LLR.. The LLR was set to zero in case a map was compared with itself. The calculated LLRs between maps were shown as color-encoded matrix. For clustering, average linkage was used and the distance was defined as $1/LLR$. The exact values of LLR_{ij} can be found in Appendix table A.4.

3.3.3 Functional assessment

Besides small overlap, previous assessments of yeast protein interaction maps have revealed that most methods for generating interaction maps have their own characteristic biases. Especially, their composition regarding functional classes is influenced by the mapping procedure chosen. On the other hand, the tendency of interacting proteins to have common functions has previously been utilized for assessing the quality of interaction maps as well as for prediction of function (Schwikowski *et al.*, 2000; von Mering *et al.*, 2002). In order to detect potential sampling and selection biases in the human PPI maps, I also carried out an analysis similar to mentioned for yeast studies (Mrowka *et al.*, 2001; Bader and Hogue, 2002; von Mering *et al.*, 2002). To determine the functional composition and coherency of maps compared, proteins were classified based on their annotation in Gene Ontology (GO) database (Harris *et al.*, 2004).

Functional composition

I started by examining whether the interaction maps cover a wide range of different functions or are rather restricted to proteins of certain functionality. I found that all interactions maps showed generally a similar composition on the first (i.e. most general) level of the GO hierarchy (Appendix figure A.1-A.3). Since this composition is also similar to the overall composition for human genes currently annotated in GO, the interaction maps are in first approximation representative samples of the human genome. At first glance, the observed distribution of all three categories looks similar for all eight maps. Detailed analysis of molecular function class showed that the largest category is 'binding' for all maps, representing ~50% of proteins included. This is to be expected in an analysis of interaction maps. The second largest functional category in all maps (~29% - 35%) is 'catalytic activity' including kinases known to be major regulators of cellular pathways. Literature-based maps show an over-representation of signaling proteins (~20% - 24%) as compared to homology-based and Y2H-based maps. Nearly ~5% proteins were assigned to the category 'molecular function unknown' (MFU) from Y2H-based and homology-based (ORTHO, HOMMINT) maps. Thus, these maps might be especially useful for *de novo* functional annotation of hitherto uncharacterized proteins.

Functional coherency

To assess the tendency that proteins of the same function interact, the number of interactions between GO categories was counted and the resulting interaction matrix was visualized for each map. Focusing here on the third level of the GO hierarchy, I observed that maps show enrichment of interactions between proteins of the same GO category to a varying extent (Appendix figure A.4-A.6). Literature-based maps display generally the strongest enrichment of interactions between proteins sharing function or location, and the largest depletion of interaction between proteins of different function or location. To a lesser extent, orthology- or Y2H-based networks maps follow this pattern. Such observation can be expected as GO annotations are frequently derived from literature. Interestingly, proteins of some classes show strong tendency to interact with proteins of the same class independently of the map. For example, proteins of the cytoskeleton are highly likely to bind to each other in all interaction maps.

Alternatively, functional coherency of maps can be assessed by examining the similarity of GO annotations of interacting proteins. As similarity measure of two GO terms, the length of their shared paths from the root term was calculated. The larger the shared path length is, the more related the GO terms are expected to be. Thus, one would observe a larger shared path compared to a random network if interacting proteins correlate in function. Indeed, such patterns were detected for all networks (Appendix figure A.7). Generally, this tendency is most apparent for categories on a high level within the GO hierarchy with the root term at the bottom. This implies that interactions are more beneficial for prediction of specific functions of proteins, but less informative for prediction of more general functions. While the overall tendencies for coherency are similar for different interaction maps with respect to process and location, differences exist regarding molecular function. MDC-Y2H and OPHID showed the least coherent structure, whereas COCIT displays the largest coherency. This observation suggests that a differentiation between interactions maps might be favorable for future prediction of protein function. The attachment of larger weights to interactions from maps of large coherency might improve the prediction accuracy.

3.3.4 Graph-theoretical comparison

Using graph theoretical measures, fundamental topological properties of protein interaction maps can be compared and characterized. After converting all interaction maps to graphs, I analyzed their internal connectivity (table 3.2). For all graphs, the

vast majority of proteins were connected in a main network, which appears to be a general feature of protein-protein interaction networks being also observed in other species (Uetz *et al.*, 2000; Giot *et al.*, 2003; Li *et al.*, 2004). The remaining proteins formed predominantly smaller networks of less than 10 proteins. Only for BIND, COCIT, OPHID and ORTHO, medium sized networks (including 10-100 proteins) emerged. If such separated network islands are artifacts reflecting the fragmentary state of proteins maps or functionally separated units remain subject for further research.

Table 3.2: Graph-theoretical analysis of the interaction maps. The second column includes the numbers of disconnected networks according to their size. The measures listed in all following columns are based on the largest connected network. Mean path length refers to the average shortest path length between all possible pairs of protein in the network. The diameter is defined by the maximal shortest path length found. The degree exponent γ of the power-law distribution ($P(k) \sim k^{-\gamma}$) was calculated using linear regression of the corresponding dependencies (Appendix figure A.12).

Network	Nr of Networks >1000/ 101-1000/ 11-100/ 1-10	Mean path length	Diameter	Degree Exponent	Clustering coefficient
MDC-Y2H	1/0/38/4	4.9	13	1.63	0.01
CCSB-H1	1/0/90/27	4.4	12	1.46	0.05
HPRD	1/0/135/140	5	15	2.44	0.13
BIND	1/3/169/256	5.9	16	1.90	0.17
COCIT	1/7/545/0	5.9	20	2.18	0.43
OPHID	1/3/95/0	4.8	15	1.36	0.23
ORTHO	1/2/183/9	6.5	17	2.14	0.19
HOMOMINT	1/0/85/45	5.1	12	2.76	0.07

Small-world property

A main conclusion of previous studies was that protein interaction networks display 'small world' properties having a small mean path length. This is also the case for the networks compared here. Their mean path length is similar and ranges from 4.4 (CCSB-H1) to 6.5 (ORTHO) (table 3.2). For most networks, it is smaller than expected for corresponding random graph (Appendix figure A.8). For all networks, however, the mean path length is larger than expected for corresponding random scale-free networks pointing to internal structures.

Degree-distribution

An important determinant of a network's structure is the degree distribution $P_{(k)}$. I found that all networks display power law distribution (Appendix figure A.9). However, some deviations can be observed. Networks derived from the BIND, OPHID or Y2H-assays followed most closely the power law distribution in contrast to remaining ones that show a relative depletion of interaction-poor proteins. Notably networks obeying closely the power law distribution also tend to have smaller mean path lengths.

Modularity

Cellular networks have been proposed to exhibit modular structure (Ravasz *et al.*, 2002; Rives and Galitski, 2003). A commonly used measure for modularity is the clustering coefficient reflecting the cohesiveness of the neighborhood of network nodes (Watts and Strogatz, 1998). In my analysis the average clustering coefficient ranges considerably by a factor of almost 50 from 0.01 to 0.45 (table 3.2) (Appendix figure A.10). The smallest coefficients were found for Y2H-based networks. These values were similar to the expected values for random scale-free networks leading to the conclusion that the Y2H-based maps do not display particularly strong neighborhood cohesiveness. A possible reason could be a large number of undetected interactions (false negatives). In contrast, clustering coefficients for literature- and orthology-based networks were considerably larger than for corresponding random networks implying that these networks are highly modular.

Hierarchical structure

Besides for assessment of modularity, the clustering coefficient has been employed to study hierarchical structures of networks. Ravasz and co-workers associated a decrease of clustering coefficient for highly connected nodes to a hierarchical organization of metabolic network (Ravasz *et al.*, 2002). Such decrease can also be observed for most networks compared (Appendix figure A.11). For orthology-based networks, however, this pattern is absent suggesting the lack of a hierarchical structure in these networks. For yeast, it was found that interaction-rich proteins have a propensity to avoid direct interaction with each other (Maslov and Sneppen, 2002). Maslov and Sneppen conjectured that such structural feature would be beneficial to decouple cellular modules in networks and would lead to an increased robustness against perturbation. The disassortativity of network hubs is exhibited by a tendency of hubs being linked to interaction-poor proteins. However, I observed such pattern only for Y2H-based maps and to a lesser extent for BIND (Appendix figure A.12). In contrast, the majority of networks showed an increase of links between highly connected proteins. Thus, the disassortativity of network hubs cannot be generally confirmed for human protein interaction maps and further investigation is needed.

Connectivity

After comparison of global network topologies, I turned towards examination of local topological properties. Here, the question was addressed whether the connectivity of proteins is conserved across interaction maps. To measure the conservation of connectivity between pairs of networks, I correlated the number of interactions in the two networks using Spearman correlation for the set of common proteins. An overall weakly positive correlation (0.20) ranging from -0.07 to 0.57 was recorded (Appendix table A.5). Only 6 out of 28 pair-wise comparisons resulted in correlation coefficients larger than 0.3. Notably, all of these 6 moderately positive correlations were found between maps were generated by the similar approaches. Connectivity is less conserved between maps derived by different methods. This is also reflected in a subsequent cluster analysis based on the Spearman correlation. The interaction maps group according to their method of generation (figure 3.5).

3.3.5 Analysis of network hubs

Earlier graph-theoretical analysis showed that nodes of high degree, so-called hubs, are generally of crucial importance for scale-free network structure (Albert *et al.*,

2000). In protein interaction networks, such hubs are given by highly connected proteins. This led to the conjecture that these interaction-rich proteins should be essential for correct functioning of cellular networks. To investigate the potential role of hubs in human protein-protein networks, I examined first whether hubs tend to be assigned to specific functions, processes or locations using GO annotations. Proteins were defined as hubs, if their number of interactions is within the top 10% in the corresponding network. The significance of enrichment by hubs relative to the overall composition of the network was assessed using Fisher exact test.

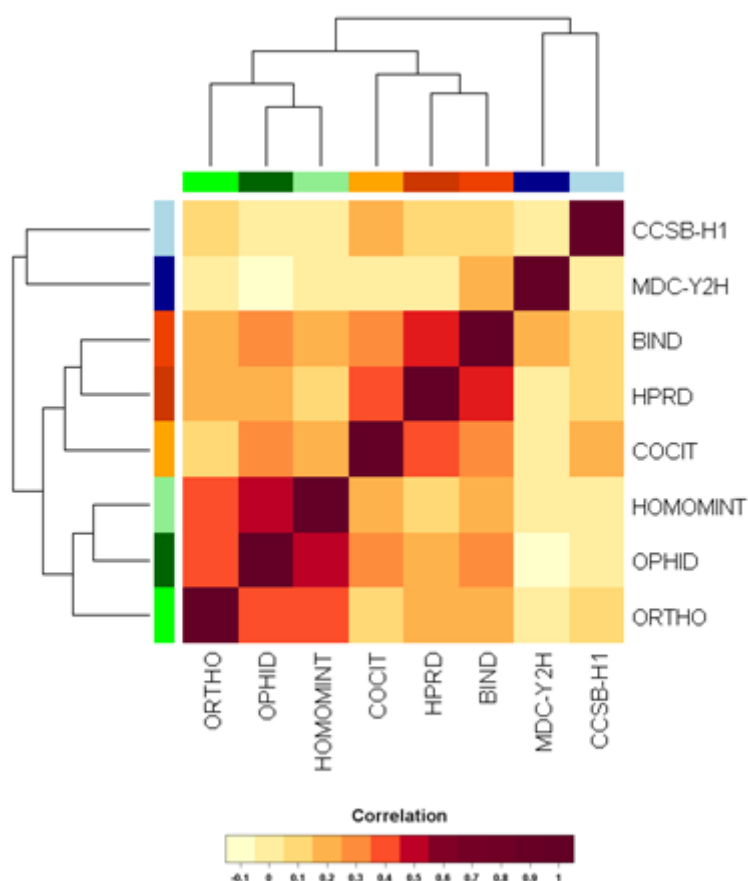


Figure 3.5: Hierarchical clustering of interaction maps assessing similarity of connectivity of proteins. The connectivity was measured by Spearman rank correlation coefficient r_{ij} between the degrees of shared proteins of maps i and j . The correlation matrix is shown color-encoded. For clustering, average linkage was used and the distance Δ was defined as $\Delta=1-r$. The calculated values of r_{ij} can be found in Appendix table A.6.

Although the explicit sets of hubs were distinct in different maps, common trends were observed for literature- and orthology-derived networks whereas no significant enrichment was recorded for hubs of Y2H-based networks (Appendix table A.6-A.8).

Regarding molecular function, hubs in literature-based maps tend to be associated with protein binding, whereas hubs in orthology-derived maps tend to be assigned to RNA binding. For both literature- and orthology-based networks, the set of hubs was enriched by proteins assigned to metabolism. Interestingly, only literature-based network hubs were found enriched by proteins linked to regulation.

A similar picture emerged when I explicitly compared the list of proteins that constitute hubs in different networks. To facilitate this comparison, the previous definition of hubs was modified. Proteins were defined as network hubs if they belong to the 100 proteins having the most interactions in the corresponding network. Remarkably, no protein was selected as hub in more than 4 networks. If I consider only hubs selected in at least three networks, three different classes of hubs become apparent (table 3.3). The first class consisted of proteins selected as hubs in networks generated by different approaches. Example of these protein hubs are members of the canonical MAP-kinase pathway (MAPK1, MAPK3) and TP53. This class also includes the only hub (VIM) shared between Y2H-based networks. The second class comprises protein hubs selected only in orthology-based networks. These hubs apparently tend to belong to complexes of cellular core machineries such as the proteasome (e.g PSMB2, PSMD14) or polymerases (e.g. POLR2C). The third class contains proteins representing hubs only in literature-based networks. This last class tends to be enriched by signaling transducers and regulators transcription factors. A possible explanation for this observation could be that these types of proteins are popular research targets, while their interactions are difficult to measure by high-throughput methods or orthology-based approaches. However, an intriguing alternative explanation is that the selected signal-processing proteins have become only late in evolution hubs to meet the increased demand for coordination of molecular functions in higher organisms. Therefore, they do not represent hubs in orthology-based networks as these networks are based on evolutionary conserved interactions.

3.4 Discussion and Conclusions

I presented here a first comparative assessment of eight currently available large-scale human protein-protein interaction maps. The maps were derived from the three main approaches: Y2H-assays, literature search or orthology-based predictions of

interactions. The aim of this study was to examine critically the coherency within maps as well as the concurrency between maps. The comparison showed that current maps have only a small, but statistically significant overlap. Whereas the majority of proteins can be found in multiple maps, this is only the case for less than 10% of the interactions making the maps largely complementary. This study showed that maps were generally more concurrent if they were based on the same method. This demonstrates the method of generation has considerable impact on the resulting interaction map.

Table 3:3: Comparison of network hubs. Proteins were listed if they belong to the 100 proteins having the largest number of interactions in at least three maps. The number of interactions is listed and was underlined by color if the corresponding protein belongs to to the hubs in the network. Missing values indicate that no interactions were found for the corresponding protein in the interaction map. Hubs can be divided into three classes: hubs found in interactions maps of different origin (red color); hubs which were found in orthology-based networks only (green color) and hubs in literature-based networks only (yellow color). Counts refer to the number of networks in which the protein is a hub.

	MDC-Y2H	CCSB-H1	HPRD	BIND	COCIT	OPHID	ORTHO	HOMOMINT	Counts
MAPK3			45	17	29	2	29	2	4
MAPK1			51	9	37	19	29	16	4
YWHAZ	2		52	11	33		9	19	4
GRB2			149	31	43		20	28	4
ITGB4BP	24		4	1		68	30	27	4
TP53	43	2	126	49	74	9		1	4
GNB2L1	5	1	38	19				25	3
UBE2I	1	21	42	12			7	6	3
VIM	43	16	20	17	1	2		1	3
PSMD11	24					69	16	24	3
DKC1	4					46	38	18	3
PSMB2			1		12	47	25	17	3
PSMD14	3					48	26	22	3
CRFG						58	34	18	3
PSMD4		1	7			59	32	41	3
PE\$1						69	35	22	3
KPNA6			2			108	39	59	3
POLR2C	6			11		61	63	13	3
AKT1			36	16	20		1	1	3
MYC			45	18	62	3	9	5	3
STAT3			48	11	38	8	12	1	3
RAF1			49	28	27	11	6		3
PCNA		1	53	28	19	24	20	12	3
SP1			59	14	59			1	3
RB1			70	54	36	11		2	3
CREBBP			76	18	20	2		2	3
SHC1		1	77	13	43	7		5	3
CTNNB1			77	16	31	5	2	3	3
SRC	1		91	19	18				3
EGFR	2		94	21	22	8	3	3	3

Several selection and detection biases could be identified, indicating the role of the methods of generating the maps. For example, RNA binding proteins are overrepresented in orthology-based maps, whereas signal transducers are overproportionally sampled in literature-based maps. A significant depletion of membrane proteins was observed in all networks compared and not only in Y2H-based maps as expected. The existence of such sampling and detection biases is a main reason for the small number of common interactions in different maps. Notably, this small

overlap limits the capacity for pooling approaches assigning higher confidence to interactions observed in multiple experiments (von Mering *et al.*, 2002). On the other hand, as interaction maps are highly complementary, thus, integration of maps promises to be greatly beneficial.

Using GO for assessment, literature-based maps displayed generally larger coherency than orthology- or Y2H-based maps. However, the coherency of literature-based maps might be overestimated, as GO annotations are frequently based on literature reviews and, thus, do not represent a truly independent benchmark set. In this case, the apparent lack of coherency in some maps might indicate that these maps may provide more novel information about observed interactions. Nevertheless, the detected differences in coherency should be taken into consideration for future prediction of protein functions based on the interaction networks. As some networks include highly coherent data for certain protein classes, differentiation between networks could improve the prediction accuracy. For example, it may be favourable to put more weight on interactions derived networks of high functional coherency for a chosen protein class.

All interaction networks showed small world properties and correspond to scale-free networks. As both features have commonly been observed in earlier studies of interaction maps, they are likely to be general valid for protein interaction networks. However, I also observed that several previous conclusions for network structures in lower eukaryotes cannot be generally reproduced for human interaction networks. For example, protein hubs are only separated in some of the compared maps. Thus, some previous results of network analyses might not refer to the underlying biological network structure, but rather reflect features specific to the approach taken for the generation of the network. It also suggests that the present view of modularity in protein interaction networks may have to be modified. The structure of interactomes of higher eukaryotes might differ substantially from those for lower organisms and, thus, general re-evaluation of concepts regarding network structure and evolution may be warranted.

A more dynamic view of network evolution is also proposed by a comparison of hubs in different maps. It shows that hubs can be divided into different evolutionary categories. Ancient hubs are constituted by proteins of core machineries as proteasome and polymerases whereas evolutionary novel hubs can be associated

with signal transducers and regulators. This classification suggests that the current theory of simple preferential attachment may be not sufficient, but that network hubs have arisen to meet the particular requirements of an organism (Barabasi and Oltvai, 2004). The existence of distinct classes of network hubs was also previously reported for the yeast interactome (Spirin and Mirny, 2003; Han *et al.*, 2004). However, these divisions were based on the expression dynamics or network structures, but did not involve an evolutionary component. Here, I conjecture that design principles for network follow the requirements of robustness, adaptability and effectiveness. Differences in these requirements for different organisms are likely to be reflected in the network architecture. For example, previous studies showed that the number of signal transducer is considerable larger for humans than for fly and worm (Lander *et al.*, 2001; Pires-daSilva and Sommer, 2003). These results suggest further that the increased demand of regulation in human organism is also reflected by a strongly increased number of interactions of signaling molecules.

In summary, this comparison presents the status quo of the mapping of the human interactome. It shows that protein interaction maps are in their current status incomplete and considerably biased. Hence, they constitute only a limited representation of the human interactome. Results of network analyses are strongly influenced by chosen approach to create the networks. At this point of time, thus, caution should be taken for over-interpretation of such results especially if general conclusions are derived based on single maps. Any result should be critically assessed regarding potential bias in the underlying interaction map and possibly verified for other maps.

4 UniHI: Integration of Human Interactome

This chapter is an extended version of following three papers:

References:

Chaurasia, Gautam; Malhotra, Soniya; Russ, Jenny; Schnögl, Sigrid; Hänig, Christian; Wanker, Erich;

and Futschik, Matthias (2009), UniHI 4: New tools for query, analysis and visualization of the human protein-protein interactome, *Nucleic Acids Res.* 37; Database issue: D657–D660, 2009

Chaurasia, Gautam,; Iqbal Yasir; Hänig, Christian; Hanspeter Herzel; Wanker, Erich; and Futschik, Matthias. UniHI: an entry gate to the human protein interactome, *Nucleic Acids Res.* 35; Database issue:D590-4, 2007.

Chaurasia, Gautam,; Iqbal Yasir; Hänig, Christian; Hanspeter Herzel; Wanker, Erich; and Futschik, Matthias; Flexible web-based integration of distributed large-scale human protein interaction maps. *Journal of Integrative Bioinformatics*, 4(1):51, 2007.

In chapter 3, I focused my analysis on the comparative evaluation of analysis of currently available large protein-proteins interaction networks generated by different approaches. My analysis showed that the current human PPI maps are unsaturated, highly divergent and showed a very small overlap between them. Therefore, integration of these maps could be very beneficial. In this chapter, I will elaborate those existing limitations of currently available human PPI maps in more detail, and outline the steps required for the successful integration of these maps. This chapter is

organized is as follows: Section 4.1 addresses the current problems in human PPI data. Section 4.2 provides details on the database architecture. In section 4.3, I will describe the steps required for the successful integration of protein interaction data from different sources, and its further integration with other biological omics data. Furthermore this section also introduces the features of web-interface, visualization tool and updates and extensions of UniHI database. Finally, I will conclude this chapter in section 4.4 by presenting the discussion and conclusions.

4.1 Introduction

PPI data are of great potential for the biomedical research. Recent advances in high-throughput methods have resulted in a rapid accumulation of human protein interaction networks on a global, genome-wide level. But before these PPI networks can become a cornerstone in disease research, considerable challenges are still to overcome. The restrictions involved are manifold:

4.1.1 Highly divergent and distributed PPI networks

Previous studies have also shown that PPI maps are highly divergent (Bader and Hogue, 2002; von Mering *et al.*, 2002). My analysis, discussed in Chapter 3, also showed the similar picture for the human PPI networks. Further, I observed only less than 19% of all interactions occur in multiple maps, indicating a low degree of saturation (figure 3.1B). The small number of shared interactions is remarkable considering the large number of proteins common to different datasets. More than 50% of all proteins are included in two or more maps (figure 3.1A), suggesting the highly divergent nature of the current human PPI networks and their unification can be therefore a useful step towards completeness of human interactome.

A subsequent problem of human PPI networks is that they are distributed across multiple locations (table 2.1). To find comprehensive information on human proteins of interest, scientists may have to perform repeated searches in many databases. Such efforts are evidently very time-consuming as various query formats and identifiers have to be used in different databases. Another major limitation in currently available interaction databases is that frequently only interactions for single protein can be queried. However, modern system biology requires complex network-oriented

search for interaction of multiple proteins.

4.1.2 Quality of human PPI networks

High quality PPI networks are essential for the biomedical research (de Silva *et al.*, 2006; Stelzl and Wanker, 2006). But, the current large-scale human PPI networks are quite noisy and many interactions are conjectured to be false positives (Chaurasia *et al.*, 2006; Futschik *et al.*, 2007a). To address this problem, various confidence scoring schemes using omics data have been developed (Kierner *et al.*, 2007; Li *et al.*, 2008). Additionally, few of current PPI networks provide their own confidence scoring schemes (Rual *et al.*, 2005; Stelzl *et al.*, 2005). Integration of PPI maps with such confidence score can help experimentalists to assess quality of interactions found in the databases.

4.1.3 Regular updates

Further challenges are regular updates and extensibility of PPI databases. As human interactome is still very far from the completion, interaction data will grow continuously. Therefore, it is necessary to implement flexible architecture that can keep the existing interaction data updated, and also enable easy inclusion of newly discovered interactions in future.

4.1.4 Functional interpretation of PPI networks

Final, but a very important issue is the biologically meaningful interpretation of PPI maps. Although advances in recent genome-wide interactome projects have generated a wealth of PPI data, this also poses new challenges for researchers mainly due to the complexity of interaction networks. In order to understand this complexity, it is necessary to gain meaningful information in the context of physiological systems, which requires identification of not only the function of individual proteins but also the physical interactions and biological process in which they participate. To this end, PPI networks have to be integrated with other functional data to derive the substantial information from them. Previous studies have also shown that integrating PPI networks with expression or pathway data can lead us to characterize biological processes or potential disease modifiers (Oti *et al.*, 2006; Chuang *et al.*, 2007; Ergun *et al.*, 2007; Baranzini *et al.*, 2009).

The main challenges discussed here demand i) comprehensive integration of the currently available human interaction maps; ii) performing network-oriented complex queries; iii) quality assessment of the data; iv) regular updates and the extensibility of the integrated database; and v) integration of PPI networks with other functional and genomic data. To address these challenges, I have constructed a flexible web-based database, termed UniHI, which integrates human protein interaction data from diverse sources, including several quality schemes. In the following sections, I will provide the details on the data sources currently included in the UniHI database, its architecture and the list of currently implemented features of the UniHI web-interface.

4.2 Materials and Methods

4.2.1 PPI data sources

Currently, protein interactions in UniHI are derived from twelve large-scale human protein-protein interaction maps (table 2.1, see in Chapter 2). These maps were generated using yeast-two-hybrid (Y2H) assays, literature review or orthology-based approaches. Currently, UniHI includes five literature-based interaction maps (BIND (Bader *et al.*, 2003), HPRD (Prasad *et al.*, 2009), COCIT (Ramani *et al.*, 2005), DIP (Salwinski *et al.*, 2004), BioGrid (Breitkreutz *et al.*, 2008), IntAct (Aranda *et al.*, 2010) and REACTOME (Matthews *et al.*, 2009)), two Y2H-based interaction maps (MDC-Y2H (Stelzl *et al.*, 2005), and CCSB-H1 (Rual *et al.*, 2005)) and three orthology-based maps (HOMOMINT (Persico *et al.*, 2005), OPHID (Brown and Jurisica, 2005) and ORTHO (Lehner and Fraser, 2004)). Further details on these maps are given in the Chapter 2, section 2.2.1 and table 2.1.

4.2.2 Gene and protein identifiers

Since PPI data were collected from different sources, one of the major problems was to find common identifiers for their aggregation. For this purpose, lists of different protein identifiers were downloaded from the websites of NCBI (Maglott *et al.*, 2007), HUGO Gene Nomenclature Committee (HGNC) (Eyre *et al.*, 2006), and Ensembl EnsMart (Kasprzyk *et al.*, 2004).

4.2.3 Gene annotation

Additional information on proteins (e.g. functional annotations and description of

proteins, chromosomal location and potential disease association of corresponding genes) were imported from National Center for Biotechnology Information (NCBI) (Maglott *et al.*, 2007), Online Mendelian Inheritance in Man (OMIM) (McKusick, 1998) and Gene Ontology (GO) (Ashburner *et al.*, 2000) databases.

4.2.4 Gene expression data

Gene expression data were collected from Gene Atlas database (Su *et al.*, 2004). The data set comprises of expression profiles for 79 human tissues measured in replicates using Affymetrix HG-U133A and custom-designed GNF1H arrays. Expression summaries for the 44,775 transcripts (corresponding to ~15000 genes uniquely identified by their EntrezGene ID) were derived utilizing the MAS5 algorithm (Pepper *et al.*, 2007).

4.2.5 Pathway information

Association of genes with pathways and the information about the relations between them such as regulatory or physical interactions (e.g. phosphorylation, dephosphorylation, activation, inhibition, and binding association) were collected from the KEGG pathways database (Kanehisa and Goto, 2000). KEGG database constitutes a collection of manually drawn pathway maps representing accumulated knowledge on molecular interaction and reaction networks for metabolism, cellular processes, and human diseases as well as for genetic and environmental information processing.

4.3 Results

UniHI was developed to provide an integrated platform for finding comprehensive information on human proteins and their potential interaction partners. In its latest version, UniHI integrates interaction data from twelve major sources, establishing it as the largest catalog for human PPIs worldwide (table 2.1, and figure 4.1). Currently, UniHI houses over 250,000 distinct interactions between 22,300 unique proteins (table 2.1). For addressing the final problem “functional interpretation of PPI

networks”, I recently performed a major update of UniHI database (Chaurasia *et al.*, 2009), in which, PPI data was integrated with biological pathway data from KEGG (Kanehisa and Goto, 2000) and gene expression data from Human Gene Atlas (Su *et al.*, 2004). Moreover, several new tools have been included in UniHI to offer more convenient web interface for the network analysis, especially for researchers less acquainted with bioinformatical analyses. Besides a basic search for interactions, it offers advanced tools that allow user to construct the tissue-specific interaction networks or annotate edges with the specific type of interaction. UniHI provides scientists with a user friendly web-interface available at <http://www.unihi.org>.

4.3.1 Architecture of the UniHI

The architecture of the UniHI database has been designed to integrate interaction data obtained from different sources, and also to incorporate future human interaction maps, if they become available. The advantage of the UniHI architecture is its modularity and portability by introducing a multi-tier architecture with four separated layers: i) integration; ii) database; iii) persistence; and iv) application. The integration layer is responsible for downloading, parsing, and pre-processing of the data (Appendix B). The database layer is a relational database which stores and manages the information on proteins and their interaction partners from different sources into one common schema (Appendix figure B.1). The persistent layer is used for inserting and retrieving data from the database. The application layer provides a web-interface and a visualization tool with many interactive features for accessing and viewing the interaction data. Figure 4.2 shows the architecture of the UniHI database.

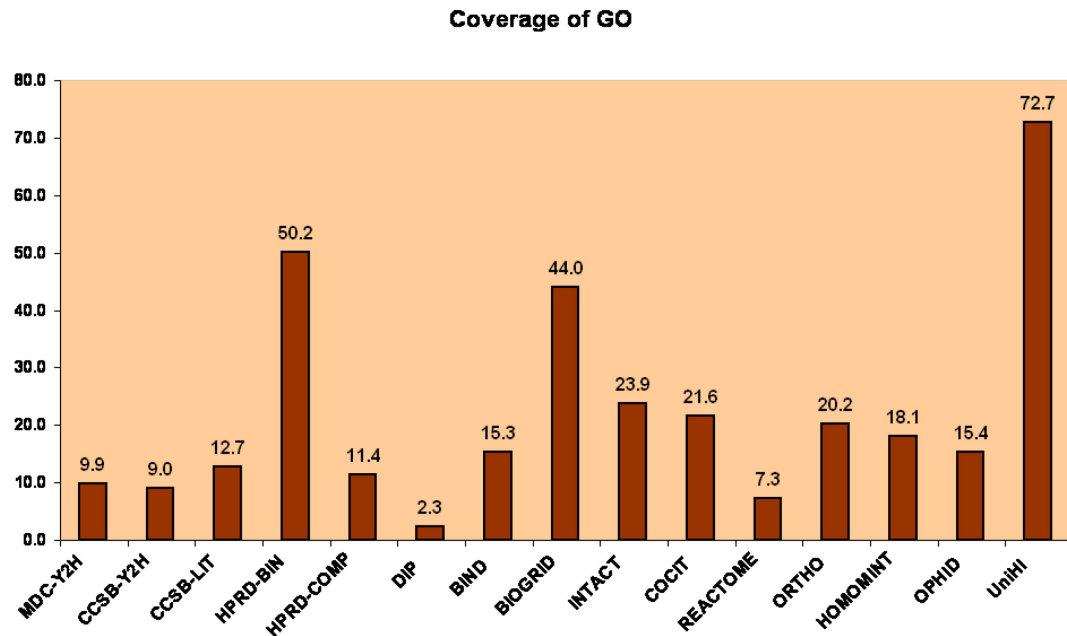


Figure 4.1: Coverage of the functionally annotated human genome by PPI resources. For annotation, Gene Ontology was utilized. Coverage rates were derived after mapping of proteins to corresponding Entrez Gene IDs. Notably, the coverage of UniHI is considerably larger than of the individual PPI resources.

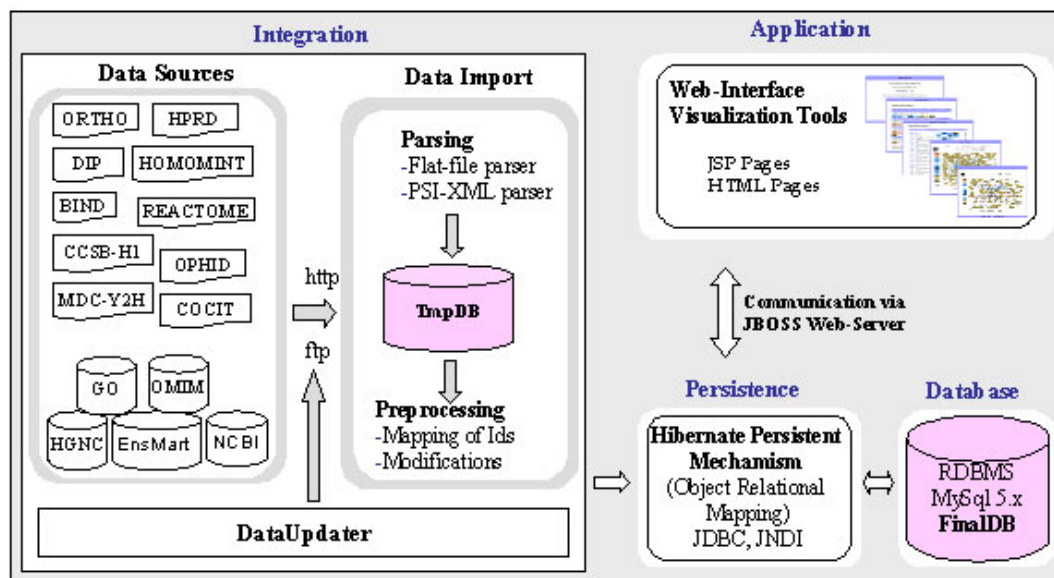


Figure 4.2: Architecture of the UniHI, consists of four main architectural layers: i) integration, responsible for the data downloading, parsing, preprocessing and updating; ii) database, consists of a relational database which stores and manages the information on proteins and their interaction partners from different sources using one common schema; iii) persistence, used for inserting and retrieving the data from the database via Hibernate persistent mechanism; iv) application, providing a web-interface and a visualization tool for accessing and viewing the interaction data.

4.3.2 Mapping of proteins

As UniHI integrates a large number of different PPI resources, aggregation of heterogeneous interaction maps and building a unique identifier indexing system is a foremost task. For unification of primary data, I first computed complete lists of proteins for each interaction map separately. Subsequently, these lists were compared employing information from NCBI (Maglott *et al.*, 2007), HGNC (Bruford *et al.*, 2008) and EnsMart (Kasprzyk *et al.*, 2004) to map their corresponding identifiers in other interaction datasets. After mapping, identical protein identifiers were merged together in a horizontal manner where each protein is a unique entry in a table. A unique identifier was assigned to each protein entry of this table. These unique identifiers were further used for grouping of the redundant interactions from all interaction datasets.

4.3.3 Data quality assessment

For data quality, I have computed two different measurements i.e. co-expression and co-annotation. Co-expression provides the indication whether two proteins are correlating with each other on transcript level, and thus are likely to have interaction between them after translation occur. For computing the co-expression, Gene Expression Atlas dataset was used (Su *et al.*, 2004). To measure co-expression for interacting proteins, Spearman rank correlation of expression levels was calculated. Additionally, corresponding quantiles for correlation coefficient was also derived. For example, a quantile of 0.05 means that the corresponding correlation coefficient is within the top 5% of the total distribution of observed coefficients. Similarly, co-annotation supports the hypothesis that two proteins are likely to interact if they share same functions or involved in same biological process. For computing the co-annotation, I assessed the similarity of GO categories assigned to interacting proteins. The similarity of GO categories was approximated by calculating the length of the shared path from the root category. Large shared path lengths indicate that the GO categories are in proximity to each other within the GO graph and, thus, can be considered similar (Jansen *et al.*, 2003). In case of multiple GO assignments for proteins, the largest shared path length is counted.

4.3.4 Data query, analysis and visualization

For data query and visualization, I developed a web-based query interface and a visualization tool, offering several features described in the next sections.

Web-based query interface

My primary aim was to provide an easy and intuitive, but nevertheless efficient and comprehensive access to the integrated data. Thus, the UniHI web-interface provides the user with options to perform multiple proteins search, in a network-oriented manner, where they can supply a list of proteins to find the network between them. Interactions can be queried using following protein identifiers: EntrezGene ID, Uniprot ID, Ensembl ID, Unigene ID, NCBI Geneinfo ID, OMIM ID, Gene Symbol, Biogrid and HPRD IDs. Retrieved interactions can be displayed either in textual (figure 4.3) or graphical form (figure 4.4a and 4.4b).

Visualization tool

Visualization of the retrieved interaction networks remains to be crucial for the evaluation of query results. The complexity of retrieved networks, however, requires highly flexible graphical tools. Thus, I have implemented a visualization tool for interaction data which offer many attractive features for rapid analysis and adjustment of the extracted information. For example, nodes (i.e. proteins) can be anchored or hidden allowing filtering of the network and manual adjustment of the layout. Also, information about proteins and interactions can be accessed directly in the network graphics, thereby avoiding cumbersome comparisons with the textual output.

To permit users a highly targeted search, UniHI offers several tools to specify the displayed interactions. The display can be restricted to direct interactions between query proteins or extended to include bridging proteins. Such procedure can narrow down the context of a chosen set of proteins and can help to identify putative modifiers of physiological processes.

GS: HD	EntGID: 3064	Total interacting partners: 88						Info
PRPF40A	PRP40 pre-mRNA processing factor 40 homolog A (yeast)	MDC-Y2H	HPRD-BIN	HPRD-COM1	BIND	BIOGRID	INTACT	i
SH3GL3	SH3-domain GRB2-like 3	MDC-Y2H	HPRD-BIN	BIND	BIOGRID	INTACT		i
TCERG1	transcription elongation regulator 1	MDC-Y2H	HPRD-BIN	BIND	BIOGRID	INTACT		i
CREBBP	CREB binding protein (Rubinstein-Taybi syndrome)	HPRD-BIN	BIND	BIOGRID	INTACT			i
HIP1	Huntingtin interacting protein 1 (HIP-1)	MDC-Y2H	HPRD-BIN	BIOGRID	INTACT			i
KIAA1377	KIAA1377	MDC-Y2H	HPRD-BIN	BIND	INTACT			i
PIAS4	protein inhibitor of activated STAT, 4	MDC-Y2H	HPRD-BIN	BIND	INTACT			i
MED31	mediator of RNA polymerase II transcription, subunit 31 homolog (yeast)	MDC-Y2H	HPRD-BIN	BIND	INTACT			i
GIT1	G protein-coupled receptor kinase interactor 1	MDC-Y2H	HPRD-BIN	BIND	INTACT			i
UTP14A	UTP14, U3 small nucleolar ribonucleoprotein, homolog A (yeast)	MDC-Y2H	HPRD-BIN	BIND	INTACT			i
ZDHHC17	zinc finger, DHHC-type containing 17	HPRD-BIN	BIND	BIOGRID	INTACT			i
CRMP1	collapsin response mediator protein 1	MDC-Y2H	HPRD-BIN	BIND	INTACT			i
XRCC6	70K thyroid autoantigen	MDC-Y2H	HPRD-BIN	BIND	INTACT			i
PFN2	profilin 2	MDC-Y2H	HPRD-BIN	BIND	INTACT			i
TP53	cellular tumor antigen p53	HPRD-BIN	BIND	BIOGRID	INTACT			i

Figure 4.3: Textual representation of a query result for protein interactions in UniHI. Multiple links indicate identification of the interaction by different methods. For easy discrimination between maps, specific colors have been assigned. Shades of blue have been used for datasets derived by literature search, shades of green for orthology-based maps, shades of red for maps derived from Y2H screens.

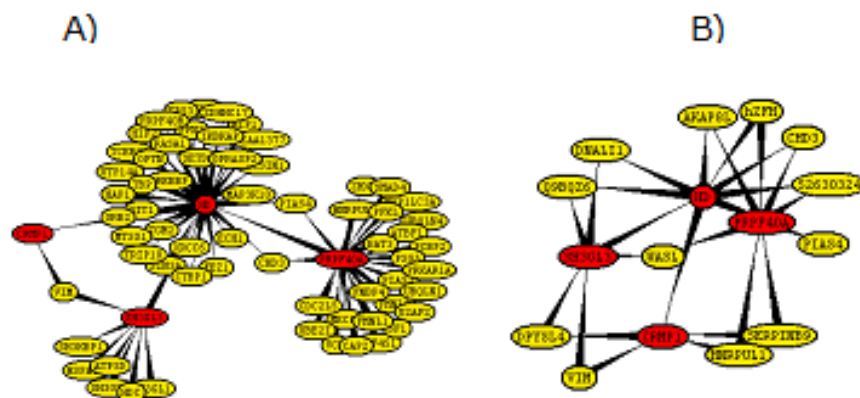


Figure 4.4: Graphical representation and analysis of PPI networks using UniHI Search visualization tools. Display of the interaction partners (yellow or grey) of the query proteins (red) HD, CRMP1, PRPF40A and SH3GL3. (A) Using the number of Pubmed as a criteria, only those interaction partners are shown, which have been reported in more than two publications. (B) Only common or direct interactions between query proteins are displayed.

For quality control, users can specify the PPI resource interactions should be retrieved. This feature allows user to view only those interactions which have been identified by different approaches that may be used to gain confidence in interactions retrieved. On the other hand, user can exclude those interactions which have been detected by less validated mapping approaches such as computational prediction. As additional criteria, interactions can be filtered based on a minimum number of PubMed references in which they have been reported (figure 4.4a and 4.4b).

4.3.5 Integration of PPI data with Gene Expression and Pathway Data

Protein interactions are known to be highly dynamic and to greatly depend on conditions. Current protein interaction maps, however, only represent a static view of the human interactome. Experimentally validated protein interactions are generally identified under a variety of conditions in numerous cell and tissue types. Thus, the output of current interaction maps may represent a compilation of all possible protein interactions throughout the human body. In practice, however, biomedical research is focused on the specific tissues which are involved in pathogenesis.

Beside the study the dynamics of interactome, functional interpretation of interaction network is another important field of interactomics. Integration of PPI data with biological pathway information can provide highly useful cues about the functions and dynamics of interactions. Especially for the elucidation of local network structures, knowledge about interrelated pathways can be of crucial importance.

Addressing the need of a more dynamical interactome and functional interpretation, I developed UniHI Express and UniHI Pathway Scanner as two new tools in UniHI database (figure 4.5). UniHI Express allows the filtering of PPI based on the expression in a selected tissue and thus enables the construction of tissue-specific networks. First preliminary studies show that the usage of UniHI express can be highly efficient to prioritize interactions. On the other hand, UniHI pathway Scanner provides the possibility to examine the intersection of canonical pathways from KEGG with the extracted networks (Spirin and Mirny, 2003). Thus, it enables researchers to detect possible modifiers of pathways as well as proteins involved in the cross-talk between different pathways.

UniHI Express: generation of tissue specific networks

Biomedical researchers frequently study processes that occur in specific tissue or cell types e.g. degeneration of neuronal tissue. In contrast, current collections of protein interactions are derived from experiments using various cell and tissue types. Thus, PPI networks retrieved from these resources represent rather a gross summary of possible interactions, thereby neglecting the actual conditions in specific tissues. Particularly, since only a small percentage of proteins correspond to ubiquitously-expressed house-keeping genes, the probability may be high, that many proteins included in retrieved networks, are not expressed in a chosen tissues. Hence, researchers are required to examine carefully the presence of proteins in their model system of interest. This is a considerable task considering the high number of interaction partners that even a small number of query proteins can produce.

I therefore integrated gene expression data with PPI data to allow researchers the construction of tissue-specific networks. As tissue expression data set, the Human Gene Expression Atlas was utilized (Su *et al.*, 2004). To enable the integration with PPI data, microarray probes were mapped to their corresponding Entrez Gene IDs using the annotation by the curators of the Gene Expression atlas. Expression values were averaged over probes which correspond to the same Entrez Gene IDs. To facilitate the use of UniHI Express, the samples were assigned to 19 main tissue classes (i.e. adrenal gland, brain, heart, kidney, liver, lung, prostate, pancreas, placenta, muscle, salivary gland, thymus, thyroid, tonsil, lymph node, testis, trachea, uterus, and uterine corpus). To obtain a unique tissue expression profile, I averaged expression values of tissues samples belonging to the same class.

Using UniHI Express, users can filter the interacting proteins by requiring a minimum expression in the selected tissue class (figure 4.5A and 4.5B). Note that the cut-off value is not applied to the query protein. By adjusting the expression threshold, the PPI network retrieved from UniHI can be reduced to include only highly expressed or extended by include also lowly expressed proteins. Additionally, the PPI resource to be queried can be specified.

Clearly the use of gene expression as proxy for protein abundance has its limitation. However, it can give researchers valuable first indications regarding the prioritization of interactions for follow-up studies. For future releases, I expect that the inclusion of

quantitative measurement of protein abundance in tissues – once such data becomes available on a proteome-wide scale – will lead to considerably more accurate tissue-specific PPI networks. Nevertheless, I consider the current release of UniHI as a first important step towards a dynamic representation of the human interactome.

UniHI Scanner: pathway-focused interaction networks

For the functional interpretation of PPI networks, I have implemented a new web-based tool termed UniHI Pathway Scanner which integrates the UniHI protein interaction data with pathway data from Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). To find the functional relation between query proteins and also between their interaction partners, user can provide a list of query proteins, and a list of pathways to be scanned against identified network. Additionally, user can also select the source of interactions. UniHI Scanner performs three step functions. In the first step, interaction partners of search proteins are identified and a PPI network is created. Subsequently, a pathway network is created from the KEGG pathway IDs provided by user. Edges of this pathway network contain the explicit information about the mode of interactions (such phosphorylation, activation or inhibition). Finally, the two networks (PPI and pathway) are intersected. Nodes and edges in the PPI network are annotated if they are found in the pathway network. For viewing both types of networks, identified and annotated PPI, I have developed a visualization tool (figure 4.5C and 4.5D) which facilitates their interactive and dynamic visualization. At present this tool offers two alternative viewing options. Users can choose between the display of the full PPI network (using the option “Show All”), or the annotated intersection (using the option “Only mapped pathway”). Directed edges in an annotated network represent the molecular relations found in pathways between pathway proteins. I expect that UniHI Pathway scanner will be highly valuable tool to find pathways modifiers of pathways or to detect proteins involved in cross-talk between pathways or for the large community of researchers working in cell signaling.

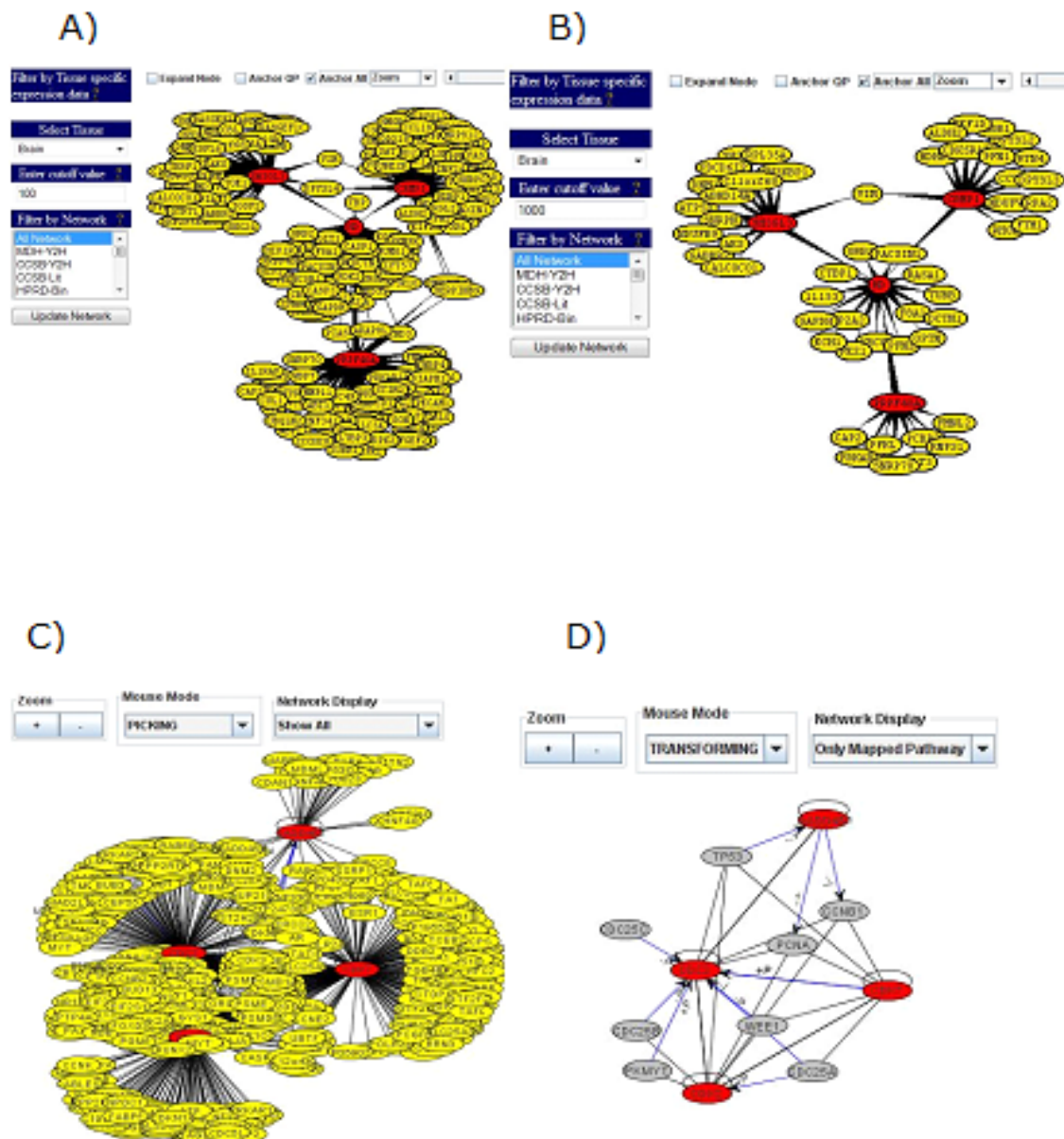


Figure 4.5. Graphical representation of PPIs using UniHI Express and UniHI Scanner visualization tools. Figure 4.5A and 4.5B show the interaction partners (yellow) of HD, CRMP1, SH3GL3 and PRPF40A query proteins (red) in brain tissue using 100 and 1000 as cutoff values. User can additionally filter the interaction partners based on the source of interaction. Visualized network can be further analyzed using options like hideNode, expandNode and collapseNode. Further information of each node and edge can be extracted by selecting the features “functional information” and “link to reference”. Figure 4.5C and 4.5D represent the interaction partners (yellow and grey) of query proteins (red) GADD45, CDK1, CDK2 and CDK7. Grey nodes are annotated proteins found in “cell cycle” KEGG pathway. Feature “Only mapped pathway” allow user to view only annotated proteins and their signals. Directed edges are the flow of the signals like phosphorylation or dephosphorylation, activation or inhibition. Displayed networks can be zoomed in or out using “+” and “-” signs. Whole network or individual nodes can be moved using the features “transforming” or “picked” respectively.

4.4 Discussion and Conclusions

Increasing numbers of human PPI datasets provide enormous amounts of valuable, but frequently unconnected information whose application in biology and medicine is still limited (Goh *et al.*, 2007; Yildirim *et al.*, 2007; Braun *et al.*, 2008; Ideker and Sharan, 2008). Lack of integration and overlap need to be addressed more strongly with experimental and bioinformatical strategies. There is clear necessity to have integrated tools, which provide direct access to distributed interaction data at one common platform. Therefore, I developed a flexible web-based database that integrates the human interaction data from twelve major sources. UniHI is aimed to constitute an integrated platform allowing users to perform simultaneous querying of the major human protein-protein interaction maps.

The architecture and implementation of UniHI aims to overcome the major challenges, i.e. rapid growth, fragmentation and complexity of data. It is intended to assist researchers seeking to utilize human protein-protein interaction data from various sources. Several features included in UniHI enable researchers to perform network-oriented and global analysis of the human interactome. To examine the constitution of UniHI, several statistical analyses were performed regarding network structure and functional annotations of maps integrated (see Chapter 3, Section 3.3.3). Since the scope of UniHI can be expected to be continuously expanding, these analyses will be regularly repeated and presented on the UniHI webpage.

To assess the quality of interaction data, information about co-expression and co-annotation is presented for each interaction pair. I also list for every integrated PPI dataset how protein interactions were validated. The new tools included in UniHI allow researchers a more rapid inspection and prioritization of extracted interactions. Tissue-specific networks can help to focus on biologically relevant interactions, whereas use of pathway information can give important hints about functional modules of interacting proteins.

UniHI does not replace already available interaction maps, but facilitates single portal access to the larger part of the human interactome analyzed so far. Its importance to the scientific community is that it facilitates the assembly of comprehensive lists of protein interactions and that it enables flexible network-orientated querying of interaction maps. Simultaneous querying of multiple interaction maps also promise to

allow identification of network structures which would not be detectable if single maps are analyzed separately. Thus, UniHI provides a highly desirable basis for the systematical utilization of the human interactome in biomedical research. I hope that this unified database and its integration with omics data can provide a convenient platform to support scientists undertaking large-scale systems biology.

5 Functional and Transcriptional Coherency of Modules in the Human Protein Interaction Network

This chapter is an extended version of the following paper:

Matthias E. Futschik, **Gautam Chaurasia**, Jenny Russ and Hanspeter Herzel, (2007), Functional and Transcriptional Coherency of Modules in the Human Protein Interaction Network, *Journal of Integrative Bioinformatics*, 4(3):76.

In chapter 3 and 4, I presented the analysis and integration of current human PPI networks. After integration, this consolidated dataset can be utilized for different aspects of network analysis (as discussed in chapter 2 in Section Applications of Interactomics). In this chapter, my aim was to study the modular structures of the human interactome. This chapter is organized as follows. Section one 5.1 introduces the fundamentals of modularity and discusses few related studies and the motivation for this work. Section 5.2, describes the material and methods employed. Section 5.3 presents the various results of identified modules and their integrated analysis with functional annotation and gene expression data. Finally, section 5.4 summarizes this chapter with discussions and conclusions.

5.1 Introduction

In a cell, a protein complex or module can be defined as a set of genes or proteins which are related by one or more genetic or cellular interactions, e.g. involved in same biological process, or localized in same cellular compartment. In other words, modules can be described as a group of cellular components whose interactions can be attributed to a specific biological function (Hartwell *et al.*, 1999). Several other studies of PPI networks showed that modularity reflects both the tight interaction between proteins to perform a specific functions as well as the need for separation of interfering processes (Bader and Hogue, 2003; Spirin and Mirny, 2003). It is therefore important to identify those modules or complexes of interacting proteins to enhance our current understanding of organization of human PPI networks. Especially, it may help us in characterizing the functions of an unknown protein, by the functions of its related known proteins. To date, a number of studies have been performed to

observe the modular organization in several biological networks, ranging from metabolic (Ravasz *et al.*, 2002), transcriptional (Ihmels *et al.*, 2002; Segal *et al.*, 2003) to PPI (Rives and Galitski, 2003; Spirin and Mirny, 2003; Pereira-Leal *et al.*, 2004) networks.

Whereas majority of previous studies have been performed for the yeast interaction network, for which data have become abundant, the systematic examination of the human protein interaction network, however, was still in an early phase. Therefore, in this chapter, I aimed to gain an overview of the modular structures in the human protein interaction network. For this purpose, first I created an integrated set of interaction network by merging several large literature-based interaction networks. Next, I applied Cfinder algorithm, based on Clique perlocation method developed by Palla *et al.* (Palla *et al.*, 2005), to the integrated set of interaction network for identifying tightly connected clusters of interacting proteins. Whereas previous studies (Rives and Galitski, 2003; Spirin and Mirny, 2003) concentrated on specific subsets of modules, my aim was the systematic assessment of coherency of function, localization and expression of the proteins in the identified modules. For this, identified modules were integrated with functional and localization information from Gene Ontology databases, and gene expression data from Human Gene Atlas database. Details on the datasets and the analysis are provided in following sections.

5.2 Materials and Methods

5.2.1 Human protein-protein interaction data

Data on the human protein interaction network were collected from the Unified Human Interactome database (UniHI) (Chaurasia *et al.*, 2007; Chaurasia *et al.*, 2009). For my analysis, I extracted interactions included in the Human Protein Reference Database (HPRD), Biomolecular Interaction Network Database (BIND) and Database of Interacting Proteins (DIP) (Bader *et al.*, 2003; Salwinski *et al.*, 2004; Prasad *et al.*, 2009). These interactions were derived from the review of published literature. To ensure non-redundancy, I considered only interactions between proteins which could be mapped to their respective EntrezGene identifiers in the UniHI database. Altogether, over 35,000 interactions were extracted. Self- and redundant interactions were excluded from the obtained data leaving a total of over 31,000 interactions between more than 8,400 unique proteins for further analysis.

Note that I only considered binary interactions, i.e. direct interactions between proteins to ensure the reliability of the detected complexes. Complex interactions were excluded as they could otherwise interfere with the computational approach taken here for detection of modules.

5.2.2 Identification of modules in the protein interaction network

The identification of modules was based on the detection of k-cliques, i.e. fully connected subgraph of k vertices. Such k-cliques can form densely connected structures termed as k-clique communities. These communities are the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques, where cliques sharing k-1 nodes are defined as adjacent. Palla and co-authors previously developed a powerful tool Cfinder based on clique percolation method (CPM) for detecting overlapping k-cliques communities in networks (Adamcsek *et al.*, 2006). CPM first locates all k-cliques in a network and then identifies communities by carrying out standard component analysis of the clique-clique overlap. This method has been successfully applied to uncover the complex structure of overlapping communities in several types of networks (Palla *et al.*, 2005). For my analysis, I applied Cfinder to identify highly connected modules in the human protein interaction network.

5.2.3 Generation of random graphs

To assess the significance of the identified cliques, I generated 100 random networks containing the same number of nodes and edges as in original network but with repeated random exchange of interactions. For instance, in such a procedure, two pairs of interacting proteins are randomly picked. The link between the nodes A and B (A-B) and between the nodes C and D (C-D) were changed to A-C and B-D, if such edges are not present in the original network. Note that since this is an undirected network, swapping of edges could happen between any pair of non-interacting nodes in the original network. Though there are several procedures to generate random networks, the current procedure, which I adopted, allows me to generate random networks with the same degree distribution as the original network. These random networks were used to obtain the expected number of cliques and were compared to the number of cliques obtained in the original interaction network.

5.2.4 Protein annotation

For the annotation of proteins, I utilized the Gene Ontology (GO) database supplying information about the assigned molecular function, biological process and cellular location (Ashburner *et al.*, 2000). I assessed the significance whether the detected modules are enriched for proteins of certain functions, processes or locations by application of Fisher's exact test. Since multiple testing was applied, the significance was adjusted by the Benjamini-Hochberg procedure delivering false discovery rates (Benjamini and Hochberg, 1995).

The coherency of modules with respect to cellular location was examined by an assessment of average pair-wise similarity of annotation of the participating proteins. To capture the similarity between two proteins, the induced GO graphs were compared. Subsequently, the size of their intersection divided by the size of their union was taken as a similarity measure (simCC). The values can range between 0 and 1 with larger values indicating larger similarity.

To facilitate the examination of localization of modules, I reduced the set of possible GO terms to so called informative categories. This previously introduced scheme selects GO categories which contain more than 100 genes while each of their children contains less than 100 genes (Zhou *et al.*, 2002). The GO analysis was carried out using the R/Bioconductor package GO and GOstats (Balasubramanian *et al.*, 2004).

5.2.5 Expression data

To assess co-expression of proteins, I utilized a large human tissue expression dataset derived by 158 microarray measurements of 79 different tissue samples (Su *et al.*, 2004). Altogether, the expression level of over ~15,000 genes was measured using Affymetrix HG-U133A and GNF1H arrays. Corresponding transcript levels were derived using Microarray Analysis Suite (MAS5) (Pepper *et al.*, 2007). To improve the data consistency, I additionally applied quantile normalization. Using EntrezGene IDs, I could assign expression levels to approximately 8,000 proteins in our network. Co-expression was measured by the Spearman's rank correlation.

5.3 Results

5.3.1 Identification of modular structures in the human interaction network

For the identification of modules, the Cfinder algorithm was applied to the assembled human interaction network. Altogether, 671 distinct k -clique communities were detected with k ranging from 3 to 11 (figure 5.1). Most of the communities were based on 3- and 4-cliques ($k = 3$: 355; $k = 4$: 200). To assess the statistical significance, I constructed 100 random graphs with the same number of nodes and degree distribution and scrutinized them for the existence of cliques. Figure 5.1 shows the distribution of individual protein communities for different k in the original and random interaction networks.

For $k = 3, 4$ and 5 , similar numbers of cliques were found in random networks. However, for $k = 6$, only an average of 0.1 cliques were detected in the random networks, which is in sharp contrast to the 23 cliques found in the original network. Remarkably, no cliques of size larger than six were found in the random networks indicating the presence of a highly statistically significant modular structure in the human protein interaction network. This also confirms the findings in a previous study of the yeast interaction network that highly interconnected enriched communities did not emerge by chance (Spirin and Mirny, 2003).

5.3.2 Community size distribution

Next, I analyzed the size of the individual communities for all k -cliques. As shown in figure 5.2, I found 267 communities which have less than 5 proteins, most of them belonged to $k = 3$ and 4 cliques. The largest interconnected community containing nearly ~3200 proteins were found at $k = 3$. As shown in the previous studies that best communities' structure are obtained when k -value is between 5 and 6. I also detected several interesting clusters containing proteins ranging from 5 to 15 for k size between 5 and 6. Some of these clusters include transcription initiation, transcription factor TFIIID complex, signalosome complex, intracellular signaling cascade. Only few clusters were detected containing proteins between 15 and 100. These were signalling transduction and regulation of transcription.

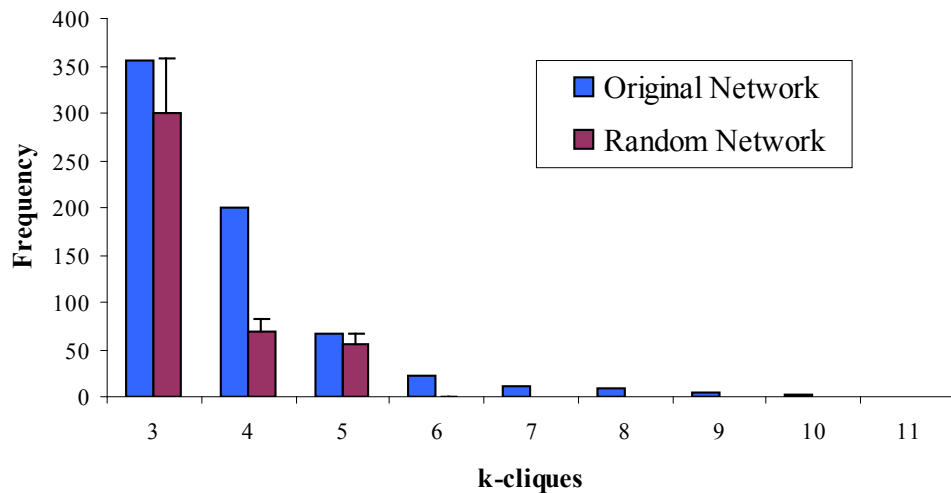


Figure 5.1:

Identification of k-clique communities. The number of identified k-clique communities is shown for the original and random networks.

5.3.3 Distribution of proteins

I further analyzed the distribution of each protein in found clusters and detected several important proteins involved in many complexes. The number of communities in which a protein participates is highly variable (figure 5.3). Nearly ~2000 protein were found to be associated only in single community. Transcription factor TP53 was found to be involved in maximum number of communities (21 communities), also one of most important protein in molecular biology studied so far, and involved in several complex biological processes. Other proteins such as HDAC1, TBP, EGFR, CREBBP, TAF1, CTNNB1, BRCA1, GRB2, and PCNA were found to be involved in more than 10 communities. Most of these proteins are known as transcription factors, and involved in the many signalling and regulatory processes (Table 5.1).

5.3.4 Functional annotation of the detected modular structures

I detected a large number of protein clusters based on k-cliques. But do these cluster structures reflect functional modules in the protein interaction network? To address this question, I used annotation information supplied by the Gene Ontology. Each detected modules was subsequently tested for enrichment of proteins assigned to specific GO categories. Examples of detected modules with annotation information are shown in table 5.1.

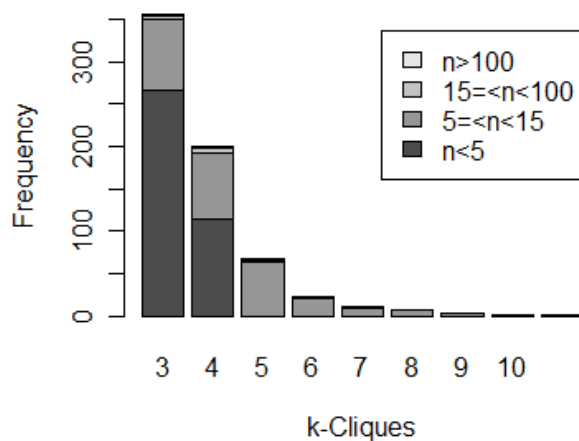


Figure 5.2: Communities distribution original and random interaction networks

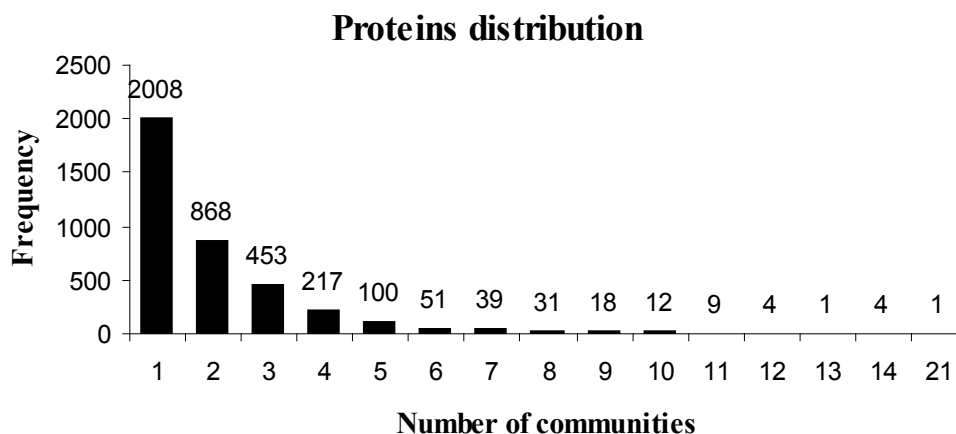
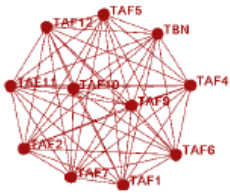


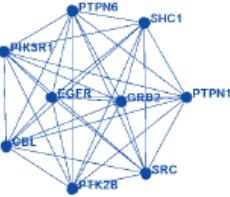
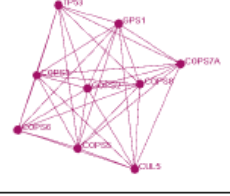


Figure 5.3: Proteins distribution in all communities for all k-cliques.

To facilitate the interpretation, only GO categories are shown that were both significant and representative. Many detected modules could be linked to known physical protein complexes. The largest identified module contained the TATA-binding protein (TBP) and multiple evolutionarily conserved TBP-associated factors (TAFs). The eleven included proteins are all known members of the transcription factor TFIID. Notably, this was also the largest fully connected clique discovered by Spirin and Mirny in the yeast interaction network (Spirin and Mirny, 2003).

Table 1: Table 5.1: Examples of detected protein modules: k - size of cliques, N - number of proteins included in the module, cor - average correlation of expression. The false discovery rates are shown for representative biological processes and cellular components.

<i>k</i>	<i>N</i>	<i>Graph</i>	<i>Proteins</i>	<i>Biological process</i>	<i>Cellular component</i>	<i>Cor</i>
11	11		TAF1 TAF2 TAF10 TAF11 TAF12 TAF4 TAF5 TAF6 TAF7 TAF9 TBN	transcription initiation $6.14 \cdot 10^{-12}$	transcription factor TFIID complex $7.95 \cdot 10^{-26}$	0.39
10	10		BRMS1 BRMS1L HDAC1 HDAC2 ING1 RBBP4 RBBP7 RBP1 SAP30 SIN3A	chromatin modification $5 \cdot 10^{-4}$	histone deacetylase complex $2.84 \cdot 10^{-05}$	0.26
9	10		EXOSC2 EXOSC4 EXOSC5 EXOSC6 EXOSC7 EXOSC8 EXOSC9 KIAA1008 MPP6 SKIV2L2	rRNA processing $3.10 \cdot 10^{-13}$	exosome $8.58 \cdot 10^{-19}$	0.30
9	9		CBL EGFR PIK3R1 PTK2B PTPN11 PTPN6 SHC1 SRC TRKB	transmembrane receptor protein tyrosine kinase signalling $3.71 \cdot 10^{-9}$		0.11
7	9		COPS2 COPS3 COPS5 COPS6 COPS7A COPS8 CUL5 GPS1 TP53		signalosome complex $6.86 \cdot 10^{-20}$	0.44

Similarly, I can confidently link detected modules to the rRNA processing exosome complex and the COP9 signalsome, a highly conserved protein complex whose functions however are poorly understood. In contrast, modules were difficult to relate to known complexes if no prominent association with a specific cellular location existed.

5.3.5 Localization of modules

Previous analyses for yeast indicated that modules in interaction networks can be subdivided into protein complexes and dynamic functional modules (Spirin and Mirny, 2003). Protein complexes consist of tightly interconnected proteins which bind each other at the same time and location. In contrast, proteins in dynamic modules can interact at different times and locations despite being highly connected. To analyse the co-location of proteins in the detected modules, I utilized information about their assigned cellular component in the GO. I reduced the set of possible GO terms to 20 informative categories to facilitate interpretation. Four categories comprised more than 1,000 proteins: 'nucleus' (3,895 proteins), 'intracellular' (1,931), 'cytoplasm' (1,169) and 'integral to plasma membrane' (1,017).

Subsequent analysis showed a remarkably high degree of co-localization of proteins in modules. Of the 316 modules based on k-cliques (with $k > 3$), more than half (170) contained proteins allocated exclusively to a single cellular location. For over 75% of the modules, a majority of the included proteins were assigned to a single location.

Figure 5.4 displays the distribution of coherent locations of the modules. Most of the coherent modules were assigned to the nucleus (65%). Since proteins in steady complexes are necessarily co-localized, this observation may indicate an enrichment of protein complexes located in the nucleus.

5.3.6 Co-expression of modules

Besides the coherency of location, stable protein complexes might be distinguished from dynamic modules based on expression. I would expect that proteins in complexes underlie the same regulatory mechanism and thus would show co-expression. Of specific interest here is the question whether such co-expression correlates with the other distinct feature of complexes namely the co-localization of included proteins.

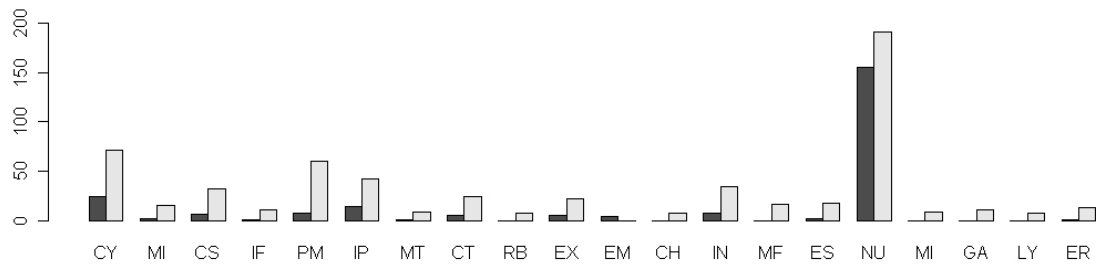


Figure 5.4: Sub-cellular localization of the detected modules. The number of modules is shown for which the majority (black bars) or a fraction of the included proteins (gray bars) was assigned to the corresponding cellular compartment. The distribution is based on the analysis of 316 modules which have a clique size $k > 3$. The following abbreviations are used: CY-cytoplasm, MI-mitochondrion, CS-cytoskeleton, IF-intermediate filament, PM-plasma membrane, IP-integral to plasma membrane, MT-microtubule, CT-cytosol, RB-ribosome, EX-extracellular region, EM-extracellular matrix, CH-chromosome, IN-intracellular, MF-membrane fraction, ES-extracellular space, NU-nucleus, MI-microsome, GA-Golgi apparatus, LY-lysosome and ER-endoplasmic reticulum.

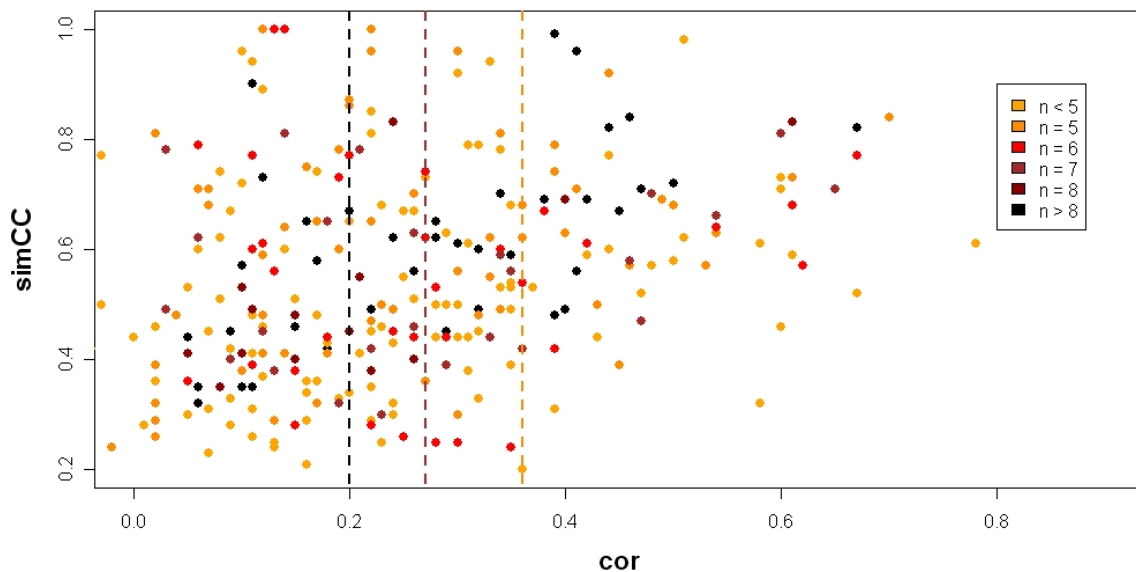


Figure 5.5: Coherency of co-expression and location within modules. The similarity of cellular localization (simCC) is plotted against the Spearman correlation. The size of the detected modules is colour-coded. The number n in the figure legend denotes the number of proteins included in the modules. Dashed lines indicate thresholds for different modules sizes where 99% of the correlation values in random samples are smaller for $n = 5$, $n = 8$ and $n > 8$.

To examine this issue, I calculated the correlation of expression within detected modules. The significance was assessed based on the expected correlation between randomly sampled proteins. Additionally, the similarity of cellular location based on GO annotation was derived (see section 5.3.5). Figure 5.5 displays both co-expression and similarity of location within modules. Comparison of the co-expression with co-localization of proteins within modules yields only a modest correlation of 0.27. This may indicate that a substantial percentage of the detected clusters in the interaction network are dynamic modules.

Inspection of this plot reveals that a majority of the modules containing 10 or more proteins is significantly co-expressed. In fact, 34 out of 51 modules (i.e. 66%) show a correlation coefficient larger than 0.20 for which 99% of equally sized random samples have smaller coefficients. Modules of smaller size are generally less significantly co-expressed due to a higher threshold for significance.

5.3.7 Overlap between modules and identification of linking proteins

Protein interaction networks are organized in multiple levels. Their lowest level is constituted by binding proteins to each other. These binding patterns can lead to the emergence of modular structures as I observed. Furthermore, the modules themselves can be interconnected by functional relationships. One major advantage of the applied algorithm for the detection of modules is that it allows modules to overlap. Thus, identified modules may constitute a higher level network. I exploited this possibility by creating a network of modules to analyse their functional relationship. Selecting modules based on 6-cliques, a highly connected network of 16 modules was detected (figure 5.6).

The largest module within this network contained over 80 proteins of which many are involved in signal transduction. Examples of the included proteins are members of the epidermal growth factor (EGF) receptor family (EGFR, ERBB2), janus kinases (JAK1, JAK2) and signal modifiers such as SOCS1. The second largest module of 51 proteins was enriched by various transcription factors such as the CREB-binding protein, forkhead box O1 (FOXO1), MYC, RB1 and TP53.

The association of the signal transduction module to the plasma membrane and the transcription module to the nucleus was highly significant ($\text{FDR} = 6.00 \cdot 10^{-5}$ and $6.57 \cdot 10^{-21}$, respectively). Notably, these large modules are linked by four proteins (STAT1, STAT3, MAPK1, ESR1) which are known to shuttle between cytoplasm and nucleus.

In contrast, several modules were linked to the transcription module by single proteins. Examples of such sparse interconnections are the linkage of the transcription module to the COP9 signalosome complex by TP53 and to the TFIID complex by TBP.

5.4 Discussion and Conclusions

System-wide interaction network analysis offers the possibility to study cellular mechanisms in a comprehensive manner. However, there are numerous challenges to overcome. Interaction data are still sparse and might be compromised by a large number of false positives and by various experimental biases. In fact, I have recently demonstrated that the approach used for assembling protein interactions networks has severe effects on the resulting networks. For example, signalling proteins tend to be overrepresented in networks based on review of literature (see Chapter 3, section 3.3.3). Thus, it is not surprising that the largest module was associated with cell signalling since our network was constructed using only literature-based interactions maps. I utilized here only such interaction maps to facilitate the interpretation of the results. However, this restriction is likely to limit the number and type of possible modules that can be identified. Nevertheless, this study demonstrates clearly that the constructed human interaction network comprises a large number of functional modules.

My analysis shows that many modules can be assigned to cellular processes. It also indicates that protein complexes and dynamic functional modules can be distinguished based on co-localization and co-expression, although there exists no rigorous threshold to distinguish them.

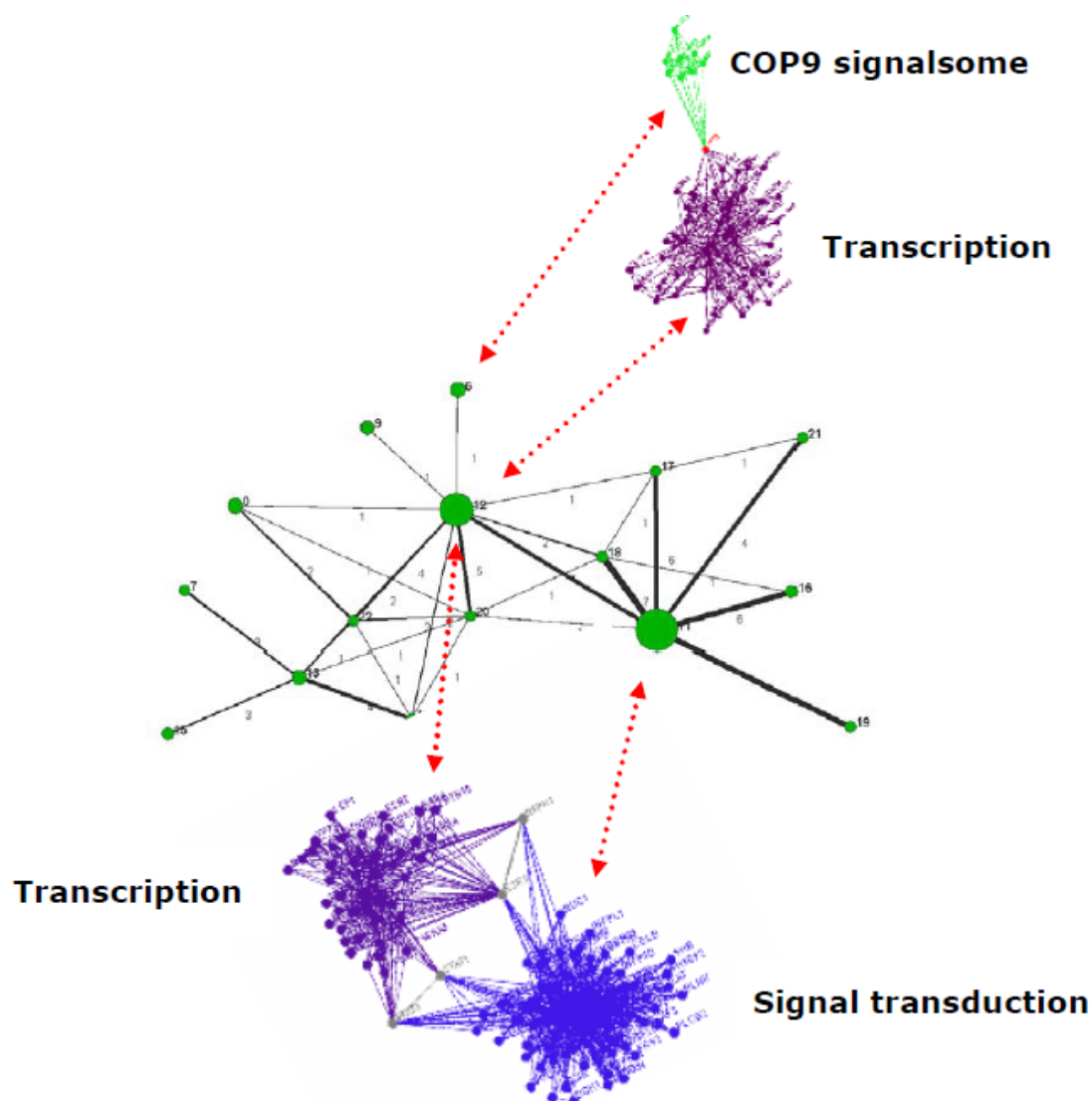


Figure 5.6: Network of modules. Nodes signify detected modules based on 6-cliques. The size of the nodes represents the number of proteins included in the corresponding modules. Edges between nodes indicate the existence of overlap. The width of the edges correlates with the number of linking proteins.

Note that the applied method for detection of modular structures is restrictive, since it requires fully connected cliques. Alternative methods may therefore be favourable to detect less densely connected modules. It should be noted that such restrictive definition of modules leads to an increased robustness of the detected modules regarding false positive interactions. Even if a substantial percentage of interactions are removed, the identified modules will still form highly connected clusters (Spirin and Mirny, 2003). A further major advantage of the applied method is that an overlap between modules is allowed. This enabled to identify potential key proteins linking

different cellular processes. The constructed 'meta-network' of modules gives a first intriguing image of the complex interplay between different components of the cellular machinery.

6 Network-based characterization of brain specific Huntington's disease modifiers

This chapter is an extended version of the following paper:

Martin Stroedicke, Yacine Bounab, **Gautam Chaurasia**, Shuang Li, Stephanie Plaßmann, Jenny Russ, Cecilia Nicoletti, Jan Bieschke, Sigrid Schnoegl, Rona Graham, Josef Priller, Michael Hayden, Stephan Sigrist, Maciej Lalowski, Matthias Futschik and Erich E. Wanker (2010), Brain-specific interaction partners control polyglutamine-mediated huntingtin misfolding and neurotoxicity, (*in review*)

In chapter 2, I discussed the role of PPI networks in biomedical research. In particular, I reviewed several studies applying network-based approaches for predicting disease genes modifiers and the dysregulated biological processes. In this chapter, I focus my analysis to study the disease modifiers involved in a specific disease Chorea Huntington. Furthermore, I introduce a multi-step filtering approach for integrating PPI network with gene expression data and other available HD-relevant pathological data, to predict the tissue-specific dysregulated protein interaction network in Huntington disease. This predicted network is further scrutinized with regard to its function enrichment and validated using several statistical methods. This chapter is organized as follows. Section 6.1 introduces the fundamentals of Huntington disease and reviews few published network-based approaches to predict disease genes. Section 6.2, provides details on the used material and methods. In section 6.3, I will present the various results of this study and *in silico* validation of the predicted network. Finally, section 6.4 discusses the findings of bioinformatic analysis, followed by conclusions.

6.1 Introduction

Chorea Huntington is an autosomal late-onset, monogenic neurodegenerative disorder, characterized by progressive movement disturbances, cognitive dysfunction and psychiatric abnormalities. It is caused by the presence of a dominant mutation in the polyglutamine tract at the N terminus of huntingtin (Htt) protein, resulting in formation of a mutant copy of huntingtin protein (mHtt). This mutation causes the protein to misfold and aggregate, and finally leading to the disturbances in movement and behavior control. Although the exact functions of Htt and mHtt are unknown, it appears that Htt is essential for neuronal development, while mHtt causes the toxic effects on certain types of cells, particularly in the striatum, composed of caudate nucleus and putamen of brain region (Walker, 2007). But as the disease progresses, other areas of the brain such as globus pallidus, thalamus, subthalamic region, pons, medulla, amygdale, hippocampus, spinal cord, superior olive, claustrum and cerebellum (Vonsattel and DiFiglia, 1998) are also significantly affected, causing the symptoms associated with the functions of these damaged cells.

Htt is a ubiquitously expressed multidomain protein, with glutamine/proline-rich domain at the N terminus (HDCTG, 1993), also known as expanded polyglutamine

tract (PolyQ region) or CAG repeats. In healthy patients, length of the PolyQ region ranges from 11 to 34 glutamine residues. However, an expansion of this region above 36, a pathologic threshold, results in the formation of a mutant huntingtin (mHTT). Individuals with 36 to 40 CAG repeats are rarely associated with Huntington disease, but due to the meiotic instability during paternal transmission, the successive generation may inherit an expanded disease gene causing increased severity of neuropathological changes (Myers 2004). HD symptoms are visible when length of PolyQ region crosses over 40. The expanded proteins undergo a conformational change and form protein aggregates. The accumulation of aggregation has been found to be closely associated with disease progression and psychomotor disturbances (Davies *et al.*, 1997; Sanchez *et al.*, 2003).

In HD patients, strong correlations have been reported between the age of the onset and the length of the CAG repeats (HDCTG, 1993). It has been observed that longer PolyQ region may result in an earlier age of onset and more severe symptoms (HDCG, 1993). For example, in most cases, disease symptoms have been found at the age of 35-50 years, with 40-55 CAG repeats (Vonsattel and DiFiglia, 1998), or even in extreme cases, disease symptoms have been reported at the early age of onset between 20-30 years, with number of CAG repeats reaching over 70 (Vonsattel and DiFiglia, 1998). However, contrasting results have also been reported, demonstrating that two individuals with identical CAG repeat lengths are unlikely to have neuropathological changes at exactly the same age (Gusella and Macdonald, 2009). In a recent study, it has been shown that the number of CAG repeats accounts for about 60% of the variation in age of onset, whereas reminder is attributed to the environmental factors or the presence of genetic modifiers (Walker, 2007).

To understand the underlying mechanism behind disease and the role of these, so called gene modifiers, it is crucial to identify them. More importantly, identifying network between the products of these genes modifiers may provide us the list of proteins and biological processes which are altered during disease pathogenesis. The rationale behind this assumption is that interacting proteins are likely linked to the same or similar phenotype. In other words, proteins triggering the same or similar disease phenotypes may interrelate with or be part of the same pathway. Therefore, identification of such disease specific pathways may help us to find novel drug target for complex diseases and to provide a basis for the new treatments (Oti *et al.*, 2006;

Goh *et al.*, 2007; Kann, 2007; Ideker and Sharan, 2008).

The first such disease-specific PPI networks have already been created for the Huntington disease (Goehler *et al.*, 2004; Kaltenbach *et al.*, 2007), identifying many proteins interacting directly or indirectly with Htt (Harjes and Wanker, 2003; Goehler *et al.*, 2004; Li and Li, 2004; Kaltenbach *et al.*, 2007). A subset of these proteins was also found to colocalize to insoluble htt inclusions in the brain and enhance or suppress the mutant htt phenotype (Goehler *et al.*, 2004; Kaltenbach *et al.*, 2007). However, the effect of the modifiers on mutant huntingtin *in vivo*, in the brain-specific context, and at which stage in the pathogenic process they act is largely unknown. Proteins influencing this process by enhancing, e.g. GIT1 (Goehler *et al.*, 2004) or suppressing (e.g. chaperones Hsp40/70, TRiC, CHIP (Jana *et al.*, 2000; Miller *et al.*, 2005; Behrends *et al.*, 2006; Tam *et al.*, 2006) the mutant Htt aggregation process represent potential Huntington's disease cellular modulators. These studies also clearly demonstrate the importance of PPI networks in disease research. However, a major limitation is that these studies are conducted at small-scale level and may miss important proteins in network.

Recent advances in high-throughput approaches enabling the comprehensive studies of PPI networks resulted in large, highly connected networks. However, these networks are only static picture of the complex networks occurring within the cell, and do not provide direct opportunity to study the complexities and dynamics of disease pathways (Barabasi and Oltvai, 2004). One of the possible ways to pinpoint biologically relevant local networks, and decipher e.g. vital functional modules is to integrate PPI network with other types of information e.g. expression, localization or genetic data (de Lichtenberg *et al.*, 2005; Ergun *et al.*, 2007; Baranzini *et al.*, 2009). Calvano & colleagues integrated transcription profiling data with a protein interaction networks to portray time-dependent endotoxin responses in human blood leukocytes (Calvano *et al.*, 2005). More recently a similar strategy was used by Pujana *et al.* (Pujana *et al.*, 2007), where gene expression profiling was combined with functional genomic and proteomic data from various species to generate breast cancer related network in humans. Using a similar approach, Baranzini & colleges integrated PPI maps with genome-wide SNP markers data for indentifying sub-networks involved in multiple sclerosis, a neurodegenerative disorder (Baranzini *et al.*, 2009).

Here, I have designed a network-based multi-level filtering-out strategy to uncover brain-specific genetic modifiers of Htt-function, by combining the focused Htt-based PPI network with available microarray expression data from HD patient. In subsequent steps I ranked the proteins significantly dysregulated in HD and predicted a caudate-nucleus specific HD network. In the following sections, this approach and subsequent analyses are presented in details.

6.2 Materials and Methods

6.2.1 PPI data source

Information on protein interaction was collected from different sources. Besides protein interactions identified in previous HTT modifier Y2H screens (Goehler *et al.*, 2004) I extracted additional interactions from the UniHI database, which at present represents one of the most comprehensive sources for the human protein-protein interactions (Chaurasia *et al.*, 2009). To obtain experimentally well characterized interactions, only literature-curated (Aranda *et al.*, ; Bader *et al.*, 2003; Salwinski *et al.*, 2004; Breitkreutz *et al.*, 2008; Prasad *et al.*, 2009) interactions from UniHI were added, which are mainly derived from small-scale studies. To ensure non-redundancy, I considered only interactions between proteins that could be confidently mapped to EntrezGene identifiers in the UniHI database.

6.2.2 Microarray data analysis

Microarray data were extracted from two different sources: Gene Atlas by Su *et al.* (Su *et al.*, 2004), which includes expression data for ~15.000 human genes in 79 different tissues, and Huntington's disease versus control expression data by Hodges *et al.* (Hodges *et al.*, 2006) that consists of ~18.000 genes in four different tissues (Caudate, FCBA4, FCBA9 and Cerebellum). Gene expression analysis were carried out with open source R software packages, available as part of the BioConductor project (Gentleman *et al.*, 2004).

Gene Atlas data were processed using MAS5 algorithm (Pepper *et al.*, 2007); and adjusted p-values (i.e. false discovery rates) were calculated based on local pooled error approach to identify the significantly expressed genes in the normal brain. Genes that show significant differential expression (adj. p-value < 10^{-5}), were considered for the further analysis.

For detecting the genes that were significantly differentially expressed in caudate nucleus region of normal and HD patient, I utilized microarray dataset generated by Hodges et al. (Hodges *et al.*, 2006) that compares gene expression levels in 44 human HD brains with those from 36 unaffected controls in 4 different brain regions (caudate nucleus, motor cortex, prefrontal cortex and cerebellum). For the comparison of expression data in both cases (caudate normal vs versus motor cortex, prefrontal cortex and cerebellum) and HD versus normal aged matched controls, I computed empirical Bayes moderated t-statistics with the Limma package, correcting gene expression for the collection site (Boston or New Zealand), gender and age (45, 45–60, 60–70 and 70+ years) using a modified version of the analysis script provided by the Hodges & colleagues (Hodges *et al.*, 2006).

6.2.3 Functional enrichment analysis using Gene Ontology database

Functional analyses of the networks were performed using Gene Ontology database (Ashburner *et al.*, 2000). I utilized GO and GOStats package available at Bioconductor platform, and applied Fisher's exact test to find functionally enriched genes (Gentleman *et al.*, 2004; Falcon and Gentleman, 2007). All genes were tested simultaneously for multiple GO categories, and obtained *p*-values were converted to false discovery rates applying the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

6.2.4 Functional analysis by manual curation

Biological functions were assigned to proteins using the PubMed literature database (<http://www.ncbi.nlm.nih.gov/pubmed/>). Literature searches were achieved by screening titles, abstracts and keywords of publications with the official name and aliases (full name and symbol) of the respective gene/protein. Further, I applied OMIM (Ref) and annotated HD therapy targets (HDTTs) databases to find the enrichment of known disease genes in predicted HD network. HDTTs were obtained from the Crossroad database (<http://www.hdresearchcrossroads.org>).

6.3 Results

The aim of this study was to systematically identify HD modifiers within the cellular context of the disease causing protein Htt. A first model of such context can be assembled from the set of proteins that directly or indirectly interact with Htt. I

assembled therefore an Htt-focused protein interaction (in short Htt-network) that I subsequently scrutinized for disease modifiers. This approach has been previously applied for the successful identification of GIT1 as potent modulator of htt aggregation (Goehler *et al.*, 2004). As Htt, however, represents one of the network hubs, the generated htt network is of considerable size. Thus, systematic general screens for potential HD modifiers within this Htt-network - as applied previously - are demanding in resources and time. For more efficient and rapid identification of modifiers, I therefore designed a novel prioritization scheme based on multi-level filtering and network analysis (figure 6.1).

This scheme can be divided into two phases: In the first phase, multi-level filtering using genome-wide expression data was employed to obtain a highly concise tissue- and disease-specific Htt-network. Each filtering step increased the specificity of the Htt-network and lead to a reduction of the number of included proteins. The filtering steps are driven by the knowledge about the pathogenesis of HD. In specific, they are based on the observations that *i*) HD is a neurodegenerative disease affecting primarily the central nervous system, *ii*) medium spiny projection neurons in the caudate nucleus are especially vulnerable to the affects of mutated Htt and *iii*) considerable expression changes occur in the caudate nucleus during HD (Cowan and Raymond, 2006; Walker, 2007). The first two observations address the issue of selective neuronal vulnerability i.e. why neurons in the brain and especially in the caudate nucleus are damaged by mutant Htt despite its ubiquitous expression. As previously proposed, one reason could be the tissue-specific presence or absence of Htt interaction partner in vulnerable cells (Harjes and Wanker, 2003). This motivated me to inspect the Htt network for tissue specific expression. As a first step, I filtered the network for proteins that are differentially regulated in the brain compared to the rest of the human body (Brain-specific Htt network). This was followed by a second filtering step for differentially expressed proteins in the caudate nucleus (Caudate nucleus Htt network). To obtain insights in the HD pathogenesis, subsequent filtering was applied to reduce the Htt network further only to proteins dysregulated during HD. The final network termed as caudate-nucleus (CN)-specific HD network includes therefore, only proteins that are tissue specific differentially expressed and dysregulated during HD pathogenesis (figure 6.1b). This scheme is described in detail in next section.

In a subsequent second prioritization phase, the proteins in the resulting CN-specific HD network were analysed in detail with respect to their molecular functions, transcriptional regulation and involvement in other neuronal diseases to identify interesting candidates for experimental validation.

6.3.1 In silico construction and analysis of a Huntingtin focused protein interaction network.

The Htt network was assembled using data from UniHI and or detected in a previous modifier Y2H screen (Goehler *et al.*, 2004). A large number of direct interactions (N=62) were found reflecting the role of Htt as a hub in the human protein interaction network and its potential function as a scaffold protein. The set of direct interactors constituted the core neighbourhood of Htt. To establish a more comprehensive image of the cellular context of Htt, I expanded the initial core neighbourhood by proteins that have at least two interactions with the direct Htt-interacting partners. The inclusion of such bridging proteins provided me with an extended neighbourhood and a dense Htt-focused network comprising in total of 509 proteins linked by 1319 interactions. Besides Htt, several other proteins display large number of interactions within the network. Notably, the *GRB2*, *TP3*, *EGFR*, *CREB* binding protein and *CASP3* form highly connected hubs in the Htt network. This indicates that the extended network not only captures the molecular context of Htt itself, but also those of important direct interacting partners. Functional analysis using Gene Ontology criteria demonstrated that the constructed Htt-network agrees well the current knowledge about the molecular role of Htt. I found a highly significant overrepresentation of network proteins participating in transcription, metabolism and signal transduction. This supports that Htt serves as a multi-functional scaffold protein acting in various cellular processes (Harjes and Wanker, 2003; Li and Li, 2004). Notably, I could identify a clear enrichment in proteins involved to cell death cascades, which might correspond to previous observations that Htt is involved in anti-apoptotic activities. Earlier studies also showed that Htt is present in different cellular compartments. I found that proteins in the Htt network are indeed distributed across various cellular locations with approximately half of them assigned to the nucleus and 30% to the cytoplasm. After the construction of the Htt network, I proceeded with the application of several filter steps for detection of potential modifiers.

6.3.2 Prioritization by multi-level filtering using gene expression data

The first filtering step reflects that primarily neurons of the central nervous systems are damaged during HD progression. Thus, I reasoned that potential modifiers could be specifically up- or down-regulated in the brain. To determine differentially expressed genes in the brain, I utilized the gene expression data from Gene Atlas database (Su *et al.*, 2004). After selection of tissues derived from different brain regions, differential gene expression between the human brain and the remaining body was determined. The results of the statistical analysis were subsequently used to construct a first filter for the Htt network. Excluding proteins that did not show significant differential expression (adjusted p-value $< 10^{-5}$), a 'brain-specific' Htt network was derived, connects 56 proteins via 67 interactions.

The second filtering step is based on the observation that the caudate nucleus is the most severely affected brain region in HD patient brains (Cowan and Raymond, 2006; Walker, 2007), suggesting that alterations in gene expression levels in this brain region are crucial for development of HD pathology (Vonsattel *et al.*, 1985). A recent study by Hodges *et al.* (Hodges *et al.*, 2006) comparing gene expression levels in 44 human HD brains with those from 36 unaffected controls confirmed this finding. Besides, in the caudate nucleus, gene expression was measured in the motor cortex, prefrontal cortex and cerebellum. The authors indeed observed the greatest magnitude of differential expression in the caudate nucleus. The availability of expression measures for different brain regions from HD and normal cases allowed me in the following to construct two complimentary filters. First, I filtered the brain-specific Htt network for genes that were differentially expressed in the caudate nucleus compared to the cerebellum and the two cortical regions. Utilizing the same threshold for significant differential expression (adjusted p-value $< 10^{-3}$) as did Hodges *et al.*, resulting in a 'caudate-specific' Htt network, containing 38 proteins and 44 interactions.

Finally, I filtered the caudate nucleus-specific PPI network, created in previous step, using a threshold (adjusted p-value $< 10^{-3}$) for genes that were differentially expressed in the caudate nucleus of HD patient brains compared to healthy individuals. This resulted in a caudate nucleus-specific (CN-specific HD) network of proteins potentially dysregulated in HD pathogenesis (figure 6.1b). It contains 14 proteins that are directly or indirectly linked to htt.

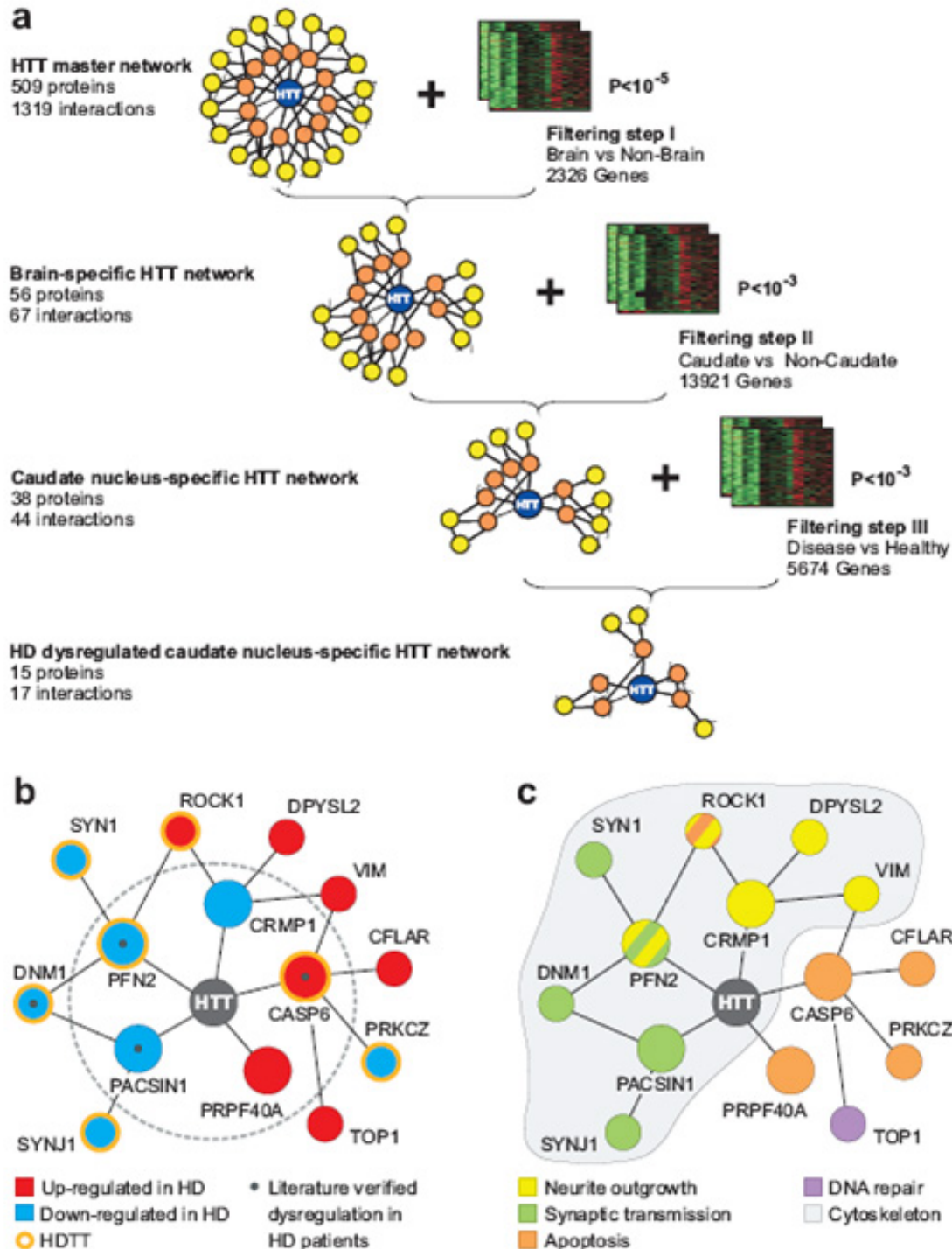


Figure 6.1. Network-based prediction of brain-specific, dysregulated HTT associated proteins.
a, Data integration strategy for interaction network filtering by differentially expressed genes. By systematic integration of protein interaction and gene expression data (three filtering steps) a caudate nucleus-specific HD network with potentially dysregulated HTT associated proteins was predicted. **b**, **c**, Schematic representation of HTT associated proteins. **b**, The predicted dysregulated, caudate nucleus-specific HTT network links 14 proteins directly or indirectly to the disease protein HTT. The proteins marked with a black dot were found previously to be dysregulated in brains of HD patients. The orange ring indicates known annotated targets for

HD therapy development (HDTTs, <http://www.hdresearchcrossroads.org>). c, Many predicted proteins have synaptic functions critical for processes such as endo/exocytosis, neurite outgrowth and synaptic transmission. In addition, proteins involved in apoptosis and DNA repair processes were identified.

Further analysis predicts that 7 of the direct and indirect htt interaction partners are abnormally up-regulated under disease conditions, while 7 proteins are down-regulated (figure 1b). This predicted CN-specific HD network was analysed and characterized in detail in the second phase of prioritization scheme.

6.3.3 Functional analysis of dysregulated HD network

Functional analyses of CN-specific HD network were performed using Gene Ontology (GO) database, OMIM database and by manual curation of published literature. GO analysis showed over-representation of highly enriched biological processes such as apoptosis, cell growth, and endocytosis (adj. p-value= 0.037). These findings were further supported by literature analysis. Four apoptotic proteins, namely CFLAR, PRKC, PRPF40A and CASP6 were predicted by this approach supporting previous observations that cell death pathways are selectively activated in HD brains (Owen *et al.*, 2005; Graham *et al.*, 2006; Caldecott, 2008). Six of the HD network proteins are expressed predominantly in neurons (neuronal polarity regulator CRMP1, actin cytoskeleton component PFN2, synaptic proteins SYN1 and SYNJ1, endocytosis regulators PACSIN1 and DNM1) and play a role in neuronal development. This suggests that processes like synaptic transmission or neurotransmitter release pathways are altered in HD pathogenesis (figure 1c). DNA topoisomerase 1 (TOP1), an enzyme that controls and alters the topologic states of DNA during transcription was also identified (figure 1c), demonstrating the involvement of DNA damage repair during HD pathogenesis (Hodgson *et al.*, 1999). OMIM analysis of CN-specific HD network proteins further supported this approach. Notably, the majority of the network proteins (80%) are implicated in diverse neurodegenerative disorders, e.g. Alzheimer's disease (CASP6, CRMP1, DNM1, DPYSL2, PFN2, ROCK1, and SYN1), Parkinson's disease (CRMP1, DPYSL2), Rett syndrome (PRPF40A, SYN1), mood disorders and schizophrenia (DPYSL2, SYN1, and SYNJ1).

6.3.4 Enrichment analysis using annotated targets for HD therapy development

In order to evaluate the results, I computed the enrichment of proteins from HD therapy development targets (HDTT) database (<http://www.hdresearchcrossroads.org>) in the predicted CN-specific HD network proteins. HDTT is a literature-based manually-curated database, housing 692 genes/proteins, which are considered “important targets” for HD therapy development on rational grounds. I found that the annotated HDTT were significantly enriched ($p = 10^{-4}$ using the Fisher’s exact test) compared to a control human interactome data set obtained from UniHI database (Chaurasia *et al.*, 2009). 7 (PFN2, ROCK1, CASP6, DNM1, SYNJ1, SYN1 and PRKCZ) out of 14 proteins from caudate-specific HD network are found to be annotated as HD therapy targets in HDTD database.

6.3.5 Precision of predicted HD network

Next, I investigated whether the precision of predicted HDTTs in CN-specific HD network is better as compare to the precision obtained using other HD pathology-related datasets (for detail of the dataset, see Appendix C.1). The precision of HDTTs prediction in the different data sets (A-F) was determined using the formula: $TP/(TP + FP)$. Where TP is the number of true positives (number of HDTTs found in the analyzed data set), and FP is the number of false positives (number of genes/proteins in the analyzed data set which are not annotated as HDTTs).

The results of this analysis are shown in Figure 6.2. Analysis showed that HDTTs are predicted with a very low precision in data sets of differentially expressed genes (A-C), while they are predicted with higher precision in a HTT centered PPI data set (D, HTT master network). Analysis of a PPI data set (E) obtained after a 1-step filtration with differentially expressed genes did not significantly increase the precision of HDTT prediction.

However, prediction of HDTTs was about 2-fold higher compared to the HTT PPI data (D) when the PPI data predicted by this approach (F) were analyzed. Thus, the chance to identify HDTTs increases considerably when PPI data are filtered multiple times with gene expression data (3-step data integration strategy). In comparison PPI data that were not filtered or data that were only filtered once with gene expression data (1-step strategy) are less suitable for prediction of annotated HDTTs.

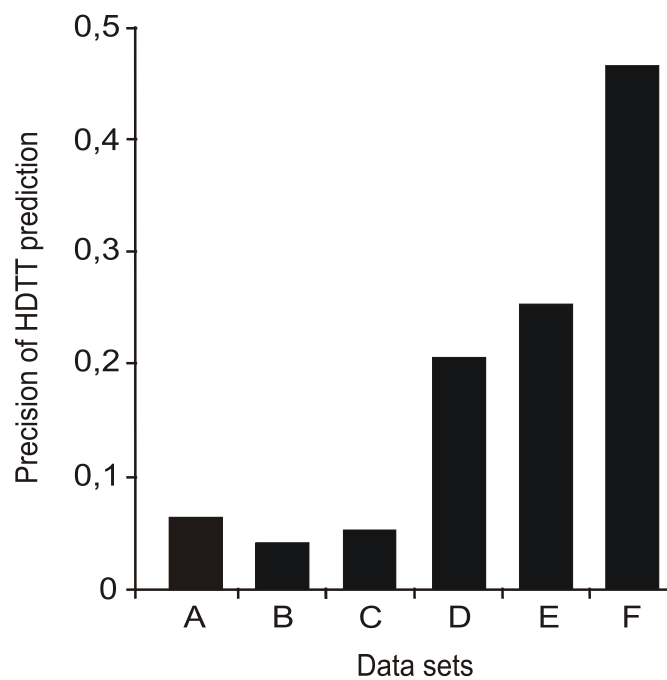


Figure 6.2. Estimating the precision of HDTT prediction (A:0.06, B:0.04, C:0.05, D:0.21, E:0.25, F:0.50) using different HD relevant gene expression and PPI data sets.

6.3.6 Specificity of predicted HD network

I also examined whether the sequential integration of different tissue specific gene expression data sets influences the outcome of the prediction strategy. I systematically compared the caudate nucleus (CN)-, cerebellum (CE)-, motor cortex (MC)- and prefrontal cortex (PFC)-specific integration of PPI and gene expression data in order to elucidate whether similar or dissimilar dysregulated PPI networks are obtained when data sets of different brain regions are combined with this method (figure 6.3). I found that the CN-specific integration of PPI and gene expression data reveals a dysregulated HD PPI network with 14 direct and indirect HTT associated proteins [PPI₄(CN-HD)]. While such a network was not obtained when CE-, MC- and PFC-specific gene expression data were step-wise integrated with PPI data, suggesting that the predicted, dysregulated HD PPI network (figure 6.1b) is CN-specific and cannot be generated with gene expression data of other brain regions (figure 6.4). These findings suggest that the results are relevant for HD pathogenesis, which exhibits a selective neuropathology in the caudate nucleus.

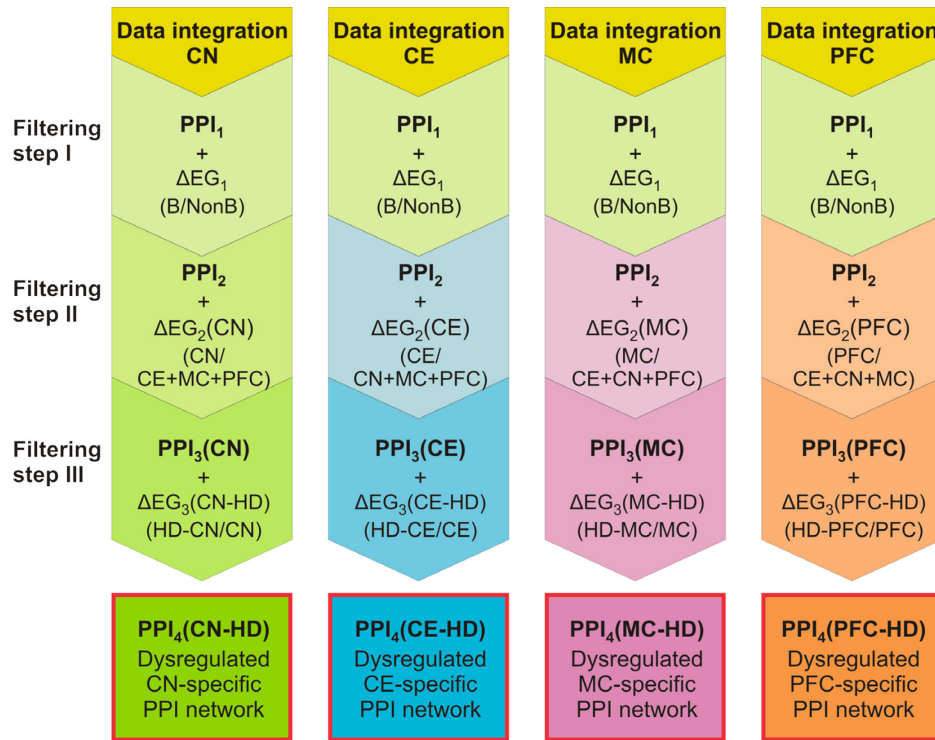


Figure 6.3: Prediction of potentially dysregulated caudate nucleus (CN)-, cerebellum (CE)-, motor cortex (MC)- and prefrontal cortex (PFC)-specific HTT PPI networks. PPI_1 , HTT master network; ΔEG_1 , differentially expressed genes obtained by comparing brain versus non-brain tissues; PPI_2 , brain-specific HTT PPI network; $\Delta EG_2(CN)$, caudate nucleus-specific differentially expressed genes defined by comparing gene expression profiles of the caudate nucleus (CN) with gene expression profiles of the cerebellum (CE), motor cortex (MC) and prefrontal cortex (PFC); $\Delta EG_2(CE)$, cerebellum-specific differentially expressed genes (CE versus CN+MC+PFC); $\Delta EG_2(MC)$, motor cortex-specific differentially expressed genes (MC versus CE+CN+PFC); $\Delta EG_2(PFC)$ prefrontal cortex-specific differentially expressed genes (PFC versus CE+CN+MC); Brain tissue-specific PPI networks: $PPI_3(CN)$, $PPI_3(CE)$, $PPI_3(MC)$, $PPI_3(PFC)$; Differentially expressed genes obtained from caudate nucleus, cerebellum, motor cortex or prefrontal cortex of HD patients and healthy individuals: $\Delta EG_3(CN-HD)$, $\Delta EG_3(CE-HD)$, $\Delta EG_3(MC-HD)$, $\Delta EG_3(PFC-HD)$; Brain-tissue-specific dysregulated HTT PPI networks: $PPI_4(CN-HD)$, $PPI_4(CE-HD)$, $PPI_4(MC-HD)$ and $PPI_4(PFC-HD)$.

6.3.7 Grade-associated analysis of predicted HD modifiers

This multi-level filtering approach identified dysregulated proteins in the immediate molecular neighbourhood of Htt. The dysregulation of those might directly contribute to the dysfunction of mutant Htt enhancing pathogenic mechanisms or interfere with the normal (e.g. neuroprotective) Htt function. Such possible mechanisms would be supported if a significant correlation between the dysregulation and the disease progression could be observed (Vonsattel *et al.*, 1985).

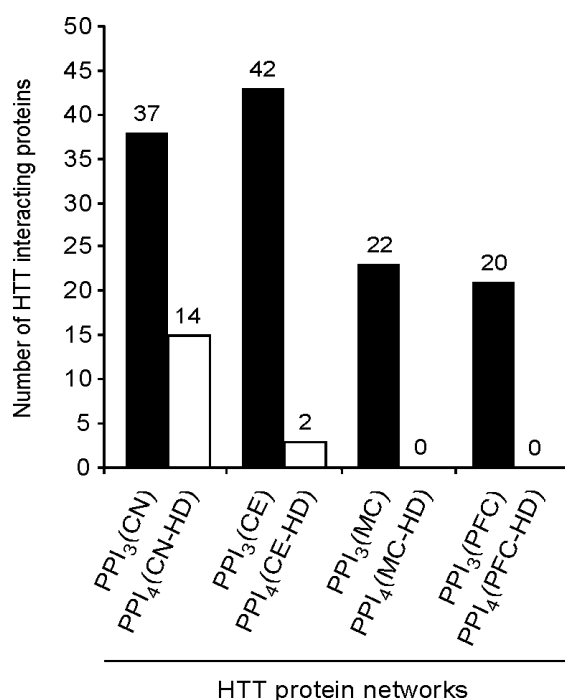


Figure 6.4: Integration of PPIs with brain region specific gene expression data reveals different dysregulated HTT PPI networks. A dysregulated HD network with 14 HTT associated proteins was only obtained when caudate nucleus-specific gene expression data were step-wise integrated with PPI data using this method. After each filtering step proteins that could not be directly or indirectly linked to HTT were excluded from further analysis.

As gene expression data for the caudate nucleus of HD brains with different neuropathological changes (grades 0-4) are available (Hodges *et al.*, 2006), I assessed whether the observed dysregulation of predicted CN-specific HD genes is correlated with the disease grade. I grouped caudate nucleus-specific gene expression profiles of HD brains with mild (grades 0-1, 16 profiles) and more severe neuropathological changes (grades 2-4, 22 profiles) and compared the data with gene expression profiles of healthy controls (36 profiles).

I observed that expression of predicted genes is not only altered in brain tissues with severe neuropathological changes but also in tissues with mild pathological alterations (figure 6.5). Specially, for CRMP1, DNM1, PACSIN1, PRKCZ, SYN1 and SYNJ1, repression for more severe disease grade was apparent (figure 6.5). In

contrast, CASP6, CFLAR, DPYSL2, PRPF40A, ROCK1, TOP1 and VIM showed increased expression for higher grades (figure 6.5).

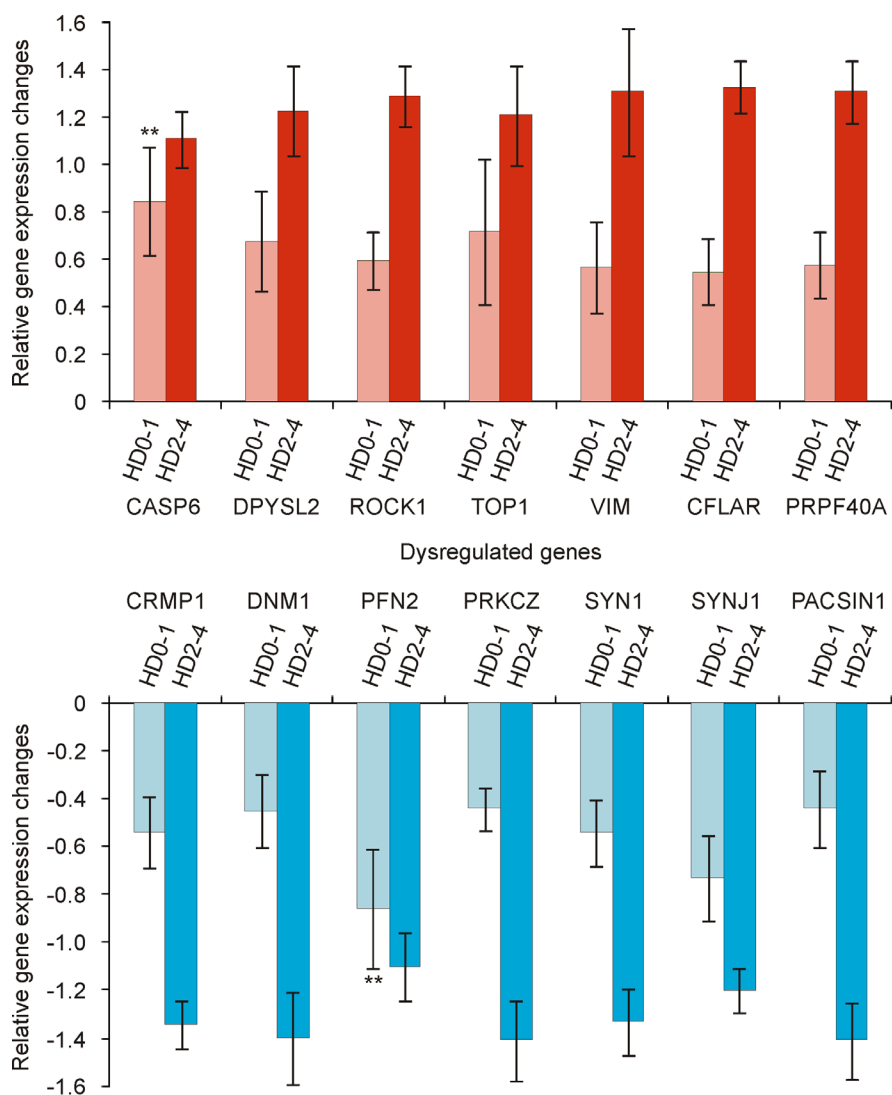


Figure 6.5: Genes predicted by this approach are dysregulated in the caudate nucleus of HD brains with mild neuropathological changes. Expression profiles of HD brains with mild (grades 0-1, 16 profiles) and severe neuropathological changes (grades 2-4, 22 profiles) were compared with expression profiles of healthy individuals (36 profiles). Two stars indicate significant dysregulation ($p^{} \leq 0.001$).**

6.4 Discussion and Conclusions

Interaction networks linking human disease proteins to cellular pathways and functional modules are valuable resources that allow the identification and

characterization of potential disease modifiers. However, such networks often do not provide immediate clues about the pathogenesis and potential disease mechanism. Besides the molecular relationships of a disease protein, additional information about dysregulated proteins is required to construct networks that reflect altered disease processes and permit the prediction of key initiators of a disease cascade. Here, I have developed a generic bioinformatic strategy to create tissue-specific interaction networks that link disease proteins to potentially dysregulated interaction partners in Huntington disease (Figure 6.1). By step-wise integrating microarray gene expression data from clinical case-control studies and specific brain tissues with protein-protein interaction data, a caudate nucleus-specific interaction network of proteins dysregulated in HD was generated. Strikingly, this unbiased, bioinformatic approach allowed the elucidation of known as well as novel modulators of HD pathogenesis.

Functional analysis of predicted CN-specific HD network using GO and published literature indicated a significant enrichment of brain-specific biological process such as apoptosis synaptic transmission, neurotransmitter related pathway, or neuronal development. Furthermore, I found a large fraction of the identified Htt interaction partners (7 proteins) are indeed dysregulated in brains of HD patients and transgenic animals, supporting the value of this unbiased bioinformatic network modeling approach. Analysis of the available literature information confirmed the dysregulation of 4 HTT associated proteins (DMN1, PACSIN1, PFN2 and CASP6) in brains of HD patients (DiProspero *et al.*, 2004; Hermel *et al.*, 2004; Burnett *et al.*, 2008) (Fig. 1b). Further, I also observed a significant enrichment of HD therapy targets (HDTT) in predicted HD network. Seven proteins (PFN2, ROCK1, CASP6, DNM1, SYNJ1, SYN1 and PRKCZ) out of fourteen were found to be annotated as HD therapy targets in HDTT database, demonstrating the power of this predictive approach.

Observation from precision and specificity analysis further supported my approach. Using HDTT database, I observed a 2-fold higher precision value for a network predicted by multi-step filtering approach such as compare with one, computed by only one-step filtering approach (figure 6.4). Results from specificity analysis showed that predicted HD network is CN-specific and cannot be generated when applying in other brain compartments such as cerebellum, prefrontal and motor cortex. Caudate nucleus has been reported as mostly affected brain region in HD patients, almost with 95% loss of neurons, my findings shows that the predicted genes are specific to HD

pathogenesis.

Grade-associated analysis showed that predicted genes were found to be strongly correlated with both type of pathological changes in HD patients i.e. with mild and increased severity changes. Specially, in HD patients with grade 2-4, significant down-regulation ($p\text{-value}<0.1$) for following proteins CRMP1, DNM1, HD, PACSIN1, PRKCZ, SYN1 and SYNJ1 was observed, as compare to the up-regulation for proteins CASP6, CFLAR, DPSYL2, PRPF40A, ROCK1, TOP1 and VIM (figure 6.5). Strikingly, grade-associated changes in expression levels of CRMP1 achieved the highest significance ($p=0.00007$) hinting a significant role of CRMP1 during HD. Based on this initial reasoning, this finding supports the hypothesis that CRMP1 could play a crucial role in the known specific tissue vulnerability (i.e. of the caudate nucleus) during HD. To further check the role of CRMP1 during HD-pathogenesis, several experiments were performed using different model systems in a separate study. Few of the findings from that study are discussed below.

CRMP1 belongs to the collapsin response family of proteins, formed by five members of high sequence similarity in humans (Charrier *et al.*, 2003). While, the other members of the family are more widely expressed, CRMP1 expression is narrowed to the distinct neuronal populations of the central nervous system (Charrier *et al.*, 2003; Bretin *et al.*, 2005). Beyond its signalling function in axon outgrowth and guidance at early developmental stages, CRMP1 is also present in mature neurons both in axons and in dendrites where the function of it largely unknown (Bretin *et al.*, 2005). CRMP1 has been described as a part of Semaphorin3A (Deo *et al.*, 2004; Schmidt and Strittmatter, 2007) and Wnt signaling pathway (Stelzl *et al.*, 2005), important for development of dendritic spines in the brain (Yamashita *et al.*, 2007). Aberrations in the dendritic spines represent early neuropathological changes in HD brain (Guidetti *et al.*, 2001). Recently it was suggested that both CRMP1 and its closely structurally related CRMP2 protein might function as biomarkers in Parkinson's disease (Stauber *et al.*, 2008). Interestingly, hyperphosphorylation of CRMP2 delineates early events in Alzheimer's disease (Cole *et al.*, 2006; Cole *et al.*, 2007).

To investigate the role of CRMP1 during HD pathogenesis, Bounab & colleagues performed functional characterization of CRMP1 utilizing *in vitro* and *in vivo* model systems (Bounab, 2010). Initial analysis showed that the expression level of CRMP1 was significantly decreased in striatal tissues of transgenic mice compared to

controls, also confirming one of my results from this approach. They also observed that overexpression of CRMP1 in a *Drosophila* model of HD, suppressed polyQ-mediated Htt aggregation and improved the photoreceptor degeneration and motor impairment phenotypes as well as survival, indicating potential role of increased level of CRMP1 during HD pathogenesis. Further analysis using *in vitro* model system with purified recombinant human protein (HttQ51) demonstrated addition of the fusion protein GST-CRMP1 to reactions diminished polyQ-mediated Htt aggregation dramatically, while an equal concentration of the control protein GST did not, suggesting that CRMP1-mediated suppression of Htt misfolding and aggregation is highly concentration-dependent. Results from this experimental validation study indicate a potential role of CRMP1 during HD pathogenesis, and thereby clearly demonstrating the usefulness of this integrated bioinformatic approach.

In summary, I have shown that network-based integrated approaches are a powerful strategy, predicting many known and novel modifiers for neurodegenerative diseases. Experimental validations using different model systems have demonstrated that predicted genes are relevant to HD pathogenesis. I hope that this integrative network strategy should be overall useful for the discovery of dysregulated proteins in disease processes, and can be applied to predict modifiers in other diseases as well, given that suitable PPI and gene expression data are available for network modeling.

7 SUMMARY AND OUTLOOK

In this thesis, I presented a unique framework for analysing and integrating the currently available human PPI networks. This framework was further applied in two different studies, resulting in the predictions of protein complexes and genetic modifiers for Huntington disease. In this chapter, I will review the findings of each of the performed studies, and conclude it by discussing the biological importance and the future prospects of the current work.

7.1 Review of findings

Large-scale maps of protein interactions aim to constitute a scaffold for comprehensive models of molecular processes. Similarly to fully sequenced genomes serving nowadays as fundament for genetics, complete maps of protein-protein interactions could serve as a solid basis for a systematic modeling approach of cellular processes. In contrast to the highly successful mapping genome projects, however, the progress in revealing interactomes has been much slower, especially for the human interactome. Only recently, there have been a growing number of both experimental and computational efforts to gain systematical maps of human protein interactome. Although, these maps are likely to provide a better understanding of human biology, careful evaluation of these maps is needed, since each of network generation approaches has its own strengths and weaknesses, which could lead to experimental biases and high rate of false positives interactions in individual maps (Mrowka *et al.*, 2001; Bader and Hogue, 2002; von Mering *et al.*, 2002). Here, I provide a review of my findings from comparative assessment, integration and applications of these maps and discuss their impact on biological research.

7.1.1 Analysis and integration of human Protein-Protein interaction networks

In chapter 3, I addressed the problem of reliability of human PPI networks. To this end, I conducted a comparative assessment of eight different large scale human protein-protein interaction networks (Aranda *et al.*, ; Bader *et al.*, 2003; Hoffmann and Valencia, 2004; Lehner and Fraser, 2004; Salwinski *et al.*, 2004; Brown and Jurisica, 2005; O'Brien *et al.*, 2005; Pagel *et al.*, 2005; Persico *et al.*, 2005; Ramani *et al.*,

2005; Rual *et al.*, 2005; Stelzl *et al.*, 2005; Chatr-aryamontri *et al.*, 2007; Ewing *et al.*, 2007; Berglund *et al.*, 2008; Breitkreutz *et al.*, 2008; Matthews *et al.*, 2009; Prasad *et al.*, 2009). These maps were derived either from Y2H-assays, literature reviews or extrapolated on the basis of homologous interactions in other organisms. The analysis showed that the current maps have only a small, but a significant overlap. Whereas the majority of proteins can be found in multiple maps, this is only the case for less than 10% of the interactions making the maps largely complementary. I detected strong sampling and detection biases linked to the method of generating the maps. For example, RNA binding proteins were overrepresented in orthology-based maps, whereas signal transducers were over-proportionally sampled in literature-based maps. A significant depletion of membrane proteins was observed in all networks and not only in Y2H-based maps as expected. Moreover, maps were generally more concurrent if they were based on the same method. These findings will be necessary to consider in future application of these maps. I also observed that some previous conclusions for network structures in lower eukaryotes cannot be reproduced for humans. For example, protein hubs may not be separated as previously reported indicating that present view of modularity in networks may have to be modified (Maslov and Sneppen, 2002). The results of my analysis suggest that the structure of interactomes of higher eukaryotes might differ substantially from those for lower organisms and, thus, general re-evaluation of concepts regarding network structure and evolution may be warranted. A more dynamic view of network evolution is also indicated by a comparison which I performed for hubs in different maps. It proposes that hubs can be divided into different evolutionary categories. Ancient hubs include proteins of core machineries as the proteasome and the polymerase whereas evolutionary novel hubs are mainly involved in signal transduction and regulation. This classification suggests that the current theory of simple preferential attachment may be not sufficient, but that network hubs have arisen to meet the particular requirements of an organism (Barabasi and Oltvai, 2004).

Learning's from chapter 3 suggested that current human PPI networks share complementary information, and integration of them, therefore, could be very beneficial. However, integration of data from heterogeneous sources is not an easy task, as data was basically generated using various experimental conditions, applying different identifiers, and moreover is stored frequently in different formats. It required,

therefore, a careful analysis of current challenges existing in human PPI data and steps needed for the successful integration. To meet these challenges, I designed and implemented a database for integrating human PPI networks from different sources. This integrated framework was termed as UniHI. In its latest version UniHI houses over 250,000 interactions between more than 22,000 unique proteins collected from twelve major PPI sources. For the quality assessment of the each interacting pairs, UniHI provides several measurements such as co-expression, co-annotation. It, additionally, provides information how the interactions were validated. UniHI offers several tools to perform biologically meaningful and focused analysis. For example, it allows users to construct tissue-specific networks or to map pathway information on extracted network.

7.1.2 Analysis of modular structure of human PPI networks

In chapter 5, I presented the study of modular structure of human interactome. For my analysis, I extracted the interaction data from UniHI database, creating a literature-based large protein network consisting of over 30,000 interactions. Subsequent analysis identified more than 670 modules based on the detection of cliques using a module finder tool “Cfinder”. Inspection showed that these modules included numerous known protein complexes. The extracted modules were scrutinized for their coherency with respect to function, localization and expression, thereby allowing me to differentiate between stable and dynamic modules. Finally, the examination of the overlap between modules identified key proteins linking distinct molecular processes.

7.1.3 Prediction of Huntington disease modifier

Finally, in chapter 6, I developed a network-based prediction method for identifying the genetic modifiers for Huntington disease, an autosomal neurodegenerative disease. This method was based on the integration of huntingtin-specific PPI network and gene expression data from HD patients in a multiple steps. Using this approach, a brain-specific Htt protein-protein interaction (PPI) network was created, linking 14 potentially dysregulated proteins directly or indirectly to the disease protein. Comprehensive literature analysis suggested the role of many predicted modifiers in apoptotic and cell growth pathways and in neurodegenerative diseases. Follow-up analysis of identified network indicated the potential role of CRMP1 during HD

pathogenesis. CRMP1 is a neuronal specific collapsin response mediator protein 1 (CRMP1), important for axonal growth, cell survival and adult brain plasticity. Experimental validation study (Bounab, 2010) has shown that CRMP1 down regulates the formation of insoluble aggregates and reduces mutant Htt toxicity in HD models. My approach demonstrated that perturbed, disease-relevant human PPIs are predictable by network modelling strategies.

7.2 Future Directions

Human interaction maps are rapidly increasing in size and have proven to be highly valuable for the study of human health and disease. The wealth of interaction data, however, poses also new challenges in the follow-up analysis for researchers. I have also learnt from my experiences that implementing such analysis could be very time-consuming and requires expertise from several domains. A possible solution will be to develop workflows which can be applied for certain type of analysis, for example, performing large-scale network analysis to study the topological properties of disease genes, or even to predict disease modifiers or dysregulated biological processes. However, implementation of such framework from scratch would not be an easy task, and would require lot of work. With latest version of UniHI, I have already implemented first important tools in this direction to provide biologists a user-friendly platform to perform integrated systems biology analysis. In following section, I will discuss the possible future research directions based on my work.

7.2.1 Scope and extension of UniHI

The primary goal of UniHI was to provide the comprehensive information on human interactome at one integrated platform. To date, it has been very successful in fulfilling its intention. Latest citations of UniHI are very encouraging and show that UniHI data has been applied in many studies (Futschik *et al.*, 2007a; Goodman *et al.*, 2007; Yue *et al.*, 2008; Ammann and Goodman, 2009; Kamburov *et al.*, 2009; Keshava Prasad *et al.*, 2009; Navratil *et al.*, 2009). Therefore, UniHI will continue to extend its scope by the incorporation of newly available PPI resources and to consolidate the frequently divergent data.

However, there are few areas, which can be improved to make UniHI even more convenient System Biology platform. For example, graph-based and functional

analyses of PPI networks have been applied to study the structure of PPI networks, and to check quality the PPI data. Additionally, topological properties can be used to study the role of disease proteins in PPI networks (Jonsson and Bates, 2006; Platzer *et al.*, 2007; Goni *et al.*, 2008). UniHI interface provides information on topological and functional analyses done for the PPI data included in UniHI. But, these results are static, whenever a newly PPI data is added to UniHI, it is a plenty of manual work to update these results. However, it would be sensible to automate this process; therefore, a possible extension of UniHI could be to implement workflows to perform the network analysis and update the results on its web page. Furthermore, it would be useful to extend those workflows in a way where a biologist can also easily analyze the topological properties of his interest of proteins in a searched PPI network.

Another issue of UniHI DB is the multiple search interfaces. Currently, UniHI offers three different search tools to analyze and visualize PPI networks. All these tools provide integrated information on PPI either with expression or pathways. However, systems biology analysis demands integration of all kind of relevant information in one platform. Therefore, it would be useful to integrate three different UniHI applications into a single platform.

7.2.2 Quality of PPI maps

Although the size of human PPI network is growing rapidly, but the quality of the data still remains a challenge (Bader and Hogue, 2002; von Mering *et al.*, 2002; Chaurasia *et al.*, 2006; Futschik *et al.*, 2007a). My analyses have shown that current PPI maps are scanty and likely to include a considerable number of false positives (Chaurasia *et al.*, 2006; Futschik *et al.*, 2007a; Futschik *et al.*, 2007b). An unbiased estimate of the quality of interaction data sets would require the availability of 'gold standard' for true positive and true negative interactions (Jansen and Gerstein, 2004). In contrast to yeast, such sets do not exist for human protein interactions. In fact, it is doubtful if such sets will ever exist: Many interactions depend on accurate post-translational modification of proteins and occur in a tissue-specific manner. Thus, different sets of true positive interactions might have to be defined for different tissues. Even more challenging is the construction of a set of pairs of proteins that do not interact independently of a chosen physiological condition. Thus, to date, one popular approach to estimate the quality of human interactions is to examine whether the

proteins are co-localized, co-expressed or associated with the same function. In UniHI database, I have currently implemented several schemes using genome-wide study of human gene expression under normal condition and Gene Ontology data to validate the protein interaction, or to create tissue-specific networks (Chaurasia *et al.*, 2009). But, the limitation of this scoring scheme is that it is currently based on expression data only from one study, and if expression information of particular gene is missing in the dataset, then confidence score for this gene with its interacting partners cannot be computed. Therefore, this needs to be further consolidated by integrating expression data from other studies also. Additionally, this PPI data may be also integrated with protein sequence and domain information, due to the fact that interaction proteins are likely to have similar domain. I hope that integrating PPI data with all possible functional information will surely help us to reduce the number of false positive and subsequently to get good quality of PPI data.

7.2.3 Implementation of the network-based strategy for the prediction of disease genes in UniHI

In last decade, network-based approaches have been very popular for predicting disease modifiers and the dysregulated biologic processes (Calvano *et al.*, 2005; Oti *et al.*, 2006; Ergun *et al.*, 2007; Pujana *et al.*, 2007; Baranzini *et al.*, 2009). In this thesis, I have also developed a network-based approach to predict the dysregulated genes in Huntington disease. As explained in the previous chapter (Chaper 6, section 3.1), my approach was based on the multi-step interaction network filtering steps, in which PPI data was integrated with expression data from healthy and diseased HD patients using several steps. This approach successfully predicted many known and novel gene disease modifiers. This method is very scalable, and can also be extended to study other the human diseases, provided required data is available. However, this integration demands manual collection of data and lots of programming efforts for performing such type of analysis. Therefore the future aim of UniHI would be to provide such a facility within UNIH framework, to automate the process of finding disease modifiers and analyzing them in combination with disease-relevant biological data. Especially, user can upload his/her own expression data to filter the searched network for a particular condition or disease, and this filtered network can then be integrated further with other biological pathway or functional data to identify the dysregulated biological processes and functions.

7.3 Conclusions

To conclude this thesis, early applications have indicated the large potential of network biology in many research areas. Progress in experimental techniques and computational methods will continue to improve the coverage and sensitivity of interaction networks. A focus of interactomics - especially in its application to disease research – will be on the combination of different types of networks, such as protein-protein, transcriptional regulatory and metabolic networks, to enable the creation of detailed molecular models of critical diseases such as neurodegenerative disorders and oncogenesis. Furthermore, the integration of interactions networks with the rich datasets generated by on-going disease-related sequencing, microarray or imaging projects is likely to provide us with molecular maps of unprecedented detail for the human organism in health and disease. Thus, network biology promises to substantially contribute to a better understanding of the complexity of disease and eventually to its cure.

Appendix A

Table A.1: Enlarged overview of interaction maps compared. ‘Proteins’ refer to the number of proteins included in the corresponding interaction map before mapping to EntrezGene ID. ‘Protein mapped’ refer to the number of proteins that could be mapped to EntrezGene ID and, thus, were included in the comparative analysis. Substantial loss occurred for OPHID and HOMOMINT, where proteins were primarily referenced by their UniProt IDs.

Network	Proteins	Proteins mapped	Interactions	Interactions mapped	Percentage of self-interactions	Average Degree	Number of Networks >1000/>100/>10/ 1	Method
MDC-Y2H	1703	1703	3186	3186	1.1	1.9	1/0/38/4	Y2H-ASSAY
CCSB-H1	1549	1549	2754	2754	5.1	1.8	1/0/90/27	Y2H-ASSAY
HPRD	6206	5908	20940	15658	4.2	2.7	1/0/135/140	LITERATURE
BIND	4275	2677	5872	4233	13.5	1.7	1/3/169/256	LITERATURE
COCIT	3737	3737	6580	6580	0	1.8	1/7/545/0	LITERATURE
OPHID	4787	2284	24993	8962	0	3.9	1/3/95/0	ORTHOLOGY
ORTHO	3870	3503	11651	9641	2.0	2.8	1/2/183/9	ORTHOLOGY
HOMOMINT	4129	2556	10182	5582	8.1	2.3	1/0/85/45	ORTHOLOGY

Table A.2: Number of proteins shared between interaction maps.

	MDC-Y2H	CCSB-H1	HPRD	BIND	COCIT	OPHID	ORTHO	HOMOMINT
MDC-Y2H	1613							
CCSB-H1	221	1307						
HPRD	741	553	5446					
BIND	363	239	1513	1941				
COCIT	267	156	1591	639	2187			
OPHID	335	219	1133	626	507	1978		
ORTHO	429	350	1219	567	421	857	2838	
HOMOMINT	430	268	1264	574	454	1213	1092	2293

Table A.3: Normalized interaction overlap. The derivation is defined in the Methods section. The percentage corresponds to the shared proportion of interactions within the set of shared proteins. The columns indicate the reference data sets. Thus, the table can be read as follows: 16.4% of the interactions of CCSB-H1 are also included in MDC-Y2H after restriction to the proteins common to both maps. Likewise, 18.0% of the interactions in MDC-Y2H are also included in CCSB-H1. The differences are caused by the different number of interactions in the two maps within their protein overlap.

	MDC-Y2H	CCSB-H1	HPRD	BIND	COCIT	OPHID	ORTHO	HOMO MINT
MDC-Y2H	100.0	18.0	2.4	3.9	1.7	2.7	2.9	4.0
CCSB-H1	16.4	100.0	9.4	15.7	13.5	6.1	12.1	18.2
HPRD	3.4	15.8	100.0	26.6	17.0	17.9	14.8	15.0
BIND	6.4	17.6	45.5	100.0	19.2	36.7	18.0	20.9
COCIT	1.5	19.3	17.7	13.6	100.0	15.8	16.0	10.4
OPHID	1.1	6.0	14.7	33.5	18.2	100.0	21.9	23.9
ORTHO	2.4	12.5	11.3	16.4	21.7	61.6	100.0	41.4
HOMOMINT	3.4	14.2	13.1	19.1	12.3	63.9	37.7	100.0

Table A.4: Log likelihood ratio for concurrence of interactions within the protein overlap between maps. The columns refer to the reference data sets for the pair-wise comparisons. The values for self-comparison were set to zero.

	MDC-Y2H	CCSB-H1	HPRD	BIND	COCIT	OPHID	ORTHO	HOMOMINT
MDC-Y2H	0	4	3.3	3.3	2.7	2.1	3	3.2
CCSB-H1	4.6	0	5	5.6	5	3.8	4.7	5.1
HPRD	3	4.7	0	5.7	5.3	4.3	4.6	4.6
BIND	3.8	5.3	7.1	0	5.7	5.4	5.1	5.2
COCIT	2.4	4.6	5.5	4.7	0	4	5.1	4.4
OPHID	1.5	3.5	5.3	5.6	5.3	0	5.1	5.2
ORTHO	3	4.4	5	5.1	5.4	6.2	0	6
HOMOMINT	3.2	4.7	5.3	5.3	5.1	6.8	6.1	0

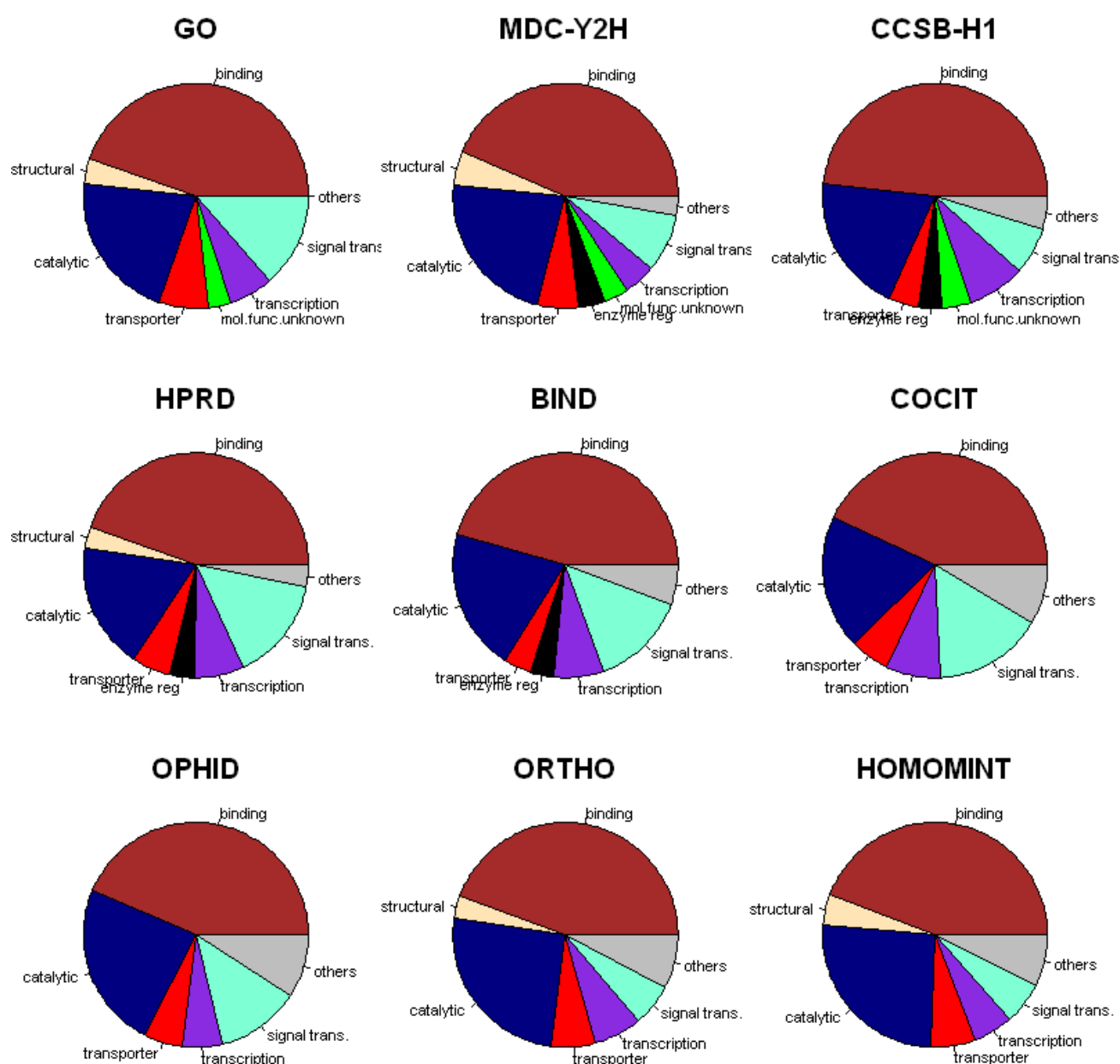


Figure A.1: Composition of interaction maps regarding the molecular function of proteins. The pie plots display the portions of proteins assigned to molecular functions of the first level in the Gene Ontology database. Categories populated with less than 2% of the annotated proteins in a map were merged in the category “others”.

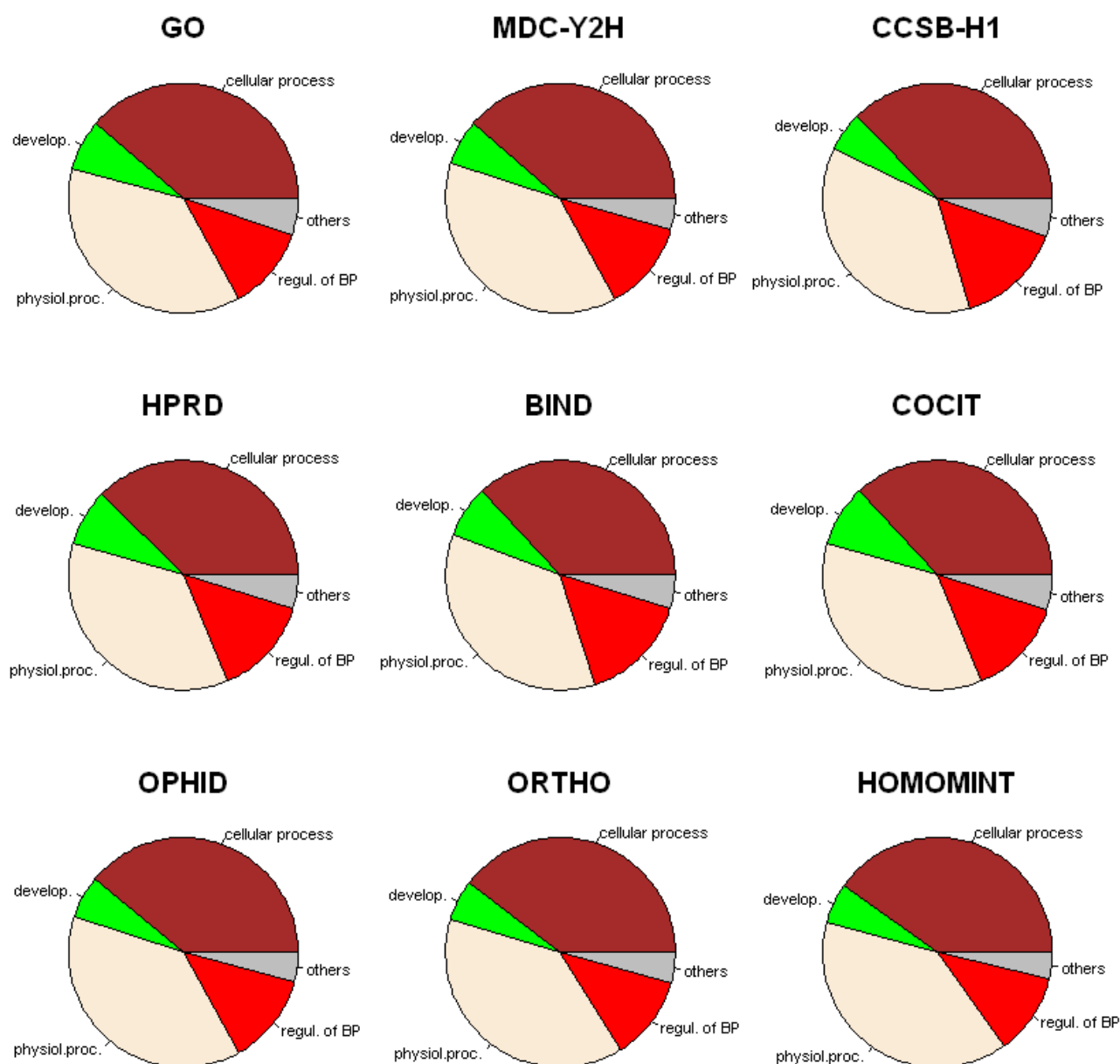


Figure A.2: Composition of interaction maps regarding the biological process to which proteins are linked. The pie plots show the portions of proteins assigned to biological process of the first level of the Gene Ontology. Categories populated with less than 2% of the annotated proteins in a map were merged in the category “others”.

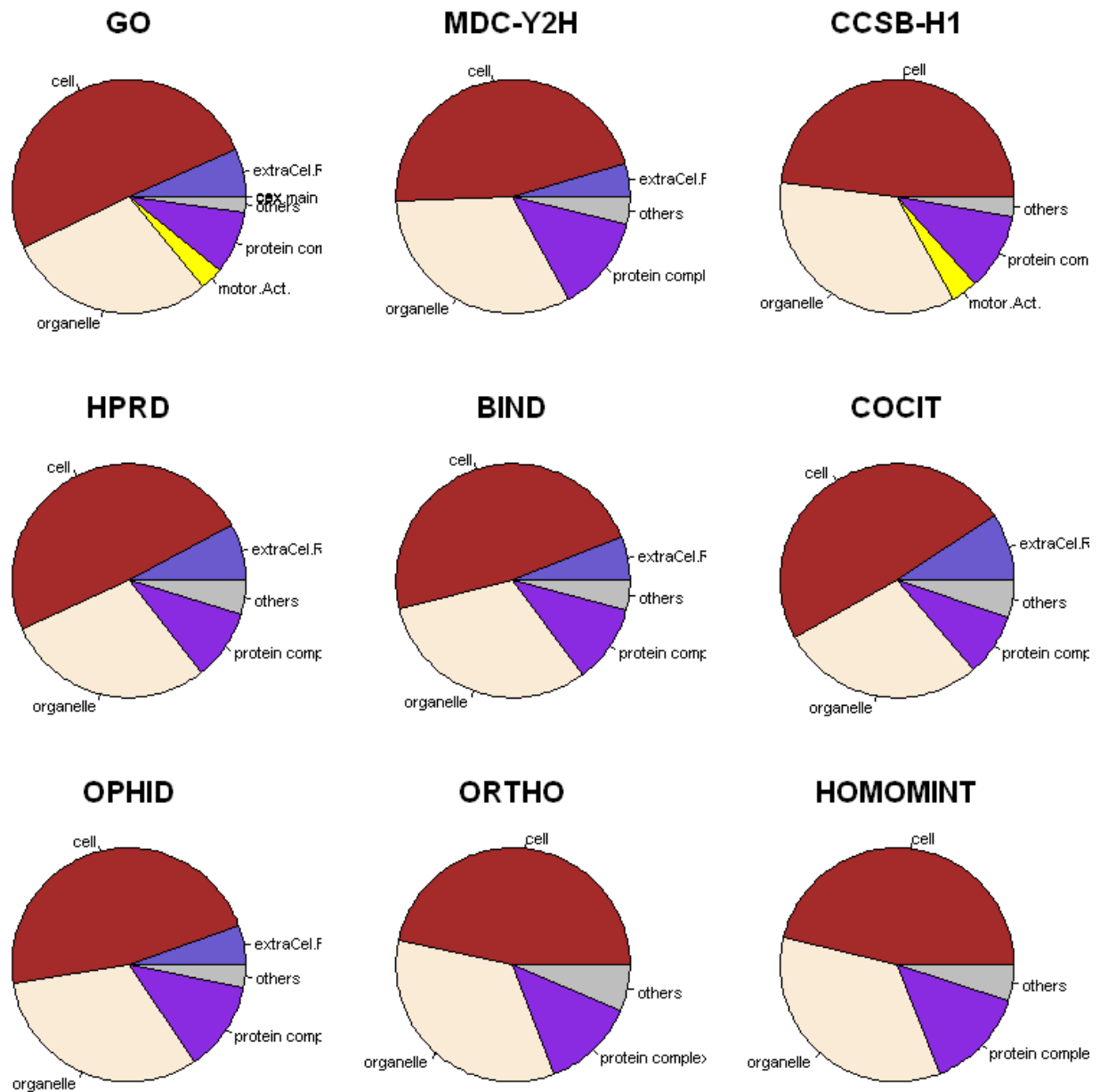


Figure A.3: Composition of interaction maps regarding the cellular component to which proteins were allocated in Gene Ontology. Pie plots show the composition on the first level of the Cellular Component ontology in Gene Ontology. Categories populated with less than 2% of the annotated proteins in a map were merged in the category “others”.

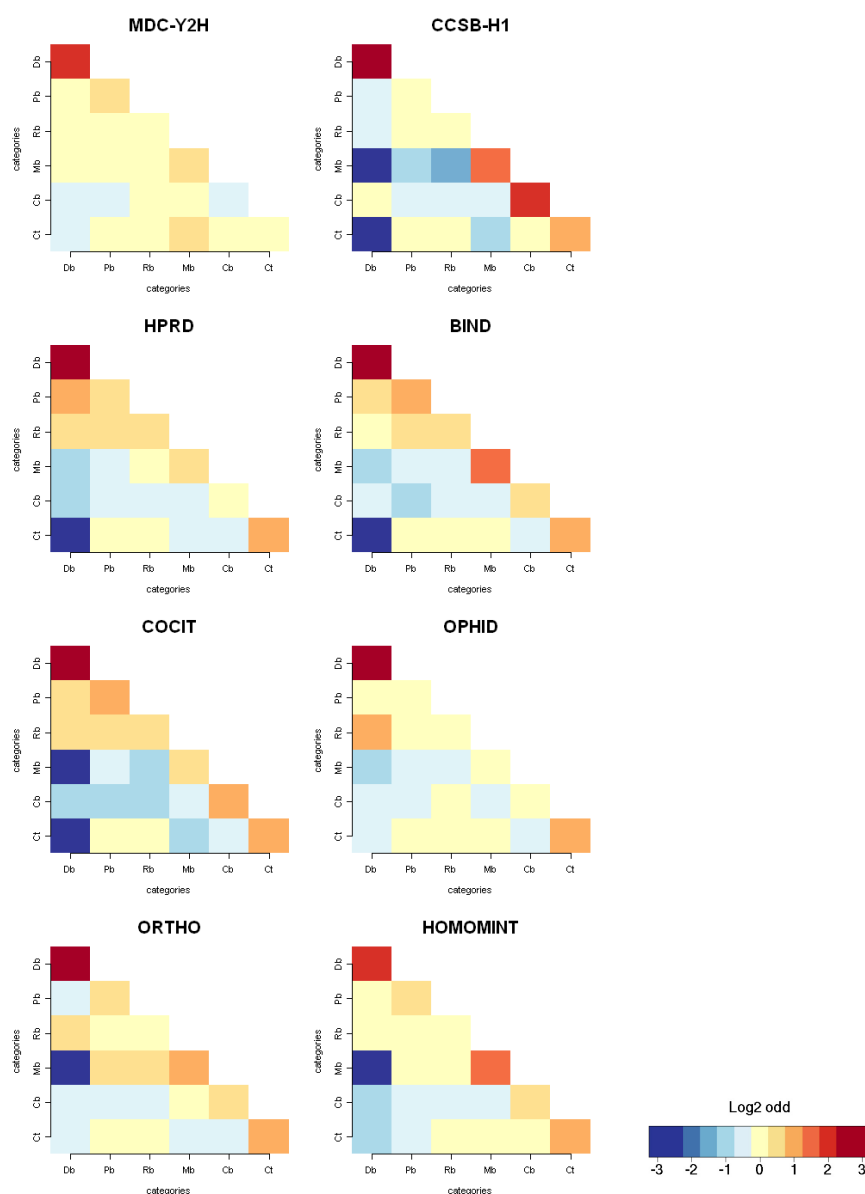


Figure A.4: Molecular function of interacting proteins. Pairs of interacting proteins were mapped to pairs of molecular function terms to which the proteins were assigned in Gene Ontology. The figures display the log odds ratios of the observed distribution compared to distribution obtained from randomized networks with conserved degree distribution. Categories of the third level of the Molecular Function ontology were chosen and the labels are displayed in the figures. For clarity, only GO terms are shown including more than 2% percent of total number of proteins are displayed. The following abbreviations were used: Db- *DNA binding*, Pb - *Purine nucleotide binding*, Rb- *Receptor binding*, Mb – *Metal ion binding*, Cb – *Cation binding* and Ct – *Cation transporter activity*.

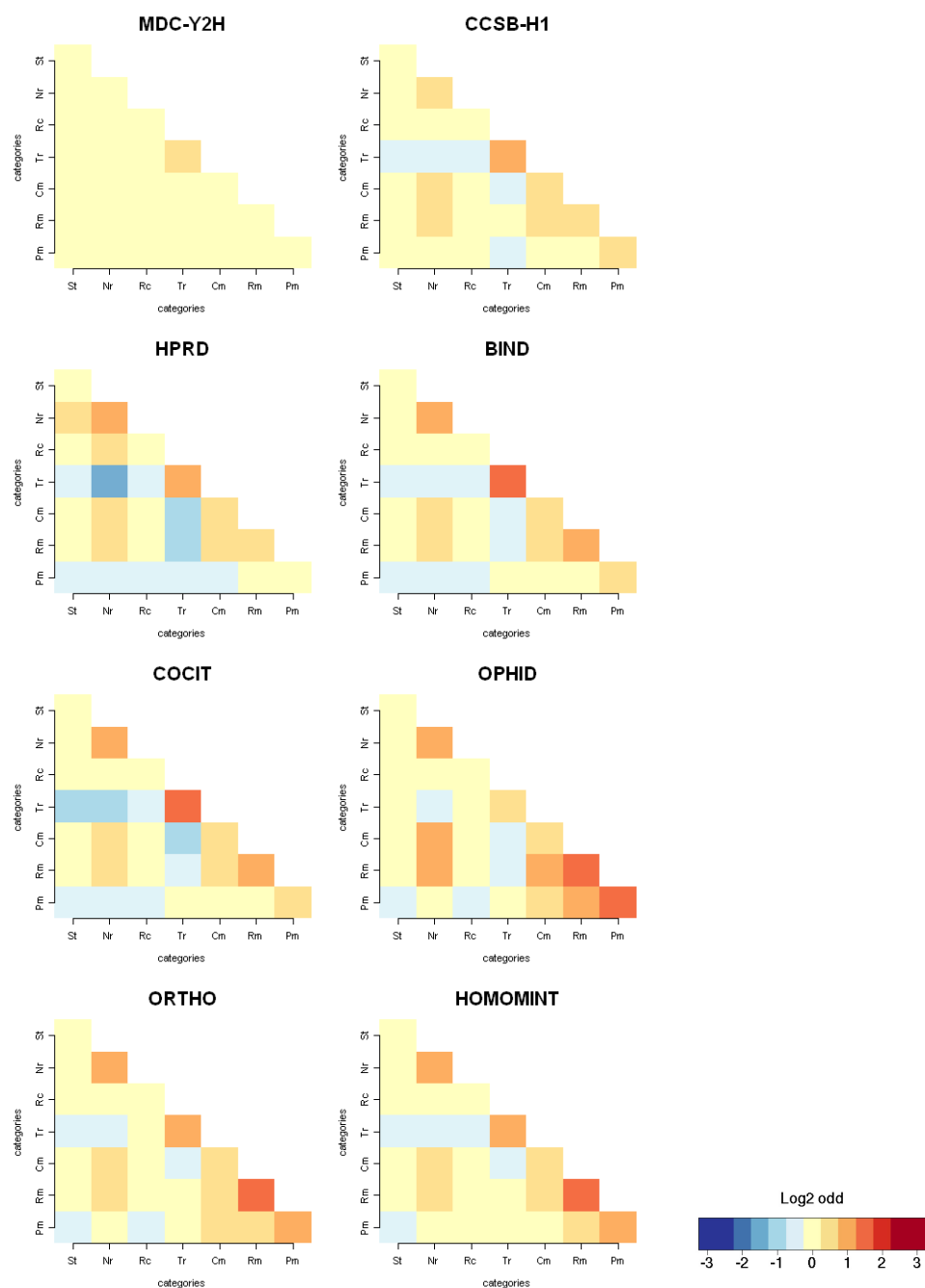


Figure A.5: Biological processes assigned to interacting proteins in Gene Ontology. Similarly to figure S6, the observed concurrence of annotations of interacting proteins was compared to the concurrence expected for corresponding random networks. The following abbreviations were used: St – *Signal transduction*, Nr – *Negative regulation of cellular process*, Rc – *Regulation of cellular physiological process*, Tr – *Transport*, Cm – *Cellular metabolism*, Rm – *Regulation of metabolism* and Pm – *Primary metabolism*.

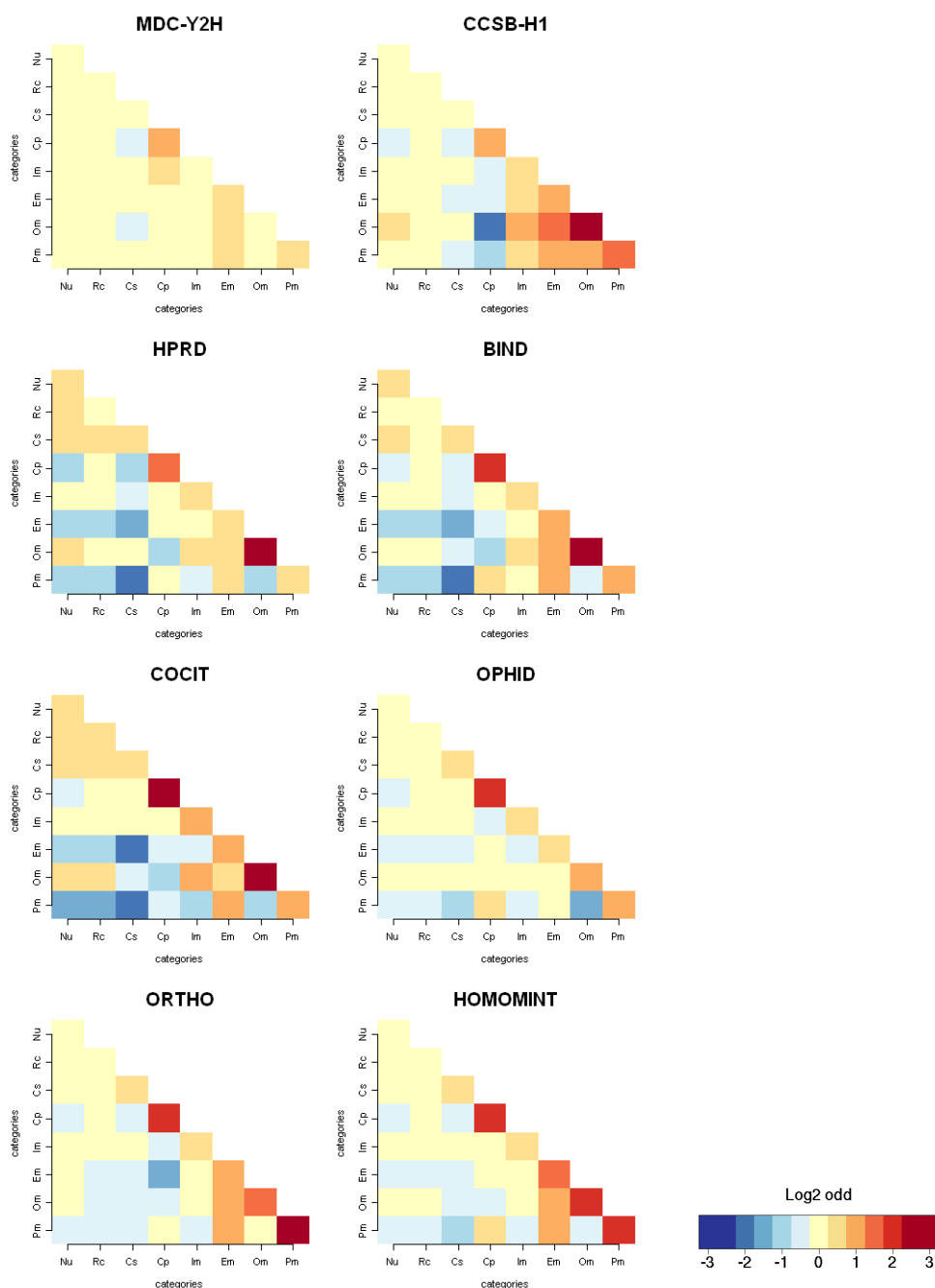


Figure A.6: Cellular component to which interacting proteins were allocated in Gene Ontology. Similarly to figure S6, the observed concurrence of annotations of interacting proteins was compared to the concurrence expected for corresponding random networks. The following abbreviations were used: Nu – *Nucleus*, Rc – *Ribonucleoprotein complex*, Ck – *Cytoskeleton*, Cp – *Cytoplasm*, Im – *Intrinsic to membrane*, Em – *Endomembrane system*, Om – *Organelle membrane* and Pm – *Plasma membrane*

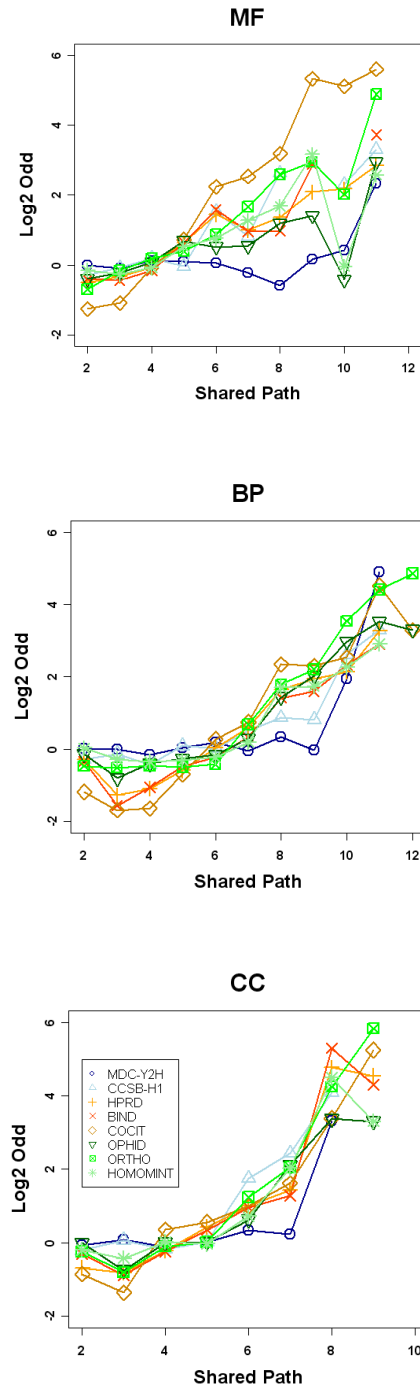


Figure A.7: Concurrent annotation of interacting proteins. The figures display the observed log odds for frequency of shared length within Gene Ontology categories for molecular function (MF), biological process (BP) and cellular component (CC) compared to expected path lengths for randomized networks.

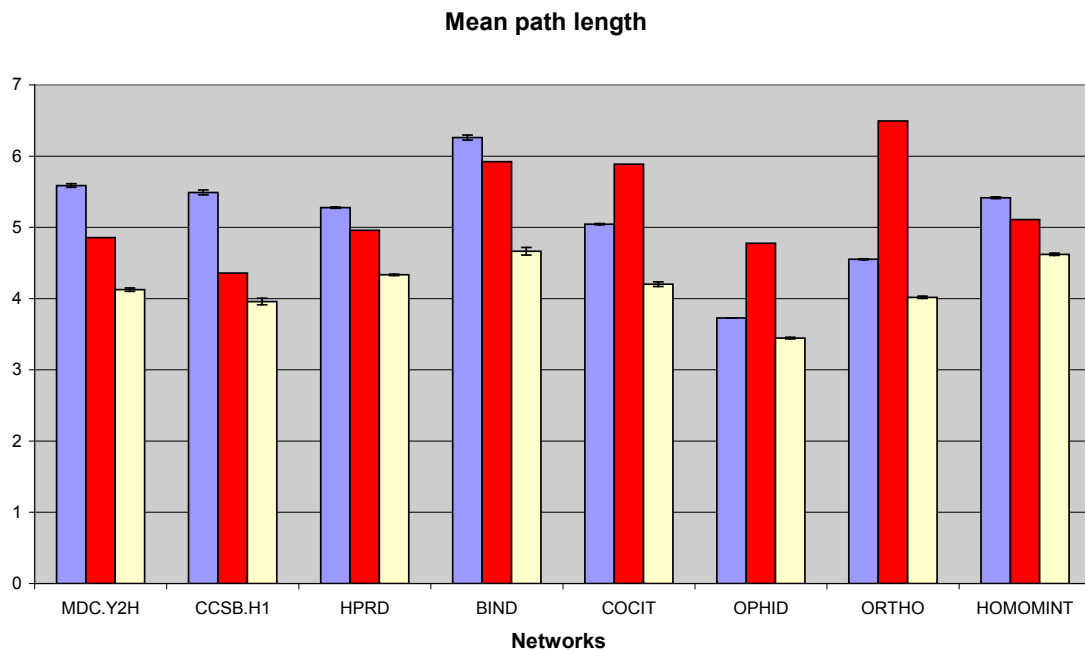


Figure A.8: Mean path lengths of interaction networks: Red bars correspond to original graphs, blue bars correspond to random graph with the same number of proteins and interactions and yellow bars correspond to random networks with conserved degree distribution. All calculations were based on the largest connected graph within the network to avoid artifacts. Errors bars show the standard deviations derived for three independent randomizations.

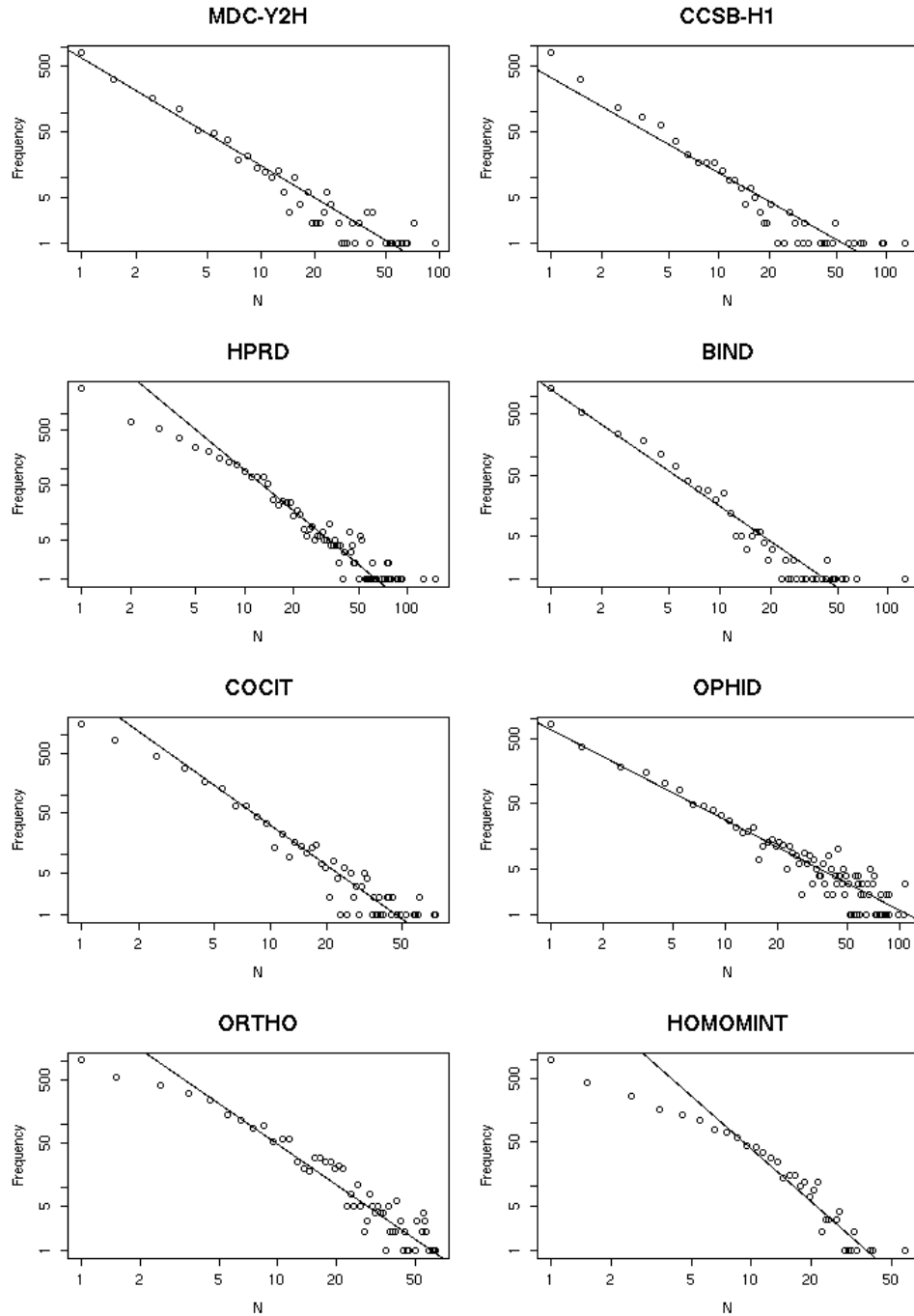


Figure A.9: Degree frequencies. The number of proteins was plotted as a function of the number of neighbors that proteins in the interaction maps have. For all maps, the degree frequencies follow a power-law $P(k) \sim k^{-\gamma}$ with some derivations for HPRD, COCIT, ORTHO and HOMOMINT. The exponent γ was derived by linear regression.

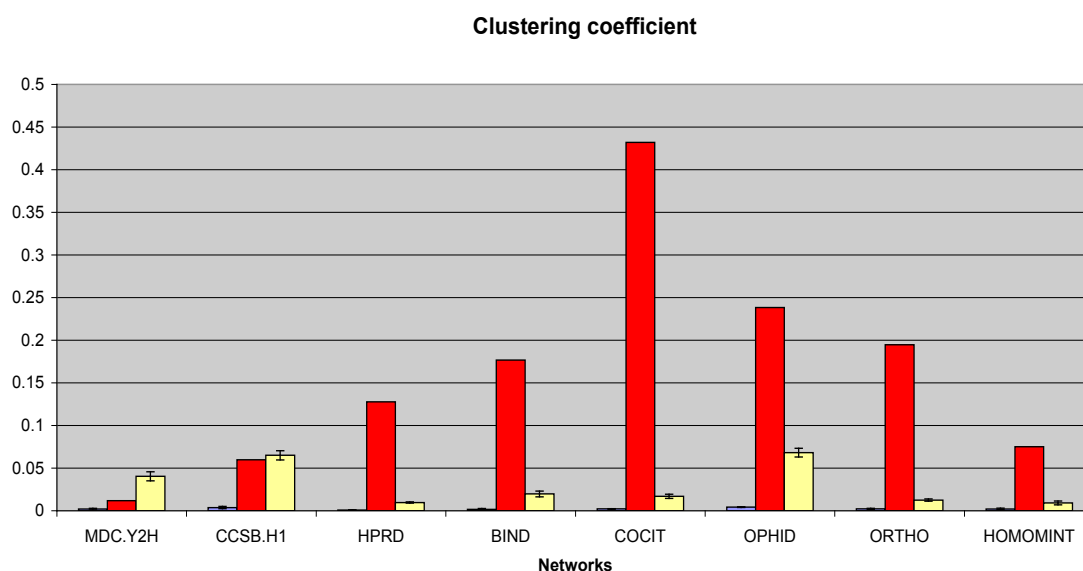


Figure A.10: Mean clustering coefficient of interaction networks. Red bars correspond to original graphs, blue bars correspond to random graph with the same number of proteins and interactions and yellow bars correspond to random networks with conserved degree distribution. Calculations were based on the largest connected graph within the network to avoid artifacts. Errors bars display the standard deviations derived for three independent randomizations.

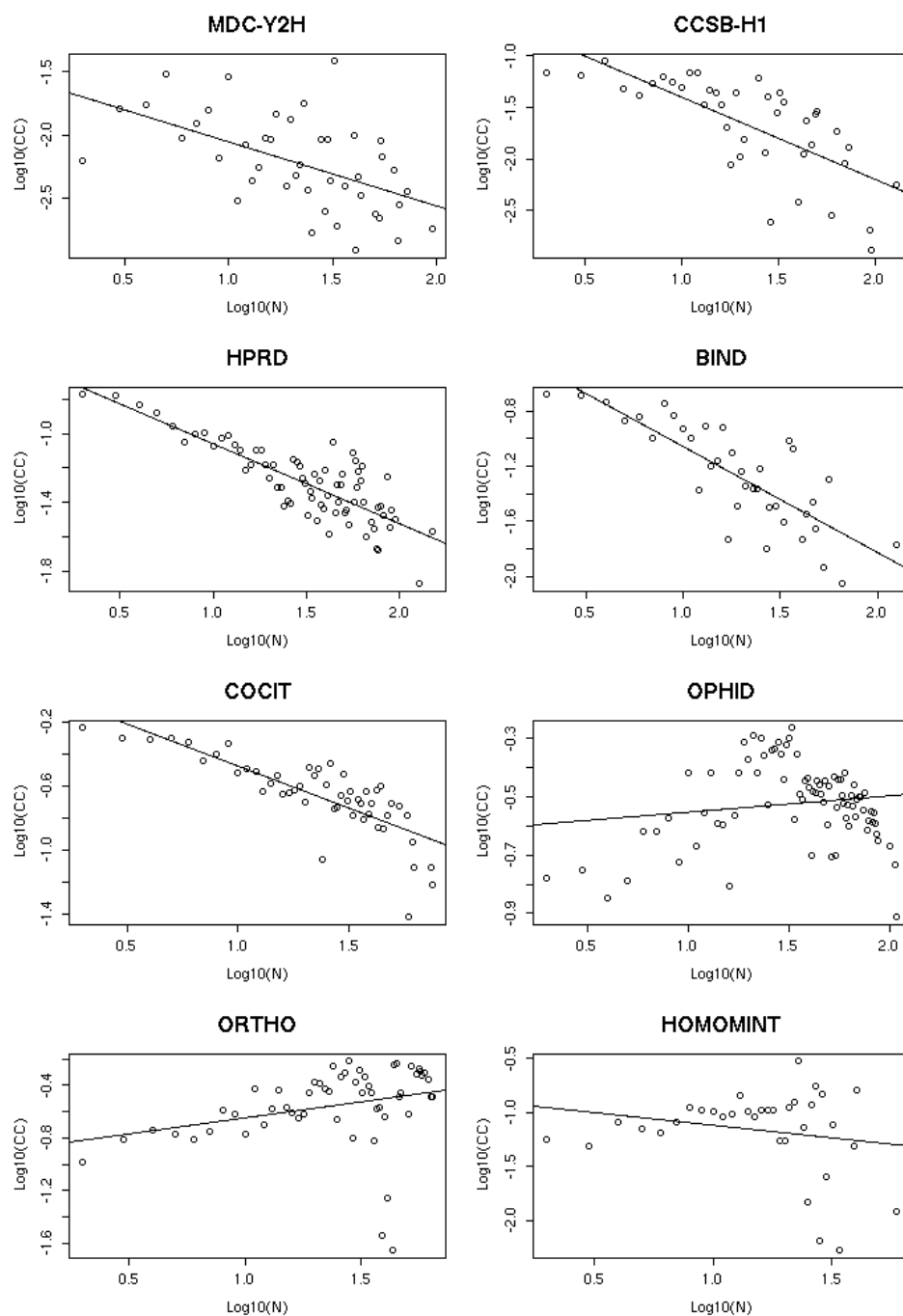


Figure A.11: Clustering coefficient. Plots show the dependence of the clustering coefficient on the degree of proteins. The clustering coefficients shown were derived by averaging over all proteins having the same degree. The solid line shows the linear fit.

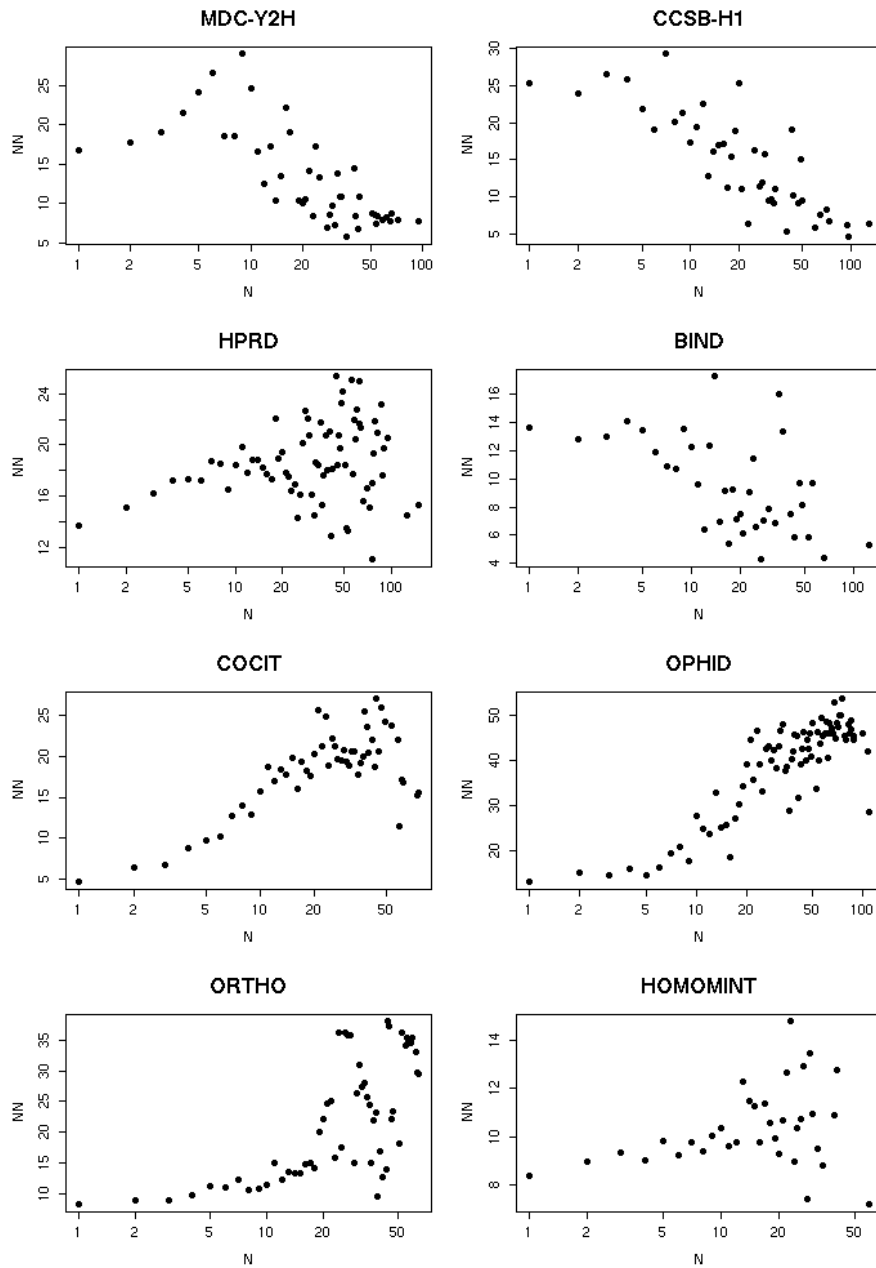


Figure A.12: Average degree of neighbors in the interaction networks. The average degree of a protein's neighbor is plotted as a function of the degree of the proteins. The conjecture that proteins hubs tend to avoid direct interaction should result in a decrease of the average degree of neighbors for interaction-rich proteins. This can be only observed for Y2H-based networks and for BIND.

Table A.5: Spearman correlation coefficients for degree of proteins.

	MDC-Y2H	CCSB-H1	HPRD	BIND	COCIT	OPHID	ORTHO	HOMOMINT
MDC-Y2H	1.00	0.04	0.03	0.18	0.07	-0.07	0.05	0.04
CCSB-H1	0.04	1.00	0.15	0.17	0.25	0.04	0.16	0.07
HPRD	0.03	0.15	1.00	0.47	0.37	0.18	0.22	0.17
BIND	0.18	0.17	0.47	1.00	0.34	0.28	0.18	0.23
COCIT	0.07	0.25	0.37	0.34	1.00	0.28	0.17	0.20
OPHID	-0.07	0.04	0.18	0.28	0.28	1.00	0.42	0.57
ORTHO	0.05	0.16	0.22	0.18	0.17	0.42	1.00	0.40
HOMOMINT	0.04	0.07	0.17	0.23	0.20	0.57	0.40	1.00

Table A.6: Significance of overrepresentation of protein hubs in molecular functions GO categories. Proteins hubs were defined here as the set of proteins within the top 10% of proteins having the largest number of interaction in an interaction map. The same procedure and threshold was applied as for table S7 except that the baseline distribution is the set of all proteins in the corresponding map. Categories are displayed if significant overrepresentation occurred in more than two maps.

MDC- Y2H	CCSB- H1	HPRD	BIND	COCIT	OPHID	ORTHO	HOMOMINT	N	GO Term
		2.09E-13	9.64E-05	0.00043			0.000618	4	binding
			0.009782		8.41E-05	3.41E-05	0.000581	4	RNA binding nucleic acid
			7.59E-05		0.000375		7.90E-10	3	binding
		6.60E-47	1.81E-13	1.37E-17				3	protein binding translation initiation factor
					0.0009	0.001633	7.00E-06	3	activity purine nucleotide
		3.02E-07	0.002466		0.000674			3	binding
		9.88E-14	0.000798	0.00103				3	kinase activity
		4.65E-05	0.00234		0.000228			3	ATP binding adenyl nucleotide
		7.02E-05	0.002356		0.000109			3	binding protein kinase
		7.97E-16	0.000798	4.67E-05				3	activity transferase activity, transferring
		3.57E-12	7.97E-05	0.001734				3	phosphorus-cont phosphotransfer ase activity, alcohol group as
		3.96E-14	0.000634	0.00026				3	acce threonine endopeptidase
				3.46E-06	8.41E-05	1.34E-07		3	activity receptor signaling protein
		3.08E-17	0.00018	0.000408				3	activity DNA-directed RNA polymerase
			0.00998		0.002208	0.000639		3	activity

Table A.7: Significance of overrepresentation of protein hubs in interaction maps for biological processes. The same procedure and threshold was applied as for table S6.

MDC- Y2H	CCS B-H1	HPRD	BIND	COCIT	OPHID	ORTHO	HOMO- MINT	N	GO-Term
		3.69E-05	0.000478		2.78E-08	2.07E-06		4	cellular metabolism
									enzyme linked receptor
		1.07E-17	0.000747	9.38E-07				3	protein signaling pathway
									regulation of progression
		1.35E-11	1.53E-07	8.35E-09				3	through cell cycle
		8.80E-10	1.53E-07	2.21E-07				3	cell cycle
									regulation of cellular
		2.18E-14	4.44E-07	2.98E-08				3	physiological process
									negative regulation of
		1.56E-05	1.84E-06	1.34E-05				3	cellular process
									positive regulation of
		4.82E-09	0.005581	2.88E-12				3	cellular process
									macromolecule
		2.59E-06			0.000185	0.006314		3	metabolism
		8.91E-14	0.008232	0.006616				3	protein modification
		7.31E-25	0.000548	3.54E-15				3	signal transduction
		1.30E-15	0.002081	9.36E-11				3	apoptosis
									protein amino acid
		4.21E-16	0.007241	2.42E-05				3	phosphorylation
		1.43E-15	0.002081	9.36E-11				3	programmed cell death
									intracellular signaling
		1.30E-15	6.37E-05	5.91E-08				3	cascade
		1.25E-14	0.009887	0.000404				3	phosphate metabolism
									ubiquitin-dependent
				0.004379		0.00013	0.000871	3	protein catabolism
									modification-dependent
				0.004379		0.00013	0.000871	3	protein catabolism

Table A.8: Significance of overrepresentation of protein hubs in interaction maps for cellular component. The same procedure and threshold was applied as for table S6.

MDC	CCSB						HOMO-		
-Y2H	-H1	HPRD	BIND	COCIT	OPHID	ORTHO	MINT	N	GO Term
		0.004942	4.78E-07		0.000242	0.000206	3.01E-16	5	nucleus
		0.004805	0.000108		1.19E-12	1.06E-08	2.70E-06	5	intracellular proteasome complex (sensu Eukaryota)
				8.09E-05	7.44E-13	2.00E-14	5.89E-09	4	membrane-bound organelle
		0.001491			0.000756	0.000403	1.13E-06	4	intracellular membrane-bound organelle
			0.001491		0.000756	0.000403	1.13E-06	4	organelle
			5.36E-06		0.001355	0.000188	0.000581	4	nucleolus
					8.26E-10	3.09E-09	4.48E-12	3	protein complex
					0.000811	1.74E-07	1.13E-06	3	organelle intracellular organelle
					0.000811	1.74E-07	1.13E-06	3	organelle
					3.00E-09	3.34E-10	5.07E-05	3	cytosol endoplasmic reticulum
					0.00045	0.00431	0.0079	3	ribonucleoprotein complex
					0.002881	1.13E-09	0.002938	3	proteasome core complex (sensu Eukaryota)
				8.42E-06	8.99E-06	3.98E-08		3	RNA polymerase complex
					0.003284	0.002238	0.0079	3	

Appendix B

B.1 Design and Implementation

As the amount of data on protein interactions is growing rapidly, there is ongoing demand for integrated platforms with a high degree of flexibility. Such platforms should not be only easily accessible but also be consistently updated. Data should be accurately integrated from different sources and queries should be processed in minimal time. The structure of the platform should be extensible to new data without changing its data structure. Thus, a careful design and implementation of the system and the selection of computational approaches to assemble heterogeneous data sources are crucial. Traditional computational approaches like object-oriented software and relational databases can be cumbersome and time-consuming. Typically, persisting data objects from SQL tables with a JDBC (Java Database Connectivity API) connection and prepared SQL statements may be easy for simple objects, but is very complicated for objects with many properties such as proteins and their interaction partners, since they have to be mapped to different domains of similar and complementary information. Thus, I decided to implement UniHI with an object/relational mapping (ORM) methodology (<http://www.hibernate.org/5.html>). ORM tools provide an easy-to-use framework for mapping an object-oriented domain model to a traditional relational database. This technique helps me to reduce the implementation costs of complex SQL queries. ORM takes plain Java objects used in the application and process them using a persistent mechanism which automatically generates all the SQL command needed to store and retrieve the object. Applications built with an ORM tool are cheaper to design, better performing, highly portable and resilient in the face of changes to internal objects or underlying relational models.

B.2 Data Integration

Data integration from different data sources imposes major tasks. They include careful assembly of similar and complementary information from heterogeneous data sources and deletion of duplicated data. To handle this problem, I implemented an integration layer which provides different parsers for importing interaction data coming in different formats, and mechanism for several steps of data preprocessing. Details on the different data sources and their integration mechanism are given in the following sections (Bader and Hogue, 2002; Balasubramanian *et al.*, 2004).

Data Downloading and Parsing

Interaction data are downloaded from different distributed sources (see Chapter 2, Table 2.1) via file transfer protocol (ftp) or hypertext transfer protocol (http). The data are generally available in two different formats either as flat-files or XML-files (eXtensible Markup Language). Most of the interaction databases now release their datasets following the XML-based PSI-MI (Proteomics Standards Initiative - Molecular Interaction) convention (Kerrien S, 2007). Separate parsers using Java application programming interfaces (APIs) have been implemented for extracting information from the XML- and flat-files. SAX (Megginson, 2005) and DOM (Hors, 2004) parser were used for processing the XML files. The extracted information is imported into a temporary database.

Data Preprocessing

As the interaction maps use different identifiers, one of the main challenges in integrating the data is the construction of a unique identifier indexing system. For unification, first, complete lists of proteins for each interaction map were compiled separately. Subsequently, these lists were compared employing information from NCBI (Maglott *et al.*, 2007), HGNC (Bruford *et al.*, 2008) and EnsMart (Kasprzyk *et al.*, 2004) to map their corresponding identifiers in other interaction datasets. After mapping, identical protein identifiers were merged together in a horizontal manner where each protein is a unique entry in the Protein table (see details DB schema). A unique identifier was assigned to each protein entry of this table. These unique identifiers were further used for grouping of the redundant interactions from all interaction datasets. Information on the source of proteins or interactions were merged vertically and inserted into two different tables ProteinSource and InteractionProperties. (Appendix figure B.1).

After integration, some modifications on interaction datasets were also performed. First, I wanted to distinguish between interactions of binary and complex type. For the binary interaction type proteins interact directly, whereas for complex interaction type proteins belong to the same protein complex but do not necessarily interact directly with others. Most interaction networks include either binary or complex type enabling easy distinction. An exception is HPRD, which provides both binary and complex type of the interaction data. To facilitate differentiation between these two categories, I

have split interaction data from HPRD into two sets (HPRD-BIN, HPRD-COMP).

Secondly, large-scale interaction networks are generally derived by literature-reviews, Y2H-assays or are based on observed interactions between orthologous proteins in other organisms. To indicate users the approach taken those interaction datasets were modified where interaction data were assembled by multiple approaches. For example, OPHID contains orthology-based interactions as well as interactions imported from other databases. I extracted only orthology-based derived interactions from OPHID as UniHI already includes the remaining interactions. Similarly, HPRD contains data from large-scale experiments that are separately incorporated in UniHI. Hence, these data were filtered from HPRD interaction map.

Finally, networks based on multiple approaches were divided according to the method used. CCSB-H1 data were split into Y2H- and literature-based interaction maps (CCSB-Y2H, CCSB-LIT). The processed and non-redundant data are inserted into a common relational database using the persistent layer, described below.

Database Scheme

Data stored in UniHI are administered by a relational database using an open source MySQL relational database management system (RDBMS) (<http://mysql.org/doc/#manual>). It consists of nine key tables. The *Protein* table contains a complete list of proteins from all interaction maps. Each protein in this table is stored with its different identifiers (EntrezGene ID, Uniprot ID, Ensembl ID, UniGene ID, OMIM ID) as well as its gene symbol, description, cross-reference database identifiers from HPRD, BIND and DIP, if known. Each protein in this table is a unique entry. The *ProteinAliases* table lists the information of different symbols assigned to the corresponding proteins. The *ProteinSource* table houses information about the occurrence of each protein in different maps. The *GOAnnotation* table stores the information about GO-environment of each protein. The *Interaction* table contains information on interactions. Each interaction in this table is a unique entry. The *InteractionProperties* table gives additional information about interactions such as source of interaction and quality score. For example, interactions of the MDC-Y2H map were categorized as low, medium and high confidence by the authors (Stelzl *et al.*, 2005). The *InteractionScore* table includes information about co-expression and co-annotation of interacting proteins. The *ExperimentDetail* and *DetectionMethod*

tables store the information about the Pubmed IDs and the methods used to detect the interactions. The schema for relational database is presented in Appendix figure B.1.

Persistence

The persistence layer is the core of the whole system and works as middleware for inserting and querying data. All objects implemented within this layer are mapped to tables in the SQL-based relational database. The role of all objects and their event classes are described in Hibernate mapping properties files. These mapping files are used for the communication with

Application

The application layer is implemented using web-services of the J2EE architecture. The main purpose of this layer includes communication with clients via a JBOSS web-server and retrieval of the data from the database using hibernate persistent mechanism. Data retrieval is carried out using the Structured Query Language (SQL) which is implemented in a set of Java APIs. Further, this layer consists of a web interface and a visualization tool. Functions included in this interface enable users to perform not only simple searches for interactions of single protein but also complex network-oriented queries for multiple proteins. It provides additionally several features for refined search and selective use of interaction maps. Validation schemes provided with each interaction map are also included to assess the quality of each interaction.

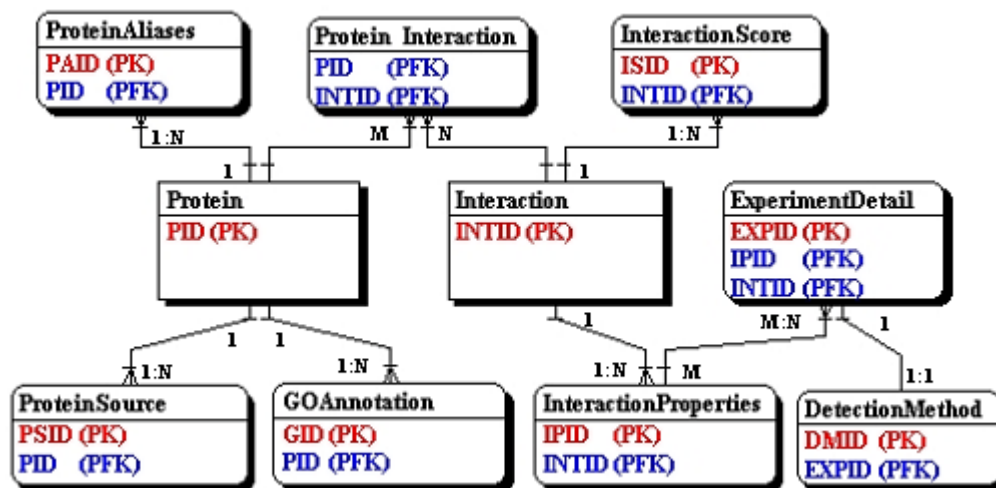


Figure B.2: An entity-relationship diagram of UniHI, showing key tables (rectangles) and relations (lines). PK (red) and PFK (blue) denote primary and foreign keys.

Appendix C

C.1 Details of the datasets used for the Precision analysis

For comparison, I computed the precision values of predicted HDTTs in following HD related datasets:

A: 2,326 differentially expressed genes; data was obtained by comparing gene expression profiles of human brain with non-brain tissues with a threshold of $p < 10^{-5}$ (Su *et al.*, 2004) .

B: 13,921 differentially expressed genes; Data was obtained by comparing gene expression profiles of the caudate nucleus with gene expression profiles of the motor cortex (MC), prefrontal cortex (PFC) and the cerebellum (CE). A threshold of $p < 10^{-3}$ was used for the analysis (Hodges *et al.*, 2006).

C: 5,674 differentially expressed genes; Data was obtained by comparing gene expression profiles of the caudate nucleus of HD patients and healthy controls using a threshold of $p < 10^{-3}$ (Hodges *et al.*, 2006).

D: 509 proteins of the HTT master network. Data was generated in small- and large-scale interaction studies and is available in the interaction database UniHI (www.unihi.org).

E: 222 proteins obtained after filtering 509 proteins (HTT master network) using information from expression dataset, defined by comparing gene expression profiles of the caudate nucleus of HD and healthy brains using a threshold of $p < 10^{-3}$ (Hodges *et al.*, 2006).

F: 15 proteins obtained after a 3-step filtration of HTT PPIs (HD master network) with gene expression data from (Hodges *et al.*, 2006).

Bibliography

- Adamcsek, B., Palla, G., Farkas, I. J., Derenyi, I., and Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* **22**, 1021-1023.
- Albert, R., Jeong, H., and Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature* **406**, 378-382.
- Ammann, L. P., and Goodman, S. R. (2009). Cluster analysis for the impact of sickle cell disease on the human erythrocyte protein interactome. *Exp Biol Med (Maywood)* **234**, 703-711.
- 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. The IntAct molecular interaction database in 2010. *Nucleic Acids Res* **38**, D525-531.
- 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**, 25-29.
- Bader, G. D., Betel, D., and Hogue, C. W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-250.
- Bader, G. D., and Hogue, C. W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* **20**, 991-997.
- Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2.
- Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Res* **28**, 304-305.
- Baitaluk, M., Sedova, M., Ray, A., and Gupta, A. (2006). BiologicalNetworks: visualization and analysis tool for systems biology. *Nucleic Acids Res* **34**, W466-471.
- Balasubramanian, R., LaFramboise, T., Scholtens, D., and Gentleman, R. (2004). A graph-theoretic approach to testing associations between disparate sources of functional genomics data. *Bioinformatics* **20**, 3353-3362.
- Barabasi, A. L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509-512.
- Barabasi, A. L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113.
- Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B. M., Kappos, L., Polman, C. H., Matthews, P. M., Hauser, S. L., Gibson, R. A., Oksenberg, J. R., and Barnes, M. R. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet* **18**, 2078-2090.
- Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., Marshall, K. A., Phillippy, K. H.,

- Sherman, P. M., Muertter, R. N., and Edgar, R. (2009). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **37**, D885-890.
- Batagelj, V., and Mrvar, A. (2003). Pajek - Analysis and Visualization of Large Networks. In *Graph Drawing Software* (M. Jünger, Mutzel, P., Ed.), pp. 77-103. Springer, Berlin.
- Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res* **32**, D138-141.
- Behrends, C., Langer, C. A., Boteva, R., Bottcher, U. M., Stemp, M. J., Schaffar, G., Rao, B. V., Giese, A., Kretzschmar, H., Siegers, K., and Hartl, F. U. (2006). Chaperonin TRiC promotes the assembly of polyQ expansion proteins into nontoxic oligomers. *Mol Cell* **23**, 887-897.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 289-300.
- Berg, J., Tymoczko, J., and Stryer, L. (2002). Protein Structure and Function *Biochemistry*. San Francisco: W. H. Freeman.
- Berglund, A. C., Sjolund, E., Ostlund, G., and Sonnhammer, E. L. (2008). InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res* **36**, D263-266.
- Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* **35**, D301-303.
- 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**, 292-299.
- Bounab, Y. (2010). CRMP1 Protein Complexes Modulate PolyQ-Mediated Htt Aggregation and Toxicity in Neurons. *PhD Thesis HU Berlin*.
- Braun, P., Rietman, E., and Vidal, M. (2008). Networking metabolites and diseases. *Proc Natl Acad Sci U S A* **105**, 9849-9850.
- Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P., and Sansone, S. A. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **31**, 68-71.
- Breitkreutz, B. J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bahler, J., Wood, V., Dolinski, K., and Tyers, M. (2008). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**, D637-640.
- Bretin, S., Reibel, S., Charrier, E., Maus-Moatti, M., Auvergnon, N., Thevenoux, A., Glowinski, J., Rogemond, V., Premont, J., Honnorat, J., and Gauchy, C. (2005). Differential expression of CRMP1, CRMP2A, CRMP2B, and CRMP5 in axons or dendrites of distinct neurons in the mouse brain. *J Comp Neurol* **486**, 1-17.
- Brohee, S., Faust, K., Lima-Mendez, G., Sand, O., Janky, R., Vanderstocken, G.,

- Deville, Y., and van Helden, J. (2008). NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res* **36**, W444-451.
- Brown, K. R., and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics* **21**, 2076-2082.
- Bruford, E. A., Lush, M. J., Wright, M. W., Sneddon, T. P., Povey, S., and Birney, E. (2008). The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res* **36**, D445-448.
- Burnett, B. G., Andrews, J., Ranganathan, S., Fischbeck, K. H., and Di Prospero, N. A. (2008). Expression of expanded polyglutamine targets profilin for degradation and alters actin dynamics. *Neurobiol Dis* **30**, 365-374.
- Caldecott, K. W. (2008). Single-strand break repair and genetic disease. *Nat Rev Genet* **9**, 619-631.
- Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., and Lowry, S. F. (2005). A network-based analysis of systemic inflammation in humans. *Nature* **437**, 1032-1037.
- Carey, V. J., Gentry, J., Whalen, E., and Gentleman, R. (2005). Network structures and algorithms in Bioconductor. *Bioinformatics* **21**, 135-136.
- Charrier, E., Reibel, S., Rogemond, V., Aguera, M., Thomasset, N., and Honnorat, J. (2003). Collapsin response mediator proteins (CRMPs): involvement in nervous system development and adult neurodegenerative disorders. *Mol Neurobiol* **28**, 51-64.
- 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**, D572-574.
- Chaurasia, G., Herzel, H., Wanker, E. E., and Futschik, M. E. (2006). Systematic functional assessment of human protein-protein interaction maps. *Genome Inform* **17**, 36-45.
- Chaurasia, G., Iqbal, Y., Hanig, C., Herzel, H., Wanker, E. E., and Futschik, M. E. (2007). UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res* **35**, D590-594.
- Chaurasia, G., Malhotra, S., Russ, J., Schnoegl, S., Hanig, C., Wanker, E. E., and Futschik, M. E. (2009). UniHI 4: new tools for query, analysis and visualization of the human protein-protein interactome. *Nucleic Acids Res* **37**, D657-660.
- Chiti, F., and Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* **75**, 333-366.
- Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3**, 140.
- Cole, A. R., Causeret, F., Yadirgi, G., Hastie, C. J., McLauchlan, H., McManus, E. J., Hernandez, F., Eickholt, B. J., Nikolic, M., and Sutherland, C. (2006). Distinct priming kinases contribute to differential regulation of collapsin response mediator proteins by glycogen synthase kinase-3 in vivo. *J Biol Chem* **281**, 16591-16598.

- Cole, A. R., Noble, W., van Aalten, L., Plattner, F., Meimaridou, R., Hogan, D., Taylor, M., LaFrancois, J., Gunn-Moore, F., Verkhatsky, A., Oddo, S., LaFerla, F., Giese, K. P., Dineley, K. T., Duff, K., Richardson, J. C., Yan, S. D., Hanger, D. P., Allan, S. M., and Sutherland, C. (2007). Collapsin response mediator protein-2 hyperphosphorylation is an early event in Alzheimer's disease progression. *J Neurochem* **103**, 1132-1144.
- Cowan, C. M., and Raymond, L. A. (2006). Selective neuronal degeneration in Huntington's disease. *Curr Top Dev Biol* **75**, 25-71.
- Davies, S. W., Turmaine, M., Cozens, B. A., DiFiglia, M., Sharp, A. H., Ross, C. A., Scherzinger, E., Wanker, E. E., Mangiarini, L., and Bates, G. P. (1997). Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell* **90**, 537-548.
- de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724-727.
- de Silva, E., Thorne, T., Ingram, P., Agrafioti, I., Swire, J., Wiuf, C., and Stumpf, M. P. (2006). The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol* **4**, 39.
- Deo, R. C., Schmidt, E. F., Elhabazi, A., Togashi, H., Burley, S. K., and Strittmatter, S. M. (2004). Structural bases for CRMP function in plexin-dependent semaphorin3A signaling. *EMBO J* **23**, 9-22.
- DiProspero, N. A., Chen, E. Y., Charles, V., Plomann, M., Kordower, J. H., and Tagle, D. A. (2004). Early changes in Huntington's disease patient brains involve alterations in cytoskeletal and synaptic elements. *J Neurocytol* **33**, 517-533.
- Ergun, A., Lawrence, C. A., Kohanski, M. A., Brennan, T. A., and Collins, J. J. (2007). A network biology approach to prostate cancer. *Mol Syst Biol* **3**, 82.
- Euler, L. (1736). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae* **8**, 128-140.
- 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, II, 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.
- Eyre, T. A., Ducluzeau, F., Sneddon, T. P., Povey, S., Bruford, E. A., and Lush, M. J. (2006). The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res* **34**, D319-321.
- Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257-258.
- Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246.
- Figeys, D., McBroom, L. D., and Moran, M. F. (2001). Mass spectrometry for the study of protein-protein interactions. *Methods* **24**, 230-239.
- 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**, 1011-1025.
- Futschik, M. E. (2003). Methods for Knowledge Discovery in Microarray Data. *PhD thesis, The University of Otago, Dunedin New Zealand*.
- Futschik, M. E., Chaurasia, G., and Herzel, H. (2007a). Comparison of human protein-protein interaction maps. *Bioinformatics* **23**, 605-611.
- Futschik, M. E., Tschaut, A., Chaurasia, G., and Herzel, H. (2007b). Graph-theoretical comparison reveals structural divergence of human protein interaction networks. *Genome Inform* **18**, 141-151.
- Gandhi, T. K., 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**, 285-293.
- Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147.
- Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80.
- Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., Jr., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727-1736.
- Glenner, G. G. (1980). Amyloid deposits and amyloidosis. The beta-fibrilloses (first of two parts). *N Engl J Med* **302**, 1283-1292.
- 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., Bussow, 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**, 853-865.
- Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A. L. (2007). The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-8690.
- Goni, J., Esteban, F. J., de Mendizabal, N. V., Sepulcre, J., Ardanza-Trevijano, S., Agirrezabal, I., and Villoslada, P. (2008). A computational analysis of protein-protein interaction networks in neurodegenerative diseases. *BMC Syst Biol* **2**, 52.
- Goodman, S. R., Kurdia, A., Ammann, L., Kakhniashvili, D., and Daescu, O. (2007). The human red blood cell proteome and interactome. *Exp Biol Med (Maywood)* **232**, 1391-1408.
- Graham, R. K., Deng, Y., Slow, E. J., Haigh, B., Bissada, N., Lu, G., Pearson, J., Shehadeh, J., Bertram, L., Murphy, Z., Warby, S. C., Doty, C. N., Roy, S., Wellington, C. L., Leavitt, B. R., Raymond, L. A., Nicholson, D. W., and Hayden, M. R. (2006). Cleavage at the caspase-6 site is required for neuronal dysfunction and degeneration due to mutant huntingtin. *Cell* **125**, 1179-1191.
- Guidetti, P., Charles, V., Chen, E. Y., Reddy, P. H., Kordower, J. H., Whetsell, W. O., Jr., Schwarcz, R., and Tagle, D. A. (2001). Early degenerative changes in transgenic mice expressing mutant huntingtin involve dendritic abnormalities but no impairment of mitochondrial energy production. *Exp Neurol* **169**, 340-350.
- Guimaraes, K. S., Jothi, R., Zotenko, E., and Przytycka, T. M. (2006). Predicting domain-domain interactions using a parsimony approach. *Genome Biol* **7**, R104.
- Gusella, J. F., and Macdonald, M. E. (2009). Huntington's disease: the case for genetic modifiers. *Genome Med* **1**, 80.
- Han, J. D., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* **23**, 839-844.
- Harjes, P., and Wanker, E. E. (2003). The hunt for huntingtin function: interaction partners tell many different stories. *Trends Biochem Sci* **28**, 425-433.
- Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258-261.
- Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol* **7**, 120.
- Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular

- to modular cell biology. *Nature* **402**, C47-52.
- HDCTG (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983.
- He, X., and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genet* **2**, e88.
- Hermel, E., Gafni, J., Propp, S. S., Leavitt, B. R., Wellington, C. L., Young, J. E., Hackam, A. S., Logvinova, A. V., Peel, A. L., Chen, S. F., Hook, V., Singaraja, R., Krajewski, S., Goldsmith, P. C., Ellerby, H. M., Hayden, M. R., Bredesen, D. E., and Ellerby, L. M. (2004). Specific caspase interactions and amplification are involved in selective neuronal vulnerability in Huntington's disease. *Cell Death Differ* **11**, 424-438.
- Hernandez, P., Huerta-Cepas, J., Montaner, D., Al-Shahrour, F., Valls, J., Gomez, L., Capella, G., Dopazo, J., and Pujana, M. A. (2007). Evidence for systems-level molecular mechanisms of tumorigenesis. *BMC Genomics* **8**, 185.
- Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180-183.
- Hodges, A., Strand, A. D., Aragaki, A. K., Kuhn, A., Sengstag, T., Hughes, G., Elliston, L. A., Hartog, C., Goldstein, D. R., Thu, D., Hollingsworth, Z. R., Collin, F., Synek, B., Holmans, P. A., Young, A. B., Wexler, N. S., Delorenzi, M., Kooperberg, C., Augood, S. J., Faull, R. L., Olson, J. M., Jones, L., and Luthi-Carter, R. (2006). Regional and cellular gene expression changes in human Huntington's disease brain. *Hum Mol Genet* **15**, 965-977.
- Hodgson, J. G., Agopyan, N., Gutekunst, C. A., Leavitt, B. R., LePiane, F., Singaraja, R., Smith, D. J., Bissada, N., McCutcheon, K., Nasir, J., Jamot, L., Li, X. J., Stevens, M. E., Rosemond, E., Roder, J. C., Phillips, A. G., Rubin, E. M., Hersch, S. M., and Hayden, M. R. (1999). A YAC mouse model for Huntington's disease with full-length mutant huntingtin, cytoplasmic toxicity, and selective striatal neurodegeneration. *Neuron* **23**, 181-192.
- Hoffmann, R., and Valencia, A. (2004). A gene network for navigating the literature. *Nat Genet* **36**, 664.
- Hors, A. L. (2004). DOM: Document Object Model: http://www.w3schools.com/dom/dom_parser.asp.
- Hu, Z., Hung, J. H., Wang, Y., Chang, Y. C., Huang, C. L., Huyck, M., and DeLisi, C. (2009). VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res* **37**, W115-121.
- Ideker, T., and Sharan, R. (2008). Protein networks in disease. *Genome Res* **18**, 644-652.
- Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002).

- Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31**, 370-377.
- Ispolatov, I., Yuryev, A., Mazo, I., and Maslov, S. (2005). Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res* **33**, 3629-3635.
- 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**, 4569-4574.
- Jana, N. R., Tanaka, M., Wang, G., and Nukina, N. (2000). Polyglutamine length-dependent interaction of Hsp40 and Hsp70 family chaperones with truncated N-terminal huntingtin: their role in suppression of aggregation and cellular toxicity. *Hum Mol Genet* **9**, 2009-2018.
- Jansen, R., and Gerstein, M. (2004). Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol* **7**, 535-545.
- Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453.
- Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411**, 41-42.
- Jonsson, P. F., and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291-2297.
- Kaltenbach, L. S., Romero, E., Becklin, R. R., Chettier, R., Bell, R., Phansalkar, A., Strand, A., Torcassi, C., Savage, J., Hurlburt, A., Cha, G. H., Ukani, L., Chepanoske, C. L., Zhen, Y., Sahasrabudhe, S., Olson, J., Kurschner, C., Ellerby, L. M., Peltier, J. M., Botas, J., and Hughes, R. E. (2007). Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genet* **3**, e82.
- Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Res* **37**, D623-628.
- Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30.
- Kann, M. G. (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform* **8**, 333-346.
- Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. (2004). EnsMart: a generic system for fast and flexible access to biological data. *Genome Res* **14**, 160-169.
- Kelley, R., and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* **23**, 561-566.
- Kerrien S, O. S., Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-Aryamontri A, Oesterheld M, Stümpflen V, Salwinski L, Nerothin J,

- Cerami E, Cusick ME, Vidal M, Gilson M, Armstrong J, Woollard P, Hogue C, Eisenberg D, Cesareni G, Apweiler R, Hermjakob H (2007). Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. . *BMC Biol* **5**.
- Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-772.
- Kiemer, L., Costa, S., Ueffing, M., and Cesareni, G. (2007). WI-PHI: a weighted yeast interactome enriched for direct physical interactions. *Proteomics* **7**, 932-943.
- Koegl, M., and Uetz, P. (2007). Improving yeast two-hybrid screening systems. *Brief Funct Genomic Proteomic* **6**, 302-312.
- Lage, K., EO., K., Størting, Z., Ólason, P., Pedersen, A., Rigina, O., Hinsby, A., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 309 - 316.
- Lee, I., Date, S. V., Adai, A. T., and Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science* **306**, 1555-1558.
- Lehner, B., and Fraser, A. G. (2004). A first-draft human protein-interaction map. *Genome Biol* **5**, R63.
- Li, D., Liu, W., Liu, Z., Wang, J., Liu, Q., Zhu, Y., and He, F. (2008). PRINCESS, a protein interaction confidence evaluation system with multiple data sources. *Mol Cell Proteomics* **7**, 1043-1052.
- Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540-543.
- Li, S. H., and Li, X. J. (2004). Huntingtin-protein interactions and the pathogenesis of Huntington's disease. *Trends Genet* **20**, 146-154.
- Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabo, G., Rual, J. F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., Barabasi, A. L., Vidal, M., and Zoghbi, H. Y. (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* **125**, 801-814.
- Lin, C. Y., Chin, C. H., Wu, H. H., Chen, S. H., Ho, C. W., and Ko, M. T. (2008). Hubba: hub objects analyzer--a framework of interactome hubs identification for network biology. *Nucleic Acids Res* **36**, W438-443.

- Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **35**, D26-31.
- Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science* **296**, 910-913.
- Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* **37**, D619-622.
- 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**, 2120-2126.
- McKusick, V. A. (1998). Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. . *Baltimore Johns Hopkins University Press*.
- Meggison, D. (2005). SAX: A simple API for XML Document <http://sax.sourceforge.net/>.
- Miller, V. M., Nelson, R. F., Gouvion, C. M., Williams, A., Rodriguez-Lebron, E., Harper, S. Q., Davidson, B. L., Rebagliati, M. R., and Paulson, H. L. (2005). CHIP suppresses polyglutamine aggregation and toxicity in vitro and in vivo. *J Neurosci* **25**, 9152-9161.
- Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. *Science* **303**, 1538-1542.
- Mrowka, R., Patzak, A., and Herzel, H. (2001). Is there a bias in proteome research? *Genome Res* **11**, 1971-1973.
- Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Pagni, M., Ponting, C. P., Quevillon, E., Selengut, J., Sigrist, C. J., Silventoinen, V., Studholme, D. J., Vaughan, R., and Wu, C. H. (2005). InterPro, progress and status in 2005. *Nucleic Acids Res* **33**, D201-205.
- Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-540.
- Navratil, V., de Chassey, B., Meyniel, L., Delmotte, S., Gautier, C., Andre, P., Lotteau, V., and Rabourdin-Combe, C. (2009). VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res* **37**, D661-668.
- O'Brien, K. P., Remm, M., and Sonnhammer, E. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33**, D476-480.
- Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stumpflen, 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., De Las Rivas, J., 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**, 894-898.
- Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *J Med Genet* **43**, 691-698.
- Owen, B. A., Yang, Z., Lai, M., Gajec, M., Badger, J. D., 2nd, Hayes, J. J., Edelmann, W., Kucherlapati, R., Wilson, T. M., and McMurray, C. T. (2005). (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. *Nat Struct Mol Biol* **12**, 663-670.
- Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H. W., Ruepp, A., and Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832-834.
- Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814-818.
- Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J., and Orengo, C. (2005). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res* **33**, D247-251.
- Pepper, S. D., Saunders, E. K., Edwards, L. E., Wilson, C. L., and Miller, C. J. (2007). The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics* **8**, 273.
- Pereira-Leal, J. B., Enright, A. J., and Ouzounis, C. A. (2004). Detection of functional modules from protein interaction networks. *Proteins* **54**, 49-57.
- Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A., and Cesareni, G. (2005). HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics* **6** Suppl 4, S21.
- Platzer, A., Perco, P., Lukas, A., and Mayer, B. (2007). Characterization of protein-interaction networks in tumors. *BMC Bioinformatics* **8**, 224.
- Prasad, T. S., Kandasamy, K., and Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol Biol* **577**, 67-79.
- Przulj, N., Wigle, D. A., and Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics* **20**, 340-348.
- Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Seraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218-229.

- Pujana, M. A., Han, J. D., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W. M., Rual, J. F., Levine, D., Rozek, L. S., Gelman, R. S., Gunsalus, K. C., Greenberg, R. A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Sole, X., Hernandez, P., Lazaro, C., Nathanson, K. L., Weber, B. L., Cusick, M. E., Hill, D. E., Offit, K., Livingston, D. M., Gruber, S. B., Parvin, J. D., and Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet* **39**, 1338-1349.
- Ramani, A. K., Bunesco, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol* **6**, R40.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-1555.
- Rives, A. W., and Galitski, T. (2003). Modular organization of cellular networks. *Proc Natl Acad Sci U S A* **100**, 1128-1133.
- Ross, C. A., and Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nat Med* **10 Suppl**, S10-17.
- 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., Llamasas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178.
- Ruder, K., and Winstead, E. A Quick Guide to Sequenced Genomes. http://www.genomenewsnetwork.org/resources/sequenced_genomes/.
- 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**, D449-451.
- Sanchez, I., Mahlke, C., and Yuan, J. (2003). Pivotal role of oligomerization in expanded polyglutamine neurodegenerative disorders. *Nature* **421**, 373-379.
- Schmidt, E. F., and Strittmatter, S. M. (2007). The CRMP family of proteins and their role in Sema3A signaling. *Adv Exp Med Biol* **600**, 1-11.
- Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol* **18**, 1257-1261.
- Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**, 166-176.
- Selkoe, D. J. (2003). Folding proteins in fatal ways. *Nature* **426**, 900-904.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**,

- Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol Syst Biol* **3**, 88.
- Spirin, V., and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* **100**, 12123-12128.
- Stauber, J., Lemaire, R., Franck, J., Bonnel, D., Croix, D., Day, R., Wisztorski, M., Fournier, I., and Salzert, M. (2008). MALDI imaging of formalin-fixed paraffin-embedded tissues: application to model animals of Parkinson disease for biomarker hunting. *J Proteome Res* **7**, 969-978.
- Stelzl, U., and Wanker, E. E. (2006). The value of high quality protein-protein interaction networks for systems biology. *Curr Opin Chem Biol* **10**, 551-558.
- 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., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957-968.
- Stumpf, M. P., 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**, 6959-6964.
- Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-6067.
- Tam, S., Geller, R., Spiess, C., and Frydman, J. (2006). The chaperonin TRiC controls polyglutamine aggregation and toxicity through subunit-specific interactions. *Nat Cell Biol* **8**, 1155-1162.
- Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A* **101**, 2981-2986.
- Tyner, S. D., Venkatachalam, S., Choi, J., Jones, S., Ghebranious, N., Igelmann, H., Lu, X., Soron, G., Cooper, B., Brayton, C., Hee Park, S., Thompson, T., Karsenty, G., Bradley, A., and Donehower, L. A. (2002). p53 mutant mice that display early ageing-associated phenotypes. *Nature* **415**, 45-53.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-627.
- van den Berg, B., Wain, R., Dobson, C. M., and Ellis, R. J. (2000). Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell. *EMBO J* **19**, 3870-3875.
- 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**, 399-403.

- Vonsattel, J. P., and DiFiglia, M. (1998). Huntington disease. *J Neuropathol Exp Neurol* **57**, 369-384.
- Vonsattel, J. P., Myers, R. H., Stevens, T. J., Ferrante, R. J., Bird, E. D., and Richardson, E. P., Jr. (1985). Neuropathological classification of Huntington's disease. *J Neuropathol Exp Neurol* **44**, 559-577.
- Wachi, S., Yoneda, K., and Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**, 4205-4208.
- Walker, F. O. (2007). Huntington's disease. *Lancet* **369**, 218-228.
- Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**, 440-442.
- Wuchty, S. (2004). Evolution and topology in the yeast protein interaction network. *Genome Res* **14**, 1310-1314.
- Yamashita, N., Morita, A., Uchida, Y., Nakamura, F., Usui, H., Ohshima, T., Taniguchi, M., Honnorat, J., Thomasset, N., Takei, K., Takahashi, T., Kolattukudy, P., and Goshima, Y. (2007). Regulation of spine development by semaphorin3A through cyclin-dependent kinase 5 phosphorylation of collapsin response mediator protein 1. *J Neurosci* **27**, 12546-12554.
- Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L., and Vidal, M. (2007). Drug-target network. *Nat Biotechnol* **25**, 1119-1126.
- 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., Svrikapa, 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**, 104-110.
- Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3**, e59.
- Yue, Q. X., Cao, Z. W., Guan, S. H., Liu, X. H., Tao, L., Wu, W. Y., Li, Y. X., Yang, P. Y., Liu, X., and Guo, D. A. (2008). Proteomics characterization of the cytotoxicity mechanism of ganoderic acid D and computer-automated estimation of the possible drug target network. *Mol Cell Proteomics* **7**, 949-961.
- Zhou, X., Kao, M. C., and Wong, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U S A* **99**, 12783-12788.

List of Publications

Peer-review articles in international journals

Martin Stroedicke, Yacine Bounab, **Gautam Chaurasia**, Shuang Li, Stephanie Plaßmann, Jenny Russ, Cecilia Nicoletti, Jan Bieschke, Sigrid Schnoegl, Rona Graham, Josef Priller, Michael Hayden, Stephan Sigrist, Maciej Lalowski, Matthias Futschik and Erich E. Wanker (2010), Brain-specific interaction partners control polyglutamine-mediated huntingtin misfolding and neurotoxicity, (*in review*)

Elisabetta Marras, Antonella Travaglione, **Gautam Chaurasia**, Matthias Futschik, Enrico Capobianco, (2010), Inferring Modularity from Human Protein Interactome Classes, *BMC Systems Biology* 4: 102

Chaurasia, Gautam; Malhotra, Soniya; Russ, Jenny; Schnögl, Sigrid; Hänig, Christian; Wanker, Erich; and Futschik, Matthias (2009), UniHI 4: New tools for query, analysis and visualization of the human protein-protein interactome, *Nucleic Acids Res*, 37(Database issue): D657–D660.

Gautam Chaurasia, Yasir Iqbal, Christian Hänig, Hanspeter Herzel Erich E. Wanker and Matthias E. Futschik, (2007), UniHI: an entry gate to human protein interactome, *Nucleic Acid Research*, 35(Database issue): D590–D594.

Matthias E. Futschik, **Gautam Chaurasia** and Hanspeter Herzel, (2006), Comparison of Human Protein-Protein Interaction Maps, *Bioinformatics*, Mar 1;23(5):605-11, 2007.

Conference Papers and Book Chapters

Gautam Chaurasia and Matthias E. Futschik (2010) The integration and annotation of the human interactome in the UniHI database, *Two Hybrid Technologies: Methods and Protocols*, eds. B. Suter, **Springer Protocols** (in press)

Gautam Chaurasia and Matthias Futschik (2010), Interactomics and Cancer; An Omics Perspective on Cancer Research, Springer Verlag, chapter 10.

Matthias E. Futschik, **Gautam Chaurasia**, Jenny Russ and Hanspeter Herzel, (2007), Functional and Transcriptional Coherency of Modules in the Human Protein

Interaction Network, *Journal of Integrative Bioinformatics*, 4(3):76.

Matthias E. Futschik, Anna Tschaut, **Gautam Chaurasia** and Hanspeter Herzel, (2007), Graph-theoretical comparison reveals structural divergence of human protein interaction networks, *Genome Inform.* 2007;18:141-51.

Gautam Chaurasia, Yasir Iqbal, Christian Hänig, Hanspeter Herzel Erich E. Wanker and Matthias E. Futschik, (2006), Flexible web-based integration of distributed large-scale human protein interaction maps, *Journal of Integrative Bioinformatics*, 4(1):51, 2007.

Gautam Chaurasia, Hanspeter Herzel, Erich E. Wanker and Matthias E. Futschik, (2006), Systematic Functional Assessment of Human Protein-Protein Interaction Maps, *Genome Informatics*, 17(1), 36-45.

Matthias E. Futschik, **Gautam Chaurasia** and Hanspeter Herzel, (2006), Comaparative Comparison of Human Protein-Protein Interaction Maps, Lecture Series on Informatics Notes.

Presentations

Gautam Chaurasia, *et. al.*, Comparison and Integartion of Human Protein Interaction Maps, *Contributed Talk*, YSF 2007, Vienna, Austria.

Gautam Chaurasia, *et. al.*, Flexible web-based integration of distributed large-scale human protein interaction maps, *Contributed Talk*, Data Warehousing Technologies in Bioinformatics (2006), Wittenberg, Halle, Germany.

Gautam Chaurasia, *et. al.*, Systematic Functional Assessment of Human Protein-Protein Interaction Maps, International workshop on Bioinformatics and Systems Biology (2006), *Contributed Talk*, Boston, USA.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Arbeit selbständig ohne fremde Hilfe verfaßt zu haben und nur die angegebene Literatur und Hilfsmittel verwendet zu haben.

Gautam Chaurasia

Berlin, den 22. Februar 2011