More complete and more accurate interactomes for elucidating the mechanisms of complex diseases

Atanas Kamburov

August 2011

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

1. Gutachter: Prof. Dr. Martin Vingron
2. Gutachter: Prof. Dr. Ron Shamir

Tag der Promotion: 17. Februar 2012

Acknowledgments

First of all, I would like to express my deepest gratitude to my advisors Ralf Herwig and Ulrich Stelzl for the crucial input in form of expertise, scientific advice, guidance, and for creating a wonderful scientific atmosphere I had the greatest pleasure of working in during my time as a PhD candidate at the Max Planck Institute for Molecular Genetics in Berlin. Our numerous discussions were very fruitful as they have led to many of the ideas presented in this thesis.

My sincere thanks go to my supervisor Martin Vingron and to Ron Shamir for reviewing my thesis. Martin Vingron and Alexander Bockmayr were my mentors and guided my research during the past four years, for which I am very grateful. I also thank Hans Lehrach for his useful comments.

I deeply acknowledge Trey Ideker for giving me the great opportunity to work in his group at the University of California, San Diego. I appreciate the time he spent in discussions with me despite his busy schedule.

I am deeply grateful to Hannes Luz and Kirsten Kelleher for their very friendly support in many organizational issues. Sadly, Hannes left us too early and is greatly missed.

Special thanks go to Konstantin Pentchev and Hanna Galicka for their support, as well as to Arndt Grossmann, Christopher Hardt, Felix Dreher, Lukas Chavez, Marcus Albrecht for their critical remarks on different parts of the thesis manuscript. I am grateful to the whole Bioinformatics group, as well as to all co-operation partners (notably Rachel Cavill and Hektor Keun) for the pleasure of doing research together.

There are a lot more people whom I am very thankful to, including my girlfriend and all my friends, and most of all my parents and grandparents, whom I owe and dedicate all my achievements to.

My time as a PhD candidate was financially supported through a scholarship from the Max Planck Society under its International Max Planck Research School for Computational Biology and Scientific Computing program, as well as by the European Union under its APO-SYS program.

Publications related with this thesis

Parts of this thesis have been published in several peer-reviewed journals. ConsensusPathDB (presented in Chapter 2) was initially published in 2009 in *Nucleic Acids Research*. A paper reporting its significant development in terms of functionality (Chapter 4) and content was published in 2011, again in *Nucleic Acids Research*. The Cytoscape plugin for ConsensusPathDB (outlined in Chapter 2) was published in *Bioinformatics*. The plugin was conceived during the analysis of data from a new interaction screen which was published in *Cell*. A revised paper manuscript describing the CAP-PIC method (Chapter 3) is currently under review by *Bioinformatics*. A revised application note manuscript about the IMPaLA tool (Chapter 4) is currently under review by *Bioinformatics*. The IMPaLA tool implements our approach initially published in *PLoS Computational Biology*. Further manuscripts that are not included in this thesis are listed in the curriculum vitae.

Contents

1 Introduction **1**

1.1 Molecular interactions 1

1.2 Interaction data 2

1.2.1 Detection and prediction of interactions 2

1.2.2 Storing and representing interaction data 4

1.2.3 Noise in interaction data 5

1.2.4 Graphical modeling of interaction data 6

1.2.5 Structure of interaction networks 8

1.3 Interactions and pathways in aid of expression data analysis 10

1.3.1 Whole-genome expression profiling 10

1.3.2 Integration of expression data with interaction and pathway knowledge 11

1.4 Aims and organization of the thesis 12

2 Toward more complete interactome maps **15**

2.1 Introduction to interaction databases 16

2.2 Data model of ConsensusPathDB and data integration 20

2.2.1 Database structure 20

2.2.2 Integration of interaction data from multiple sources 22

2.3 A global view on the integrated content of ConsensusPathDB 24

2.3.1 Complementarity of interaction data resources 24

2.3.2 Topological properties of the human protein interaction network 29

2.4 Interfaces of ConsensusPathDB 30

2.4.1 The ConsensusPathDB web interface 32

2.4.2 The ConsensusPathDB plugin for Cytoscape 35

2.5 Discussion 36

3 Cluster-based assessment of protein-protein interaction confidence 39

3.1 Introduction to protein-protein interaction confidence assessment 39

3.2 CAPPIC: A novel approach for interaction confidence assessment 44

3.2.1 Assessing protein interaction confidence by random walk interaction clustering 44

3.2.2 Optimal clustering granularity is reliably determined through partial network rewiring 46

3.3 Comparative assessment of the performance of CAPPIC on yeast networks 49

3.3.1 True positive interactions are assigned higher confidence than false positives 49

3.3.2 Cluster-based confidence scores corroborate experimental interaction evidence 53

3.3.3 High-confidence interactions are more consistent in biological process and cellular compartment annotation 54

3.3.4 Construction of a high-quality yeast physical interactome 57

3.3.5 CAPPIC as a web tool for interaction confidence assessment . . 59

3.4 Discussion 59

4 Elucidating disease mechanisms with integrated interaction networks and expression data 63

4.1 Introduction: the benefits from integrating interaction and expression data 64

4.2 Network-based functional gene sets in aid of causative gene identification 66

4.2.1 Functional gene sets based on integrated network neighborhood (NESTs) 68

4.2.2 Statistical approaches for identifying dysregulated NESTs 71

4.2.3 Application 1: Network-based meta-analysis of prostate cancer pinpoints known causative genes 74

4.2.4 Application 2: NEST enrichment analysis with numerical data unveils cancer-related genes and highlights the hallmarks of cancer 79

4.3 Extending the pathway analysis paradigm: joint pathway analysis with transcriptomics and metabolomics data 82

4.4 Discussion 85

5 Conclusion **89**

Bibliography **93**

Appendix **117**

Abbreviations **123**

Software availability **125**

Curriculum vitae **126**

Publications **128**

Zusammenfassung **131**

Ehrenwörtliche Erklärung **133**

List of Figures

1.1 Insulin signaling pathway 3
1.2 Modeling interaction data as graphs or hypergraphs 7

2.1 Entity-relationship diagram visualizing the structure of the ConsensusPathDB interaction meta-database 21
2.2 Histogram of the number of database sources per interaction in ConsensusPathDB 26
2.3 Overlap and complementarity of interactions of p53 in four major protein interaction databases 27
2.4 Overlap of pathway composition across databases 29
2.5 Distributions of protein degree and clustering coefficient in the integrated human physical interactome map 32
2.6 Overview of the funtionality of the ConsensusPathDB web interface . . 33
2.7 Overview of the funtionality of the ConsensusPathDB plugin for Cytoscape 36

3.1 Outline of our interaction confidence assessment method 45
3.2 Estimating optimal granularity for clustering through partial random rewiring of input networks 48
3.3 Interaction co-clustering matrices 50
3.4 ROC analysis measuring the performance of CAPPIC in comparison to the methods by Goldberg and Roth and Kuchaiev *et al.* 52
3.5 Histogram of confidence scores for interactions in Tarassov-all calculated by our method 54
3.6 Correlation of CAPPIC interaction confidence with semantic similarity of Gene Ontology co-annotations 55

LIST OF FIGURES

3.7 Interaction cluster refinement 57
3.8 Confidence scores and literature evidence for the CPDB-yeast network . 58
3.9 CAPPIC as a web tool at http://cpdb.molgen.mpg.de/cappic 59

4.1 Pathway annotation of human genes and its relation with protein interactions 67
4.2 Construction of neighborhood-based entity sets (NESTs) 69
4.3 Characteristics of neighborhood-based entity sets 70
4.4 Over-representation and enrichment analyses 72
4.5 Agreement of different studies focused on the same phenotypes in respect of differentially expressed genes 76
4.6 Overlap between differentially expressed (DE) genes, nest center (NC) genes, and the Cancer Gene Census 78
4.7 Neighborhood-based entity set (NEST) centered around SUV39H2 with gene/protein nodes colored according to expression fold change. 81
4.8 Pathway-level integration of transcript and metabolite data: a schematic overview of the study design 84
4.9 Pathways associated with platinum resistance based on transcriptomic, metabolomic, and combined evidence for phenotype association 85
4.10 IMPaLA: a web tool for integrated pathway-level analysis of transcriptomics and metabolomics data 86

A.1 Growth of ConsensusPathDB's unique interaction content since its initial publication 118
A.2 NESTs where cancer metastasis-associated genes are significantly over-represented 119

List of Tables

2.1 Interaction databases integrated in ConsensusPathDB 17
2.2 Interaction database overlaps 25
2.3 Topological properties of physical interactome maps 31

3.1 Reference interaction networks 43

4.1 Studies comparing whole-genome expression profiles of metastatic prostate cancer against primary prostate carcinoma 75
4.2 DE (differentially expressed) genes 77
4.3 NC (nest center) genes 77

A.1 NESTs significantly associated with metastatic prostate cancer, based on data by Yu et al. 120
A.2 Pathways significantly associated with metastatic prostate cancer, based on data by Yu et al. 121

LIST OF TABLES

Chapter 1

Introduction

As a crucial step in the quest of understanding the functioning of the cell on the molecular system level, the genomes of many species, including human, have already been largely decoded. The knowledge of the list of human genes that has been obtained by these efforts is essential but insufficient for elucidating cellular processes in health and disease. It is clear that the separate genes execute their functions through interactions between each other as well as with different other biomolecules. Knowledge of the complex functional interplay between all biomolecules in the cell promises to take us a step further toward understanding the molecular mechanisms governing life. This chapter gives a coarse summary of several types of biomolecular interactions and the most prevalent ways they are detected, stored, modeled, and utilized for the interpretation of gene expression data. A particular focus is put on current problems in the field that motivated this thesis.

1.1 Molecular interactions

The interplay between two or more biomolecules that has a specific biological effect is called an interaction. Interactions, rather than the separate physical entities (genes, proteins, metabolites, etc.), are the key drivers of biological processes. Deviations from the normal interaction patterns in the cell can lead to disease, thus it is not surprising that they constrain genome evolution (56). Interactions are commonly divided into several classes depending on the type of the interacting molecules, the mechanism, specificity, duration, and our understanding of their biological effects. Examples for such

interaction classes are gene regulatory interactions, metabolic reactions, signaling reactions, and protein-protein interactions: 1) Gene regulatory interactions are executed by the products of certain genes called transcription factors that bind specifically to the DNA at certain regions within or near other genes to enhance or repress their expression. 2) Metabolic reactions are biochemical reactions that convert metabolites from one type to another under the catalysis of specific enzymes (mainly proteins or protein complexes). 3) Signaling reactions are another type of biochemical reactions, typically involving proteins being modified (e.g. phosphorylated or cleaved) by other physical entities in order to initiate or transmit a biological signal. 4) Protein-protein interactions are a general class of physical interactions between proteins. They may have different stability (depending on the biochemical properties of the interactors), e.g. the formation of protein complexes usually results from protein-protein interactions that are stable over time, while modification reactions, for example, are administered by more transient interactions between the modifier and the protein being modified.

A biological pathway can be seen as a compilation of interactions sharing participants and constrained in space and time, which together concert a biologically relevant transformation of mass or a conduction of a biological signal. It is a key point that biological processes are usually composed of many different types of interactions. As an example, the diagram in Figure 1.1 depicts the Insulin signaling pathway that simultaneously involves protein-protein interactions (e.g. between insulin and its receptor on the cell surface), biochemical reactions (e.g. hydrolysis of GTP by RAS, modulated by GAP), and gene regulatory interactions (e.g. regulation of target genes by the C-JUN:C-FOS complex).

1.2 Interaction data

1.2.1 Detection and prediction of interactions

Obtaining knowledge of all interactions in the cell promises mechanistic insight into cellular biology in health and disease as it reveals the molecular circuitry behind biological processes. This motivates contemporary biologists around the globe to apply immense efforts in designing and applying various techniques to discover the different interactions in the cell of human and of other species. Direct gene regulatory interactions, for instance, are commonly predicted by ChIP-chip (21) or ChIP-seq (120) –

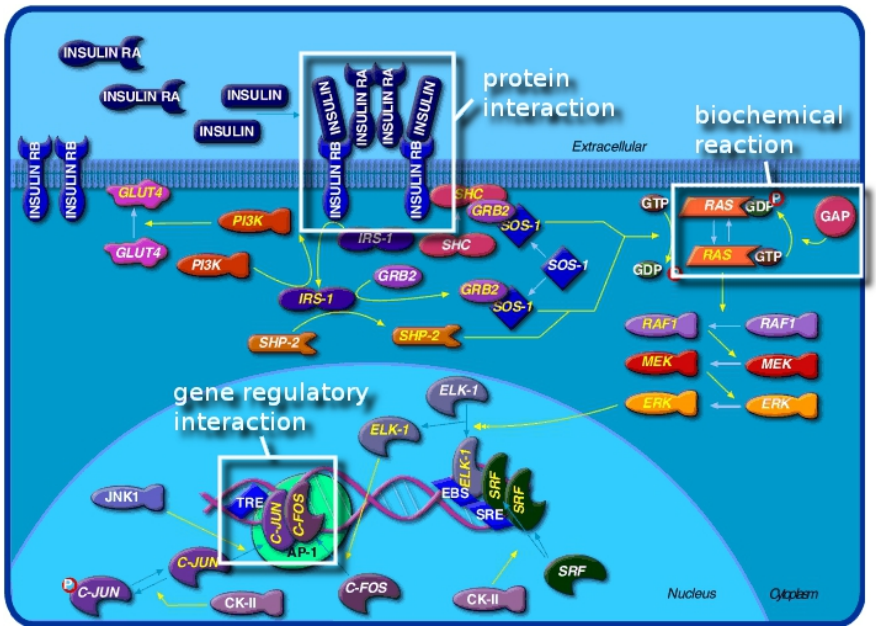

Figure 1.1: Insulin signaling pathway. Like all other biological processes, the insulin signaling pathway comprises different types of interactions including protein-protein interactions, biochemical reactions and gene regulation. Reproduced from BioCarta (http://www.biocarta.com) and modified.

experimental techniques which involve chromatin immunoprecipitation (ChIP) coupled to DNA binding site identification through hybridization or sequencing, respectively. On the other hand, metabolic and signaling reactions are typically detected through narrow-focused biochemical experiments such as enzyme assays (64), or are deduced from specific crystallographical measurements (22). Finally, protein-protein interactions are detected mostly with yeast-two-hybrid (Y2H) screening (52), affinity purification coupled to mass spectrometry (AP-MS) (1), or protein-fragment complementation assay (PCA) (59). The advantages and disadvantages of these and of other protein interaction detection techniques are discussed in (137). In addition to experimental techniques for the direct detection of interactions, different computational methods have been developed for the prediction of such. For example, many methods exist that can predict potential protein-protein interactions based on genomic sequence or homology data (88, 145, 147).

Interaction data resulting from the application of experimental or computational methods may have binary or complex nature. Binary interactions have exactly two

participants, while complex interactions may involve an arbitrary number of physical entities. For instance, protein-protein interactions detected by Y2H and PCA strictly involve two participants (which may be identical in the case of self-interactions) because these methods test pairs of proteins for direct associations. On the other hand, interactions detected with AP-MS are generally complex while the direct physical interactions between the separate proteins are not revealed by this technique. Biochemical reaction data have a complex nature as well since such interactions may involve multiple substrates that are converted to multiple products.

1.2.2 Storing and representing interaction data

Interactions that have been detected or predicted are usually assembled in specialized interaction knowledge bases through literature mining or direct data submissions (41). Most of these knowledge bases offer public access to their content through querying and visualization of interactions. Currently, interaction data are scattered among more than three hundred such databases (11). Due to their specific focus, each of the databases contains a limited number of interaction types (mostly one to two types). Moreover, even databases with similar focus on interaction type have limited overlap with each other. This is mainly because the creators of each database tend to capture interactions from a unique subset of sources (e.g. literature publications) according to their own curation rules. Thus, in a sense, our knowledge of a specific biological process is dispersed among many interaction resources, which constrains a system-level view on that process (35). The same is true from the perspective of a specific gene: The detected protein-protein interactions of its products are scattered across protein interaction databases; data on its enzymatic functions resides in metabolic databases; and its gene regulatory interactions are assembled in databases on gene regulation. The complete picture of the gene's different roles in the cell can be obtained only after integrating all of these interaction resources. Such a comprehensive picture is crucial for example in drug development to predict the possible impact from drug target binding on the human body (35). Unfortunately, the task of interaction data integration is hindered particularly by the vast heterogeneity of current databases in respect to data models and data exchange formats: each database has its own way of representing, storing, and providing access to the interaction data. The problem has been partially solved by defining standard file formats for representing molecular interactions. The

most widely used formats include BioPAX, PSI-MI and SBML (154). They differ in the representable types of interactions and the level of detail they can provide for the individual interactions. For example, PSI-MI is specialized to represent physical interactions, while SBML is designed to describe biochemical reactions and their kinetics. Probably the most descriptive of these three is BioPAX, which is able to represent a wide range of interaction types between a variety of physical entity types. Nevertheless, despite the efforts spent in the development of standard interaction representation formats, the data models of many interaction databases are often incompatible with these formats. Therefore, many databases have either adopted none of the standard formats, or the standardized data are often incomplete or inaccurate with respect to the original database content. This is why database-specific data formats are still primarily used by interaction resources as the means to distribute their content. The bottom line is that to integrate data from the existing highly complementary databases in order to obtain a more complete picture of cellular processes, one still has to deal mainly with database-specific data formats that are incompatible with each other, or with several standard formats representing the database content in a possibly imprecise manner.

1.2.3 Noise in interaction data

Interaction data are not only incomplete from the perspective of each individual database, but they may also be noisy. Above all, existing large-scale protein-protein interaction data have been shown to contain a considerable portion of false positives (72, 114), i.e. reported interactions that do not take place in reality. All techniques for detecting protein interactions generate false positives, for example due to experimental errors or bias (technical false positives). Adding to this, some of the interactions measured *in vitro* do not actually take place *in vivo*, for example because the proteins are separated in different cellular compartments (biological false positives) (105). Interactions collected from the literature are additionally prone to curation errors that may also reach striking magnitudes (41).

The integration of interaction datasets by considering their union increases the coverage of the real interactome (thus decreasing the false negative rate, i.e. the proportion of missing true interactions), albeit this is often achieved at the expense of a higher false positive rate (that is, the proportion of spurious interactions) in the integrated data compared to the separate datasets. The reason is that true interactions, being a

very small subset of all possible tuples of physical entities, have a much higher probability to be found simultaneously in two independent interaction datasets than false interactions. Thus, the number of true interactions saturates much faster in the process of data integration compared to the number of false positives. As a consequence, false positive interactions are accumulated at high rates in integrated datasets.

1.2.4 Graphical modeling of interaction data

Interaction data are usually modeled as network graphs or hypergraphs (2, 100). Such modeling benefits from the existing palette of graph-theoretical methods aiding the analysis of interaction data.

An interaction graph is a pair $G = (V, E)$ where V is a set of nodes, conventionally representing physical entities like genes, proteins, complexes, metabolites, etc., and E is a set of edges (or pairs of nodes), each edge usually representing an interaction between two nodes. A graph is connected if any pair of its nodes are linked with each other through a finite path of edges in the graph. Otherwise, the graph consists of multiple connected components. A graph is directed if its edges have a specified orientation, i.e. one of the nodes is designated the edge source and the other node is the edge sink. If edges have no orientation, the graph is called undirected. Binary protein-protein interaction data (i.e. interactions involving pairs of proteins) are usually modeled as undirected network graphs because of the symmetrical nature of protein-protein interactions (Figure 1.2). On the other hand, gene regulatory interaction graphs are directed because for every interaction, one of the genes is the regulator and the other is the regulated gene, but generally not the other way around (Figure 1.2). Special classes of graphs are multigraphs and bipartite graphs. In multigraphs, more than one edge can connect the same pair of nodes, for example to indicate different types of relations between these nodes. A bipartite graph $G = (V_1, V_2, E)$ has two disjoint sets of nodes denoted V_1 and V_2, with edges of E connecting nodes from V_1 with nodes from V_2 while no edges connect nodes within V_1 or within V_2. Unlike simple graphs with uniform nodes, which are able to represent only binary relations, bipartite graphs can be used to model complex interactions (i.e. interactions with an arbitrary number of participants). For instance, biochemical reactions are often modeled as directed bipartite graphs where V_1 is the set of physical entities, V_2 is the set of reactions, and directed edges connect reactions with their participants that can be any number (Figure

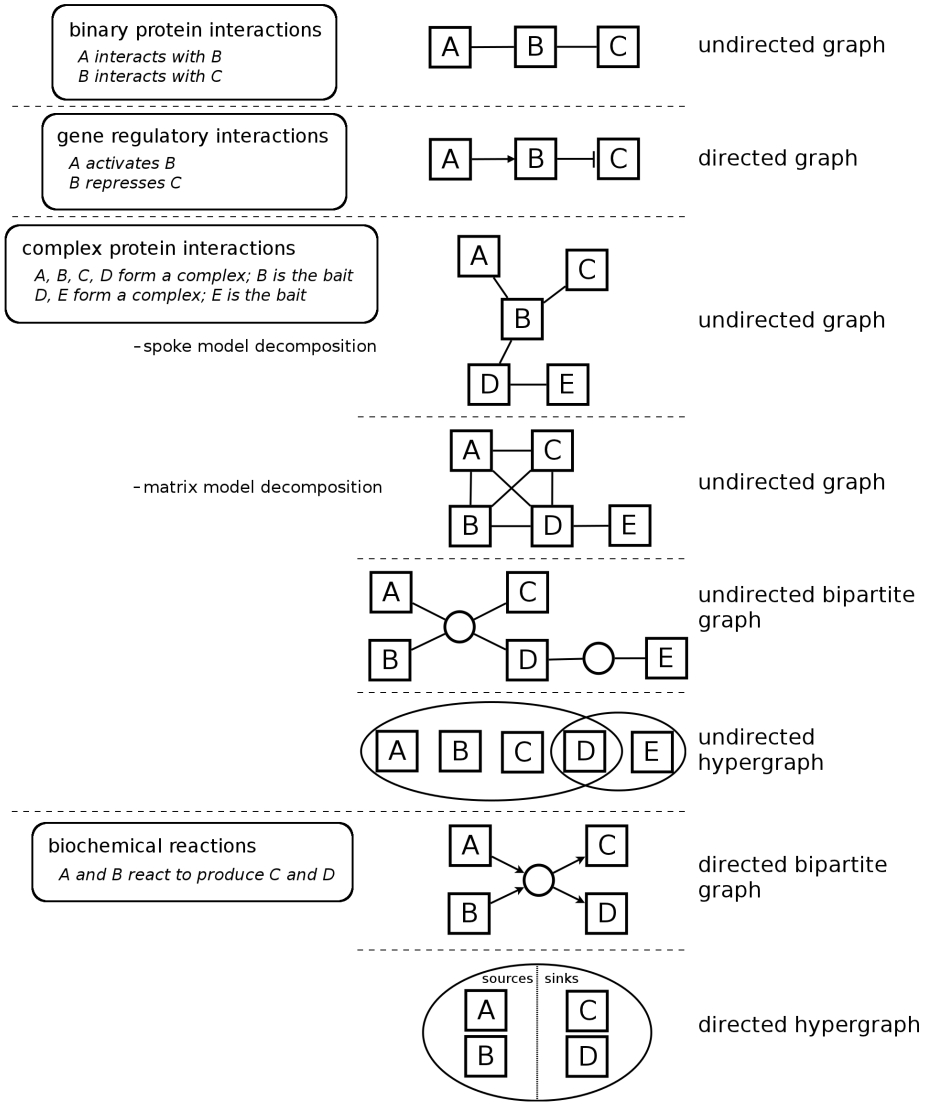

Figure 1.2: Modeling interaction data as graphs or hypergraphs. Binary interactions like physical interactions between pairs of proteins and gene regulatory interactions are modeled as undirected and directed graphs, respectively. Complex protein interactions are often decomposed into binary interactions following the spoke or matrix models, or are modeled as undirected bipartite graphs or undirected hypergraphs. Biochemical reactions are modeled as directed bipartite graphs or directed hypergraphs.

1.2). Here, edge orientation conventionally indicates whether an entity is a reaction substrate or product.

An alternative to bipartite graphs for modeling complex interactions are hyper-

graphs (100). A hypergraph $H = (V, E)$ consists of a set of nodes V and a set of hyperedges E. Unlike edges in graphs, which connect exactly two nodes (or one node with itself in case of self-interactions), hyperedges may connect an arbitrary number of nodes. Hypergraphs may be directed, in which case a subset of the nodes in each hyperedge are designated source nodes and the rest are sink nodes. Directed hypergraphs are sometimes used for representing metabolic reaction systems (Figure 1.2). Undirected hypergraphs rather than network graphs are sometimes utilized to model protein interactions detected with AP-MS because the direct pairwise interactions between the components of the complexes detected by this technique are generally unknown (Figure 1.2). However, because hypergraph operations are often more challenging computationally, and because graphs are somewhat more intuitive for manual interpretation, hypergraphs are not as widely used as network graphs for representing biological relations. Accordingly, complex interaction datasets are often transformed into binary data. For example, a protein complex detected with AP-MS can be represented as a set of binary interactions between all components (matrix model decomposition). Alternatively, since AP-MS involves isolation of complexeses formed around a specific protein called bait, these complexes can be represented as a set of binary interactions between the bait and the rest of the complex members called preys (spoke model decomposition) (Figure 1.2). Nevertheless, both strategies for complex decomposition are inevitably associated with loss of information about the detected complexes; moreover, they reportedly generate false positive and/or false negative interactions (9).

1.2.5 Structure of interaction networks

Not only the single elements of interaction networks – the separate physical entities and their interactions – have been studied extensively during the last decade, but so has been the overall structure of interaction circuits. Seminal studies (many of which are reviewed in (2, 14)) have shown with graph-theoretical measurements that such networks, like many other types of real-world networks, are highly structured.

For example, certain local patterns of interconnections between nodes, called network motifs, are found significantly more often in real-world networks than expected by chance (5, 115). In the biological context, such motifs are suggested to reflect universal biological functions: For instance, feedback and feed-forward loops abundant in

gene regulatory networks are proposed to act as amplifiers or filters of biological signals (112, 132).

Many biological networks show characteristic organization not only on the local, but also on the global level. Among the most extensively studied properties is the connectivity of network nodes. The connectivity, or degree, of a node is defined as the number of counterparts it is connected with in the network. For several types of biological networks, including metabolic (86) and protein-protein (87) interaction circuits, node degree has been shown to follow a power-law distribution: $P(X = k) = k^{-\gamma}$, where k denotes the degree and γ is a constant. This means that the vast majority of nodes in such networks have a small number of interaction partners, while there are a small number of so-called 'hubs' that have many interaction partners. Due to the absence of a characteristic node degree, or scale, in such networks, they are called scale-free (13). The emergence of the scale-free property in interaction webs is suggestedly associated with the cell's tolerance to random errors such as gene mutations (3). Another common property of many biological circuits that arises from their scale-free nature is the small-world effect (13, 51, 175). It is essentially associated with very short average numbers of interactions separating pairs of nodes, and could be a factor aiding fast reactions of the cell to stimuli (14, 51). The small-world property of a network is often quantified with the average shortest path length, defined as the average number of edges one has to follow in order to reach one node starting from another. A related measure is the network diameter, which is defined as the maximum of all shortest paths between pairs of nodes. A further phenomenon seen in interaction networks is that the set of direct network neighbors of a node share more interactions between each other than expected by chance (that is, network neighborhoods of nodes are more densely connected than expected by chance). It is traditionally quantified with the clustering coefficient, defined for each node as the fraction of existing links among its network neighbors from the number of all possible links among them (175). In other words, for a node with n network neighbors (node degree = n) among which k edges exist, the clustering coefficient is $C = k/\binom{n}{2} = 2k/n(n-1)$. A high average clustering coefficient is an indicator of the network's modular structure, since interactions form densely connected communities, or modules. Furthermore, it has been shown that modularity in biological networks is often organized in a hierarchical manner, leading to the concept

of 'network of networks' – that is, nodes connect to form modules, modules connect to form higher-level network structures, and so on (127, 128).

The bottom line from studies analysing global as well as local properties of interaction networks is that the topology of these networks is far from random. Rather, the high degree of internal order that governs the cell's molecular organization is consistently reflected in the networks' architecture (14).

1.3 Interactions and pathways in aid of expression data analysis

Interaction knowledge can be exploited in many ways and contexts. To give some examples: First, manual inspection of the circuitry of certain genes can provide clues about why a biological process develops as it does and how gene disruptions (such as mutations, or applied drugs) may lead to a certain phenotype. Second, the functions of a protein can be predicted based on the interaction network neighborhood of that protein (146). Third, interaction networks are routinely used as the basis of mathematical models aiming to simulate and predict the systems-level behavior of biological systems (99). Fourth, large-scale interaction data and manually curated pathway models are increasingly applied as a basis for interpreting whole-genome expression data.

1.3.1 Whole-genome expression profiling

Gene expression profiling involves techniques that measure the expression levels of many genes simultaneously. It is often used to assess the gene expression response of a biological system to external or internal stimuli like environmental factors or disease. Such stimuli often provoke changes in the expression of many genes, reflected by alterations in the according messenger-RNA concentrations, and ultimately by changes in the concentration of the protein products of those genes in the biological system. While protein concentrations are relatively difficult to measure on a large scale, the abundance of tens of thousands of messenger-RNA molecule types can be easily determined simultaneously. This is usually done with hybridization- or sequencing-based techniques, such as microarrays or RNA-seq (139, 174). An expression profile of a cell or a tissue obtained by such techniques is a unique snapshot of the expression activity of thousands of genes. Using statistical tests, expression profiles of an experimental

condition of interest are typically compared to expression profiles of a control phenotype to highlight a list of genes that show significant change of expression between the two phenotypes (24). For example, if the phenotype of interest is a disease, the genes that are differentially expressed compared to the control are typically considered to be related to the disease and could be effective or even causative of it. The list of genes differentially expressed in a disease is commonly termed the gene signature of the disease. Gene expression profiling has an enormous potential in molecular medicine as it can aid diagnosis by pointing to new, or assessing the expression of known disease biomarkers, and can help generate hypotheses about potential drug targets and therapies.

1.3.2 Integration of expression data with interaction and pathway knowledge

A major concern is that gene signatures found by different studies analyzing the same phenotypic condition are often barely overlapping (47). The lack of agreement may arise from differences in the experimental techniques and settings used in the according studies. Mainly, however, it is attributed to the inherent variability of biological systems including variations in the genetic background, environmental effects, tissue heterogeneity, etc. A more concrete hypothesis in the context of disease phenotypes such as cancer is that changes in the expression of genes causing the disease may be subtle compared to expression changes of the downstream effectors, which can vary largely from patient to patient (32, 47, 164). In this context, it is of highest interest that the coherence across different expression measurements of the same phenotype is often found to be significantly higher at the level of interaction subnetworks and pathways (33, 40, 80). This is primarily because changes in biochemical pathways leading to certain phenotypic conditions such as disease can often arise from a range of different alterations in the genes participating in these pathways (65, 118). Beside the better reproducibility, a further advantage of a pathway-centric perspective on expression data over the gene-centric one is that pathways provide a better mechanistic insight into the molecular mechanisms of disease. Last but not least, pathways and sub-networks may contain genes that play a major role in disease but are not captured through differential expression analysis. For example, a gene that is not differentially expressed but functionally interconnects many counterparts showing differential expression may

be causative of their dysregulation because mutations of the central gene may be disrupting its regulatory relations with its counterparts. Thus, adding a pathway layer to expression profiles can aid the discovery of molecular processes leading to complex diseases and of genes that may cause them.

1.4 Aims and organization of the thesis

A major goal of systems biology is the integration of available biological knowledge within and between different levels like gene expression, biomolecular interactions, etc. to obtain a better understanding of cellular processes. This thesis addresses three connected problems in contemporary systems biology research: **1)** Current interaction knowledge is dispersed across hundreds of heterogeneous, complementary databases, which hampers a system-level view on biomolecular relationships in the cell; **2)** Current physical interactome maps (in particular integrated interaction data) contain many false positives that may lead to the generation of false hypotheses in interaction-based research; **3)** Gene signatures are often insufficient for understanding the causes and molecular mechanisms of complex diseases, without taking into account the relationships between genes. These key points are addressed in the next three chapters, followed by a general conclusion.

Interaction data integration. Chapter 2 provides a solution to the problem that current biomolecular interaction knowledge is scattered in hundreds of heterogeneous and complementary databases, hampering a system-level view on human cellular biology. We have designed and developed ConsensusPathDB (89, 92), an interaction meta-database aiming to integrate the interactome pieces together into a seamless network comprising different types of relations between physical entities. ConsensusPathDB collates multiple functional aspects of human genes like protein interactions, catalysis, signal transduction, and gene regulation, yielding a more complete and less biased picture of cellular processes than the separate interaction resources. In Chapter 2, we outline the design and content of the meta-database as well as its web interface offering many ways to exploit the integrated network, for instance in the context of gene expression data. The necessity of data integration is demonstrated with several examples.

Confidence scoring of protein interactions. Chapter 3 tackles the problem that current protein-protein interaction data often contain considerable amounts of false positives. We propose a novel, non-parametric interaction confidence assessment approach called CAPPIC (91). It exploits solely network topology and does not depend on any reference sets or additional knowledge about the network's elements. Because such reference sets and additional information are not always available or may be ambiguous, they are a limiting factor for other interaction confidence assignment methods relying on them. We assess the performance of CAPPIC on a comprehensive set of yeast interaction networks in comparison with other topology-based methods, and demonstrate that CAPPIC reliably estimates interaction confidence and outperforms those methods. CAPPIC is used to assign confidence scores to the protein-protein interactions in ConsensusPathDB, which serve for distinguishing a high-quality physical interaction network.

Disease gene and pathway identification. Chapter 4 addresses the concept of integrating expression data with interaction or pathway knowledge to derive hypotheses about the molecular causes and mechanisms of disease. In this context, we propose the use of unbiased functional gene sets based on neighborhood of genes in the integrated interaction network. Notably, the underlying network is a result from the combination of interaction data integration (as per Chapter 2) and interaction confidence-based filtering (as per Chapter 3). The resulting gene sets can be used complementarily to curated pathways for pathway-driven expression data interpretation, and overcome several problems faced by the traditionally used manual pathway definitions. With two examples we show that the combination of collating heterogeneous interaction data, interaction de-noising, and integration of interaction and expression data could be paramount for unveiling genes causative of complex diseases such as cancer.

Further in Chapter 4 we show how integrating metabolomics data with transcriptomics/proteomics data on the level of pathways can help to generate novel hypotheses about biological processes related to a phenotype (23), and present the first available computational tool for this purpose (90). Such integration is motivated by the fact that complex diseases like cancer impact not only gene expression but also other, equally important aspects of the living cell like metabolism (77). Now that data are being generated on the large scale at several levels like gene expression, metabolism, and interaction, the time for large-scale integration of these data has come.

In summary, the key findings of the thesis are:

- design and development of the recognized interaction meta-database ConsensusPathDB (89, 92) that collates different types of interactions from currently over twenty resources into a seamless interaction network of unprecedented coverage;
- development of a novel tool for evidence mining and novelty assessment of protein-protein interactions (122);
- development of a novel, network topology-based method called CAPPIC for assessing the confidence of binary interactions (91);
- application of the integrated and de-noised human interactome map in a new approach for the identification of disease-causing genes;
- development of a novel tool for the joint analysis of large-scale transcriptomics and metabolomics data on the pathway level (23, 90).

Chapter 2

Toward more complete interactome maps

Currently, a systems view on molecular biology of the cell is severely hampered by the way interaction data are handled. The available interaction knowledge is dispersed across hundreds of databases, each of which has a specific interaction type focus, detail level, and data model and supports a different subset of the available data exchange formats. Most databases are focused on a single type of functional relations between biomolecules, while in reality, biological processes comprise many different types of interactions. Furthermore, even databases specialized on the same interaction types are often complementary than overlapping (35, 41). We designed and developed an interaction integration database called ConsensusPathDB (89, 92) to address these problems and close the gap between insular interaction data repositories. ConsensusPathDB collates the pieces of the human interactome puzzle found in these repositories into a seamless network to create a more complete snapshot of the interactions that take place in the cell. With approximately 160,000 unique interactions of different type obtained by the integration of currently 26 interaction and pathway resources, ConsensusPathDB represents the most comprehensive human interactome map available. This chapter deals with the design and content of the ConsensusPathDB meta-database, as well as its interfaces enabling researchers to exploit the integrated data in different contexts via the world wide web. While ConsensusPathDB instances exist also for the model organisms mouse and yeast, only the human instance will be referred to in this chapter.

2.1 Introduction to interaction databases

The Pathguide pathway resource list – a comprehensive catalogue of existing interaction and pathway repositories (11) – currently lists 325 different interation databases divided into several categories according to their content type (protein-protein interactions, metabolic reactions, gene regulation, etc.). Most of them are primary data resources that collect interactions directly from the literature or through manual data submissions. Since it has been recognized that primary databases are rather complementary to each other (41), efforts are made to improve the communication between their developers and unify interaction curation rules and content (117). Furthermore, standard formats have been defined for interaction data exchange (154). Several meta-databases have emerged that combine interactions from several primary resources. Examples include UniHI (28), MiMI (161) and I2D (20). Nevertheless, many of these standardization efforts and meta-databases are still limited to a single interaction type (for the above examples the focus is at protein-protein interactions). STRING (159) integrates a number of different functional associations among genes including gene neighborhood on the DNA, gene fusion, compartment co-occurrence, co-expression, co-analysis in experiments, co-occurence in databases, co-citation, physical interaction of the products, and interaction homology. These association evidence channels are combined to a joint interaction score for gene pairs. Pathway Commons (26) is a common query interface to nine interaction databases that extracts interactions through standard data formats and provides interaction and pathway search functionalities. To broaden the magnitude of interaction data integration in terms of the number of different types of interactions, number of integrated resources, and data integration depth, we created the meta-database ConsensusPathDB. Currently, it contains data from twenty-six of the most popular primary resources for direct protein-protein interactions, metabolic and signaling reactions, and gene regulation (termed source databases; Table 2.1). The number of integrated databases grows by approximately one new database per release (Appendix Figure A.1).

Ten of these resources contribute biochemical reactions. Only two of them, Reactome (39) and INOH (http://www.inoh.org), contain both signaling and metabolic reactions. The rest are focused only on metabolism (HumanCyc (131) and the Edinburgh Human Metabolic Network Reconstruction - EHMN (108)), or only on signaling

database name	web page (http://)	data types	interaction/ pathway count	data format for integration
Reactome	www.reactome.org	SR, SP, MR, MP	7627/1153	MySQL dump
KEGG	www.genome.jp/kegg	SP, MR, MP	1805/239	web services
HumanCyc	humancyc.org	MR, MP	2008/294	proprietary flat files
PID	pid.nci.nih.gov	SR, SP, GR	10158/219	proprietary XML
BioCarta*	www.biocarta.com/genes/index.asp	SR, SP, GR	3490/254	proprietary XML
NetPath	www.netpath.org	SR, SP, PPI	3237/20	BioPAX
INOH	www.inoh.org	SR, SP, MR, MP	3691/93	BioPAX
EHMN	www.ehmn.bioinformatics.ed.ac.uk	MR, MP	4187/69	SBML
IntAct**	www.ebi.ac.uk/intact	PPI	33513/0	PSI-MI
DIP	dip.doe-mbi.ucla.edu	PPI	13550/0	PSI-MI
MINT	mint.bio.uniroma2.it	PPI	21379/0	PSI-MI
HPRD	www.hprd.org	PPI	40618/0	PSI-MI
CORUM	mips.helmholtz-muenchen.de/genre/proj/corum	PPI	1664/0	proprietary flat files
BioGRID	www.thebiogrid.org	PPI	62009/0	PSI-MI
MIPS-MPPI	mips.helmholtz-muenchen.de/proj/ppi	PPI	739/0	PSI-MI
BIND	www.bind.ca	PPI	21424/0	PSI-MI
SPIKE	www.cs.tau.ac.il/ spike	GR, SR, PPI	41545/0	BioPAX
PIG	molvis.vbi.vt.edu/pig	PPI	20113/0	PostgreSQL dump
PhosphoPoint	kinase.bioinformatics.tw	SR, PPI	11609/0	Excel tables
PDZbase	icb.med.cornell.edu/services/pdz/start	PPI	101/0	proprietary flat files
InnateDB	www.innatedb.ca	SR, GR, PPI	7113/0	PSI-MI
MatrixDB	matrixdb.ibcp.fr	PPI	247/0	proprietary flat files
PharmGKB	www.pharmgkb.org	SP	0/77	BioPAX
SMPDB	www.smpdb.ca	SP, MP	0/411	proprietary flat files
WikiPathways	www.wikipathways.org	SP, MP	0/324	BioPAX
SignaLink	signalink.org	SP	0/8	Excel tables

Table 2.1: Interaction databases integrated in ConsensusPathDB. Data types: SR: signaling reactions; SP: signaling pathways; MR: metabolic reactions; MP: metabolic pathways; GR: gene regulatory interactions; PPI: protein-protein interactions. The counts correspond to possibly non-unique human interactions/pathways as per each database. Listed are only data formats used for data integration in ConsensusPathDB, which we have found to be most comprehensive. Note that in many cases, the use proprietary (i.e., database-specific) interacction formats is necessary since the standard formats, if supported, are uncomplete in respect to the database content. *BioCarta data have been downloaded from the PID web site. **Interactions from IntAct derived from small-scale or large-scale experiments are considered as separate resources in ConsensusPathDB (denoted IntAct-SS and IntAct-LS, respectively).

(Pathway Interaction Database - PID (138), Signaling Pathway Integrated Knowledge Engine - SPIKE (121), BioCarta (http://www.biocarta.com/genes/index.asp), NetPath (93) and InnateDB (107)). The Kyoto Encyclopedia of Genes and Genomes - KEGG (95) is a repository for manually drawn pathway diagrams of both signaling and metabolic pathways; however, computer-readable reaction data is available only for the metabolic but not for the signaling pathways. Most of these ten databases are general-purpose repositories, that is, they attempt to chart the molecular reaction mechanisms of a palette of biological processes. Exceptions are InnateDB, which focuses on interactions involved in the innate immune response to microbial infection, and NetPath, which catalogues immune and cancer signaling pathways. Common to all biochemical reaction databases is that they are subject to manual curation. Some of the databases (e.g. HumanCyc) have resulted from computational reaction predictions and are moderately curated, while others (e.g. Reactome, SPIKE, PID) store highly curated reaction data that are most often manually extracted from the scientific literature by experts. Furthermore, the different databases provide a different level of annotation detail of the contained interactions. For instance, compartment annotation of reactions and information on post-translational modifications of their participants is available only in Reactome, INOH, PID, BioCarta, and NetPath. In almost all of the reaction databases mentioned here, the contained reactions are organized into groups representing biochemical pathways (Table 2.1).

Four of the databases mentioned above contain gene regulatory interactions; these are SPIKE, PID, BioCarta, and InnateDB. Publicly accessible gene regulatory data are still relatively sparse for human, thus these databases provide only a small number (in the order of a few hundred to a few thousand) of gene regulatory relations mined manually from the scientific literature.

Twelve of the databases integrated in ConsensusPathDB focus only on physical protein interactions. These include IntAct (7), Database of Interacting Proteins (DIP) (136), Molecular Interaction Database (MINT) (25), Human Protein Reference Database (HPRD) (98), Comprehensive Resource of Mammalian protein complexes (CORUM) (134), Biological General Repository for Interaction Datasets (BioGRID) (149), Mammalian Protein-Protein Interaction Database of the Munich Information Center for Protein Sequences (MIPS-MPPI) (119), Biomolecular Interaction Network Database (BIND) (83), Pathogen Interaction Gateway (PIG) (46), PhosphoPoint (179), PDZbase

(16), and MatrixDB (29). Data in most of these repositories are typically collected from the literature through text mining followed by manual curation to some extent, or are directly submitted by experimentalists. Some of the databases have a particular focus on interactions between certain types of proteins or taking place in specific compartments. For example, PhosphoPoint focuses on interactions of human kinases, PDZbase on interactions involving PDZ domains, and MatrixDB comprises interactions between extracellular proteins and polysaccharides on the cell surface. Similarly, the Pathogen Interaction Gateway imports interactions between human and pathogenic proteins from other databases like IntAct. The rest of the protein-protein interaction repositories integrated in ConsensusPathDB are general-purpose databases aiming to assemble a protein interactome map of human as well as of other species. Some of the protein interaction databases (like IntAct, DIP, and MINT) contain interactions involving more than two proteins (complex interactions), while others (like BioGRID and PIG) contain only binary interaction data. Some of the databases (e.g. IntAct) provide information on the modification state of interactors. This feature is particularly important as some interactions are modification-dependent, that is, they take place only if the proteins are post-translationally modified (which is often the case with interactions building up signaling cascades). Apart from the data extracted from the protein interaction-focused databases, we have explicitly defined protein interactions based on the composition of protein complexes found in some of the biochemical reaction databases. Furthermore, many physical interactions are provided by the signaling database SPIKE.

Pathway annotation of the bulk of available protein-protein interactions is still forthcoming. In contrast, most of the resources for biochemical reactions annotate all or most of the reactions to biochemical pathways as mentioned above. Several further pathway resources exist that do not provide information about pathway constitution in terms of reactions but instead depict pathways in manually drawn diagrams and list the genes participating in each pathway. Such resources provide valuable information which can be used in approaches for pathway-level analysis of gene expression data (discussed in Chapter 4). Such pathway databases integrated in ConsensusPathDB are Pharmacogenomics Knowledge Base - PharmGKB (163), Small Molecule Pathway Database - SMPDB (57), WikiPathways (125), Signalink (102), and the signaling pathway domain of KEGG.

2.2 Data model of ConsensusPathDB and data integration

2.2.1 Database structure

Developing a meta-database that holds information on interactions of different nature and annotated in a different level of detail by the source databases required a design of an adequate database schema. The schema had to be general enough to allow for representing interactions of different nature, and at the same time specific enough so that interaction details such as cellular location could be included, if available. Moreover, the schema design had to consider the fact that interaction datasets are overlapping to some extent, and had to offer an adequate way to identify overlapping information and avoid redundancy.

The design of the data repository of ConsensusPathDB follows a bipartite multigraph interaction data model (Figure 2.1), which enables it to acommodate molecular relations with arbitrary cardinality. Its central classes are *PhysicalEntity*, *Interaction* and *Edge*. Physical entities and interactions are accordingly the two different types of nodes in the bipartite graph model and are connected by edges denoting the participation of entities in interactions. There are currently three types of interactions represented by three distinct classes that inherit from the general *Interaction* class: *physicalInteraction*, *biochemicalReaction* (representing both metabolic and signaling reactions), and *geneRegulation*. Each physical entity has a type as well, which is either gene, messenger RNA (mRNA), non-coding RNA (ncRNA), peptide, protein, protein complex, family (gene or protein family), compound/metabolite, or unknown type. Physical entities are accommodated in ConsensusPathDB in a basic form – for example, physical entities of the type protein do not have post-translational modifications by themselves. Instead, edges linking physical entites to their interactions are the carriers of information about the state (such as post-translational modifications or mutations), the cellular compartment location, as well as stoichiometry information of the interaction participant in the interaction. Each edge records the role of the physical entity in the interaction (such as product, substrate, enzyme, physical interactor, regulated gene, enhancer or inhibitor). Physical entities are organized hierarchically (accomplished through the relation *has_component*, Figure 2.1), which is necessary for representing protein complexes and gene families in terms of their composition. Interactions

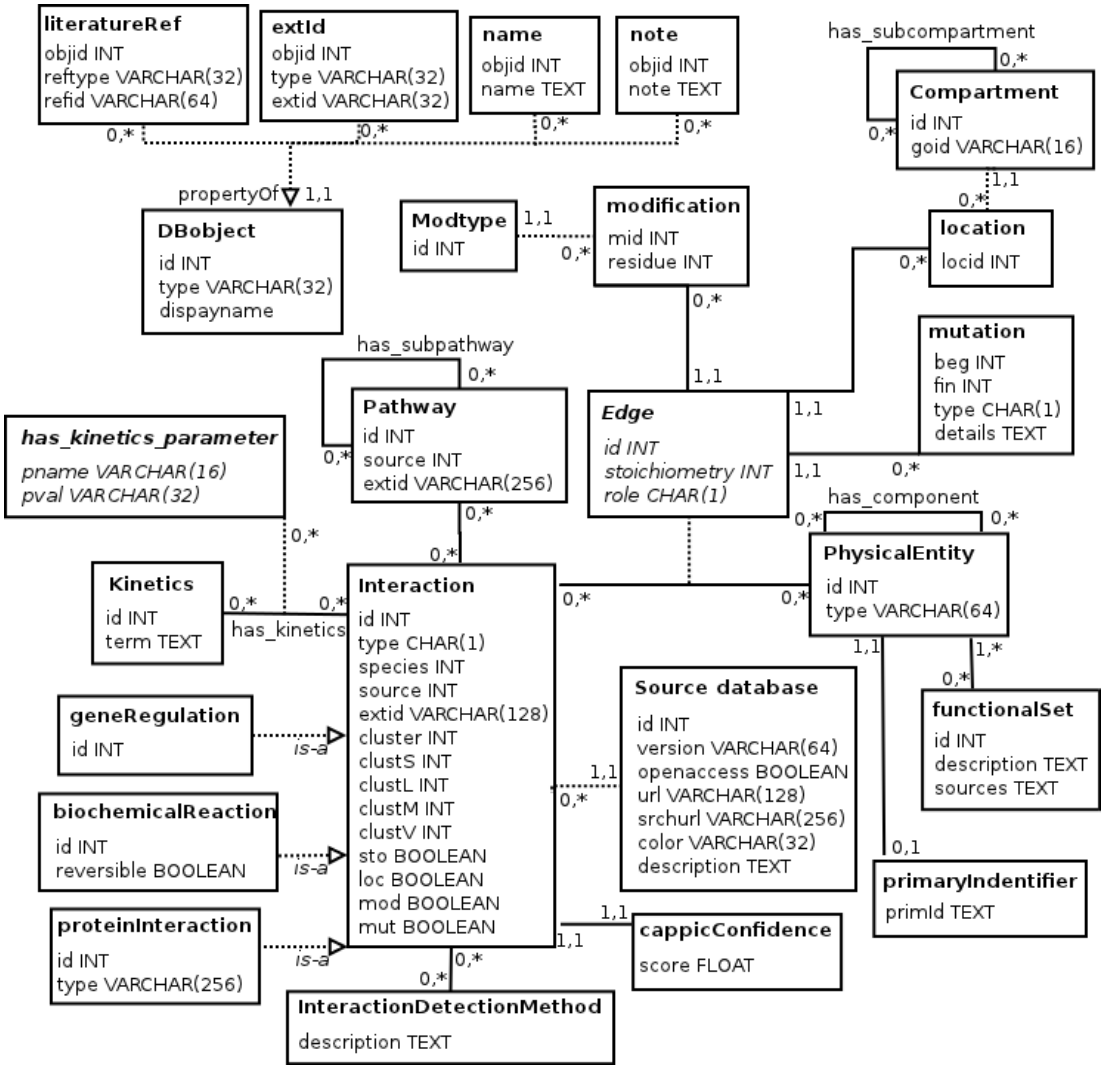

Figure 2.1: Entity-relationship diagram visualizing the structure of the ConsensusPathDB interaction meta-database. The names of classes (rectangles) which are members of the generic *DBobject* (database object) class are capitalized (the corresponding 'is-a' relationships are omitted for better visualization). The numbers correspond to relationship cardinalities. The central classes are *Entity* (representing physical entities like genes, proteins, complexes, metabolites, etc.), *Interaction* (representing interactions), and *Edge* (connecting entities with their interactions). The schema follows a bipartite graph model for representing interactions, to allow representation of biomolecular relationships of arbitrary cardinality. The *cluster*,*clustS*, *clustM*, *clustL*, and *clustV* attributes of *Interaction* serve for indicating which interactions are similar regarding the entity composition and have matching stoichiometry, modification, location, and mutation patterns of the entities (see main text).

can be organized in pathways (instances of class *Pathway* which link to *Interaction* via *has_interaction*, Figure 2.1). Similarly to physical entities, pathways are organized in a hierarchical manner (a pathway may consist of sub-pathways: relation *has_subpathway*, Figure 2.1). Many further relations exist in the database schema for storing additional data, including kinetics information (including kinetics laws and parameter values), details on mutations (site and mutation type) or modifications (residue and chemical group) (Figure 2.1). The relational schema described here was implemented as a PostgreSQL database system.

2.2.2 Integration of interaction data from multiple sources

The task of interaction data integration is hindered primarily by the heterogeneity regarding the data formats of currently available interaction resources. We retrieve the data from source databases in different ways and formats, ranging from files in standard interaction exchange formats including BioPAX, PSI-MI, and SBML to database-specific XML or tab-delimited files, Excel tables, MySQL or PostgreSQL database dumps, or SOAP web services. Table 2.1 provides information about how interaction data were retrieved from each source database. We created a separate data adapter for each database that extracts its content and translates it in compliance with the data model of ConsensusPathDB. The data are then not simply stored in the repository of ConsensusPathDB, but also compared to the information already present in it to detect similarities and consequently avoid redundancy. Simple physical entities are compared to each other based on identifiers from a unified namespace, called primary identifiers. These identifiers are UniProt (6) for proteins, Ensembl (53) for genes, and KEGG (or ChEBI (42), in case that a KEGG identifier is missing) for metabolites, because these databases annotate very extensively human proteins, genes and metabolites, respectively. We attempt to map all identifiers provided for every entity by the accorging interaction resources to one or more primary identifiers. For this purpose we created an identifier cross-map by parsing and extracting acession number mappings from eight genomic, proteomic, and metabolite databases including UniProt, Ensembl, Entrez (109), HUGO Gene Nomenclature Committee (HGNC) (142), Human Protein Reference Database (HPRD), KEGG, ChEBI and Human Metabolome Database (HMDB) (177). Simple physical entities whose set of primary identifiers match, or complex entities such as protein complexes or families with matching composition, are considered

identical and are merged in ConsensusPathDB. Annotation of merged physical entities such as external identifiers, literature references, and synonyms are stored in a complementary manner.

Interactions from different sources are also compared to each other. Biochemical reactions with matching substrates and products, physical interactions with matching interactors, as well as gene regulatory interactions with matching regulated gene are considered similar. Notably, similar interactions may differ in the modification state, location, or stoichiometry of their participants. For example, as mentioned above, the Reactome database provides information about the modification state and subcellular location of each interacting entity, whereas KEGG does not. To enable the comparison of interactions from databases with such differences in the annotation detail, we apply the following strategy: Each interaction is stored separately in ConsensusPathDB, and similar interactions (as defined above) are marked as similar. This is accomplished through equal settings of the 'cluster' attribute of the *Interaction* class (Figure 2.1) for similar interactions. It should be noted that in this context, the word 'cluster' denotes a group of interactions that have identical composition in terms of substrates and products (for biochemical reactions), physical interactors (for protein interactions) and regulated gene (in the case of gene regulatory interactions) and is not to be confused with graph clusters, for example. Interactions within the same 'cluster' are divided into sub-groups depending on whether their stoichiometry, modification, location, and mutation information match. This is done through settings of the 'clustS', 'clustM', 'clustL', and 'clustV' attributes of *Interaction*. For example, 'clustM' has the same value for interactions in the same 'cluster' that match in the post-translational modification pattern of their protein participants. The decision, which of the similar interactions are to be considered identical, depends on the concrete application and is therefore left to the end-user. If, for example, a network of reactions from ConsensusPathDB is to be used as the basis for models and computer simulations of a biological process, then interactions in different compartments should probably be differentiated. If, on the other hand, the aim is to retrieve all functional relationships of specific biomolecules in the cell, then compartment information is probably irrelevant.

For each object in ConsensusPathDB (including physical entities, interactions and pathways), we record its sources and source database identifiers to enable linking to the original data, as well as all literature references where the object is primarily described.

The data integration module of ConsensusPathDB comprises computer programs that create an empty repository following the described schema, download the latest versions of all data from the source databases, translate each dataset into a unified format consistent with ConsensusPathDB's data model, integrate the data into the data repository in a non-redundant manner, and perform post-processing on the integrated data e.g. to calculate overlap statistics. The integration module is executed fully automatically every three months to ensure the content of our meta-database is always up-to-date. Appendix Figure A.1 shows a release timeline summarizing the unique interaction count and integration of new source databases in ConsensusPathDB since its initial publication.

2.3 A global view on the integrated content of ConsensusPathDB

ConsensusPathDB is the largest interactome map for *Homo sapiens*. Currently (Release 20), it comprises 51,564 unique physical entities (32,357 proteins or protein families, 10,252 protein complexes, 120 non-coding RNA molecules, 5,040 metabolites, etc.), 157,461 unique interactions (2,270 gene regulatory interactions, 16,721 biochemical reactions, and 138,470 complex or binary protein interactions), as well as 3,161 pathways. Interaction integration enabled the assessment of the overlaps and differences between the integrated resources for interaction and pathway data, which we detail below.

2.3.1 Complementarity of interaction data resources

The interaction network in the ConsensusPathDB repository has been obtained by collating a total of 317,065 interactions from the source databases. The fact that the unique interactions in the integrated network are less than half that number indicates that the databases do overlap to some extent. We have summarized the pairwise database overlap sizes both in terms of interactions and physical entities in Table 2.2. The table essentially shows that the databases are complementary to each other and none of them is completely contained in another. Each database contributes unique interactions to the integrated network. The non-zero overlaps between biochemical reaction and protein interaction repositories is due to the fact that we explicitly defined protein interactions from protein complexes found in the former, as mentioned above.

	Reac.	*KEGG*	*Huma.*	*PID*	*Bioc.*	*NetP.*	*INOH*	*EHMN*	*Inna.*	*IA-SS*	*IA-LS*	*DIP*	*MINT*	*HPRD*	*CORUM*	*BioG.*	*M.MPPI*	*BIND*	*Matr.*	*SPIKE*	*PIG*	*Phos.*	*PDZb.*
Reac.	**7091** 12817	**305**	**221**	**334**	**99**	**67**	**216**	**352**	**105**	**187**	**20**	**125**	**102**	**448**	**121**	**278**	**9**	**210**	**1**	**269**	**0**	**109**	**0**
KEGG	1370	**1803** 3030	**295**	**0**	**4**	**0**	**381**	**1119**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Huma.	2003	1887	**1934** 5259	**5**	**9**	**0**	**151**	**338**	**5**	**7**	**2**	**7**	**4**	**33**	**8**	**18**	**0**	**20**	**0**	**8**	**0**	**1**	**0**
PID	1653	200	619	**6880** 6805	**385**	**272**	**116**	**2**	**272**	**188**	**14**	**183**	**130**	**528**	**132**	**447**	**18**	**221**	**2**	**622**	**0**	**454**	**7**
Bioc.	1032	173	454	1047	**2260** 2884	**129**	**53**	**9**	**116**	**63**	**7**	**59**	**45**	**111**	**41**	**123**	**7**	**64**	**0**	**302**	**0**	**211**	**0**
NetP.	540	50	248	626	390	**2139** 1074	**65**	**0**	**326**	**278**	**13**	**227**	**325**	**1315**	**35**	**729**	**20**	**194**	**0**	**1430**	**0**	**641**	**0**
INOH	994	1050	995	398	310	199	**2690** 4364	**396**	**61**	**36**	**5**	**37**	**33**	**79**	**23**	**76**	**1**	**49**	**0**	**103**	**0**	**64**	**0**
EHMN	1316	2269	1800	194	177	51	1005	**3872** 3885	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Inna.	1261	129	532	1160	722	542	294	129	**5890** 3344	**636**	**156**	**385**	**577**	**1798**	**77**	**1678**	**34**	**483**	**4**	**1597**	**0**	**924**	**7**
IA-SS	2587	472	1312	1691	990	701	442	434	1860	**14051** 7373	**448**	**4054**	**3616**	**3767**	**206**	**3737**	**66**	**1379**	**9**	**7728**	**0**	**1295**	**11**
IA-LS	1461	366	813	818	483	340	269	341	1020	2753	**12212** 4618	**5679**	**3931**	**5168**	**16**	**6453**	**12**	**133**	**0**	**9759**	**0**	**241**	**0**
DIP	2174	364	1018	1445	858	615	400	345	1531	4374	2780	**13457** 5409	**2417**	**2149**	**114**	**4163**	**47**	**544**	**3**	**8120**	**0**	**454**	**2**
MINT	2320	421	1098	1548	919	695	411	399	1684	4179	3241	3429	**14980** 6588	**7230**	**111**	**5644**	**58**	**1215**	**2**	**11564**	**0**	**1630**	**53**
HPRD	3554	756	1832	2183	1231	927	598	681	2213	5373	3618	4072	5161	**40410** 9824	**469**	**16658**	**250**	**4244**	**28**	**14962**	**0**	**5997**	**91**
CORUM	1513	105	444	987	641	428	211	115	916	1990	1120	1698	1820	2166	**1664** 4195	**280**	**9**	**230**	**0**	**250**	**0**	**66**	**0**
BioG.	3325	609	1654	2148	1224	840	553	570	2291	5560	3762	4460	5006	7634	2262	**40545** 10800	**203**	**2523**	**16**	**14712**	**0**	**3015**	**62**
M.MPPI	257	14	78	257	185	142	60	10	235	312	158	285	309	356	226	379	**323** 400	**160**	**2**	**152**	**0**	**39**	**3**
BIND	2829	646	1464	1717	1035	719	535	573	1816	4103	2454	3270	3671	5290	1935	5102	363	**20686** 9547	**4**	**2724**	**0**	**1211**	**16**
Matr.	74	4	20	77	44	10	13	7	72	88	31	65	71	120	46	115	13	103	**247** 159	**11**	**0**	**0**	**0**
SPIKE	3406	743	1758	2227	1283	877	626	679	2426	5829	3989	4619	5535	7566	2317	7449	373	5248	118	**38108** 11222	**0**	**3583**	**60**
PIG	1549	289	660	1036	687	482	293	277	1170	2508	1489	2025	2217	2927	1137	2905	230	2416	65	3013	**20098** 7578	**0**	**0**
Phos.	1561	169	795	1376	823	654	298	163	1297	2252	1288	1834	2205	2972	1119	2856	274	2219	50	2852	1276	**10017** 3186	**20**
PDZb.	54	9	33	50	23	20	8	8	50	78	36	54	102	113	27	109	15	79	2	110	43	74	**101** 115

Table 2.2: Interaction database overlaps. Main diagonal: number of unique human interactions (bold font) and physical entities in the according database; above the diagonal (bold font): number of shared unique interactions; below the diagonal: number of shared unique physical entities. Abbreviations: Reac.: Reactome; Huma.: HumanCyc; Bioc.: BioCarta; NetP.:NetPath; Inna.: InnateDB; IA-SS: IntAct(small-scale); IA-LS: IntAct(large-scale); BioG.: BioGRID; M.MPPI: MIPS-MPPI; Matr.: MatrixDB; Phos.: PhosphoPoint; PDZb.: PDZbase. PIG overlaps with no other database because it is a source for host-pathogenic interactions. The table refers to ConsensusPathDB Release 20(13.07.2011). Note that, unlike in Table 2.1, unique interaction counts are given.

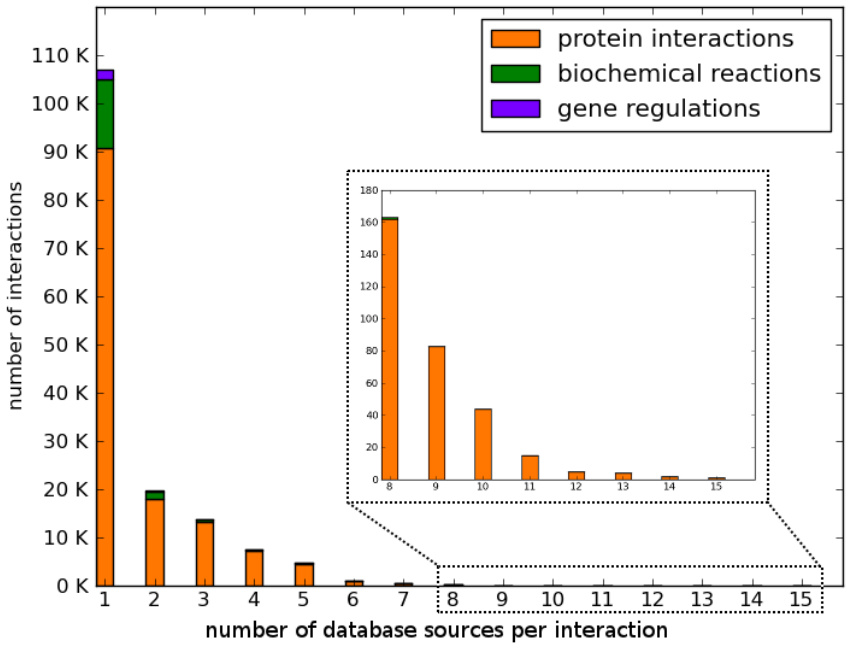

Figure 2.2: Histogram of the number of database sources per interaction in ConsensusPathDB. The vast majority of interactions (69%) are available in only one of the integrated databases. Only one interaction has 15 sources (NFKB1_HUMAN – TF65_HUMAN), two interactions have 14 sources (IF4E_HUMAN – 4EBP1_HUMAN and GRB2_HUMAN – SOS1_HUMAN), and four interactions have 13 sources (CCNB1_HUMAN – CDK1_HUMAN, EGF_HUMAN – EGFR_HUMAN, CCNE1_HUMAN – CDK2_HUMAN, and SMAD3_HUMAN – SMAD4_HUMAN); all of these are physical protein interactions.

We further dissected the interactions in the integrated network according to the number of different source databases per interaction (Figure 2.2). Strikingly, around 69% of the interactions are contained in a single source database only, while the fraction of interactions from exactly two or exactly three source databases is 13% and 9%, respectively. Only one interaction (the physical interaction between the 105p and 65p subunits of Nuclear factor NF-kappa-B: NFKB1_HUMAN and TF65_HUMAN, respectively) is present in 15 source databases, while no interactions are common to more than 15 databases.

We exemplarily looked at the distribution of protein interactions of one of the best annotated proteins, the Tumor suppressor protein p53 (P53_HUMAN), which plays a central role in the cell cycle and whose mutations are often associated with cell cycle

dysregulation leading to cancer (75). We found 745 unique protein interactions of p53 in ConsensusPathDB. Four of the most comprehensive protein interaction databases that we have integrated – IntAct, HPRD, BioGRID and DIP – contained in total 509 of these interactions, and only 12 interactions were common to all four databases (Figure 2.3). This finding evidences that the separate databases, even if focused on the same interaction types, are highly complementary in their interaction content.

In addition to providing a more comprehensive view on each physical entity's interactions of a given type like in the above example, ConsensusPathDB reveals multiple functional relationships between the entities at the same time. For instance, we found that each human gene/protein represented in our database is involved in 1.5 distinct types of interactions (gene regulatory interactions, biochemical reactions, and protein interactions) on average. The number is relatively high, considering the uneven numbers of interactions of each type found in in ConsensusPathDB. For instance, currently there are 61 times more physical protein interactions than gene regulatory interactions. If only genes/proteins participating in available gene regulatory interactions are considered, the average number of different interaction types per gene/protein is 2.6. With the elucidation of more regulatory and biochemical gene relationships in human, an ascending tendency of the number of different interaction types available for every gene in ConsensusPathDB is expected.

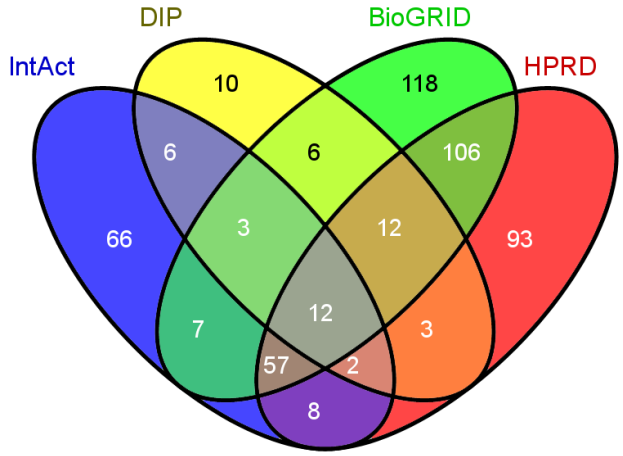

Figure 2.3: Overlap and complementarity of interactions of p53 in four major protein interaction databases. Although p53 is one of the most extensively analyzed proteins, protein-protein interaction databases contain complementary subsets of its available interactions as shown in this Venn diagram.

Next, we analyzed the degree of coherence and complementarity of the integrated source databases beyond the interaction level. Several databases are concerned with

the higher-level organization of interactions in biological processes and attempt to create molecular-level models of such processes. The resulting pathway definitions are extensively used, for example, in methods for pathway-based expression data analysis (discussed in Chapter 4). We were interested in the level of similarity between pathway definitions from different databases. For each pathway from each database, we compared its composition (in terms of physical entities or interactions) to all pathways from the rest of the source databases. The similarity of a pair of pathways P and Q was quantified with the Jaccard index, $J(P,Q) = |P \cap Q|/|P \cup Q|$, where $|P \cap Q|$ is the size of the intersection and $|P \cup Q|$ is the size of the union of the two pathways in terms of entities or interactions. $J(P,Q)$ ranges from 0, if P and Q share no items, to 1, if they completely match regarding their composition. The maximum reached Jaccard index value per pathway (i.e. the maximum similarity to any pathway from a different database) is shown for all pathways in Figure 2.4, A) and B) (for physical entities and interactions, respectively). It is evident that pathways from every database are mostly unique in their composition. Since most pathway databases attempt to chart extensively studied biological processes such as Apoptosis, TCA cycle or Glycolysis, we were interested how well the compositions of such pathways match across the databases. We exemplarily inspected the composition of the Glycolysis pathway according to four established metabolic pathway databases (Reactome, KEGG, HumanCyc and INOH). The pathway was present as "Glycolysis" in Reactome, "Glycolysis and gluconeogenesis" in INOH, "Glycolysis and gluconeogenesis" in KEGG, and "Glycolysis I" in HumanCyc. We found astonishing differences in the pathway composition across the four databases (Figure 2.4 C) and D) show their overlaps in terms of physical entities and interactions, respectively). For example, the INOH Glycolysis and gluconeogenesis instance contained 21 reactions involving a total of 45 distinct physical entities, while the homonymous KEGG instance consisted of 33 reactions and 65 entities. The overlap between all four databases comprised only 3 reactions or 17 entities (Figure 2.4 C) and D)). Results were similar for the comparably well-studied Apoptosis and TCA cycle pathways (not shown), indicating that pathway definitions are rather a matter of subjective judgment as pathway boundaries are generally unclear (169).

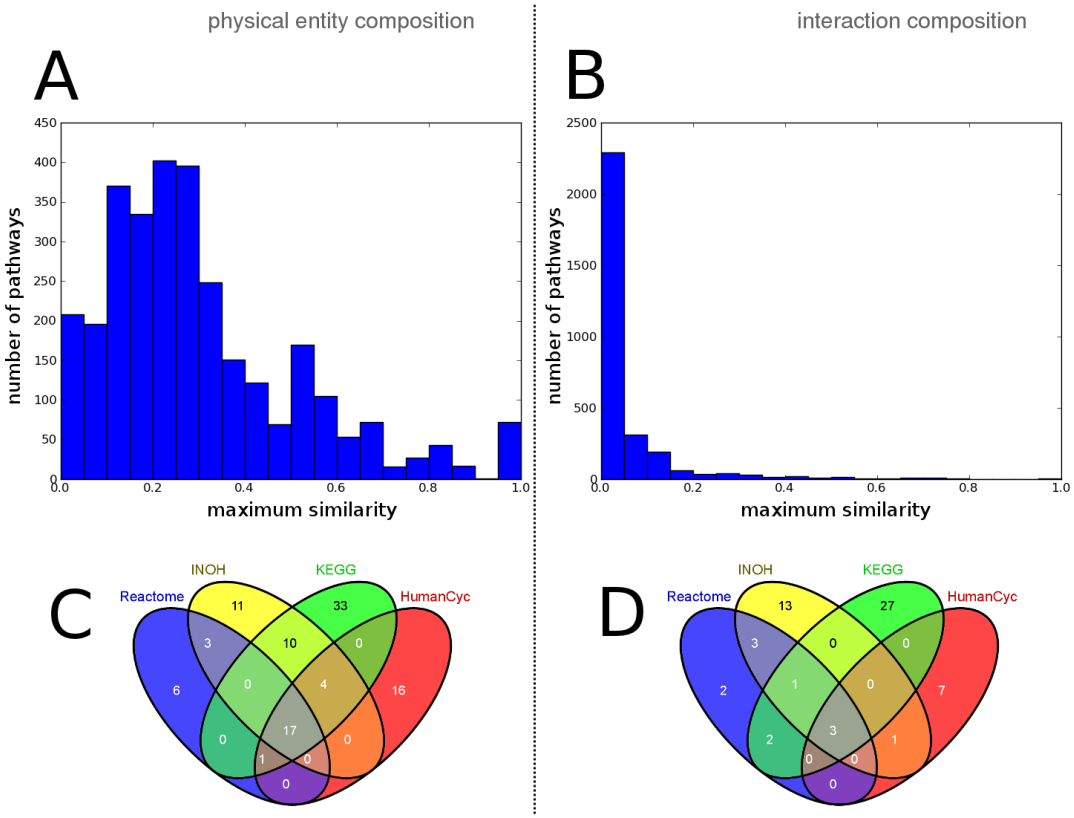

Figure 2.4: Overlap of pathway composition across databases. **A** and **B**: Histograms of the similarity (maximum Jaccard index) of pathways to counterparts from other databases in terms of physical entity composition (**A**) or interaction composition (**B**); **C** and **D**: comparison of the composition of the Glycolysis pathway in terms of physical entities **C** or interactions **D** across four major pathway databases.

2.3.2 Topological properties of the human protein interaction network

In the previous chapter we mentioned several network metrics that can be applied to characterize the structure of real-world networks. Such metrics are commonly used to derive hypotheses about the organization and evolution of functional associations of biomolecules within the cell, and are often directly related with biological phenomena (14). Because of the compositional differences of current interaction databases, however, topological analyses hide the risk of bias in the results depending on which database analyses are based on. We utilized the binary protein-protein interaction network from ConsensusPathDB, as well as the separate contributing source databases, to spot common as well as specific network structural properties. Interactions with more

than two participants were disregarded in this analysis. We modeled each interaction dataset as an undirected graph where nodes represented proteins and edges represented their interactions. The graphs were characterized in terms of the topological measures introduced in the previous Chapter (Table 2.3). Notably, the networks had different coverage of the human physical interactome, ranging from around 100 (MatrixDB) to over 96,000 (ConsensusPathDB) interactions. The average shortest path spanned around 4 interactions, and the diameter ranged between 8 and 17. This demonstrates that all networks in the analysis are the small-world, suggesting that the property is often preserved in samples of the real human physical interactome barely dependent on their size. On the other hand, the range of the average clustering coefficient across the analyzed networks was fairly big: For the large-scale dataset from IntAct (IntAct-LS) it measured only 0.05, which was more than five times smaller than for BIND (clustering coefficient = 0.26). IntAct-LS and BIND seem to represent different subsets of the interactome that are barely overlapping: Table 2.2 shows that they have only 133 interactions in common. While IntAct-LS consists of the large-scale experimental data published by Rual *et al.* (133), Stelzl *et al.* (152), and Ewing *et al.* (50), BIND comprises mostly small-scale experimental data manually curated from the literature. The clustering coefficient of the integrated network lied between the two extremes (clustering coefficient = 0.16). The average node degree ranged from less than 2 (CORUM and MIPS-MPPI) to 13.3 (ConsensusPathDB) interaction partners per protein. Overall, the results confirmed that conclusions about the topological properties of the human interactome may differ according to which database is used as a basis for the analysis.

The distributions of protein degree and clustering coefficient in the integrated human physical interactome map are shown in Figure 2.5. The evident power-law distribution of protein degree, approximated by $P(X = k) \sim k^{-1.42}$ (where k denotes protein degree) indicates the scale-free nature of the network (13). The power-law scaling of the clustering coefficient with protein degree obvious in Figure 2.5 is a direct evidence for a hierarchical organization of modularity in the network (127, 128).

2.4 Interfaces of ConsensusPathDB

To grant researchers around the globe access to the integrated content of ConsensusPahDB, we have developed a web interface and a specialized plugin for the popular

property	**BIND**	**BioGRID**	**CORUM**	**DIP**	**HPRD**	**InnateDB**	**IntAct-LS**	**IntAct-SS**
node count	7033	8984	653	4854	9596	2237	4318	5441
edge count	14495	37607	505	12873	38933	4695	11470	12454
average degree	4	8.3	1.4	5.3	7.9	4.1	5.3	4.5
number of CC	297	177	231	149	261	77	80	172
% nodes in largest CC	94.7%	96.9%	5.4%	94.2%	95.9%	93.0%	96.6%	93.9%
% edges in largest CC	97.6%	99.4%	7.3%	98.2%	99.1%	97.7%	99.2%	98.0%
avg. shortest path*	3.43	4.19	5.18	4.46	4.22	4.6	4.4	4.73
diameter*	12	12	12	17	14	10	12	13
avg. clustering coeff.	0.26	0.16	0.08	0.15	0.13	0.21	0.05	0.13
power-law exponent*	1.72	1.48	1.8	1.6	1.48	1.64	1.6	1.64

property	**MINT**	**MIPS-MPPI**	**Man. Cur.**	**MatrixDB**	**PDZBase**	**PhosphoPOINT**	**SPIKE**	**CPDB**
node count	5921	377	1323	98	115	3180	8512	14264
edge count	14473	305	5235	137	101	9022	34610	96360
average degree	4.8	1.6	7.7	2.8	1.8	5.6	8.1	13.3
number of CC	162	96	18	6	19	12	69	160
% nodes in largest CC	95.1%	14.1%	98.4%	80.6%	21.7%	99.4%	98.5%	98.5%
% edges in largest CC	98.6%	17.7%	99.7%	89.1%	27.7%	99.8%	99.7%	99.8%
avg. shortest path*	4.45	5.94	3.62	4.11	3.51	3.92	4	3.46
diameter*	12	13	9	9	8	11	11	11
avg. clustering coeff.	0.12	0.14	0.2	0.14	0	0.18	0.11	0.16
power-law exponent*	1.62	1.82	1.46	1.71	1.8	1.63	1.5	1.42

Table 2.3: Topological properties of physical interactome maps. Abbreviations: SS: small-scale; LS: large-scale; CPDB:ConsensusPathDB; CC: connected component; avg.: average; coeff.: coefficient. *The properties marked with an asterisk refer to the largest connected network component.

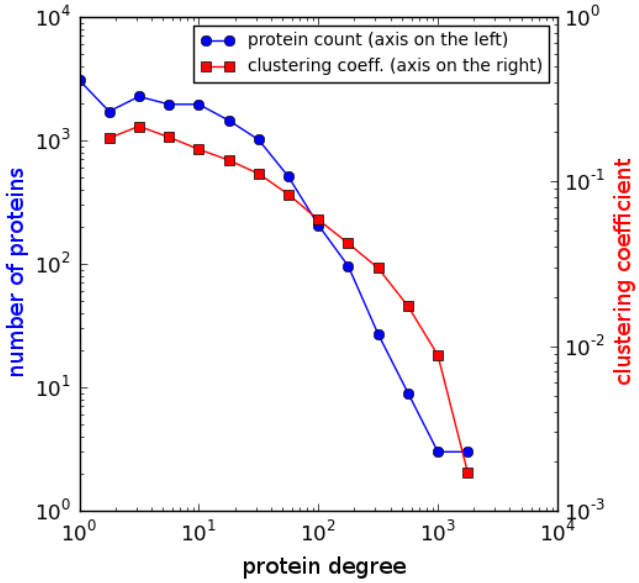

Figure 2.5: Distributions of protein degree and clustering coefficient in the integrated human physical interactome map. The number of proteins (blue line, left-hand-side y-axis) and the clustering coefficients of proteins (red line, right-hand-side y-axis) are plotted against protein degree.

network analysis and visualization software tool Cytoscape (148).

2.4.1 The ConsensusPathDB web interface

The web interface of ConsensusPathDB can be accessed with a contemporary web browser at http://cpdb.molgen.mpg.de. Its functionality is summarized in Figure 2.6 and is extensively documented in an online manual found on the ConsensusPathDB home page. The web interface offers possibilities to query the interactions of specific physical entities or pathways, or search for shortest interaction paths connecting pairs of biomolecules. Selected interactions can be visualized either in an image-based or a Java-applet-based visualization framework. Both frameworks represent interactions in an identical way. Interactions are displayed as directed bipartite multigraphs where circular nodes represent interactions and rectangular nodes represent physical entities (Figure 2.6). The color of each node encodes the type of the according interaction or entity. Entities are linked to their interactions with edges whose line style, arrow shape and orientation encode the roles of the entities in the interactions. The color

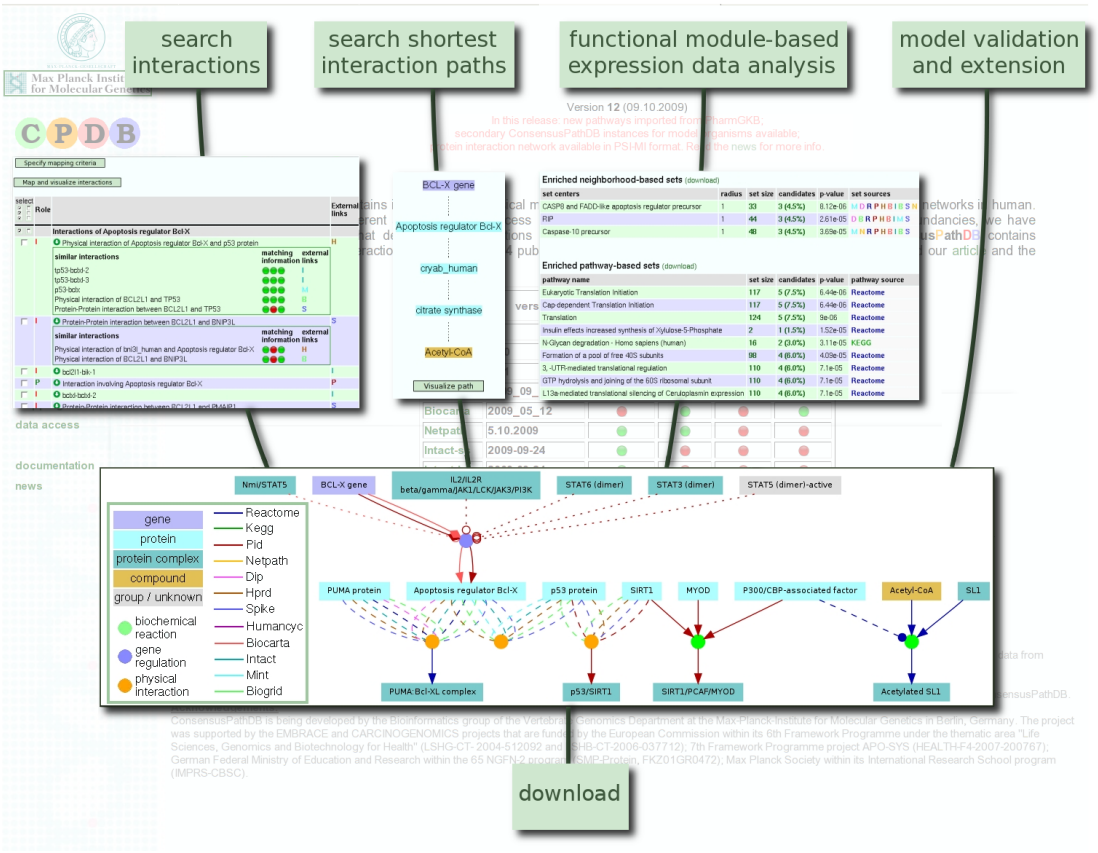

Figure 2.6: Overview of the funtionality of the ConsensusPathDB web interface. By either searching for interactions of specific physical entities or pathways, searching for shortest interaction paths connecting two physical entities, upload of expression data for gene set-based analysis, or upload of standard files containing interactions which are matched to the meta-database, custom interaction networks can be constructed and displayed in one of ConsensusPathDB's visualization environments. Consistent with the data model of the database, these networks are visualized as bipartite multigraphs where one class of nodes (shown as rectangles) represent physical entities, and the other class (shown as circles) represent their interactions. Node color shows the type of the corresponding physical entity (gene, protein, metabolite, etc.) or interaction (gene regulation, protein interaction, or biochemical reaction). Edges connect physical entities to their interactions; edge style denotes the role of the entity (regulated gene, transcription factor, physical interactor, reaction substrate, etc.) and edge color shows the source of the interaction.

of edges encodes the database source of the interaction. Multiple edges with different styles denote that an entity has multiple roles in the interaction (e.g. in gene regulatory self-interactions, the protein product will also serve as a transcription factor in the inter-

action). Multiple edges of different color, on the other hand, show that the interaction is present in multiple source databases. Figure 2.6 shows as an example a connected interaction network comprising one gene regulatory interaction, three physical interactions, and two biochemical reactions, originating from different source databases. The depicted interactions involve one gene, several proteins and protein complexes, and one compound molecule. The visualization frameworks of the ConsensusPathDB web interface allow interactive operations on the displayed networks, such as interaction removal, node expansion, node location, etc. While the Java-applet-based framework requires a Java Runtime Environment to be installed on the client computer and has higher processor and RAM requirements to the client computer than a simple computer image, it has several advantages, especially when it comes to visualizing larger networks. Network nodes (physical entities/interactions) are movable and can be rearranged automatically using different layout methods. Network viewing is further facilitated through a zoom function. Most notably, in the Java-applet-based visualization environment, custom numerical values (e.g. gene/protein expression data) can be overlaid on the displayed network. The values are shown in a red-green color gradient on the according physical entity nodes. This feature aims to enable the visual interpretation of numerical data in the context of interaction sub-networks from ConsensusPathDB such as manually curated pathways or user-generated sub-networks. Any network displayed in the visualization frameworks of ConsensusPathDB can be downloaded in BioPAX format or as a computer image. Moreover, the protein interaction part of the ConsensusPathDB network is available for download through the web page in PSI-MI and tab-delimited formats.

Apart from interaction querying and visualization, the web interface offers the possibility to verify pathway models and extend them in the context of the ConsensusPathDB content. Users can upload interaction networks in BioPAX, PSI-MI or SBML formats. Upon upload, the interactions are matched to the content of the meta-database, and are displayed along with their similar counterparts from the integrated source databases. This aids the identification of spurious interactions in the uploaded models, and easily shows evidence for each interaction from the dozens of integrated databases. The uploaded interactions and entities are enriched with annotation from the meta-database such as publications, synonym names, and database identifiers. Notably, the network can be extended by expanding its physical entities with further

interactions from the integrated repository, and downloaded for use with other software.

In the primary focus of the web interface of ConsensusPathDB are tools for interaction- and pathway-based analysis of transcriptomics or proteomics data. Such data can be uploaded either as a summary list comprising e.g. differentially expressed genes, or in the form of numerical values for every measured gene/protein. Over-representation and enrichment analyses can be carried out with these data based on predefined pathways, sub-networks, and Gene Ontology (8) categories residing in the meta-database. The goal of these functionalities is to identify pathways and hot-spots in the integrated network which exhibit a changed activity in the phenotype of interest. Results can help to unveil the molecular mechanisms leading to these phenotypes and to suggest novel phenotype-associated genes. The underlying approaches are detailed in Chapter 4.

2.4.2 The ConsensusPathDB plugin for Cytoscape

As mentioned in the previous chapter, protein-protein interactions can already be detected on a large scale, owing to the development of a multitude of biological and computational techniques for this purpose. After generating a network of detected or predicted interactions, one usually faces the task to collect evidence for every interaction from the literature, and to identify interactions that have not been published previously. This information is useful in order to estimate the performance of the interaction screen, and to assess the contribution of its results toward the completion of the protein-protein interaction map of the species under study. To accomplish this task, one typically has to search the new data against every single protein-protein interaction repository. Even more tedious is the manual mining for interactions in the scientific literature in order to collect the publication references and different detection methods for each detected or predicted interaction. Apart from that, the number of publications reporting an interaction is an often desired interaction attribute when dealing with protein-protein interaction networks, since it is a direct evidence for interaction veracity (114). To aid the process of interaction evidence mining, we have developed a ConsensusPathDB plugin for Cytoscape (122). Our plugin searches all interactions from a network loaded in Cytoscape against the interaction space of ConsensusPathDB through dedicated web services. Interactions that are not present in any of the integrated resources are highlighted, since they constitute either novel or false positive interaction predictions, likely

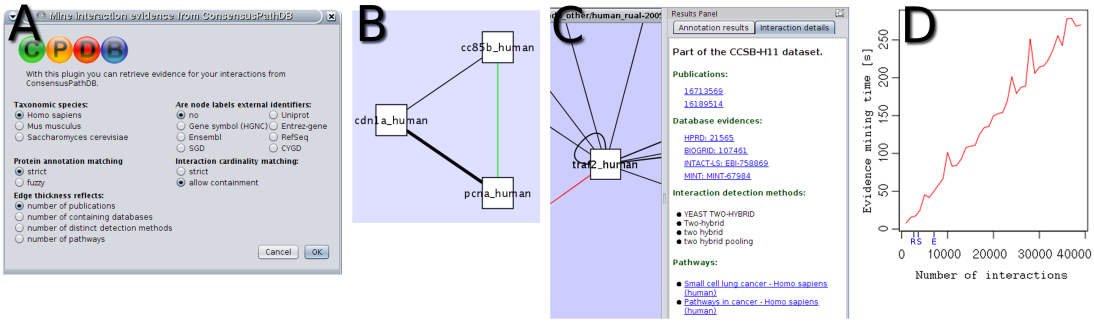

Figure 2.7: Overview of the funtionality of the ConsensusPathDB plugin for Cytoscape. **A** Menu screen of the plugin; **B** the ConsensusPathDB custom visual style where interactions with database evidence are black and weighted by the number of publications, while novel interactions are shown in green; **C** newly imported attributes of a selected interaction are shown in the 'Interaction details' tab of Cytoscape's results panel; **D** evidence mining time plot for networks of different size with default parameters (for this performance assessment, all query interactions were present in ConsensusPathDB such that the mining process took maximal time). The sizes of the networks predicted using large-scale interaction screening by Rual et al. (133) [R], Stelzl et al. (152) [S] and Ewing et al. (50) [E] are marked on the x-axis for a reference.

necessitating confirmation with complementary techniques. For the rest of the interactions, literature evidence (in the form of PubMed identifiers), interaction detection methods, interaction database references, and pathway co-ocurrence of interactors are extracted from our meta-database and can be viewed in Cytoscape. From there, these data can be exported as interaction attribute files. The plugin can be used also to spot interactions that have been missed in the screen (i.e., false negatives) by applying it on the complement of the interaction graph (which comprises all possible protein pairs that are not contained in the network). The ConsensusPathDB plugin-in is available through Cytoscape's plugin manager. Its functionality is summarized in Figure 2.7.

2.5 Discussion

Through the collation of dozens of publicly available interaction resources, we have created ConsensusPathDB: the most comprehensive interactome map available for human and for the model organisms mouse and yeast. Data integration enabled us to assess the similarities and differences between the separate resources. We found grave discrepancies regarding the interaction content of these resources even for well-studied

proteins and pathways. Our findings strongly advise against limiting to a single primary dataset in interaction- and pathway-based research, because the outcome of such analyses would be highly dependent on the particular interaction database employed. Integrated interaction data should be used instead, as they represent biological reality in a more comprehensive and unbiased way (35).

The interaction content of ConsensusPathDB can be used in many ways and contexts. 1) It offers a basis for analyses of the global and local topological properties of the human interactome. 2) It provides molecular models of biological processes for computational simulations. 3) It serves as a centralized repository for curated pathway models for pathway-driven analyses of expression data. 4) It can be used as a common interaction query interface for many databases. 5) It easily shows content overlap and discrepancies across databases, pointing molecular biologists to those best suited for their specific research, and helping database developers to spot and amend data errors. 6) Since it additionally includes many physical interactions between human and pathogenic proteins, it can serve as an explanatory basis for infectious diseases. There are many more application areas of ConsensusPathDB that are not mentioned here. In Chapter 4 we describe its applications in the context of gene expression data for identifying causative genes and interaction communities related with complex diseases such as cancer.

Although ConsensusPathDB contains several major types of direct interactions between biomolecules, there are further functional relation classes that are not yet integrated. An example are genetic interactions, referring to a phenomenon in which two or more mutations in different genes have an effect on the phenotype that is different than expected from the individual mutations (38). With the increasing generation of such data in human, a natural extension to ConsensusPathDB would be to integrate genetic interactions into the interactome map. Due to the generic design of the database, such an extension is in no way challenging. In fact, we have already integrated the DRYGIN yeast genetic interaction database (101) into the yeast instance of ConsensusPathDB that will be visible in the next database release.

Notably, all integrated interaction datasets are treated equally in ConsensusPathDB (that is, they are imported without any filtering), albeit in reality the separate datasets are of different quality. Due to the considerable manual curation efforts that have been applied to generate the currently integrated metabolic, signaling, and gene regulatory

interaction data, these data are much less error-prone than large-scale protein interaction data (72, 105). One way to deal with the high level of false positive protein-protein interactions is to consider the number of methods each interaction has been detected with, as suggested by von Mering *et al.* (114). Because literature evidences for interactions in ConsensusPathDB are assembled from many databases, their number is certainly a more reliable interaction confidence measure than the according numbers in the separate databases. However, in the next Chapter we argue that literature evidence is not an optimal criterion for interaction confidence because interactions with weaker literature evidence are not necessarily false. As a more elegant solution, we propose a novel interaction confidence scoring method. With that method we have calculated a confidence value for each binary protein-protein interaction in ConsensusPathDB. The resulting confidence scores are provided in the downloadable protein interaction data and are shown in the interaction visualization frameworks. The confidence score can be used as a criterion for interaction filtering, or can be treated as interaction weights by network-based approaches that are designed to deal with weighted network data.

Chapter 3

Cluster-based assessment of protein-protein interaction confidence

Protein-protein interaction data often contain a considerable amount of false positives originating from experimental or curation errors (41, 72). In this Chapter, we propose a novel method to assign confidence scores to interactions in a given network (91). Our method exploits solely the structure of interaction networks to assess the confidence of their individual interacitons and does not require additional information on the network elements.

3.1 Introduction to protein-protein interaction confidence assessment

Accurate physical interaction networks are fundamental to answering questions about how the biochemical machinery of cells organizes matter, processes information, and carries out transformations to perform specific functions leading to various phenotypes (73, 151). Toward this goal, a number of experimental and computational techniques, some of which were mentioned in Chapter 1, have been devised and applied to map the interactions of human proteins (50, 79, 133, 152) and those of model organisms such as yeast (61, 84, 103, 160, 167, 183). Despite their incompleteness (72, 171), physical interaction networks already serve as a basis for numerous methods aiming to

elucidate biological processes in health and disease (67, 80, 146). Current interactome maps are contaminated with false positives that can make up a considerable portion of the data (41, 43, 85, 114, 171). False positive interactions dim the explanatory light of interaction networks and also decrease the predictive value of methods using such data. Even more gravely impacted are integrated networks due to the much higher probability of overlap for true interactions than false positives from different datasets, which leads to an accumulation of false interactions in integrated data. It is thus of primary importance to derive confidence values for individual interactions, which can serve to refine current interactome maps or can be used as interaction weights.

Von Mering *et al.* (114) showed that interactions detected with multiple methods are more likely to be true than those detected with a single method, which is why literature evidence is an often used criterion for interaction confidence. Nevertheless, interactions with weaker evidence (e.g. those detected with a single method) found in interaction databases are not necessarily false: First, protein interaction detection techniques are barely comparable, and interactions consistently measured with one technique could be missed by another e.g. because the techniques tend to detect interactions with different stability. For instance, affinity purification combined with mass spectrometry captures interactions that are stable over time, while yeast-two-hybrid is able to detect more transient interactions. Second, the coverage of existing interactome screens is still limited (72), meaning that not all possible interactions have been tested even for well-studied organisms, and different large-scale studies usually test different subsets of the possible interactions. Third, a vast amount of the reported interactions are not even captured by database curators (35). This is also evidenced by the fact that databases mining interactions from the literature are mostly complementary (Table 2.2 in Chapter 2), suggesting that database curators tend to cover unique subsets of publications to extract interactions from. As a consequence, filtering out interactions with a weaker publication evidence is certainly sub-optimal as it would discard many true interactions that have been tested rarely, seen rarely by database curators, or are specific to a certain discovery technique.

Other strategies to validate protein-protein interaction data beside considering the literature evidence of interactions involve comparison of the interactions with reference datasets. For example, interactions between proteins that are often found together in

known protein complexes are more likely to be true. However, knowledge about protein complexes is still limited. Similarly, interactions between proteins that are known to participate in the same biological processes are more likely to be true, but unfortunately, pathway annotation is still lacking for many proteins (146). Several further approaches have been proposed for interaction confidence assessment, many of which are reviewed in (157) and (31). Most of these methods are meta-approaches that require the integration of additional data like interaction homology (43), co-expression of genes encoding interacting proteins (43, 44, 97), or a combination of these and other evidence features (12, 106, 145). While being certainly useful, such additional data are not always available, and may introduce additional bias and ambiguity since results depend on the particular data employed. Other methods do not require additional features and use network topology alone to predict interaction veracity (30, 66, 104, 135). Network topology-based methods are the tools of choice for interaction confidence assessment if other types of data are limited; moreover, topological features can be combined with other features to achieve better predictions. Topology-based methods are motivated by the fact that at various levels, the topology of interaction networks encodes biological properties which are largely independent of the biochemical function of the individual members of the network (5, 14, 18). Importantly, modularity of interaction networks is currently the most successful concept for addressing the dynamics of cellular processes (4, 54, 73).

Goldberg and Roth (66) proposed a network topology-based approach for interaction confidence assessment where the number of common neighbors of a pair of predicted interaction partners counts in support of the interaction. They defined interaction confidence as the level of enrichment of common network neighbors of interacting proteins. It is quantified by the hypergeometric distribution p-value given the number of common neighbors and total network neighbors of both interacting proteins. The underlying principle of the approach has been established in seminal studies demonstrating that biological networks are marked with short interaction paths separating random pairs of proteins in the network (small-world property), and densely connected local neighborhoods (neighborhood cohesiveness property) (153, 175). False positive protein-protein interactions are expected to violate the network cohesiveness property more frequently than true interactions. Recently, Kuchaiev *et al.* (104) proposed a different method that embeds interaction networks into a low-dimensional Euclidean space based on

network metrics (shortest path length) and then calculates confidence of interactions depending on the Euclidean distance between proteins within that space. The basis of the approach is the geometric graph model that was proposed to better reflect biological networks than e.g. the small-world model (74). Although its biological basis remains elusive, the authors argue that applying the geometric graph model to assess network distance should be consequently more reliable. Both of these topology-based methods assign confidence as numerical values to protein-protein interactions in a network and are additionally able to predict new interaction candidates by assigning confidence scores to non-interactions. However, both methods have certain shortcomings. The method by Goldberg and Roth is able to assess the confidence of those interactions whose participants have common neighbors only. Often, however, interacting proteins do not share neighbors. The method of Kuchaiev *et al.* appears limited in that it requires fixing six free parameters. For example, one of them is the prior probability for interactions. To calculate it, knowledge about the sizes of the proteome and interactome of the species under study is crucial. Unfortunately, even for well-stidied model organisms such as yeast, these quantities can still only be guessed at (71, 72). The rest of the parameters are algorithm-specific and barely have any biological motivation.

Here, we propose CAPPIC (cluster-based assessment of protein-protein interaction confidence) – a novel network topology-based approach that exploits the inherent modular structure of interactomes for confidence assessment of individual protein-protein interactions (91). Our method combines the basic principles of the topology-based methods described above: high neighborhood interconnectedness of a couple of proteins and short distance between them (the features exploited by Goldberg and Roth and Kuchaiev *et al.*, respectively) are indicators that both proteins participate in the same module. We apply Markov clustering (45) to the line graph (176) of an interaction network to dissect it into modules of interactions. As demonstrated in (123), this strategy can generate interaction clusters that significantly overlap with known biological pathways. Notably, the interaction clusters overlap in their protein constitution. This is biologically more meaningful than clustering the proteins into disjoint modules because pathways and protein machineries are known to overlap (61, 73). The rationale behind our approach is that proteins that are specific to certain modules are expected to have more interactions with proteins that are specific to the same modules than with other proteins (54). Intuitively, we assign low confidence to interactions that disagree

network property	Tarassov-all	Tarassov-hq	Yu-Ito-Uetz	Gavin-Krogan	CPDB-yeast	Costanzo
method	PCA	PCA	Y2H	AP-MS	multiple	genetic
node count	2238 (2293)	889 (1124)	1647 (2018)	2864 (2964)	6073 (6075)	4278 (4278)
link count	9360 (9646)	2407 (2770)	2518 (2930)	12006 (12068)	74332 (74333)	63927 (63927)
clustering coefficient	0.14	0.24	0.06	0.24	0.19	0.06
links in triangles	5861 (62%)	1761 (73%)	440 (17%)	8701 (72%)	63385 (85%)	47822 (74%)
mean shortest path length	3.7	5.6	5.6	4.3	2.7	2.9
links with ≥ 3 publications	546 (5%)	419 (17%)	782 (31%)	4090 (34%)	6324 (8%)	2546 (3%)*

Table 3.1: Yeast interactome maps used in this study for method evaluation. Interaction discovery methods: PCA: protein-fragment complementation; Y2H: yeast-two-hybrid; AP-MS: affinity purification coupled to mass spectrometry. The node and link counts correspond to the largest connected network component which is used for method evaluation; the according numbers of items in the complete network are given in brackets. The number of links in triangles corresponds to the number of interactions whose interaction partners share at least one network neighbor. *In the case of the Costanzo network, the number in the last row corresponds to the number of genetic interactions also reported in (36).

with the modular structure of biological networks and high confidence to those that comply with it. While the aim of CAPPIC is to detect false positive interactions, an approach based on the same idea of high link density within network modules has been proposed for identifying false negative interactions (182).

We applied CAPPIC and the methods by Goldberg and Roth and Kuchaiev *et al.* on a comprehensive benchmark of six interaction networks from yeast (Table 3.1) to assess and compare their performance. The six datasets were: 1) a network published by Tarassov *et al.* (160) that was generated using the protein-fragment complementation assay technology (Tarassov-all); 2) a sub-network of Tarassov-all obtained by the authors after applying several filtering steps (160) (Tarassov-hq); 3) a combined network of interactions found by yeast-two-hybrid screens (Yu-Ito-Uetz) comprising the networks published by Yu *et al.* (183), Ito *et al.* (84) and Uetz *et al.* (167) (retrieved from (183)); 4) a combined network of interactions detected by affinity purification coupled to mass spectrometry (Gavin-Krogan) published by Gavin *et al.* (61) and Krogan *et al.* (103) and downloaded from BioGRID (149); 5) a comprehensive physical interaction network from the interaction meta-database ConsensusPathDB, release 6(yeast) (92) obtained by the integration of multiple publicly accessible interaction repositories (CPDB-yeast); and 6) a genetic interaction map published by Costanzo *et al.* (38) obtained at a stringent experimental cutoff (Costanzo). The physical interaction networks constitute a representative benchmark since they result from different, most prevalent methods: yeast-two-hybrid, protein-fragment complementation, affinity purification,

and integration of interaction data from multiple methods. We applied our method additionally to the genetic interaction map by Costanzo *et al.* to provide evidence that it is not limited to physical interactome maps.

3.2 CAPPIC: A novel approach for interaction confidence assessment

3.2.1 Assessing protein interaction confidence by random walk interaction clustering

As mentioned previously, binary physical interaction data are usually modeled as graphs where nodes represent proteins or genes and edges represent interactions between them. For assessing the confidence of every interaction in such a network, we apply the following strategy (illustrated in Figure 3.1). First, the interaction graph is transformed into its line graph (176) where interactions are represented by nodes, and proteins are represented by links that connect their interactions (step 1 in Figure 3.1). In other words, while the original interaction graph is a network of proteins connected by their interactions, its line graph is a network of interactions connected by their shared proteins. Second, we employ Markov clustering – an algorithm for network clustering through random walk simulation (45) – on the line graph to dissect it into disjoint clusters of interactions (step 2 in Figure 3.1). In the third and last step of the approach (step 3 in Figure 3.1), we evaluate the distribution of interactions among the resulting clusters. It is a key point that interactions of a given protein can be clustered together, or distributed among multiple clusters. A protein is specific to a cluster if the cluster is enriched in interactions of that protein. We utilize the cumulative hypergeometric distribution to assess the enrichment of links of a given protein in a given interaction cluster. We define the fidelity $F_{p,c}$ of a protein p to cluster c as the value of the cumulative hypergeometric distribution function (Equation (3.1)) given $L_{p,c}$, the number of interactions of protein p in cluster c; $L_{p,\cdot}$, the total number of interactions of p (called the degree of p); $L_{\cdot,c}$, the total number of interactions in c; and $L_{\cdot,\cdot}$, the total number of interactions in the network:

$$F_{p,c} = P(X \leq L_{p,c}) = \sum_{k=0}^{L_{p,c}} \frac{\binom{L_{p,\cdot}}{k}\binom{L_{\cdot,\cdot}-L_{p,\cdot}}{L_{\cdot,c}-k}}{\binom{L_{\cdot,\cdot}}{L_{\cdot,c}}} \tag{3.1}$$

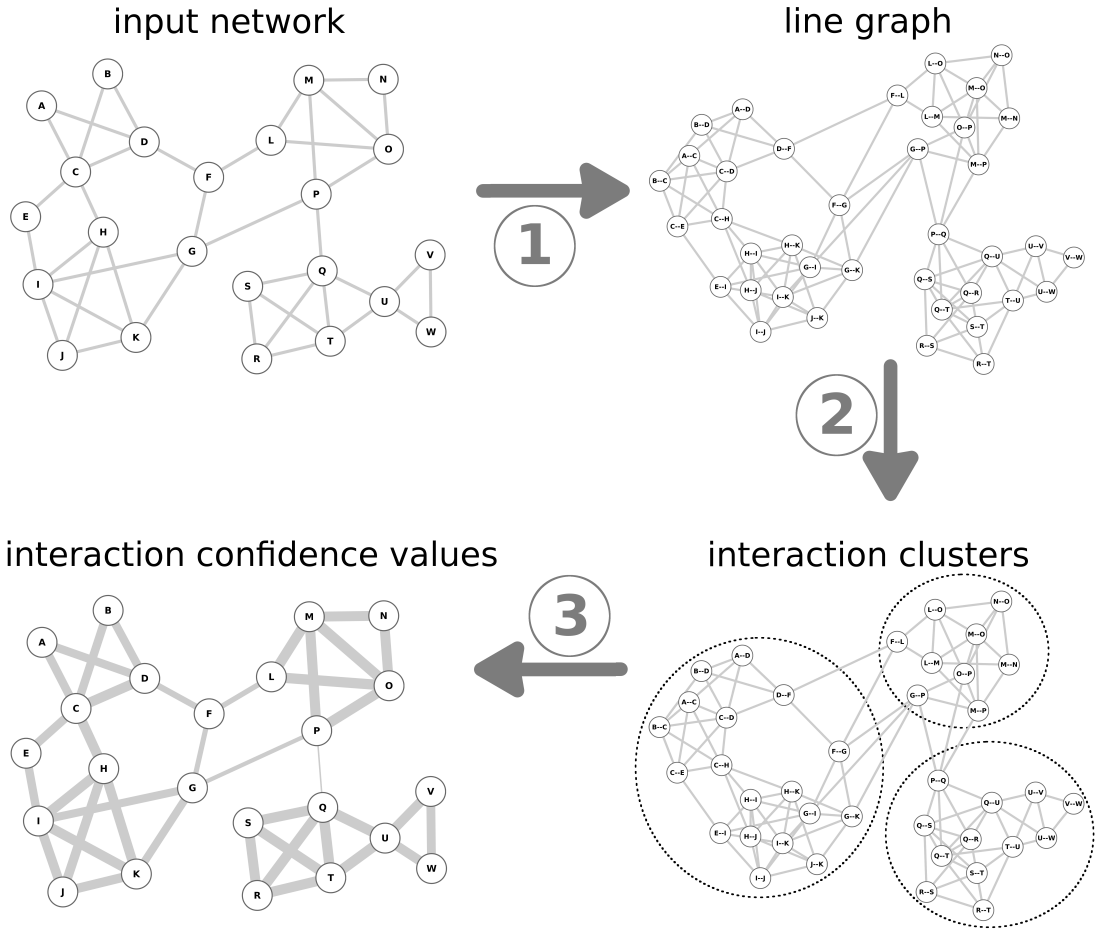

Figure 3.1: Outline of our interaction confidence assessment method. In the input interaction network (upper left picture), proteins are labeled with letters (A, B, etc.) and interactions between them are represented by edges. In the first step of the approach, we create the line graph of the given network where nodes represent interactions (labeled A–C, A–D, etc.) and edges represent shared interaction participants. In the second step, we use Markov clustering on this line graph to dissect it into interaction clusters. The clustering granularity is optimized in a previous step of the algorithm. Importantly, proteins can be part of more than one cluster. The relative number of interactions of a protein in a cluster determines how specific a protein is to that cluster. In the third step, we calculate confidence values for every interaction based on how specific both proteins are to the according clusters. The thickness of interaction links in the lower left picture corresponds to the calculated interaction confidence values for this example network.

The value of the fidelity $F_{p,c}$ lies between 0 and 1, with values near or equal to 1 if a protein p is specific to cluster c, i.e. if it has relatively many links in that cluster.

For a fixed $L_{p,c}$ it holds that the smaller the cluster (smaller $L_{\cdot,c}$), the greater the fidelity value – meaning that big clusters are less informative. Finally, if all the links of two proteins lie within a cluster, the fidelity is greater for the protein with the greater degree.

We define interaction confidence as the product of the fidelity values of both interacting proteins to the cluster c which the interaction has been assigned to:

$$\text{confidence}(l_{p_1,p_2}) = F_{p_1,c} \cdot F_{p_2,c} \tag{3.2}$$

Interactions get high confidence values if both proteins are specific to the cluster containing the interaction, and low confidence values when one or both of the proteins are not specific to the cluster.

3.2.2 Optimal clustering granularity is reliably determined through partial network rewiring

The interaction confidence scores calculated by CAPPIC are dependent on the granularity of the interaction clustering. It has been shown that modules in many complex networks, including protein interaction maps, are organized in a hierarchical manner (127). Accordingly, interaction clustering can yield protein complexes, cellular machineries (like the spliceosome), pathways, or higher-order biological processes depending on the clustering granularity. To estimate the granularity for a given network that will result in the best discrimination between true and false interactions, we first create an instance of that network where a small part of the interactions are randomly rewired to produce a set of false interactions. Our experiments have shown that rewiring 3% of the links is a good choice because this yields a false interaction set of reasonable size while keeping most of the network structure intact. In the rewiring procedure, pairs of interactions are selected at random and two of the proteins are swapped (so that no real interaction is reconstituted), thus generating two false interactions for two real ones while preserving each protein's degree. We additionally make sure that the network stays connected as a single component. Then, we apply our confidence scoring algorithm to this partially rewired network instance using different inflation values. The inflation parameter of the Markov clustering algorithm essentially controls clustering granularity (45). For every inflation value, we quantify the significance of the difference between the confidence score distributions of the rewired and the remaining

non-rewired links. This is done with the Wilcoxon rank-sum test under the alternative hypothesis that the confidence scores of the non-rewired links are greater than the confidence scores of the rewired links. The inflation value minimizing the Wilcoxon test p-value is considered optimal.

The inflation scan is carried out in two steps: a coarse scan with step size of 0.1 within a fixed range $I \in [1.1, 2.0]$ is followed by a fine scan with step size of 0.025 around the optimal inflation value ± 0.1 resulting from the coarse scan. In general, the inflation parameter takes values from the interval $I \in (1.0, 30.0]$ with higher values resulting in finer granularity. In all our experiments the Wilcoxon test p-value peaked at values far below 2.0, motivating the choice of this value as an upper boundary of the inflation scan.

Our experiments have shown that the estimated granularity value is robust to the introduced random rewiring as long as it is of reasonable extent: If the set of false interactions obtained through random rewiring is too small, the granularity estimation will lack statistical power, while if too many interactions are rewired, the network's original modular structure will be difficult to capture by the Markov clustering algorithm and the estimate may be biased. For all networks CAPPIC was applied on, random rewiring of 1%, 3%, 5%, or 10% of the interactions yielded identical or very similar optimal granularity estimates (data not shown). As mentioned above, we chose to rewire 3% in the inflation estimation step to ensure statistical power of the estimation while keeping most of the network intact.

The inflation estimation approach described above builds on the assumption that the optimal granularity value inferred from a partially rewired network instance (where both false positive and false negative rates are increased compared to the real network) is transferable to the real network. We aimed to scrutinize this reasoning and verified for all reference networks that 1) the estimated optimal granularity was rather independent of the random choice of links for rewiring; and 2) that interaction clusters were similar for the intact and the partially rewired networks clustered with the same inflation value.

To test the first hypothesis, we created 100 instances of each of the six reference networks (Table 3.1) where 3% of the links were randomly selected and rewired, and performed an inflation value search for each. For every instance and every inflation value, we calculated the Wilcoxon rank-sum test p-value reflecting the significance of the score difference between original and rewired interactions (optimality criterion).

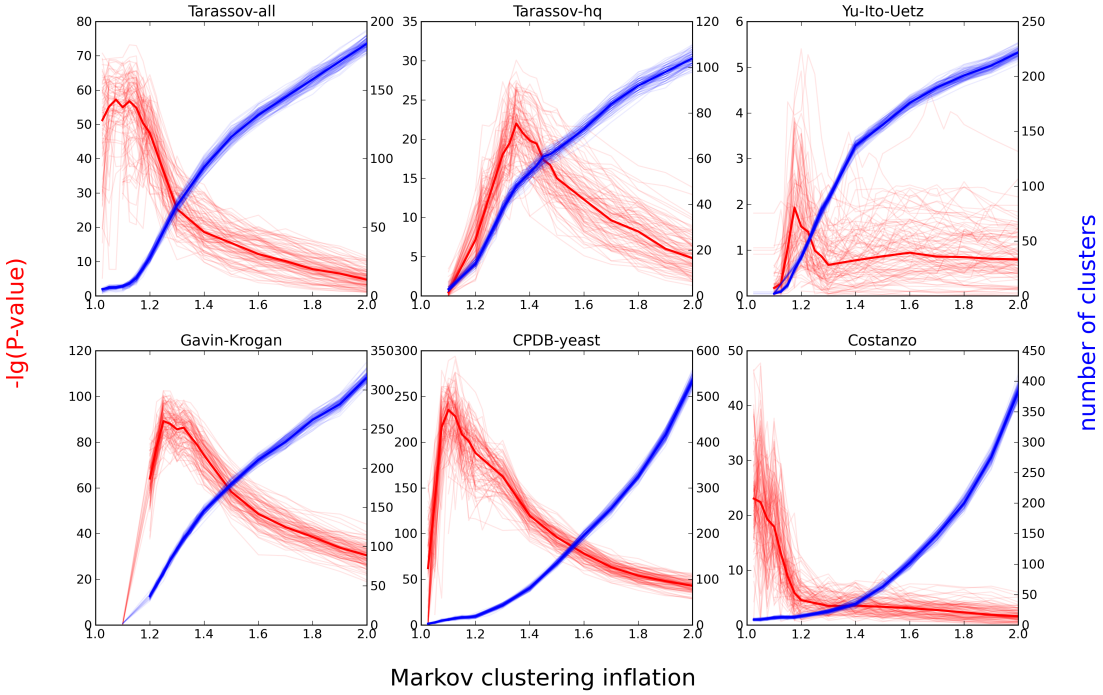

Figure 3.2: Estimating optimal granularity for clustering through partial random rewiring of input networks. 100 instances of every reference network were created where 3% of the links were randomly rewired. The negative common logarithm of the Wilcoxon rank sum test p-value reflecting the confidence score difference between rewired and non-rewired interactions (red curves, left-hand-side Y-axis) was calculated for each inflation value (X-axis). Moreover, the number of resulting clusters (blue curves, right-hand-side Y-axis) is plotted against varying inflation. Thick lines indicate the median values. We note that in the case of the Yu-Ito-Uetz network, the achieved p-value in the optimization step was one to two orders of magnitude higher than for the rest of the networks. Intuitively, the reliability of confidence scores calculated by our method can be appraised from the best achieved Wilcoxon rank sum test p-value in the inflation optimization step. If the overall performance of confidence scoring for a network is bad, then the score difference between random and real interactions in the optimization phase is less significant. However, these p-values are not suited for a strict comparison between networks.

The negative logarithm of the Wilcoxon test p-value and the number of clusters are plotted against varying inflation value in Figure 3.2. For all six networks, the 100 partially randomized instances were highly consistent regarding the estimated optimal inflation value. Figure 3.2 also shows that the number of clusters generated for the network instances did not vary much for any given inflation value within the inflation

search range.

To test the second hypothesis, namely that clusters have similar interaction composition for the intact and the partially rewired networks, we first clustered these networks with the same inflation value (resulting from granularity optimization). Based on the 100 partially rewired instances, we calculated an interaction co-clustering matrix $r_{i,j}$ which contained the relative frequencies that two interactions, i and j, end up in the same cluster for all partially rewired network instances where both interactions survive rewiring. We compared this matrix with the binary co-clustering matrix $c_{i,j}$ reflecting interaction co-clustering for the intact reference network. We defined a clustering agreement score to measure the agreement between $r_{i,j}$ and $c_{i,j}$:

$$\text{clustering agreement} = 1 - 2\frac{\sum_{\substack{i,j \\ i \neq j}} |r_{i,j} - c_{i,j}|}{\binom{L_{\cdot,\cdot}}{2}} \tag{3.3}$$

By definition, the clustering agreement equals 1, if and only if pairs of interactions that are co-clustered in the non-rewired case are also co-clustered in all rewired instances where both interactions have survived rewiring. The agreement value is around 0 if clusters in the non-rewired and rewired instances are completely independent from each other, and equals -1 if they are negatively correlated. Figure 3.3 shows the two co-clustering frequency matrices and their global mutual agreement for each reference network. In all six cases we found the cluster composition of the real network in high agreement with its partially randomized instances. We conclude that clusters are very similar for the original and the partially rewired networks clustered with the same inflation value. In other words, the link randomization we introduce to estimate the optimal granularity in the clustering step of the algorithm does not change the clustering result as such.

3.3 Comparative assessment of the performance of CAPPIC on yeast networks

3.3.1 True positive interactions are assigned higher confidence than false positives

We measured the performance of CAPPIC and compared it to the previously proposed network topology-based interaction confidence assessment methods ba Goldberg and

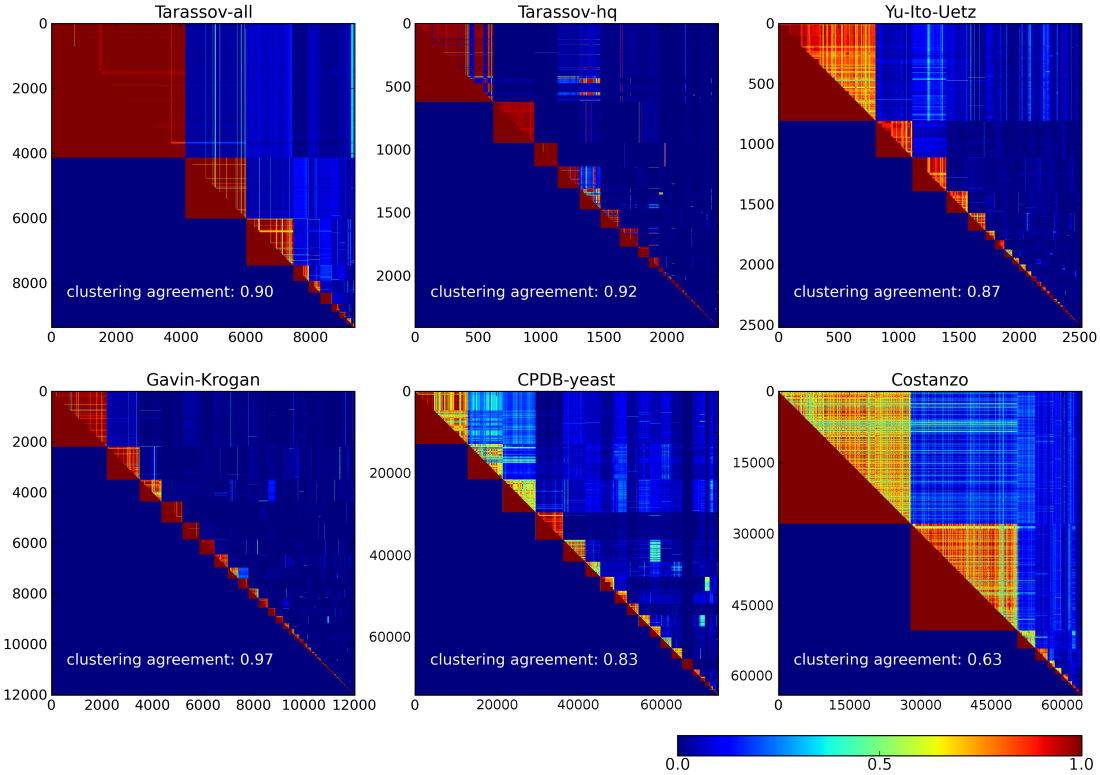

Figure 3.3: Interaction co-clustering matrices. For each reference network and its 100 partially rewired instances, we calculated interaction co-clustering matrices $r_{i,j}$ and $c_{i,j}$ for a fixed inflation corresponding to the estimated optimal value. This figure shows the co-clustering heatmaps for the non-rewired networks ($c_{i,j}$, below the diagonal) and the rewired instances ($r_{i,j}$, above the diagonal). Also provided is the overall agreement between both co-clustering matrices which, by definition, equals 1 if and only if pairs of interactions that are co-clustered in the non-rewired case are also co-clustered in all rewired instances where both interactions have survived rewiring. The agreement is around 0 if clusters in the non-rewired and rewired instances are completely independent from each other, and equals -1 if they are negatively correlated. For the six reference networks, the agreement ranges from 0.63 to 0.97.

Roth and Kuchaiev *et al.* using five yeast physical interaction networks and one yeast genetic interactome map, covering major interaction inference methods (Table 3.1).

We first constructed positive (literature interactions) and negative (random links) link sets and then evaluated the methods using receiver operating characteristic (ROC) analysis. The positive set for each network consisted of interactions that are reported at least three times in the literature (ranging from 3% to 34% for the six reference

networks, Table 1), since such interactions have been shown to be on average more reliable (114, 171). An exception was made for the Costanzo network because of the scarcity of genetic interaction data: the positive set in this case consisted of interactions that are also reported in (36). Literature evidences for all networks were retrieved with the interaction evidence mining ConsensusPathDB plugin (122) described in the previous Chapter. For each network, the negative interaction set for the ROC analysis was constructed by randomly rewiring a small subset (3%) of the interactions. For the partially rewired network networks we ranked interactions according to confidence as calculated with CAPPIC or reference methods and created receiver operating characteristic (ROC) curves.

The reference methods were applied as follows. We set the number of yeast genes to 6,000 in the method by Goldberg and Roth, which we implemented in R (82). The parameters of the method by Kuchaiev *et al.* (implementation downloaded from http://www.kuchaev.com/Denoising) were set as follows: priorEdge=0.002945 (which results when the estimated yeast interactome size of 53,000 interactions (72) is divided by the number of all possible protein pairs, $\binom{6000}{2} = 17997000$); priorNonEdge=1-priorEdge; dim=5 (default); d=3 (default); learnSetSize=min(5,000 or half the number of interactions); delta=1.0; and stopEps=0.01 (default). In the case of the Costanzo network, we set dim=3 because the program (run on a standard AMD X2 5600+ machine with 8GB of RAM running Matlab version 7.10.0.499 under Linux) did not return results within five days for a higher number of dimensions.

ROC curves were created by successively comparing the interactions ordered by confidence against the real positive (literature interactions) and real negative (random links) sets to determine the true positive and false positive rates at each step. The true positive rate is defined as the fraction of true positives from the real positives, while the false positive rate is the fraction of false positives from the real negatives. The performance of a given confidence assessment method in ranking positive interactions higher than negative ones was quantified with the area under the ROC curve (AUC). The AUC is around 0.5 if a method does not perform better than random interaction ranking, and is closer to 1 the better it ranks positive interactions higher than negative ones. Since the constitution of the negative and positive sets involves a random process (that is, the random selection of interactions for rewiring), we repeated the procedure 100 times and averaged the ROC results. In general, CAPPIC assigned higher confidence

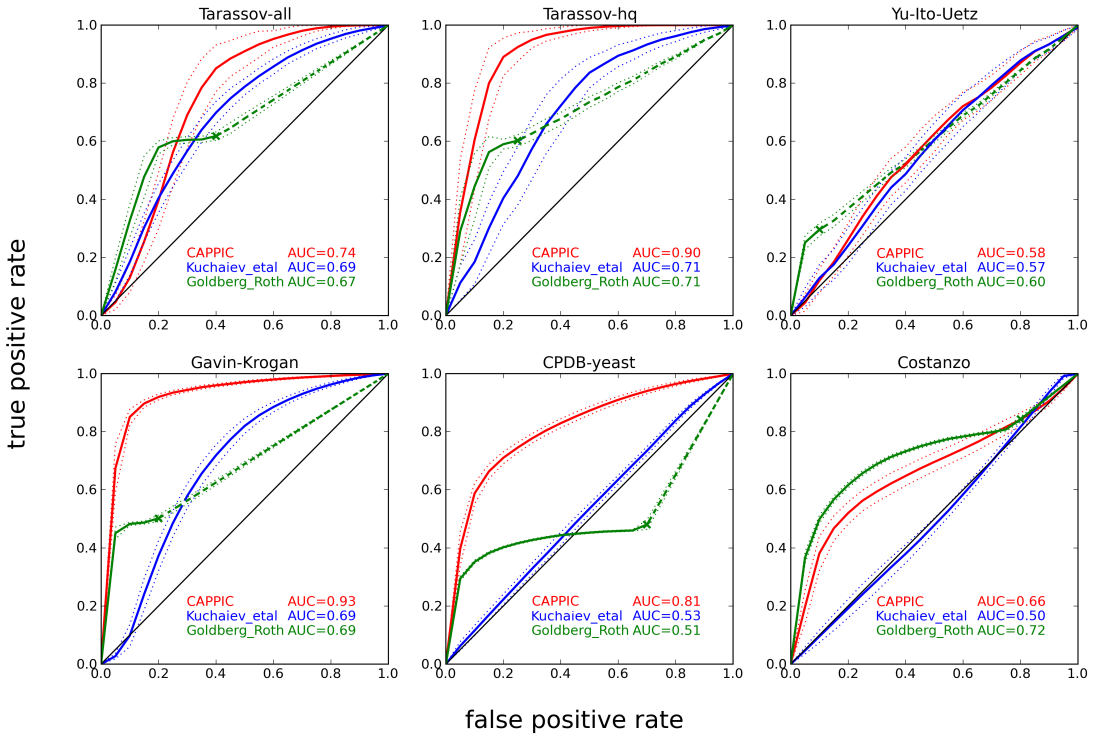

Figure 3.4: ROC analysis measuring the performance of CAPPIC in comparison to the methods by Goldberg and Roth and Kuchaiev *et al.*. False positive rate (1-specificity) is plotted against true positive rate (sensitivity) for each of the six reference networks. Since the definition of a negative interaction set in the performance assessment involves a random process, the ROC plots summarize the outcome of 100 runs. Plots show the average ROC curves (thick lines), their standard error bands (dotted lines), as well as the mean area under the ROC curve (AUC) of all runs. The 'X'-marks on the green ROC curves correspond to the fraction of true/false interactions whose proteins share network neighbors and are thus scored by Goldberg and Roth's method.

to true interactions than false interactions (Figure 3.4). The area under the ROC curve (AUC), which quantifies the confidence ranking performance, was as high as 93% for the Gavin-Krogan network. At a fixed specificity of 90% our method reached 86% sensitivity, outperforming the other topology-based methods. None of the methods in question showed convincing performance on the combined Y2H network Yu-Ito-Uetz. In this example, Goldberg and Roth's method successfully classified interactions whose proteins shared network neighbors; however, such interactions comprised only 17% of Yu-Ito-Uetz (see 'X'-mark on the green line in Figure 3.4 and row "links in triangles" in Table 1) while the rest of the interacting protein pairs did not share network neighbors.

Nevertheless, Goldberg and Roth's method outperformed CAPPIC by 6% AUC in the case of the Costanzo genetic interaction network, whereas the method by Kuchaiev *et al.* did not discriminate between true and false interactions better than random (AUC=50%). Based on the results for all six networks, we conclude that the method of Goldberg and Roth is suitable for identifying a relatively small set of high-confidence interactions but often does not provide predictions for a considerable fraction of the data. On the other hand, the method by Kuchaiev *et al.* and our approach generate confidence scores for the complete dataset, and are therefore more appropriate when the aim is to assess the confidence of all interactions or to filter out a relatively small sub-set of low-confidence interactions. In all cases, CAPPIC outperformed the method by Kuchaiev *et al.* in terms of AUC. It should be noted that in order to define a reliable negative link set, we destroyed some real interactions (increasing the false negative rate) and simultaneously introduced the same number of false positive interactions into the network, diminishing the biological signal in its structure. Thus, the AUC values reported here probably underestimate the real performance.

3.3.2 Cluster-based confidence scores corroborate experimental interaction evidence

To compare confidence values calculated by CAPPIC with experiment-based interaction scores, we exploited the fact that some of the interactions in Tarassov-all have been designated high-quality by the authors based on experimental interaction intensity (160). We checked whether our method assigned significantly higher confidence scores to high-quality interactions than to the rest of the interactions in Tarassov-all. As shown in Figure 3.5, the confidence score distributions of both interaction sub-sets were different. Using the Wilcoxon rank-sum test we confirmed that confidence values were greater for high-quality interactions than for the rest of the links in Tarassov-all (p-value $< 3 * 10^{-10}$). The high agreement between cluster-based interaction confidence scores and experimental interaction weight for the Tarassov-all network was corroborated by a significant Spearman rank correlation between both ($\rho = 0.3$, p-value $< 10^{-5}$).

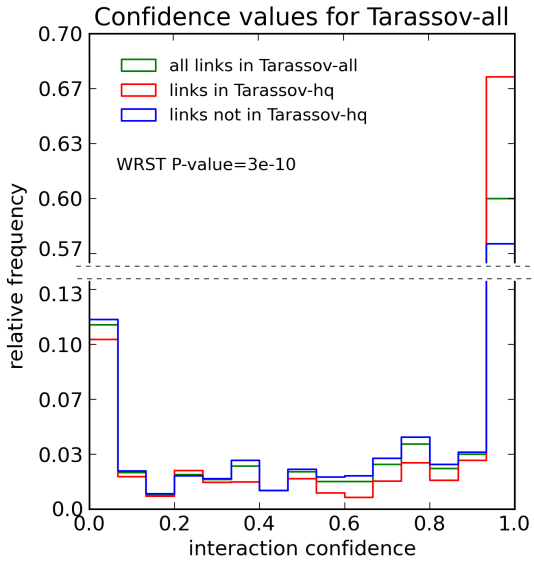

Figure 3.5: Histogram of confidence scores for interactions in Tarassov-all calculated by our method. The normalized histograms of interaction confidence scores are shown for the complete Tarassov-all network, as well as for its high-quality (Tarassov-hq) and non-high-quality parts. WRST: Wilcoxon rank sum test of the difference between confidence score distributions of both network parts. Note that the Y-axis is interrupted to better show the differences between the three datasets.

3.3.3 High-confidence interactions are more consistent in biological process and cellular compartment annotation

Interacting proteins are expected to participate in related biological processes and to be co-localized in compartments of the cell (116). Therefore, Gene Ontology (GO) (8) annotations of interacting proteins agree more often than expected by chance. We utilized the semantic similarity of GO biological process and cellular compartment annotations of proteins predicted to interact as a performance measure of our approach. If confidence values reflect the correctness of discovered interactions, we expect interactions with higher confidence score to have a higher average semantic similarity of the proteins' GO annotations.

To test this, we calculated the GO semantic similarity (GOSemSim) values for all interacting proteins in each network in respect to their biological process and cellular component annotations. This was done using the method proposed by Resnik (129) implemented in the software package GOSemSim version 1.8.0 (181). GO annotations

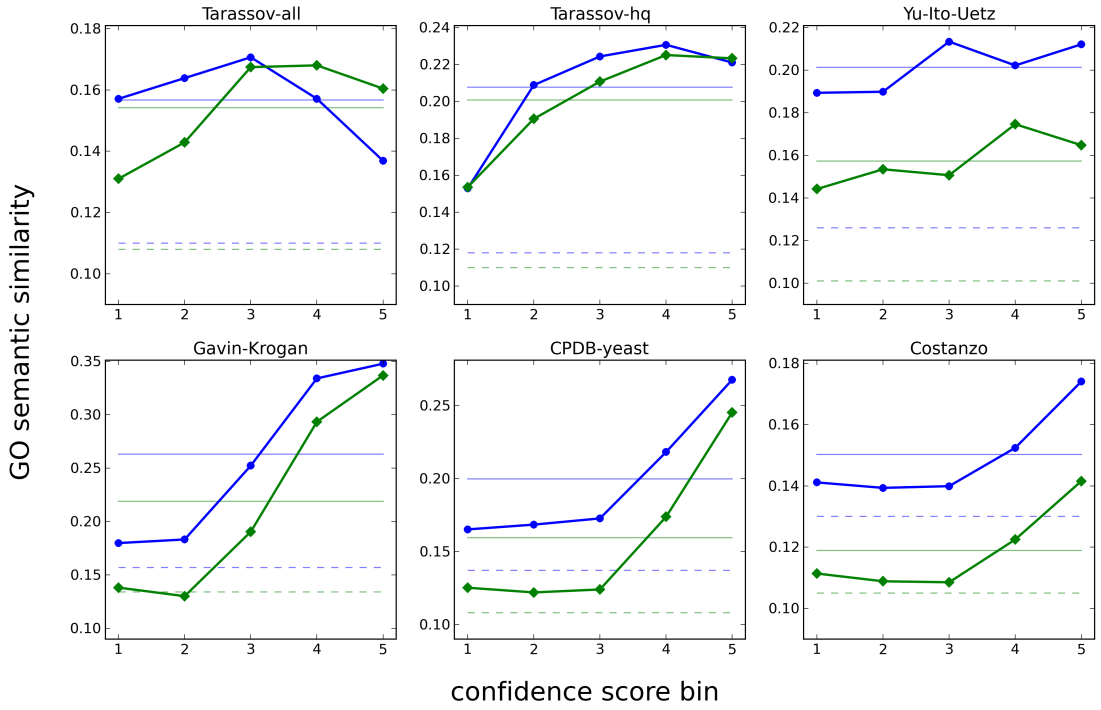

Figure 3.6: Correlation of CAPPIC interaction confidence with semantic similarity of Gene Ontology co-annotations. Interactions from every network are ranked by confidence and divided into five equal sized bins (X-axis); for each bin, the average semantic similarity of GO biological process (blue) and cellular component (green) annotations of interacting proteins is shown (Y-axis). Additionally, the pale continuous lines correspond to the mean GO semantic similarity over the complete network rather than the separate bins. The dashed lines reflect the average GO semantic similarity of random pairs of proteins from the network.

inferred from physical interaction (GO evidence code 'IPI') were excluded from the semantic similarity calculation to avoid circularity. For each network, interactions were ordered by increasing confidence score and divided into five equal sized bins. The average semantic similarity values for interacting proteins within each bin were calculated (Figure 3.6). Additionally, the mean GO semantic similarity for random pairs of proteins from the according network was assessed by completely rewiring the networks while preserving each protein's degree and then calculating the mean GO semantic similarity of links in those randomized networks (Figure 3.6, dashed lines). The GOSemSim generally correlated with interaction confidence. In several extreme cases (e.g. Gavin-Krogan), the average GOSemSim of low-confidence interactions was

barely distinguishable from the average GOSemSim of random protein pairs (dashed horizontal lines), while the higher-confidence interactions reached average GOSemSim far above the average value of all interactions in the according network (continuous horizontal lines). These results suggest that there are more false links among the lower-confidence interactions than among the higher-confidence ones.

Following this line of thought, we asked whether removing low-confidence interactions from clusters generated in our confidence assessment procedure would improve the consistency in pathway annotation of proteins remaining in the clusters. Pereira-Leal *et al.* have shown that Markov clustering applied to the line graph of a comprehensive interactome map yields clusters that are significantly consistent with KEGG (94) biological pathways (123). In the context of our method, low-confidence interactions are those involving proteins that are not specific to the according cluster, thus likely do not belong to its pathway context. We successively removed interactions ranked by confidence from clusters generated in the Markov clustering phase of our confidence assignment procedure. At each step, the resulting reduced interaction clusters were transformed into non-weighted lists of genes or proteins involved in interactions remaining in the cluster as in (123). We quantified the consistency of the gene or protein lists with KEGG pathways using the measure proposed in (123):

$$\text{consistency} = \sum_{j=1}^{C} \left(1 - \frac{-\sum_{s=1}^{n} p_{j,s} log_2 p_{j,s}}{log_2 n} \right) \tag{3.4}$$

In this measure based on Shannon's entropy (144), C is the number of interaction clusters, $p_{j,s}$ is the relative frequency of pathway s in cluster j, and n is the number of KEGG pathways. In general, the consistency value increases, the more homogeneous the clusters are regarding pathway classification of the contained proteins or genes. We found that the pathway annotation consistency of interaction clusters increased with the number of low-confidence interactions removed (Figure 3.7, continuous lines). Results were clearly different when the order of the removed interactions was reversed, i.e. when high-confidence interactions were removed first (dotted lines in Figure 3.7). This confirmed that lower-confidence protein-protein interactions do not fit in the pathway context of the according clusters as well as higher-confidence ones. Unlike the five physical interaction networks, in the case of the Costanzo genetic interaction map the consistency increased faster when interactions were removed from clusters starting with

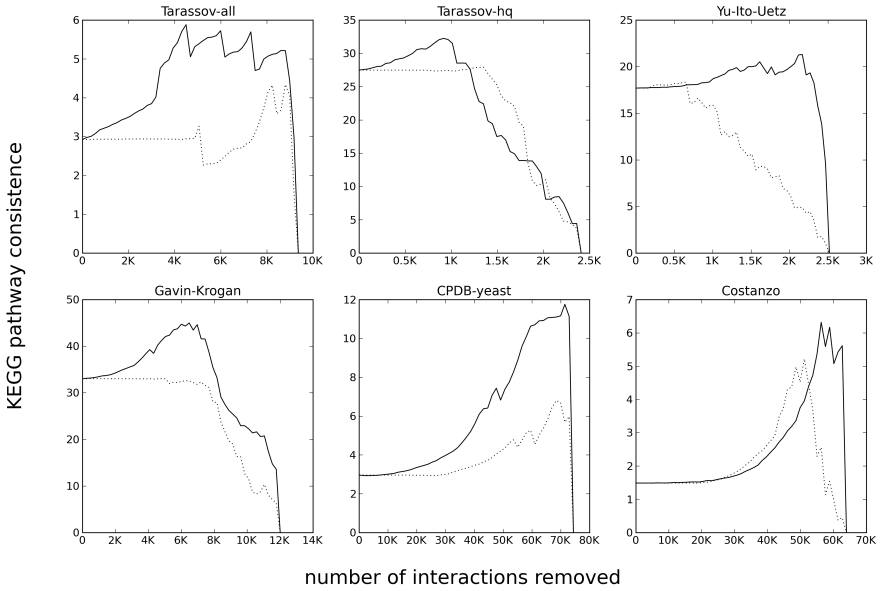

Figure 3.7: Interaction cluster refinement. Each reference network was transformed into its line graph and clustered with the estimated optimal inflation value for that network. Interactions were ranked according to confidence and successively removed from the according clusters. Pathway annotation consistency (Y-axis) was plotted against the number of interactions removed from interaction clusters (X-axis) starting with the low-confidence (continuous line) or high-confidence (dotted line) interactions.

the high-confidence interactions. The reason for this is probably rooted in the fact that most of the detected genetic interactions involve proteins in different pathways (between-pathway interactions) than proteins in the same pathway (within-pathway interactions) (70). Overall, the results suggest that our approach can be used to obtain more refined functional modules in physical interaction datasets.

3.3.4 Construction of a high-quality yeast physical interactome

We used CAPPIC confidence scores of interactions in the most comprehensive available yeast physical interaction network, CPDB-yeast, to derive a high-quality yeast physical interactome. The distribution of confidence scores for this network is shown in Figure 3.8 A). Evident from the results presented in the previous sections is that low-confidence interactions probably represent false positives, which was also confirmed by the small fraction of lower-confidence interactions reported more than twice in the literature (Figure 3.8 B). Based on the distributions of confidence scores and literature

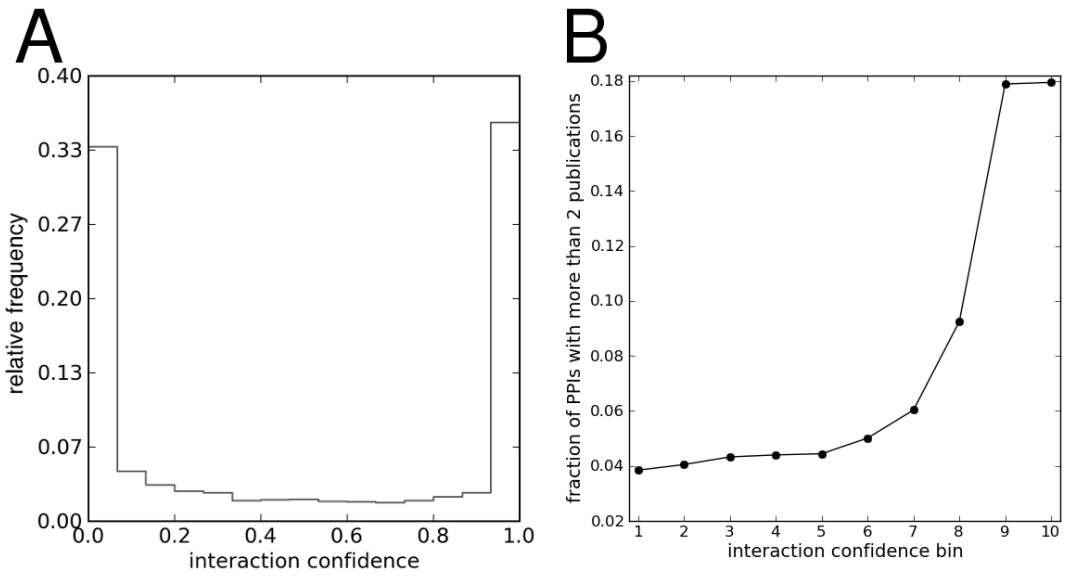

Figure 3.8: Confidence scores and literature evidence for the CPDB-yeast network. **A** Histogram of of CAPPIC confidence scores of the interactions in CPDB-yeast. Approximately 33% of the scores are very small (near 0.0), and roughly 35% are big (near 1.0). **B** Interactions of the CPDB-yeast network were ordered by CAPPIC confidence score and divided into ten bins of equal size. For each bin (x-axis), the fraction of the interactions reported in more than two publications is indicated (y-axis).

evidence (Figure 3.8 A and B), we selected the top 35% interactions with the highest confidence scores, as well as interactions found in more than two publications from the remaining 65%, to construct a high-quality yeast physical interactome. The resulting network contained 28,241 interactions between 3,779 proteins. Similar to CPDB-yeast, the high-quality interactome consisted of one large connected component and only 14 interactions were isolated. On the other hand, the high-quality interactome had a much higher average clustering coefficient than CPDB-yeast (0.33 for the high-quality interactome as opposed to 0.19 for CPDB-yeast), indicating a more pronounced modularity (175). Moreover, it possessed a longer average shortest path (4.1 for the high-quality interactome versus 2.7 for CPDB-yeast).

The high-quality interactome is available for download from the ConsensusPathDB-yeast web page (http://cpdb.molgen.mpg.de/YCPDB) beside the complete physical interactome. It should be noted that, apart from this filtered dataset, we did not use CAPPIC to discard any interactions integrated in ConsensusPathDB.

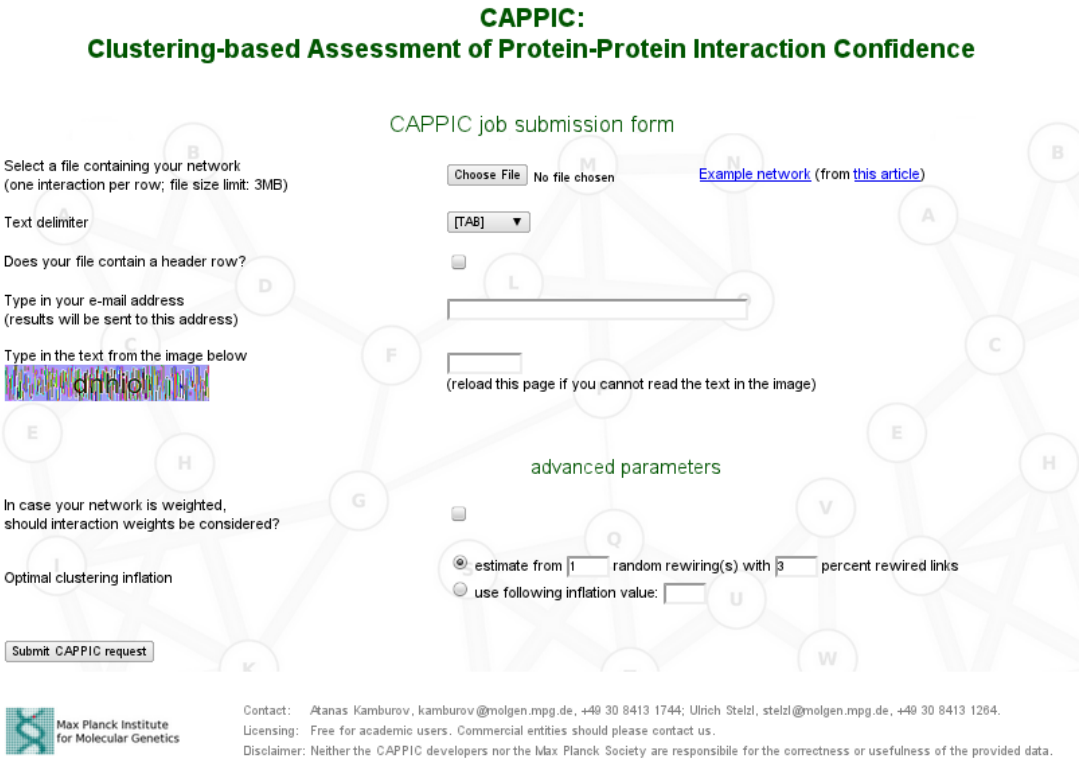

Figure 3.9: CAPPIC as a web tool at http://cpdb.molgen.mpg.de/cappic.

3.3.5 CAPPIC as a web tool for interaction confidence assessment

To provide easy and fast access to CAPPIC, we have implemented it as a web tool available at http://cpdb.molgen.mpg.de/cappic (Figure 3.9). Optional parameters enable influencing how the optimal granularity for the given network is estimated. The source code is also available from the web site.

3.4 Discussion

Network topology-based approaches are motivated by the fact that the structure of interaction networks is not random but reflects biological functionality (14). Modularity is a topological property that is inherent to protein-protein interaction networks (54, 61, 127). It is often exploited by graph clustering-based approaches aiming to find network modules reflecting pathways or protein complexes (10, 123). We propose a novel method (CAPPIC) to assess the confidence of individual protein interactions in

an interaction network. Our method exploits solely network modularity for estimating the confidence of interactions and does not require any additional knowledge about the interacting proteins. We demonstrate the power of CAPPIC in discriminating between true and false interactions on the basis of five physical protein interaction networks and one genetic interaction map.

CAPPIC outperforms previous topology-based approaches by Goldberg and Roth and Kuchaiev *et al.* in assigning continuous confidence scores to all interactions in a given physical interaction network. The method of Goldberg and Roth is dependent on shared network neighbors of interacting proteins; however, many interacting proteins do not share neighbors. Absence of shared network neighbors is especially prevalent for interactions in binary networks constructed using a bait-prey strategy, where links exist between baits and preys only. As a result, the method of Goldberg and Roth scores many interactions (83% of the interactions in the Yu-Ito-Uetz network, for example) with a confidence value of zero. This is a particularly strong drawback of that method, considering that many approaches operating on networks take as input probabilistic rather than binary data. Thus, the ultimate goal of confidence assessment is often to score all interactions in the network rather than disregard a large portion of them. In particular, all proteins with a single available interaction would be disregarded by Goldberg and Roth's method (as such proteins do not share neighbors with their interaction partner), albeit these single protein associations could give important clues about the function of these proteins. On the other hand, the method by Kuchaiev *et al.* is able to assign continuous scores to all interactions in a given network. Nevertheless, their method requires fixing six free parameters. These include e.g. the dimension of the Euclidean space for embedding, the prior probability for an interaction (which can currently only be guessed at because the interactome size of no species is known (72)), and several algorithm-specific parameters. In contrast, our method does not require any parameters or reference interaction sets. The only parameter that influences the resulting confidence scores – clustering granularity – is optimized internally for each individual input network. Our results have shown that the number of clusters obtained at the optimal granularity tends to be small for all reference networks, ranging from 10 to 60 clusters (see Figure 3.2). This condemned our initial concerns that interactions executing essential crosstalks between related pathways could be assigned low

confidence. Because the optimal granularity tends to be very coarse, closely related pathways will probably not be separated but clustered together.

CAPPIC should be applicable to any binary network (of biological or non-biological nature) with an inherent modular structure. CAPPIC fails to generate reliable confidence scores in cases where modularity is not pronounced, i.e. if many of the real links within modules are missing. This is probably the case with the Yu-Ito-Uetz reference network: here, the topological signal that our method exploits seems to be weaker, therefore it achieves only 58% AUC on average. Absence of modularity in this example is evidenced by the relatively low clustering coefficient (175) of 0.06 which is four times lower than that of the Gavin-Krogan network where CAPPIC achieves 93% AUC. Moreover, the Yu-Ito-Uetz dataset is the sparsest of all reference networks (Table 1). For the ConsensusPathDB-yeast network, which includes the rest of the reference networks used in this Chapter along with interactions from many further large-scale and small-scale experiments, our method performs well (AUC = 81%).

Due to the multiplicative nature of the interaction confidence definition, the method should be extendable (with an appropriate cardinality normalization) also to complex interaction data (i.e. non-binary interaction data). Furthermore, other mathematical functions instead of the cumulative hypergeometric distribution function can be applied for assessing the fidelity value of a protein to a cluster (which is used to derive interaction confidence). For example, in our initial experiments we defined fidelity as the number of interactions of the protein found in the cluster normalized either by the total number of interactions of the protein (that is, its degree), or by the maximum number of its interactions found together in any cluster. Also, we experimented with using as fidelity value the negative logarithm of the hypergeometric test p-value reflecting the enrichment of interactions of a protein in the cluster (that is, fidelity was defined in this experiment as $-\log(1 - F_{p,c})$ where $F_{p,c}$ corresponds to Equation (3.1)). None of these alternative fidelity definitions yielded better results in the method's performance assessment as per Section 3.3.1.

Unlike the reference methods by Goldberg and Roth and Kuchaiev *et al.*, CAPPIC is able to accommodate experimental evidence weights of interactions. Interaction detection techniques often associate such weights with predicted interactions, reflecting for example the number of times an interaction is observed in repetitions of a yeast-two-hybrid experiment (171, 183) or the reporter intensity value in the case of a protein-

fragment complementation assay (160). If available, such weights can be exploited by our method in its random walk-based interaction clustering step. This can improve the interaction clustering result and consequently increase the performance of confidence assessment. Moreover, the ability to incorporate experimental interaction weights helps to avoid interaction data pre-filtering, commonly executed to derive binary interaction networks (where pairs of proteins either interact or not). Such filtering of probabilistic interaction data is inherently associated with data loss. However, since we set out to estimate the performance of CAPPIC in comparison to other methods that cannot accommodate interaction weights, we did not make use of this advantage in this work and considered all interactions equal.

Although our approach alone is able to successfully rank true and false interactions and achieves up to 93% AUC on the reference interactomes, it can be combined with other lines of interaction evidence like protein co-expression and interaction homology (157). Aggregation of different features holds the promise of even more reliable interaction confidence assessment in specific contexts.

Chapter 4

Elucidating disease mechanisms with integrated interaction networks and expression data

Gene expression profiling is a powerful technique for measuring the activity of thousands of genes simultaneously, often applied to distinguish phenotype-specific gene signatures, i.e. lists of genes with a consistent activity change in a certain phenotype (for example a disease) compared to a control. A major concern, however, is that differentially expressed genes found in different studies analyzing the same phenotypic condition are barely overlapping (47). This is mainly attributed to the inherent variability of biological systems and of techniques for measuring gene expression. Significantly higher coherence between different studies is often found at the level of biochemical pathways where the distinguished genes function (33, 40, 80, 155). This finding has shifted analyses of expression data to a more pathway-centric perspective. This perspective can give more concrete hypotheses about the molecular mechanisms underlying the phenotype under study than simple lists of differentially expressed genes. The analysis of expression data at the level of interactions and pathways has proven useful for various purposes (discussed below). At this, integration of interactions and pathways is a crucial prior step, because it increases the coverage of the real interactome and thus enables more accurate predictions of methods based on these data. In this Chapter, we show the utility of the integrated and de-noised human interaction network from ConsensusPathDB in the context of expression data. We describe a simple approach to

identify phenotype-associated network hot-spots and causative genes. The approach is implemented as part of the expression data analysis module of the ConsensusPathDB web interface at http://cpdb.molgen.mpg.de. Furthermore, we present a method for the integrated analysis of large-scale transcriptomics/proteomics and metabolomics data at the level of known pathways (23), and introduce the first available computational tool for this purpose, accessible at http://impala.molgen.mpg.de (90).

4.1 Introduction: the benefits from integrating interaction and expression data

In the previous chapters, we emphasized that the high degree of organization of matter, information flow and energy transformations in the cell is reflected in the structure of interaction networks (14). Building on this, many approaches have been developed to extract network structures with specific topological properties from large-scale interaction data. For example, different methods exist for the detection of network modules: densely connected sub-networks which can be assigned a distinct biological function (73) from protein-protein interaction data (19, 141). In parallel, functional groups of genes are often searched in whole-genome gene expression data. This is routinely done by searching for genes with similar activation patterns with clustering techniques, with the presumption that co-regulated genes are often involved in the same biological processes (48). Motivated by the correlation between interaction and co-expression data (62), computational methods have been developed that integrate both data types (some are mentioned below). While both interaction and expression data are often incomplete and may contain large numbers of false positives, their integration should be beneficial because signals supported by both are more likely to be of biological relevance than those supported by either data type alone (62).

The integration of interaction and expression data has a successful history in a variety of contexts. One class of methods search for modules of genes corresponding to biochemical pathways or complexes supported by physical interaction and co-expression evidence simultaneously (69, 81, 143, 150, 168). Furthermore, the existence of disease-specific functional modules (65, 118) has motivated a group of related methods aimed at mining such disease-relevant modules directly from large-scale interaction data in conjunction with phenotype-associated gene expression data (32, 166, 169). Identified

modules have potential applications in molecular medicine as they have been shown to possess biomarker potential (32, 80, 162). Unlike the conventional, one-dimensional approach of selecting differentially expressed genes as disease biomarkers, network-based biomarkers constitute groups of interacting genes whose joint expression signature is discriminative for disease. Network-based biomarkers can achieve better classification accuracy and reproducibility across datasets than lists of discriminative genes (32). A further advantage over gene lists is that the identified networks can provide concrete hypotheses about the molecular mechanisms of disease in terms of interactions with altered activity (49).

While the methods cited above attempt to construct *de novo* disease modules from a whole-genome interaction network given expression data, a complementary strategy is to assess a priori defined functional gene sets to spot the ones showing an abnormal activity in a phenotype under study (40). Functional gene sets often correspond to the genes found in manually curated pathways, retrieved from pathway databases or the Gene Ontology (GO) (8). The key assumption here is that if a known pathway contains significantly many differentially expressed genes, or if the pathway genes show a jointly significant differential expression, then the pathway is dysregulated in the phenotype. Such pathways may be indicative or even causative of the phenotype, and have also been shown to possess biomarker potential (17, 165). Among the most popular approaches to identify phenotype-associated pathways are over-representation analysis (described in detail below) and gene set enrichment analysis (155).

A further research area where the integration of interaction and expression data has proven useful is the identification of genes causative of complex diseases (55, 96, 113, 158, 173). The main assumption behind such methods, and the basis for our integrative approach described in the next section, is that complex diseases like cancer are often caused by mutations in one or a few genes and the biological signal initiated by these mutations is propagated from the causative genes through their interactions to provoke differential expression of downstream genes. While the differentially expressed genes are often secondary manifestations of disease rather than its cause and thus can vary strongly from patient to patient, they are expected to lie near the mutated genes in a network of interactions (32, 34, 60, 65, 118). Following this assumption, existing methods (e.g.(96, 158)) for the identification of causative genes usually attempt to find

small sets of genes that lie near the differentially expressed ones in a physical interaction network.

4.2 Network-based functional gene sets in aid of causative gene identification

As mentioned above, manually curated pathways from public pathway databases and ontologies are routinely used in pathway-based analyses of expression data aiming to highlight biological processes associated with a phenotype of interest. Unfortunately, currently available curated pathways face several problems: First, functional annotation is still lacking for nearly half of the human genes (146)(Figure 4.1 A). Second, the composition of a biochemical pathway is a matter of subjective judgment as pathway boundaries are generally unclear (169). Even pathways of the same name found in different databases rarely agree regarding their composition (see Figure 2.4 in Chapter 2). Third, there is a widely recognized research bias toward inferring pathways associated with certain diseases like cancer; thus, currently defined pathways are predominantly disease-related. For example, several databases like NetPath and InnateDB are focused only on disease-related signaling (Chapter 2). This research bias naturally causes that disease pathways preferentially appear in the results of pathway-driven gene expression data analyses, challenging reliability of the latter. Fourth, pathway databases typically contain process-specific pathways but often miss essential pathway crosstalks which are important for pathway coordination, and whose dysregulation may play an equally important role in disease onset and progression like the pathways themselves. For example, dysregulation of the crosstalk between Wnt and Notch signaling has been implicated in cancer (37).

The availability of genome-wide interaction networks enables the definition of functional gene sets based on network neighborhood. In principle, these functional sets overcome all hurdles of manually curated pathway definitions listed above. Network-based sets are motivated by the fact that interacting genes are likely to have similar functions, shown for instance in the case of physical interaction networks in (116, 146). For example, Sharan *et al.* demonstrated the correlation of interaction network distance (defined as shortest path length) with functional distance (defined as semantic similarity of Gene Ontology annotations) of proteins, pointing out in particular that the

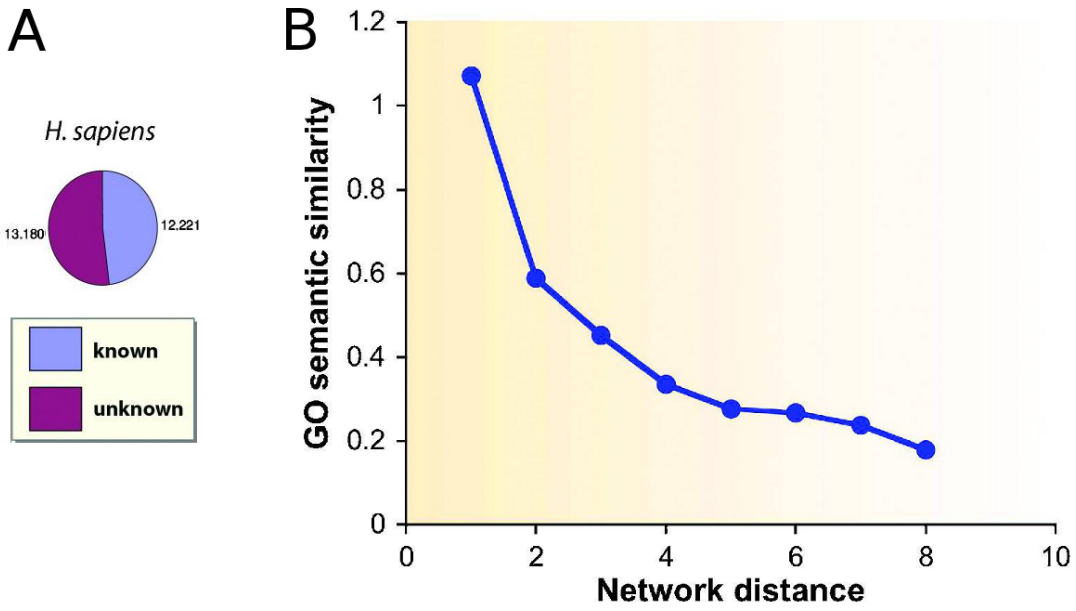

Figure 4.1: Pathway annotation of human genes and its relation with protein interactions. **A** Pie diagram showing the number of human genes with and without biological pathway annotation according to the Gene Ontology (GO). **B** Correlation of functional distance (quantified with GO semantic similarity) with interaction network distance for human proteins. Reproduced from (146).

direct neighbors of a protein in a physical interaction network often share its functions (146) (Figure 4.1 B).

We constructed network neighborhood-based functional gene sets from the integrated network content of ConsensusPathDB (Chapter 2), which was de-noised beforehand on the basis of cluster-based protein-protein interaction confidence (Chapter 3). These gene sets can be used in enrichment and over-representation analyses to highlight network hot-spots with an abnormal activity in a phenotype under study. The identified sub-networks can yield hypotheses about the mechanisms behind the phenotype in terms of disrupted interactions. Importantly, each neighborhood-based set per definition has a distinguished central gene (detailed below) and the central gene of a sub-network that is dysregulated in a phenotype is a more probable cause for the dysregulation. The reasoning behind this assumption is that a gene that is mutated does not necessarily show a change in expression, but its mutations often disturb its interactions with other genes and hence affect the expression of the interaction partners

(32, 34, 65, 118). Below, we detail the construction of network neighborhood-based gene sets from ConsensusPathDB's integrated content, and overview statistical methods that can be used to identify those which are dysregulated in a phenotype. Based on two example cases, we demonstrate that while being relatively simple, our approach is able to pinpoint known causative genes from cancer patient data.

4.2.1 Functional gene sets based on integrated network neighborhood (NESTs)

As the basis for construction of functional sets we used the integrated interaction data assembled in ConsensusPathDB from dozens of public interaction databases. Notably, the interaction network contains multiple types of interactions (gene regulations, signaling, catalysis, and physical interactions) of human genes/proteins. Prior to defining functional sets, we de-noised the binary physical protein-protein interaction content in ConsensusPathDB based on cluster-based interaction confidence and literature evidence as per Chapter 3. This was important because spurious interactions resulting from protein-protein interaction screens in principle accumulate in the meta-database and might diminish the predictive power of our approach. In contrast, the gene regulatory interactions and biochemical reactions currently contained in ConsensusPathDB have been primarily mined from the literature by experts, thus these data are expected to contain much less spurious interactions and do not necessitate filtering. The physical interaction content was de-noised by excluding 10% of the interactions with the lowest CAPPIC confidence and a single publication evidence. The procedure was analogous to Section 3.3.4 in Chapter 3; however, the confidence score and literature evidence thresholds were relaxed here as we aimed to retain all proteins in the network (whereas in Section 3.3.4, more than 1/3 of the proteins were removed at the chosen thresholds).

For every gene in the database we define a neighborhood-based entity set (NEST) including the gene itself and its network neighbors (Figure 4.2). More precisely, each NEST contains a gene as a center, as well as genes encoding proteins that interact physically with the products of the center, genes regulating or being regulated by the central gene transcriptionally, genes whose products participate in the same biochemical reaction as the products of the central gene, and genes encoding enzymes that catalyze or modulate successive biochemical reactions (in case that the central gene itself encodes an enzyme or a modulator). Two biochemical reactions are successive

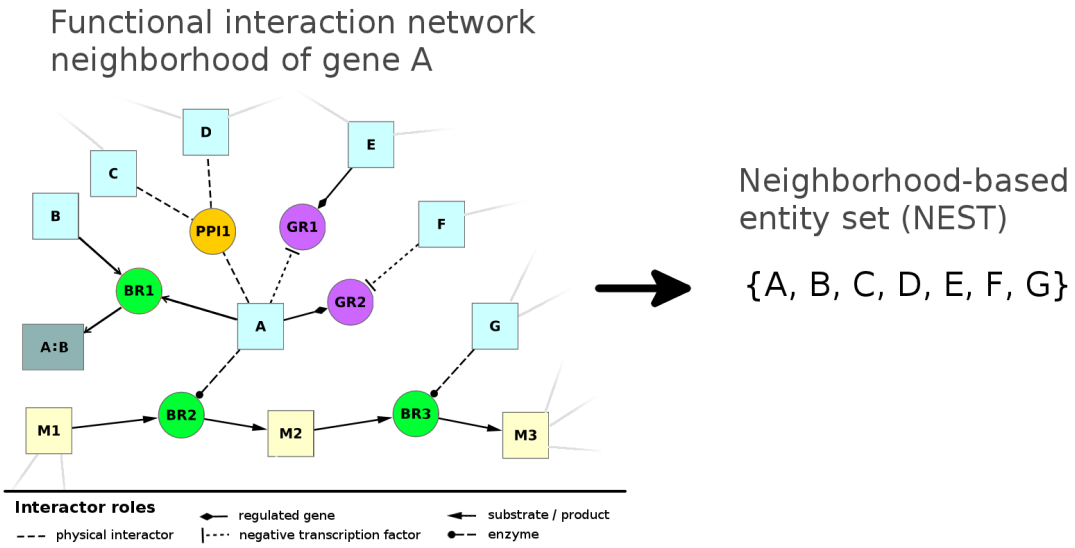

Figure 4.2: Construction of neighborhood-based entity sets (NESTs). Each NEST contains a central gene and all of its gene neighbors in the integrated interaction network. This example network comprises the interactions (circular nodes) of a gene A: one complex protein interaction (PPI1), two gene regulatory interactions (GR1, GR2), and three biochemical reactions (BR1: complex binding reaction, BR2 and BR3: metabolic reactions involving the metabolites M1, M2 and M3, catalyzed by A and G, respectively). Network neighborhood of genes is defined as either direct physical interaction of gene products, direct gene regulation (where the central gene is either the regulator or is being regulated), co-participation in a biochemical reaction, or catalysis of successive metabolic reactions (i.e. reactions sharing a non-hub metabolite).

if a product of one reaction is a substrate for the other (in Figure 4.2, BR2 and BR3 are successive reactions sharing the metabolite M2). Because many reactions are connected through non-specific metabolite hubs (for instance, ATP), we have constrained the definition of successive reactions to reactions sharing metabolites participating in five or less reactions from the whole network in total. NESTs with different centers and identical gene composition are collapsed together, resulting in NESTs with more than one center. Based on the content of ConsensusPathDB (release 19), we have created 19,666 distinct NESTs with these definitions. The number of genes per NEST, being 87 on average (Figure 4.3 A shows the NEST size distribution), is comparable to the size of manually curated pathways. However, the vast majority of NESTs are not subsumed by such pathways: the fraction of pathway-annotated NEST members found within the same manually defined pathway is 0.46 on average (Figure 4.3 B shows the

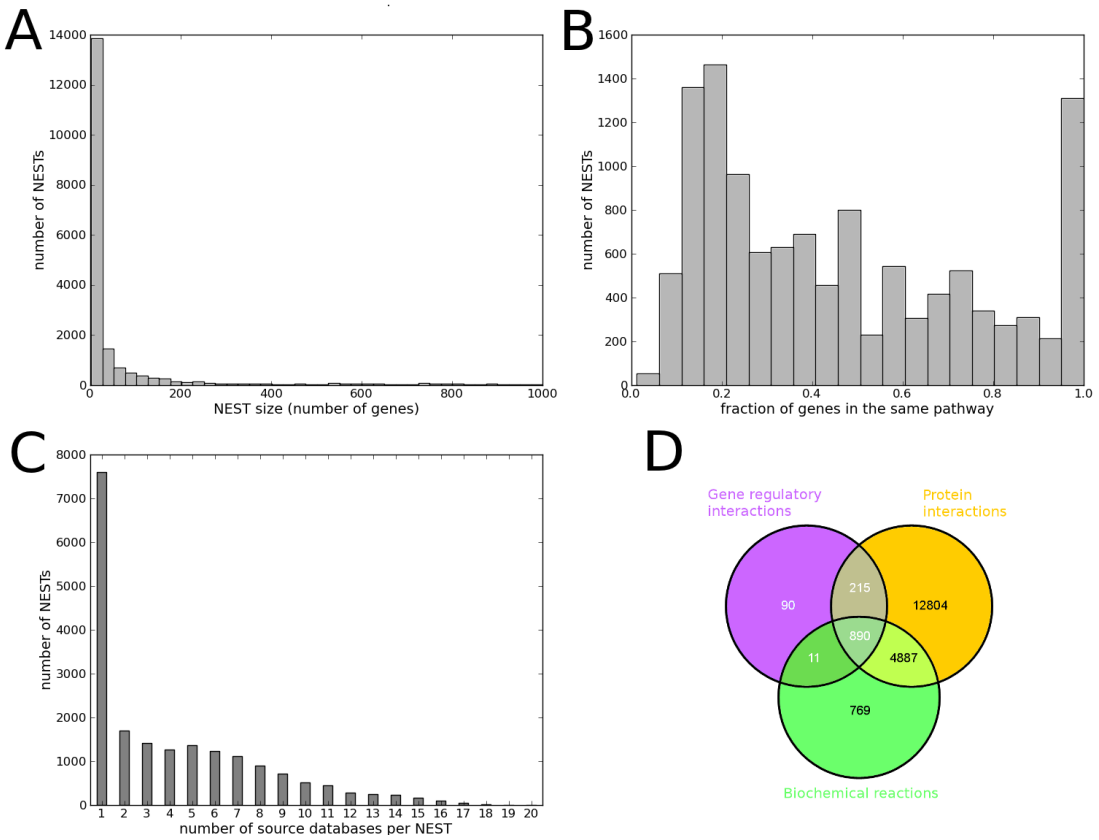

Figure 4.3: Characteristics of neighborhood-based entity sets. **A** Histogram of NEST size (number of genes per NEST); **B** Overlap between the gene compositions of NESTs and pathways; **C** Histogram of the number of different databases contributing interactions per NESTs; **D** Number of different types of interactions per NEST.

according distribution). For each NEST, this fraction was calculated as the size of the largest overlap with any pathway divided by the number of NEST members. This result means that many of the NESTs may represent pathway crosstalks. Most notably, in contrast to manually curated pathways, NESTs comprise the vast majority of human genes. Furthermore, a view on the number of sources per NEST reveals that in the majority of cases, more than one database contributes interactions for NEST composition (4.3 source databases per NEST on average, see Figure 4.3 C for the distribution). For instance, NESTs centered by SMAD4 or by members of the histone deacetylase family are composed with interaction data from as many as 20 source databases. Many NESTs are constructed from physical interactions only (Figure 4.3 D) because the currently available interaction knowledge is dominated by such interactions. This is

mainly due to the high throughput of protein interaction discovery techniques. 30% of the NESTs are contributed by both protein interaction and biochemical reaction data, while 5% include in addition gene regulatory relations, limited by the small number of gene regulatory interactions (2,270 interactions) compared to protein interactions (138,470 binary or complex interactions) in ConsensusPathDB.

4.2.2 Statistical approaches for identifying dysregulated NESTs

Given an expression dataset obtained e.g. by microarray-based or RNAseq-based profiling (139, 174) of a phenotype under study compared to a control, NESTs can be tested for differential activity using statistical methods. **Entity set over-representation analysis** is a classical approach used in gene set-based analysis to assess the significance of overlap between a predefined functional set, e.g. a NEST, and a custom list of genes that usually comprises the ones that show significant differential expression in the phenotype of interest (40). To quantify the significance of overlap, the hypergeometric test (identical to the one-tailed version of Fisher's exact test) is commonly used. Suppose that a NEST consists of n genes, the input set comprises m differentially expressed genes, and the background has N genes (Figure 4.4 A). The background typically comprises all genes whose expression has been measured and which are found in at least one NEST. The probability that exactly k entities from the input set are found by chance in the NEST is given by the probability mass function of the hypergeometric distribution:

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{4.1}$$

For an observed overlap of size K between a NEST and an input gene list, we rather aim to assess the probability that an overlap of this size or larger is obtained by chance. This probability corresponds to the hypergeometric test p-value for the observation K:

$$P(X \geq K) = 1 - P(X < K) = 1 - \sum_{k=0}^{K-1} \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{4.2}$$

The p-value is small for big overlap sizes K that are unlikely to appear by chance, supporting the alternative hypothesis that the overlap is caused by a biological effect. NESTs containing significantly many differentially expressed genes correspond to hotspots in the interaction network with altered activity between phenotypes.

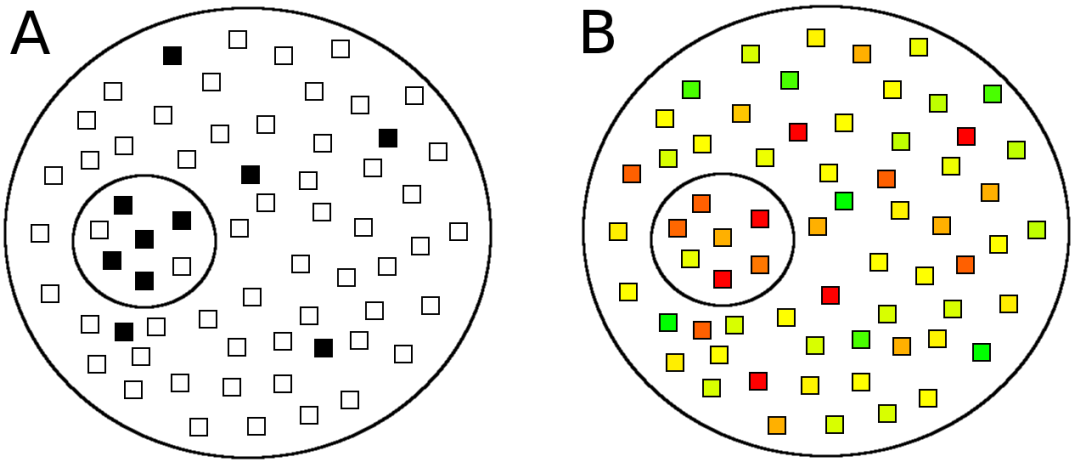

Figure 4.4: Over-representation and enrichment analyses. **A** and **B**: squares represent genes, the big ovals represent the gene background and the small ovals represent a functional gene set, e.g. a NEST. **A** For over-representation analysis, a relatively small set of genes (shown in black) is first to be distinguished from the expression data often. Conventionally, this set comprises the differentially expressed genes. Over-representation analysis then quantifies for each NEST whether it contains significantly many of these distinguished genes. **B** in contrast, enrichment analysis takes as input the complete measured set of genes rather than just the differentially expressed ones. Here, continuous values must be provided for every measured gene reflecting its expression level in the case and control phenotypes. The color of the squares in **B** scales with the change in expression of genes between the phenotypes (red: over-expression in the case compared to control; green: underexpression in the case compared to control; yellow: no expression change).

Over-representation analysis faces several practical problems associated with how the input set of genes is distinguished from the expression data. First, the list of differentially expressed genes is ambigouous and depends on the applied statistical test and the chosen significance level. Second, in order to assess the significance of differential expression with enough statistical power, repeated measurements per phenotype are necessary. Third, genes that pass the significance threshold are considered equally important for the phenotype under study, regarless of the magnitude of their expression change. The reason is that the hypergeometric test is a discrete test that cannot handle gene weights or ranks. These problems are overcome by **entity set enrichment analysis**. A key point here is that no decision is made *a priori* regarding which genes are differentially expressed and belong in the input set. Instead, enrichment analysis

takes as input all genes that have been measured in the case and control phenotypes with numerical values reflecting each gene's expression in both phenotypes. Different approaches can be applied to assess the enrichment of a functional set with up- or down-regulated genes, probably the most established one being gene set enrichment analysis (GSEA) (155). We utilize the paired Wilcoxon signed-rank test to assess the significance of joint differential expression of genes contained in a functional category such as a NEST. This test has been argued to be more suitable for enrichment analyses than e.g. Student's t-test, because its validity does not depend on a specific assumption about the distribution of expression values (e.g. Gaussian) (110). Accordingly, the Wilcoxon signed-rank test is more robust, in particular with respect to experimental outliers often found in biological measurements. Suppose that a NEST has n genes, for each of which a pair of expression values is provided. The NEST is thus represented as a set of pairs $(x_1, y_1), ..., (x_n, y_n)$, where x_k is the expression value for the k'th gene in the control phenotype and y_k is its expression value in the case phenotype. First, a vector z of expression differences is calculated such that $z_k = y_k - x_k$. Observations with no expression difference between the phenotypes, i.e. $z_k = 0$, are excluded so z has a possibly reduced size (denoted n_r) compared to the number of genes n in the NEST ($n_r \leq n$). The null hypothesis of the Wilcoxon signed-rank test is that the expression differences in the vector z are symmetric around a median of 0. To test it, the absolute values $|z_1|, ..., |z_{n_r}|$ are first sorted in ascending order and are assigned ranks such that the smallest absolute value in z gets the smallest rank $R_i = 1$. A mean rank is assigned to tied expression differences, i.e. where $|z_i| = |z_j| \neq 0$. The ranks of all $|z_k|$ where $z_k > 0$ are summed up to give R^+. Similarly, R^- is the sum of ranks of the values $|z_k|$ where $z_k < 0$. If the null hypothesis is true, then the values R^+ and R^- are expected to be similar. The Wilcoxon signed-rank test statistic is $S = min(R^+, R^-)$ and its critical value for rejecting the null hypothesis depends on the sample size n_r and the chosen confidence level. Exact p-values can be obtained from tables for small sample sizes n_r, while for bigger n_r, a normal approximation can be used because the test statistic S tends toward the Gaussian distribution. If the genes with the biggest expression difference in a NEST are overexpressed in the phenotype under study compared to the control, then R^- is very small (it equals zero if all genes in the NEST are over-expressed), and the NEST is likely to be phenotype-associated. Notably, even if no genes with individually significant differential expression are found in the NEST,

the joint expression of the group of genes within the NEST may be significantly increased or decreased. Thus, NEST enrichment analysis is able to identify interaction sub-network that are dysregulated on a low but nonetheless consistent gene level in the phenotype of interest.

It should be noted that Wilcoxon enrichment analysis is sensitive to pre-processing of the input expression data. For example, if the measured expression of all genes in one of the phenotypes is systematically higher or lower than in the other phenotype (e.g. due to experimental error), then many NESTs will be spuriously highlighted. To avoid this, expression values should be appropriately normalized such that that the expected gene expression differences between phenotypes is zero. The input expression values should also be logarithmized, in which case the Wilcoxon enrichment test deals with gene expression fold changes rather than absolute expression differences in assessing NEST de-regulation. This is generally advantageous because the dynamic range of expression activity varies strongly across the genome (63), thus absolute expression differences are barely comparable from gene to gene. As an example, transcription factors are usually found at very low concentrations in the cell and even subtle changes in their abundance often have a strong impact on the biology of the cell.

Because in over-representation and enrichment analyses many NESTs are typically tested for a given input, it is crucial to control for multiple comparisons in order to avoid a high false positive rate. Throughout our analyses we thus used the false discovery rate (FDR) method, defined as the expected proportion of falsely rejected null hypotheses (15). The FDR analogue of the p-value is commonly termed q-value.

4.2.3 Application 1: Network-based meta-analysis of prostate cancer pinpoints known causative genes

We carried out a comprehensive meta-analysis of prostate cancer patient data involving over-representation analysis of NESTs in order to unveil cancer causative genes.

Input dataset

To obtain an input list of genes that are commonly deregulated on the expression level in metastatic prostate cancer, we combined results from 9 studies (Table 4.1) providing a total of 11 datasets where samples from metastatic prostate cancer patients have

been compared against primary prostate carcinoma through microarray-based whole-genome expression profiling. All study results were retrieved from Oncomine 3.0 (130)

Study	PubMed ID
Dhanasekaran *et al.* (2005) FASEB J. **19**:243-5	15548588
Dhanasekaran *et al.* (2001) Nature **412**:822-6	11518967
Holzbeierlein *et al.* (2004) Am J Pathol. **164**:217-27	14695335
Lapointe *et al.* (2004) Proc Natl Acad Sci USA. **101**:811-6	14711987
LaTulippe *et al.* (2002) Cancer Res. **62**:4499-506	12154061
Tomlins *et al.* (2007) Nat Genet. **39**:41-51	17173048
Vanaja *et al.* (2003) Cancer Res. **63**:3877-82	12873976
Varambally *et al.* (2005) Cancer Cell. **8**:393-406	16286247
Yu *et al.* (2004) J Clin Oncol. **22**:2790-9	15254046

Table 4.1: Studies comparing whole-genome expression profiles of metastatic prostate cancer against primary prostate carcinoma. The studies by Tomlins *et al.* and Varambally *et al.* provide two different datasets each; the rest provide one dataset each.

where a p-value reflecting the significance of differential expression is provided for each measured gene in each study. From 19,500 genes measured in at least one dataset, 11,350 (58%) showed differential expression at a q-value threshold of 0.05 in one or more of the datasets. Not a single gene was found to be differentially expressed in all 11 datasets, and only five genes (CTGF, CYR61, MGP, PDLIM5, and PPP1R12B) showed differential expression in nine or ten of them (Figure 4.5). This result demonstrates the high variability of differentially expressed genes across different studies, which often hampers objective conclusions about the set of genes associated with cancer based on their expression value.

For NEST over-representation analysis, we used genes that were found to be differentially expressed at a confidence level $q < 0.05$ in more than half of the datasets in the analysis (i.e. at least 6 datasets). The input gene list consisted of 191 genes, termed DE genes (Table 4.2).

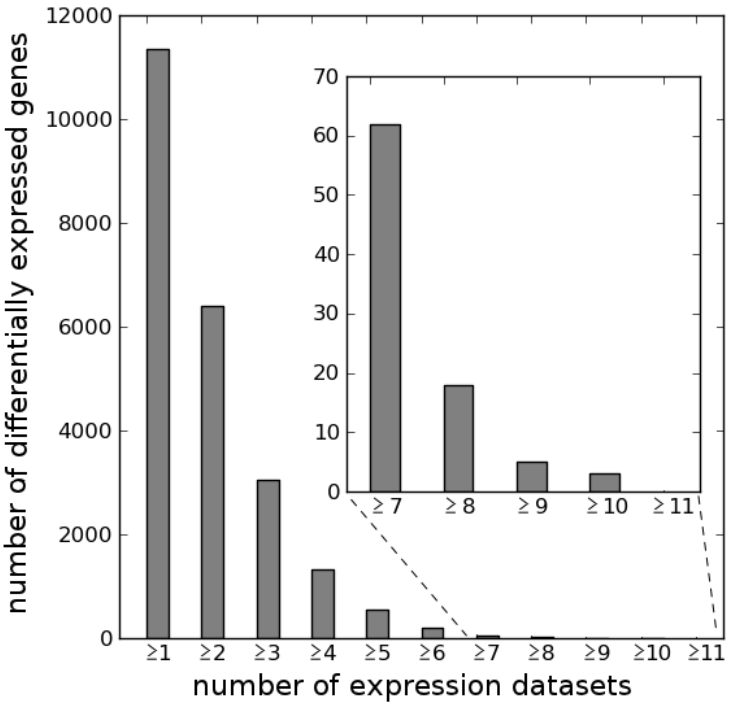

Figure 4.5: Agreement of different studies focused on the same phenotypes in respect of differentially expressed genes. The number of differentially expressed genes found in at least n of the 11 datasets from Table 4.1 is plotted against the number of datasets n.

NEST over-representation analysis results and discussion

With the DE gene list, NEST over-representation analysis highlighted 36 NESTs at an FDR level $q < 0.05$ (the NESTs are listed in Appendix Figure A.2). They yielded a total of 88 NEST center genes (termed NC genes, Table 4.3), because some of them were centered by gene families or had multiple centers as per construction (see Section 4.2.1). For example, one of the NESTs was centered by the group "cytokine receptor" comprising 39 genes, more than half of which were interleukin receptor genes [1].

We compared both the DE and the NC gene lists against the Cancer Gene Census (58). The Census catalogs 457 genes for which somatic or germline mutations have been causally implicated in cancer. We found that 11 out of the 191 DE genes (6%)

[1]The activity of the NEST centered by the "cytokine receptor" gene group points to inflammation, which is likely a secondary manifestation of cancer rather than a cause.

ABI1	ACOX1	ACTA2	ACTG2	ADAM10	ALCAM	ALDH1A2	AMOTL2	ASCL1
ATF3	ATP6V0C	ATXN3	**BMPR1A**	BTAF1	BTG2	CALD1	CALU	CCBL2
CCL2	CCNB1	**CCND2**	CCNF	CDC6	**CDH11**	**CDK6**	CHD2	CNN1
CNOT2	COX7A1	CSRP1	CTGF	CTNND1	CTSO	CYR61	DDIT4	DDR2
DIO2	DLG7	DST	DSTN	DUSP1	DUSP5	ECM2	EDNRA	EGR1
EGR2	EVI5	FAM189A2	FBLN1	FHL1	FILIP1L	FNDC3A	FOXK2	FUCA1
GABRE	GBP2	GLUD1	GOLGB1	GPD1L	GULP1	H2AFV	HEXB	HLA-DQA1
HLA-DQB1	IDE	IDS	IER2	IFNAR1	**IL6ST**	JUNB	KCNMB1	KIAA0101
KIFC1	KLF4	KLF9	LEPR	LGALS3	LMOD1	LRRFIP1	MCL1	MCM4
MED26	MEIS2	MFAP4	MGP	**MITF**	MMP7	**MYH11**	MYL9	MYLK
N4BP1	NAV1	NCKIPSD	NEAT1	**NFE2L2**	NR4A1	**NR4A3**	NT5C2	NTRK2
OSBPL8	PAGE4	PAM	PARM1	**PBX1**	PCP4	PCTK1	PDE4D	PDE5A
PDE8B	PDLIM3	PDLIM5	PELO	PGM3	PKN2	PLAGL1	PLEKHC1	PLN
PPAP2B	PPP1R12B	PPP2R1B	PRDX3	PRKACB	PRPF40A	PRPF4B	PSMA7	PTN
PTPRK	PTTG1	PYROXD1	RAB27B	RAB4A	RAD23B	RAP1A	RBM25	RBM3
RBM9	RBPMS	RCAN2	RNF141	RPS23	RPS6KB1	SELE	SFRS11	SFRS2B
SLAIN2	SLC22A3	SLC26A2	SLC2A10	SLC30A9	SLMAP	SNAP23	SOAT1	SON
SORBS1	SORBS2	SORL1	SPARCL1	SPG20	SPOP	SRD5A1	SRI	SSPN
ST13	STAT1	STC2	SYNPO2	TAGLN	TCEAL1	TCF7L2	**TFRC**	TK1
TNFSF10	TNPO1	TOP2A	TPM1	TPM2	TPX2	TSPYL1	UBE2C	UBE2J1
UBE2S	USP7	VCL	VPS39	YY1	ZFP36	ZFX	ZMYM4	ZMYND11
ZMYND8	ZNF354A							

Table 4.2: DE (differentially expressed) genes. The 191 genes listed here were found to be differentially expressed at a q-value threshold 0.05 in more than half of the prostate cancer studies (see Figure 4.5). The genes that have been causally implicated in cancer (as of the Cancer Gene Census, (58)) are underlined.

ACP5	ACTN4	ATF2	BMP2	**BMPR1A**	CCNF	CDC26	CDKN1A	**CDKN2A**
CNTFR	CSF2RA	CSF2RB	CSF3R	**CTNNB1**	CXCR1	CXCR2	EPHB2	EPOR
GHR	HCRT	HLA-DQA1	HLA-DQB1	**HOXA9**	HOXB8	IFNAR1	IFNAR2	IFNGR1
IFNGR2	IL10R	IL10RB	IL12RB	IL12RB2	IL13BP	IL13RA1	IL15RA	IL17RA
IL18R1	IL1R1	IL1RAP	IL1RB	IL2RA	IL2RB	IL2RG	IL3RA	IL4R
IL5RA	IL6R	**IL6ST**	IL7R	IL9R	ITGA1	ITGB5	**JAK1**	**JAK2**
JAK3	**JUN**	LEPR	**LIFR**	LMOD1	MAPK1	MAPK3	MGP	**MPL**
MYH11	MYL12B	MYL6B	MYL9	OSMR	PCNA	PRKG1	PRLR	PXN
SORBS1	SORBS3	STAT1	STAT2	STAT3	STAT4	STAT5A	STAT5B	STAT6
TLN1	TNFRSF10C	**TP53**	**TPM3**	**TPM4**	TYK2	VCL		

Table 4.3: NC (nest center) genes. These genes were found as centers of the 36 NESTs (Appendix Figure A.2), each of which contained significantly many of the DE genes from Table 4.2. The genes that have been causally implicated in cancer (as of the Cancer Gene Census, (58)) are underlined. For example TP53, whose mutations are known to cause cancer, was not contained in the differentially expressed gene list but was highlighted by NEST over-representation analysis because many of its network neighbors are differentially expressed. The interleukin receptor genes in the table (IL10R through IL9R) originate from the center of a single NEST: "cytokine receptors" gene group (see main text).

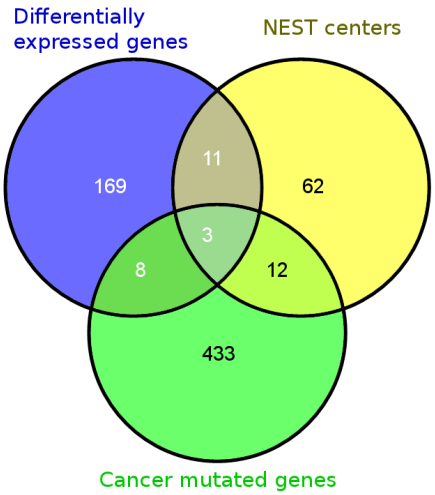

Figure 4.6: Overlap between differentially expressed (DE) genes, nest center (NC) genes, and the Cancer Gene Census. Cancer causative genes cataloged by the Cancer Gene Census are more over-represented in the NC set (p-value = 1.9e-9) than in the DE set (p-value = 5.5e-3).

and 15 out of the 88 NC genes (17%) are believed to contribute to cancer onset when mutated (Figure 4.6). The according genes are underlined in Tables 4.2 and 4.3. This means that in this example, genes connecting many differentially expressed counterparts by various interactions are roughly three times more probable to be causative of disease than the differentially expressed genes themselves. Only three disease causative genes are shared by the DE and the NC list, demonstrating the ability of NEST-based expression data analysis for finding new potential disease causes on top of the ones spotted through gene expression profiling. Notably, the protein p53 (TP53) whose mutations are known to cause cancer (75) was not found to be differentially expressed in any of the nine studies considered here, but it was identified as the center of a NEST in which differentially expressed genes were significantly over-represented. The 36 NESTs formed a connected network, supporting previous findings that disease-associated genes induce functional network modules (65).

A possible concern could be that because of the widely recognized research bias toward elucidating interactions of disease genes (80), NESTs centered by such genes would be preferentially highlighted. To address this potential issue, we assessed the expected number of known causative genes among NEST centers in a null model. We created 1,000 lists of randomly chosen genes of the same size as the DE list, and car-

ried out NEST over-representation analysis with each random list. Instead of selecting NESTs passing a fixed q-value threshold as in the analysis above (for almost all random lists, no NESTs passed the $q < 0.05$ threshold), for every random input list we selected the top 36 NESTs with the smallest q-value, and assessed the overlap of their centers with the Cancer Gene Census. The expected overlap between NC genes and the Cancer Gene Census estimated through this null model was 1.8 ± 1.6 (mean $\pm$ standard deviation) genes, resulting in a Z-score of 8.0 for the observed real overlap of 15 genes. The Z-score was defined as $Z = (K - \mu)/\sigma$ where K was the observed overlap, and μ and σ were the mean and standard deviation estimated from the null model, accordingly. In fact, the mean number of known causative genes in the NC lists in the null model μ was similar to the random expectation for the number of known cancer causative genes in the real NC list of 88 genes. Considering that the Census comprises 457 genes and there are 22,902 different entities found as NEST centers, 88 of which were highlighted in our analysis, the random expectation for the number of known cancer causative genes in the real NC list is $((457/22902) * 88 = 1.8)$. These results show that there is no recognizable effect of research bias on our NEST-based approach.

4.2.4 Application 2: NEST enrichment analysis with numerical data unveils cancer-related genes and highlights the hallmarks of cancer

As pointed out above, gene expression data can not always be summarized as lists of differentially expressed genes, for example because statistically sound conclusions about differential expression require several repeated measurements per phenotype. Often, the data consist solely of numerical values corresponding to gene expression measurements in a phenotype of interest and a in control phenotype. Here, we demonstrate the utility of the Wilcoxon enrichment analysis method in this common scenario.

Input dataset

Gene expression data were obtained from the study of Yu et al. (184) where the genome-wide gene expression of prostate carcinoma patients and metastatic prostate cancer patients has been measured with Affymetrix chips. The data were retrieved from Oncomine 3.0 (130), summarized in the form of normalized average gene expression

values for both patient cohorts. We additionally filtered the data to exclude expressed sequence tags (EST) that were not mapped to genes, as well as ambiguously identified genes. This resulted in a list of 7807 genes, for each of which the mean expression values for both patient cohorts were available.

NEST and pathway Wilcoxon enrichment analysis results and discussion

We tested for enrichment all NESTs from ConsensusPathDB release 16. Wilcoxon enrichment analysis yielded 57 significantly enriched NESTs at an FDR threshold of $q < 0.1$ (Appendix Table A.1). The most significantly enriched NEST (Wilcoxon signed-rank test p-value=8.34e-6; q-value=0.0483) had Histone H3-K9 methyltransferase 2 (gene symbol: SUV39H2) as the NEST center. It has been constructed from physical interaction and biochemical reaction information originating from overall nine different source databases. The central gene, SUV39H2, plays a role in cell cycle, transcriptional regulation and cell differentiation (Gene Ontology annotation, UniProt keywords) and its mutations have been shown to increase the risk of cancer in human and in mouse models (124, 180). It is important to mention that SUV39H2 itself has not been measured in the microarray experiment, and thus was not contained in the expression data set that we used for Wilcoxon enrichment analysis. However, many of the genes within its interaction neighborhood showed jointly significant transcriptional upregulation in metastatic prostate cancer compared to primary carcinoma. Figure 4.7 depicts the NEST as visualized by the ConsensusPathDB visualization framework, where the Yu *et al.* data were overlaid on protein nodes as logarithmized gene expression fold change.

Further significantly enriched neighborhood-based entity sets were centered by ribosomal proteins (e.g., RS4Y2_HUMAN, RS21_HUMAN, RL40_HUMAN, RL34_HUMAN) (in accordance with (170)), cell cycle proteins (e.g., CDK-activating kinase assembly factor MAT1: MAT1_HUMAN, Cyclin-H: CCNH_HUMAN, and MAP kinase p38 delta: MK13_HUMAN), and the transcription factor SP1 (SP1_HUMAN) which has been suggested to play a role in prostate cancer (185) (Appendix Table A.1).

We additionally tested for enrichment all manually created pathways from ConsensusPathDB (release 16), originating from overall nine pathway source databases. The pathways that were significantly enriched at a q-value threshold of 0.1 are provided in Appendix Table A.2. The results clearly corresponded to the hallmarks of

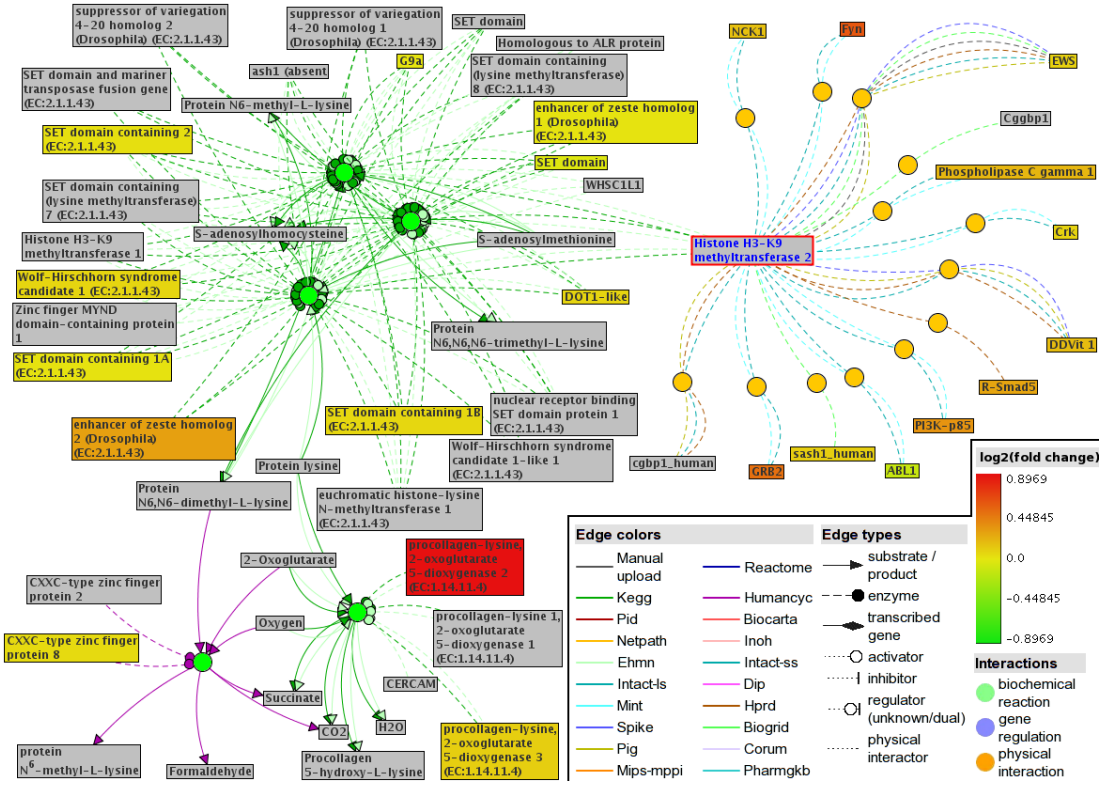

Figure 4.7: Neighborhood-based entity set (NEST) centered around SUV39H2 with gene/protein nodes colored according to expression fold change.. This interaction sub-network constitutes the NEST centered by SUV39H2 (Histone H3-K9 methyltransferase 2, highlighted with red frame in the network) and contains its direct physical interactors, as well as enzymes of successive biochemical reactions. The network consists of 13 physical interactions (orange circles) and five biochemical reactions (green circles) from nine different databases (interaction sources are encoded as edge colors). Gene expression data from (184) are overlaid as log2(fold change) values on the physical entity nodes (rectangles). Protein products of measured genes are colored according to the fold expression change (see legend), non-measured physical entities in the network are gray (note that the NEST center itself has not been measured).

human cancer (68) as they pointed to dysregulation of the cell cycle, transcription, translation, signaling, angiogenesis and immune response. For example, among the manually curated pathways whose activity is significantly changed in metastatic cancer compared to primary carcinoma were the “Ribosome pathway” (KEGG) (in agreement with (170)); “Translation” (Reactome); “Mitotic cell cycle” (Reactome); “Interleukin-5 immune pathway (IL5)” (NetPath); “VEGF, hypoxia and angiogenesis” (BioCarta); as

well as several cancer-related signaling pathways like "Signaling by GPCR" (Reactome); "PDGFR-beta signaling" (Pathway Interaction Database); "Signaling to ERKs" (Reactome); "Signaling to RAS" (Reactome); "JAK/STAT signaling" (INOH). Notably, KEGG's "Non-small cell lung cancer pathway" was also among the most enriched pathways.

4.3 Extending the pathway analysis paradigm: joint pathway analysis with transcriptomics and metabolomics data

Pathway analysis aiming to find biological processes whose executive genes are disturbed on the transcriptional level in certain phenotypes is an established technique despite the problems mentioned above. Importantly, gene expression is not the only aspect of the cell that may be altered as an effect of disease. Disease often impacts other vital processes as well, including the cell's metabolism. For instance, a classic hallmark of cancer cell metabolism is the Warburg effect (typical for proliferating cells): an increase in glucose uptake and glycolysis to lactate even under normal oxygen conditions. Furthermore, tumor cells are often found to exhibit higher rates of glutaminolysis, fatty acid and lipid metabolism, and nucleotide synthesis (23, 77). Motivated by the detectable impact of disease on metabolism, large-scale metabolomic techniques are increasingly applied to measure the whole metabolite repertoire of cells (76) to ultimately highlight metabolite biomarkers discriminative of disease (111, 172). Computational methods and tools for pathway-driven interpretation of large-scale metabolomic profiles (27, 178) are emerging in parallel to analogous utilities based on whole-genome expression profiles (78, 89, 156).

Since the cell is a complex system where gene expression and metabolism are highly coordinated not only within but also between each other, analyzing just one of these functional levels at a time is certainly sub-optimal for understanding the system's normal or abnormal functioning. With the increasing parallel generation of gene expression and metabolomics data for the same phenotypes, new methods and tools are urgently needed to allow integrated analysis of such data.

In a proof-of-principle study, we demonstrated that combining transcriptomic and metabolomic evidence for pathway association with a certain phenotype can aid path-

way biomarker discovery (23). Briefly, on the basis of a panel of 59 cell lines obtained from different types of cancer (140) we studied the associations between measured genes and metabolites and the resistance of the cells to platinum-based chemoterapeutics. The cell lines under study have been treated with four such chemoterapeutics that were carboplatin-, cisplatin-, iproplatin-, or tetraplatin-based. Our goal was to identify pathways relevant to general platinum sensitivity, as opposed to particular platinum compounds. Figure 4.8 shows a schematic outline of the study approach. As a first step, we derived a set of genes and a set of metabolites whose measured expression / concentration values were significantly correlated with sentisivity to carboplatin, cisplatin, iproplatin, or tetraplatin. For each of the four drugs, we carried out pathway over-representation analyses with the associated genes and metabolites separately. The pathways, originating from many public pathway databases, were retrieved from ConsensusPathDB. Based on the pathway over-representation analyses with genes, we identified four pathways that were coincidently over-represented at the chosen significance level for all four drugs, and thus were likely relevant to general platinum sensitivity (Figure 4.9 A). These pathways were "Rho GTPase cycle" (Reactome), "T cell receptor pathway" (NetPath), "Apoptotic dna-fragmentation and tissue homeostasis" (BioCarta), and "Integrin cell surface interactions" (Reactome). No pathways were coincident for all four drugs when over-representation analysis was performed with metabolites (Figure 4.9 B). Next, we integrated both lines (transcriptomic and metabolomic) of pathway-phenotype association evidence to identify further platinum resistance-related pathways. Essentially, we assumed that pathways highlighted when using transcriptomics or metabolomics data were independent because these data have been obtained with independent techniques. Thus, to integrate both data types at the pathway level we computed for each pathway a joint p-value $p_{i,J} = p_{i,G}p_{i,M}$ where $p_{i,G}$ and $p_{i,M}$ denote the over-representation p-values of the i^{th} pathway with respect to genes and metabolites correlated with drug chemosensitivity, respectively. The added value of the joint analysis compared to the separate gene-based and metabolite analyses was assessed through two null models, the first assuming that genes and metabolites identified as significantly associated to a phenotype were randomly selected, and the second null model assuming that pathways were selected randomly (see (23) for details). The combination of evidence for pathway association with drug resistance enabled the identification of six pathways generally related with resistance to platinum (Figure

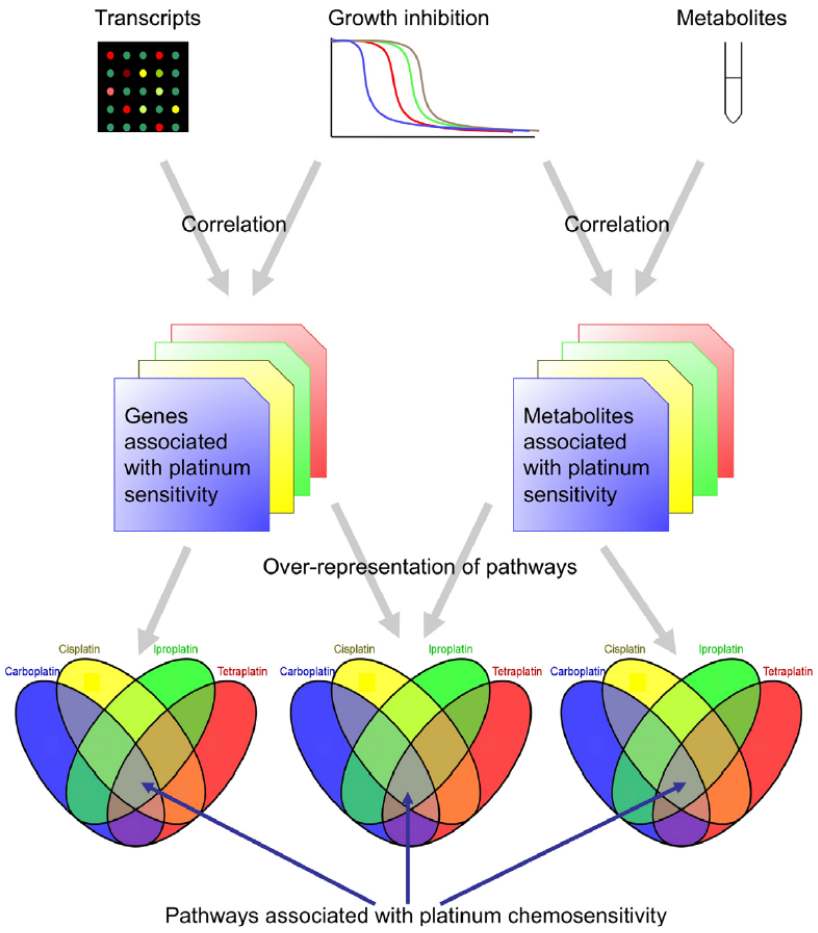

Figure 4.8: Pathway-level integration of transcript and metabolite data: a schematic overview of the study design. Large-scale gene expression, metabolomic, and drug sensitivity data obtained from the NCI60 tumor cell line panel (140) were used to distinguish genes and metabolites associated with chemosensitivity to four platinum-based cancer drugs (carboplatin, cisplatin, iproplatin or tetraplatin). For each of the four drugs, the lists of distinguished genes and metabolites were used separately and jointly for pathway over-representation analyses aiming to identify pathways associated with common chemosensitivity to platinum. Reproduced from (23).

4.9 C). For the two new candidate pathways emerging from the joint pathway analysis, "Triacylglyceride biosynthesis" (Reactome) and "Base excision repair" (Reactome), phenotype association evidence on either the transcriptomic or the metabolomic level was not significant enough; however, the two lines of evidence were agreeing and the joint p-value was significant. Details on this study can be found in (23).

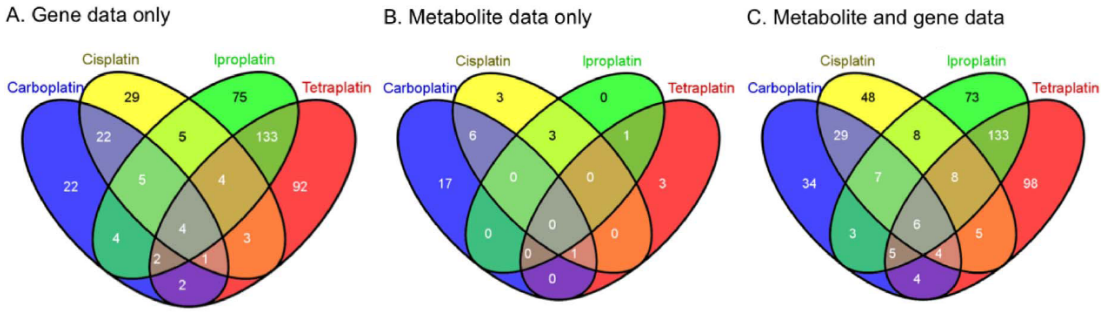

Figure 4.9: Pathways associated with platinum resistance based on transcriptomic, metabolomic, and combined evidence for phenotype association. Each Venn diagram shows the numbers of pathways related to the four drugs based on the transcriptomics data (**A**), metabolomics data (**B**), or joint metabolomics and transcriptomics data (**C**). Reproduced from (23).

To enable the scientific community to easily integrate transcriptomics/proteomics and metabolomics data at the pathway level, we developed a web tool called IMPaLA: integrated molecular pathway-level analysis (90) (http://impala.molgen.mpg.de; Figure 4.10). IMPaLA performs pathway over-representation analyses with lists of genes/proteins and metabolites (e.g. genes with differential expression and metabolites with significant concentration change in a certain phenotype), or Wilcoxon pathway enrichment analyses with numerical transcriptomics/proteomics and metabolite concentration data. As a source for predefined pathways, IMPaLA currently uses 11 freely available databases contributing over 3,000 manually curated pathways, most of which comprise both genes and metabolites. As in our proof-of-principle study, evidence of pathway association to the phenotype under study derived on the gene expression and metabolite concentration levels is combined, allowing for the identification of phenotype-associated pathways that would not be highlighted when analysis is applied to any of the separate functional levels alone.

4.4 Discussion

The identification of causative genes and pathways governing disease onset and progression is one of the major problems in contemporary molecular biology. Toward this goal, different approaches have been devised that make use of genome-wide interaction

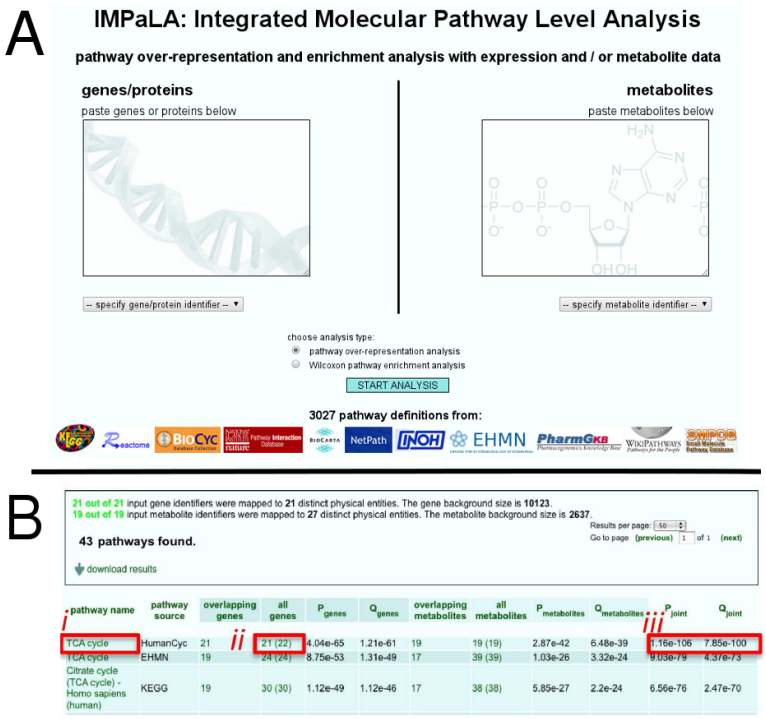

Figure 4.10: IMPaLA: a web tool for integrated pathway-level analysis of transcriptomics and metabolomics data. **A** IMPaLA input screen including the logos of the 11 pathway source databases. **B** Output screen with a ranked list of pathways showing i.a.: *i*: a link to each pathway in the according source database; *ii*: the size of each pathway in terms of entities also present in the background list, followed by the number of all pathway entities as in the source database; *iii* the p- and q-values from the joint analysis with genes and metabolites.

networks or curated pathways integrated with large-scale transcriptomics, proteomics, or metabolomics data.

We devised a simple approach to put forward the added value of prior integration and confidence-based filtering of interactions, tackled in the previous Chapters, for interpreting gene expression data. Our approach aims to identify interaction network hot-spots with altered activity and candidate causative genes in respect of a phenotype for which gene expression data are provided. Notably, it combines the basic principles of both mainstream complementary strategies for disease-related pathway identification: Similar to methods for *de novo* identification of context-specific modules (32, 166, 169), the functional modules are mined from interaction networks. However, the sub-networks are defined *a priori* and, given expression data, each sub-network

is assessed through over-representation or enrichment analysis as in classical pathway analyses (40, 155). Furthermore, the sub-networks comprise both large-scale protein interactions and manually inferred interactions from signaling, regulatory, and metabolic pathways. Thus, our approach closes the gap between the two mainstream strategies for interaction- and pathway-based interpretation of gene expression data. It yielded promising results when applied to prostate cancer patient data as it highlighted genes with known causal role in cancer, even if they were not represented in the expression data. The combination of interaction integration, de-noising, and using network neighborhood information in conjunction with gene expression data appears to be key for the identification of disease genes.

Notably, our network-based method considers only the direct interaction neighborhood of every gene separately rather than attempting to explain the entire set of observed expression effects at once with a minimal set of few causative genes like some of the previous methods (e.g. (96)). A local search for dysregulated regions in the network is motivated by the fact that disease often impacts the whole cell and is therefore reflected not only in the expression of downstream counterparts of causative genes. Rather, many effects observed at the gene expression level are not directly associated with the causative genes. In cancer, for example, the reproductive machinery of the cell is highly active due to the proliferative nature of the disease. In this light, the reported increase of ribosome production in cancer cells (170), while certainly being a hallmark of cancer, is more likely its secondary effect than its primary cause.

In a parallel line of work within the context of omics data interpretation we extended the pathway analysis paradigm to integrate transcriptomics/proteomics with metabolomics measurements of a given phenotype on the level of biochemical pathways. In a recent publication (23) summarized above we showed that combining evidence of pathway disregulation on both gene expression and metabolite concentration levels allows for the identification of phenotype-associated pathways that would be missed when pathway analysis is applied to any of these functional levels alone. We developed the first available computational tool for such integrative analyses (90).

A natural further development of the two contributions presented in this Chapter, namely 1) the definition of functional gene sets (NESTs) from a network comprising physical, regulatory, signaling and metabolic reactions in aid of disease gene identification from expression data, and 2) the integration of omics data at the pathway

level, would be to combine both concepts. Since the interaction network in Consensus-PathDB involves more than 5,000 metabolites additionally to human genes/proteins, sub-networks can be constructed such that they include metabolites. When both gene expression and metabolite measurements are present for a phenotypic condition, their integration at the level of NESTs would certainly contribute toward more accurate hypotheses.

Chapter 5

Conclusion

System biology aims to provide a mechanistic view on cellular processes in health and disease. Toward this aim, knowledge of all biomolecular interactions in the cell is crucial. Large interaction datasets for several species are already available, albeit they likely represent only parts of the underlying real interactomes. A system-level picture of cellular biology is still limited also by the quality of the available data and by the way available data are handled. In this thesis, we addressed the problems that protein-protein interactome maps often contain many false positives, and that interaction data often reside in complementary, heterogeneous databases. Furthermore, we tackled the problem that gene signatures for complex diseases are often inconsistent from experiment to experiment, and are barely sufficient for explaining the causes and mechanisms of those diseases without taking into account interaction knowledge.

First, we developed a meta-database called ConsensusPathDB (89, 92) to solve the recognized problem that existing interaction knowledge is scattered across many public repositories that are complementary and barely compatible regarding their data model and format. ConsensusPathDB integrates several different types of interactions, including gene regulatory, signaling, metabolic, and protein-protein interactions, as well as manually defined pathways from a total of 26 databases. With several examples we demonstrated the necessity of interaction data integration. For instance, many of the interactions of any particular gene would often be missed if a single primary database is used, as we showed for the well-studied p53 protein. This could have grave impact on many areas of biological research, for instance in drug development where predictions about the drug impact are based on knowledge of the target's interactions. Similarly,

pathway databases contain complementary sets of manually defined pathways, and even homonymous pathways from different sources show grave differences in their composition. Pathway-based analyses of gene expression data, however, require unbiased pathway data to ensure accurate hypotheses. Moreover, we showed that results of topological analyses of interaction networks could be different according to which databases the networks are retrieved from. Our interaction integration efforts have resulted in a human interactome map of unprecedented coverage, and have enabled a more complete view on cell biology at the molecular level. This interactome can be used in various contexts through a public interface (http://cpdb.molgen.mpg.de) offering a rich palette of functionalities for interaction query and visualization, interaction network validation and extension, and most notably, for network- and pathway-based analysis of transcriptomics/proteomics data. Moreover, a database interface plugin for Cytoscape was created to automate the process of evidence mining and novelty assessment for protein-protein interactions (122). We initially developed the plugin to assess the sensitivity of a mammalian-two-hybrid interaction screen (126). ConsensusPathDB is rebuilt automatically every three months with the newest versions of the source databases to ensure that its content stays up-to-date, and new interaction resources are added at the rate of approximately one resource per release (Appendix Figure A.1). There are further interaction types like genetic interactions that are currently missing in the database but will be added in the future.

The second problem tackled in this thesis is the high rate of false positives often found in protein-protein interaction data, arising from experimental or literature mining errors. We developed a novel approach called CAPPIC (cluster-based assessment of protein-protein interaction confidence) (91) that exploits solely the topology of a protein interaction network to assess the confidence of its individual interactions. CAPPIC requires no parameters or reference sets for confidence scoring and optimizes algorithmic parameters intrinsically. On the basis of several different yeast protein-protein interaction datasets, we showed with ROC analysis that our approach achieves a better performance than previous network topology-based methods in assigning confidence to all interactions. Confidence scores calculated by CAPPIC are affirmed by a positive correlation with Gene Ontology co-annotation of interacting proteins, and also correlate with experimental interaction evidence. We have implemented CAPPIC as

a publicly accessible web-based tool at http://cpdb.molgen.mpg.de/cappic, where the source code is also available for free download.

The third research area approached in this thesis is the elucidation of molecular causes and mechanisms of complex diseases such as cancer using interaction and pathway knowledge in conjunction with gene expression data. Toward this goal, many methods have been developed that integrate interaction or pathway data with transcriptomics or proteomics data to yield hypotheses about genes, interaction sub-networks, or known biological processes related with disease. Such approaches require comprehensive and error-free models of the cell's molecular circuitry to ensure accuracy of results. By combining interaction integration with de-noising based on interaction confidence scoring, we have created a more complete and more accurate human interactome to answer this need. We devised a simple strategy that exploits this interactome to identify centric, neighborhood-based interaction sub-networks (called NESTs) with altered activity in gene expression profiles. Our approach is similar to classical pathway analyses in that predefined gene sets are tested for over-representation or enrichment with disease-relevant genes; however, the underlying sets are defined from a genome-wide network integrating different interaction types and do not result from manual curation. Although NESTs are certainly not as likely to contain exclusively genes with the same or similar functions as manually curated pathways, they have several advantages over such pathways e.g. when it comes to genome coverage and bias. Notably, identified NESTs can be suggestive for causative genes associated with the phenotype under study. This was demonstrated with two example cases based on expression data from prostate cancer patients, where many genes were recovered whose mutations are known to cause cancer. Our approach was implemented within the gene expression data analysis module of the ConsensusPathDB web interface at http://cpdb.molgen.mpg.de.

Within the context of disease mechanism elucidation, we outlined a second integrative approach called IMPaLA that combines transcriptomics/proteomics with metabolomics data on the level of predefined pathways (23). It can be applied when both gene expression levels and metabolite concentrations have been measured in a phenotype under study and the goal is to select pathways whose association with the phenotype is supported by either or both of these datasets. IMPaLA was implemented as a web server available freely at http://impala.molgen.mpg.de (90). To our knowledge, this is

the first tool for the joint analysis of gene expression and metabolite data on the level of pathways.

Currently, gene expression, gene regulation, protein binding, protein modifications, metabolic reactions, metabolite dynamics, and other important cellular processes are still mostly studied in isolation as if they were not deeply interlinked and dependent on each other in the cell. Nevertheless, a tendency toward large-scale integration of these and other aspects is clearly recognizable in contemporary research as an important step toward better understanding of the molecular mechanisms governing life.

Bibliography

[1] Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207. 3

[2] Aittokallio, T. and Schwikowski, B. (2006) Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, **7**, 243–255. 6, 8

[3] Albert, Jeong and Barabasi (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–382. 9

[4] Alexander, R.P., Kim, P.M., Emonet, T. and Gerstein, M.B. (2009) Understanding modularity in molecular networks requires dynamics. *Science Signaling*, **2**, pe44. 41

[5] Alon, U. (2007) Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, **8**, 450–461. 8, 41

[6] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L.L. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, **32**, D115–119. 22

[7] Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S.N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K. and Hermjakob, H. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, **38**, D525–531. 18

[8] 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. *Nature Genetics*, **25**, 25–29. 35, 54, 65

[9] Bader, G.D. and Hogue, C.W.V. (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*, **20**, 991–997. 8

[10] Bader, G.D. and Hogue, C.W.V. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2. 59

[11] Bader, G.D., Cary, M.P. and Sander, C. (2006) Pathguide: a pathway resource list. *Nucleic Acids Research*, **34**, D504–506. 4, 16

[12] Bader, J.S., Chaudhuri, A., Rothberg, J.M. and Chant, J. (2004) Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, **22**, 78–85. 41

[13] Barabasi and Albert (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512. 9, 30

[14] Barabasi, A. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**, 101–113. 8, 9, 10, 29, 41, 59, 64

[15] 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. 74

[16] Beuming, T., Skrabanek, L., Niv, M.Y., Mukherjee, P. and Weinstein, H. (2005) PDZBase: a protein-protein interaction database for PDZ-domains. *Bioinformatics*, **21**, 827–828. 19

[17] Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M., Harpole, D., Lancaster, J.M., Berchuck, A., Olson, J.A., Marks, J.R., Dressman, H.K., West, M. and Nevins, J.R. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357. 65

[18] Bornholdt, S. (2005) Systems biology. Less is more in modeling large genetic networks. *Science*, **310**, 449–451. 41

[19] Brohee, S. and van Helden, J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**, 488. 64

[20] Brown, K.R. and Jurisica, I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082. 16

[21] Buck, M.J. and Lieb, J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360. 2

[22] Carey, P.R. and Dong, J. (2004) Following ligand binding and ligand reactions in proteins via Raman crystallography. *Biochemistry*, **43**, 8885–8893. 3

[23] Cavill, R., Kamburov, A., Ellis, J.K., Athersuch, T.J., Blagrove, M.S.C., Herwig, R., Ebbels, T.M.D. and Keun, H.C. (2011) Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Computational Biology*, **7**, e1001113. 13, 14, 64, 82, 83, 84, 85, 87, 91

[24] Celis, J.E., Kruhoffer, M., Gromova, I., Frederiksen, C., Ostergaard, M., Thykjaer, T., Gromov, P., Yu, J., Palsdottir, H., Magnusson, N. and Orntoft, T.F. (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Letters*, **480**, 2–16. 11

[25] Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L. and Cesareni, G. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, **38**, D532–539. 18

[26] Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D. and Sander, C. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, **39**, D685–690. 16

[27] Chagoyen, M. and Pazos, F. (2011) MBRole: enrichment analysis of metabolomic data. *Bioinformatics*, **27**, 730–731. 82

[28] Chaurasia, G., Malhotra, S., Russ, J., Schnoegl, S., Hänig, 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 Research*, **37**, D657–660. 16

[29] Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N. and Ricard-Blum, S. (2011) MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Research*, **39**, D235–240. 19

[30] Chen, J., Hsu, W., Lee, M.L. and Ng, S. (2005) Discovering reliable protein interactions from high-throughput experimental data using network topology. *Artificial Intelligence in Medicine*, **35**, 37–47. 41

[31] Chua, H.N. and Wong, L. (2008) Increasing the reliability of protein interactomes. *Drug Discovery Today*, **13**, 652–658. 41

[32] Chuang, H., Lee, E., Liu, Y., Lee, D. and Ideker, T. (2007) Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, **3**, 140. 11, 64, 65, 68, 86

[33] Chuang, H., Hofree, M. and Ideker, T. (2010) A decade of systems biology. *Annual Review of Cell and Developmental Biology*, **26**, 721–744. 11, 63

[34] Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A.R., Vailaya, A., Wang, P., Adler, A., Conklin, B.R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G.J., Ideker, T. and Bader, G.D. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, **2**, 2366–2382. 65, 68

[35] Cokol, M., Iossifov, I., Weinreb, C. and Rzhetsky, A. (2005) Emergent behavior of growing knowledge about molecular interactions. *Nature Biotechnology*, **23**, 1243–1247. 4, 15, 37, 40

[36] Collins, S.R., Miller, K.M., Maas, N.L., Roguev, A., Fillingham, J., Chu, C.S., Schuldiner, M., Gebbia, M., Recht, J., Shales, M., Ding, H., Xu, H., Han, J., Ingvarsdottir, K., Cheng, B., Andrews, B., Boone, C., Berger, S.L., Hieter, P., Zhang, Z., Brown, G.W., Ingles, C.J., Emili, A., Allis, C.D., Toczyski, D.P., Weissman, J.S., Greenblatt, J.F. and Krogan, N.J. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, **446**, 806–810. 43, 51

[37] Collu, G.M. and Brennan, K. (2007) Cooperation between Wnt and Notch signalling in human breast cancer. *Breast Cancer Research*, **9**, 105. 66

[38] Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., Prinz, J., Onge, R.P.S., VanderSluis, B., Makhnevych, T., Vizeacoumar, F.J., Alizadeh, S., Bahr, S., Brost, R.L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z., Liang, W., Marback, M., Paw, J., Luis, B.S., Shuteriqi, E., Tong, A.H.Y., Dyk, N.v., Wallace, I.M., Whitney, J.A., Weirauch, M.T., Zhong, G., Zhu, H., Houry, W.A., Brudno, M., Ragibizadeh, S., Papp, B., Pal, C., Roth, F.P., Giaever, G., Nislow, C., Troyanskaya, O.G., Bussey, H., Bader, G.D., Gingras, A., Morris, Q.D., Kim, P.M., Kaiser, C.A., Myers, C.L., Andrews, B.J. and Boone, C. (2010) The genetic landscape of a cell. *Science*, **327**, 425–431. 37, 43

[39] Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P. and Stein, L. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, **39**, D691–697. 16

[40] Curtis, R.K., Oresic, M. and Vidal-Puig, A. (2005) Pathways to the analysis of microarray data. *Trends in Biotechnology*, **23**, 429–435. 11, 63, 65, 71, 87

[41] Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A., Simonis, N., Rual, J., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D.E., Ecker, J.R., Roth, F.P. and Vidal, M. (2009) Literature-curated protein interaction datasets. *Nature Methods*, **6**, 39–46. 4, 5, 15, 16, 39, 40

[42] de Matos, P., Alcantara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S. and Steinbeck, C. (2010) Chemical Entities of Biological Interest: an update. *Nucleic Acids Research*, **38**, D249–254. 22

[43] Deane, C.M., Salwinski, L, Xenarios, I. and Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics*, **1**, 349–356. 40, 41

[44] Deng, M., Sun, F. and Chen, T. (2003) Assessment of the reliability of protein-protein interactions and protein function prediction. *Pacific Symposium on Biocomputing*, 140–151. 41

[45] van Dongen, S. (2000) *A Cluster algorithm for graphs*. Ph.D. thesis, Centrum voor Wiskunde en Informatica, Amsterdam, Netherlands. 42, 44, 46

[46] Driscoll, T., Dyer, M.D., Murali, T.M. and Sobral, B.W. (2009) PIG–the pathogen interaction gateway. *Nucleic Acids Research*, **37**, D647–650. 18

[47] Ein-Dor, L., Kela, I., Getz, G., Givol, D. and Domany, E. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* , **21**, 171–178. 11, 63

[48] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868. 64

[49] Erler, J.T. and Linding, R. (2010) Network-based drugs and biomarkers. *The Journal of Pathology*, **220**, 290–296. 65

[50] 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., Duewel, H.S., Stewart,

I.I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S., Moran, M.F., Morin, G.B., Topaloglou, T. and Figeys, D. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology*, **3**, 89. 30, 36, 39

[51] Fell, D.A. and Wagner, A. (2000) The small world of metabolism. *Nature Biotechnology*, **18**, 1121–1122. 9

[52] Fields, S. and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246. 3

[53] Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H.S., Rios, D., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y.A., Trevanion, S., Vandrovcova, J., Vilella, A.J., White, S., Wilder, S.P., Zadissa, A., Zamora, J., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J.P., Parker, A., Proctor, G., Vogel, J. and Searle, S.M.J. (2011) Ensembl 2011. *Nucleic Acids Research*, **39**, D800–806. 22

[54] Fortunato, S. (2010) Community detection in graphs. *Physics Reports*, **486**, 75–174. 41, 42, 59

[55] 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. *American Journal of Human Genetics*, **78**, 1011–1025. 65

[56] Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. and Feldman, M.W. (2002) Evolutionary rate in the protein interaction network. *Science* , **296**, 750–752. 1

[57] Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D.D., Liu, P., Gautam, B., Ly, S., Guo, A.C., Xia, J., Liang, Y., Shrivastava, S. and Wishart, D.S. (2010)

SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Research*, **38**, D480–487. 19

[58] Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) A census of human cancer genes. *Nature Reviews Cancer*, **4**, 177–183. 76, 77

[59] Galarneau, A., Primeau, M., Trudeau, L. and Michnick, S.W. (2002) Beta-lactamase protein fragment complementation assays as in vivo and in vitro sensors of protein protein interactions. *Nature Biotechnology*, **20**, 619–622. 3

[60] Gandhi, T.K.B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K.N., Mohan, S.S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J.D., Parmigiani, G., Schultz, J., Bader, J.S. and Pandey, A. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, **38**, 285–293. 65

[61] Gavin, A., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636. 39, 42, 43, 59

[62] Ge, H., Liu, Z., Church, G.M. and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature Genetics*, **29**, 482–486. 64

[63] Ghaemmaghami, S., Huh, W., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. and Weissman, J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741. 74

[64] Goddard, J. and Reymond, J. (2004) Enzyme assays for high-throughput screening. *Current Opinion in Biotechnology*, **15**, 314–322. 3

[65] Goh, K., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabasi, A. (2007) The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 8685–8690. 11, 64, 65, 68, 78

[66] Goldberg, D.S. and Roth, F.P. (2003) Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4372–4376. 41

[67] Han, J.J. (2008) Understanding biological functions through molecular networks. *Cell Research*, **18**, 224–237. 40

[68] Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70. 81

[69] Hanisch, D., Zien, A., Zimmer, R. and Lengauer, T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18 Suppl 1**, S145–154. 64

[70] Hannum, G., Srivas, R., Guenole, A., van Attikum, H., Krogan, N.J., Karp, R.M. and Ideker, T. (2009) Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genetics*, **5**, e1000782. 57

[71] Harrison, P.M., Kumar, A., Lang, N., Snyder, M. and Gerstein, M. (2002) A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Research*, **30**, 1083–1090. 42

[72] Hart, G.T., Ramani, A.K. and Marcotte, E.M. (2006) How complete are current yeast and human protein-interaction networks? *Genome Biology*, **7**, 120. 5, 38, 39, 40, 42, 51, 60

[73] Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–52. 39, 41, 42, 64

[74] Higham, D.J., Rasajski, M. and Przulj, N. (2008) Fitting a geometric graph to a protein-protein interaction network. *Bioinformatics*, **24**, 1093–1099. 42

[75] Hollstein, M., Sidransky, D., Vogelstein, B. and Harris, C.C. (1991) p53 mutations in human cancers. *Science*, **253**, 49–53. 27, 78

[76] Hollywood, K., Brison, D.R. and Goodacre, R. (2006) Metabolomics: current technologies and future trends. *Proteomics*, **6**, 4716–4723. 82

[77] Hsu, P.P. and Sabatini, D.M. (2008) Cancer cell metabolism: Warburg and beyond. *Cell*, **134**, 703–707. 13, 82

[78] Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**, 44–57. 82

[79] Hutchins, J.R.A., Toyoda, Y., Hegemann, B., Poser, I., Heriche, J., Sykora, M.M., Augsburg, M., Hudecz, O., Buschhorn, B.A., Bulkescher, J., Conrad, C., Comartin, D., Schleiffer, A., Sarov, M., Pozniakovsky, A., Slabicki, M.M., Schloissnig, S., Steinmacher, I., Leuschner, M., Ssykor, A., Lawo, S., Pelletier, L., Stark, H., Nasmyth, K., Ellenberg, J., Durbin, R., Buchholz, F., Mechtler, K., Hyman, A.A. and Peters, J. (2010) Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science*, **328**, 593–599. 39

[80] Ideker, T. and Sharan, R. (2008) Protein networks in disease. *Genome Research*, **18**, 644–652. 11, 40, 63, 65, 78

[81] Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18 Suppl 1**, S233–240. 64

[82] Ihaka, R. and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314. 51

[83] Isserlin, R., El-Badrawi, R.A. and Bader, G.D. (2011) The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database*, **2011**, baq037. 18

[84] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4569–4574. 39, 43

[85] Jansen, R. and Gerstein, M. (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current Opinion in Microbiology*, **7**, 535–545. 40

[86] Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654. 9

[87] Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42. 9

[88] Kamburov, A., Goldovsky, L., Freilich, S., Kapazoglou, A., Kunin, V., Enright, A.J., Tsaftaris, A. and Ouzounis, C.A. (2007) Denoising inferred functional association networks obtained by gene fusion analysis. *BMC Genomics*, **8**, 460. 3

[89] Kamburov, A., Wierling, C., Lehrach, H. and Herwig, R. (2009) ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Research*, **37**, D623–628. 12, 14, 15, 82, 89, 118

[90] Kamburov, A., Cavill, R., Ebbels, T.M.D., Herwig, R. and Keun, H.C. (2011) Integrated pathway- level analysis of transcriptomics and metabolomics data with IMPaLA. *Revised manuscript submitted.*. 13, 14, 64, 85, 87, 91

[91] Kamburov, A., Grossmann, A., Herwig, R. and Stelzl, U. (2011) Cluster-based assessment of protein- protein interaction confidence. *Revised manuscript submitted.*. 13, 14, 39, 42, 90

[92] Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H. and Herwig, R. (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research*, **39**, D712–717. 12, 14, 15, 43, 89

[93] Kandasamy, K., Mohan, S.S., Raju, R., Keerthikumar, S., Kumar, G.S.S., Venugopal, A.K., Telikicherla, D., Navarro, J.D., Mathivanan, S., Pecquet, C., Gollapudi, S.K., Tattikota, S.G., Mohan, S., Padhukasahasram, H., Subbannayya, Y., Goel, R., Jacob, H.K.C., Zhong, J., Sekhar, R., Nanjappa, V., Balakrishnan, L., Subbaiah, R., Ramachandra, Y.L., Rahiman, B.A., Prasad, T.S.K., Lin, J., Houtman, J.C.D., Desiderio, S., Renauld, J., Constantinescu, S.N., Ohara, O.,

Hirano, T., Kubo, M., Singh, S., Khatri, P., Draghici, S., Bader, G.D., Sander, C., Leonard, W.J. and Pandey, A. (2010) NetPath: a public resource of curated signal transduction pathways. *Genome Biology*, **11**, R3. 18

[94] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Research*, **32**, D277–280. 56

[95] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. and Hirakawa, M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, **38**, D355–360. 18

[96] Karni, S., Soreq, H. and Sharan, R. (2009) A network-based method for predicting disease-causing genes. *Journal of Computational Biology*, **16**, 181–189. 65, 87

[97] Kemmeren, P., Berkum, N.L.v., Vilo, J., Bijma, T., Donders, R., Brazma, A. and Holstege, F.C.P. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell*, **9**, 1133–1143. 41

[98] 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 Research*, **37**, D767–772. 18

[99] Kitano, H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664. 10

[100] Klamt, S., Haus, U. and Theis, F. (2009) Hypergraphs and cellular networks. *PLoS Computational Biology*, **5**, e1000385. 6, 8

[101] Koh, J.L.Y., Ding, H., Costanzo, M., Baryshnikova, A., Toufighi, K., Bader, G.D., Myers, C.L., Andrews, B.J. and Boone, C. (2010) DRYGIN: a database of quantitative genetic interaction networks in yeast. *Nucleic Acids Research*, **38**, D502–507. 37

[102] Korcsmaros, T., Farkas, I.J., Szalay, M.S., Rovo, P., Fazekas, D., Spiro, Z., Böde, C., Lenti, K., Vellai, T. and Csermely, P. (2010) Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Bioinformatics*, **26**, 2042–2050. 19

[103] Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrin-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., Onge, P.S., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. and Greenblatt, J.F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643. 39, 43

[104] Kuchaiev, O., Rasajski, M., Higham, D.J. and Przulj, N. (2009) Geometric denoising of protein-protein interaction networks. *PLoS Computational Biology*, **5**, e1000454. 41

[105] Levy, E.D., Landry, C.R. and Michnick, S.W. (2009) How perfect can protein interactomes be? *Science Signaling*, **2**, pe11. 5, 38

[106] 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. *Molecular & Cellular Proteomics*, **7**, 1043–1052. 41

[107] Lynn, D.J., Chan, C., Naseer, M., Yau, M., Lo, R., Sribnaia, A., Ring, G., Que, J., Wee, K., Winsor, G.L., Laird, M.R., Breuer, K., Foroushani, A.K., Brinkman, F.S.L. and Hancock, R.E.W. (2010) Curating the innate immunity interactome. *BMC Systems Biology*, **4**, 117. 18

[108] Ma, H. and Goryanin, I. (2008) Human metabolic network reconstruction and its impact on drug discovery and development. *Drug Discovery Today*, **13**, 402–408. 16

[109] Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, **39**, D52–57. 22

[110] Makrantonaki, E., Adjaye, J., Herwig, R., Brink, T.C., Groth, D., Hultschig, C., Lehrach, H. and Zouboulis, C.C. (2006) Age-specific hormonal decline is accompanied by transcriptional changes in human sebocytes in vitro. *Aging Cell*, **5**, 331–344. 73

[111] Mamas, M., Dunn, W.B., Neyses, L. and Goodacre, R. (2011) The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives of Toxicology*, **85**, 5–17. 82

[112] Mangan, S. and Alon, U. (2003) Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 11980–11985. 9

[113] Mani, K.M., Lefebvre, C., Wang, K., Lim, W.K., Basso, K., Dalla-Favera, R. and Califano, A. (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Molecular Systems Biology*, **4**, 169. 65

[114] Mering, C.v., 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. 5, 35, 38, 40, 51

[115] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827. 8

[116] Oliver, S. (2000) Guilt-by-association goes global. *Nature*, **403**, 601–603. 54, 66

[117] Orchard, S., Kerrien, S., Jones, P., Ceol, A., Chatr-Aryamontri, A., Salwinski, L., Nerothin, J. and Hermjakob, H. (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics*, **7 Suppl 1**, 28–34. 16

[118] Oti, M. and Brunner, H.G. (2007) The modular nature of genetic diseases. *Clinical Genetics*, **71**, 1–11. 11, 64, 65, 68

[119] Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H., Ruepp, A. and Frishman, D. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832–834. 18

[120] Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669–680. 2

[121] Paz, A., Brownstein, Z., Ber, Y., Bialik, S., David, E., Sagir, D., Ulitsky, I., Elkon, R., Kimchi, A., Avraham, K.B., Shiloh, Y. and Shamir, R. (2011) SPIKE: a database of highly curated human signaling pathways. *Nucleic Acids Research*, **39**, D793–799. 18

[122] Pentchev, K., Ono, K., Herwig, R., Ideker, T. and Kamburov, A. (2010) Evidence mining and novelty assessment of protein-protein interactions with the ConsensusPathDB plugin for Cytoscape. *Bioinformatics*, **26**, 2796–2797. 14, 35, 51, 90

[123] Pereira-Leal, J.B., Enright, A.J. and Ouzounis, C.A. (2004) Detection of functional modules from protein interaction networks. *Proteins*, **54**, 49–57. 42, 56, 59

[124] Peters, A.H., O'Carroll, D., Scherthan, H., Mechtler, K., Sauer, S., Schöfer, C., Weipoltshammer, K., Pagani, M., Lachner, M., Kohlmaier, A., Opravil, S., Doyle, M., Sibilia, M. and Jenuwein, T. (2001) Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell*, **107**, 323–337. 80

[125] Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., Conklin, B.R. and Evelo, C. (2008) WikiPathways: pathway editing for the people. *PLoS Biology*, **6**, e184. 19

[126] Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C.O., Forrest, A.R.R., Gough, J., Grimmond, S., Han, J., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C.R., Ogawa, C., Radovanovic,

A., Schwartz, A., Teasdale, R.D., Tegner, J., Lenhard, B., Teichmann, S.A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D.A., Ideker, T. and Hayashizaki, Y. (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752. 90

[127] Ravasz, E. (2009) Detecting hierarchical modularity in biological networks. *Methods in Molecular Biology*, **541**, 145–160. 10, 30, 46, 59

[128] 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. 10, 30

[129] Resnik, P. (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, **11**, 95–130. 54

[130] Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B.B., Barrette, T.R., Anstet, M.J., Kincead-Beal, C., Kulkarni, P., Varambally, S., Ghosh, D. and Chinnaiyan, A.M. (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180. 75, 79

[131] Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M. and Karp, P.D. (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, **6**, R2. 16

[132] Rosenfeld, N., Elowitz, M.B. and Alon, U. (2002) Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology*, **323**, 785–793. 9

[133] Rual, J., Venkatesan, K., Hao, T., Hirozane–Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi–Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette–Stamm, L., Cusick, M.E.,

Hill, D.E., Roth, F.P. and Vidal, M. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178. 30, 36, 39

[134] Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger–Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. and Mewes, H. (2010) CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Research*, **38**, D497–501. 18

[135] Saito, R., Suzuki, H. and Hayashizaki, Y. (2003) Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, **19**, 756–763. 41

[136] Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, **32**, D449–451. 18

[137] Sanderson, C.M. (2009) The Cartographers toolbox: building bigger and better human protein interaction networks. *Briefings in Functional Genomics & Proteomics*, **8**, 1–11. 3

[138] Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. and Buetow, K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Research*, **37**, D674–679. 18

[139] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470. 10, 71

[140] Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O. and Weinstein, J.N. (2000) A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, **24**, 236–244. 83, 84

[141] Scott, J., Ideker, T., Karp, R.M. and Sharan, R. (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **13**, 133–144. 64

[142] Seal, R.L., Gordon, S.M., Lush, M.J., Wright, M.W. and Bruford, E.A. (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Research*, **39**, D514–519. 22

[143] Segal, E., Wang, H. and Koller, D. (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **19 Suppl 1**, i264–271. 64

[144] Shannon, C. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423. 56

[145] Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M. and Ideker, T. (2005) Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 1974–1979. 3, 41

[146] Sharan, R., Ulitsky, I. and Shamir, R. (2007) Network-based prediction of protein function. *Molecular Systems Biology*, **3**, 88. 10, 40, 41, 66, 67

[147] Skrabanek, L., Saini, H.K., Bader, G.D. and Enright, A.J. (2008) Computational prediction of protein-protein interactions. *Molecular Biotechnology*, **38**, 1–17. 3

[148] Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P. and Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432. 32

[149] Stark, C., Breitkreutz, B., Chatr–Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Auken, K.V., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K. and Tyers, M. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research*, **39**, D698–704. 18, 43

[150] Steffen, M., Petti, A., Aach, J., D'haeseleer, P. and Church, G. (2002) Automated modelling of signal transduction networks. *BMC Bioinformatics*, **3**, 34. 64

[151] Stelzl, U. and Wanker, E.E. (2006) The value of high quality protein-protein interaction networks for systems biology. *Current Opinion in Chemical Biology*, **10**, 551–558. 39

[152] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. and Wanker, E.E. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968. 30, 36, 39

[153] Strogatz, S.H. (2001) Exploring complex networks. *Nature*, **410**, 268–276. 41

[154] Strömbäck, L., Hall, D. and Lambrix, P. (2007) A review of standards for data exchange within systems biology. *Proteomics*, **7**, 857–867. 5, 16

[155] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545–15550. 63, 65, 73, 87

[156] Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. and Mesirov, J.P. (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, **23**, 3251–3253. 82

[157] Suthram, S., Shlomi, T., Ruppin, E., Sharan, R. and Ideker, T. (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, **7**, 360. 41, 62

[158] Suthram, S., Beyer, A., Karp, R.M., Eldar, Y. and Ideker, T. (2008) eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular Systems Biology*, **4**, 162. 65

[159] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L.J. and von Mering, C. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, **39**, D561–568. 16

[160] Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Molina, M.M.S., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H. and Michnick, S.W. (2008) An in vivo map of the yeast protein interactome. *Science*, **320**, 1465–1470. 39, 43, 53, 62

[161] Tarcea, V.G., Weymouth, T., Ade, A., Bookvich, A., Gao, J., Mahavisno, V., Wright, Z., Chapman, A., Jayapandian, M., Ozgür, A., Tian, Y., Cavalcoli, J., Mirel, B., Patel, J., Radev, D., Athey, B., States, D. and Jagadish, H.V. (2009) Michigan molecular interactions r2: from interacting proteins to pathways. *Nucleic Acids Research*, **37**, D642–646. 16

[162] Taylor, I.W., Linding, R., Warde–Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q. and Wrana, J.L. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, **27**, 199–204. 65

[163] Thorn, C.F., Klein, T.E. and Altman, R.B. (2010) Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics*, **11**, 501–505. 19

[164] Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J.E., Shah, R.B., Pienta, K.J., Rubin, M.A. and Chinnaiyan, A.M. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648. 11

[165] Tomlins, S.A., Mehra, R., Rhodes, D.R., Cao, X., Wang, L., Dhanasekaran, S.M., Kalyana–Sundaram, S., Wei, J.T., Rubin, M.A., Pienta, K.J., Shah, R.B. and Chinnaiyan, A.M. (2007) Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics*, **39**, 41–51. 65

[166] Tuck, D.P., Kluger, H.M. and Kluger, Y. (2006) Characterizing disease states from topological properties of transcriptional regulatory networks. *BMC Bioinformatics*, **7**, 236. 64, 86

[167] 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. 39, 43

[168] Ulitsky, I. and Shamir, R. (2009) Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, **25**, 1158–1164. 64

[169] Ulitsky, I., Krishnamurthy, A., Karp, R.M. and Shamir, R. (2010) DEGAS: de novo discovery of dysregulated pathways in human diseases. *PloS One*, **5**, e13367. 28, 64, 66, 86

[170] Vaarala, M.H., Porvari, K.S., Kyllönen, A.P., Mustonen, M.V., Lukkarinen, O. and Vihko, P.T. (1998) Several genes encoding ribosomal proteins are over-expressed in prostate-cancer cell lines: confirmation of L7a and L37 over-expression in prostate-cancer tissue samples. *International Journal of Cancer*, **78**, 27–32. 80, 81, 87

[171] Venkatesan, K., Rual, J., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane–Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K., Yildirim, M.A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J.M., Cevik, S., Simon, C., Smet, A.d., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R.R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M.E., Roth, F.P., Hill, D.E., Tavernier, J., Wanker, E.E., Barabasi, A. and Vidal, M. (2009) An empirical framework for binary interactome mapping. *Nature Methods*, **6**, 83–90. 39, 40, 51, 61

[172] Vinayavekhin, N., Homan, E.A. and Saghatelian, A. (2010) Exploring disease through metabolomics. *ACS Chemical Biology*, **5**, 91–103. 82

[173] Wang, K., Saito, M., Bisikirska, B.C., Alvarez, M.J., Lim, W.K., Rajbhandari, P., Shen, Q., Nemenman, I., Basso, K., Margolin, A.A., Klein, U., Dalla–Favera, R. and Califano, A. (2009) Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nature Biotechnology*, **27**, 829–839. 65

[174] Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63. 10, 71

[175] Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442. 9, 41, 58, 61

[176] Whitney, H. (1932) Congruent Graphs and the Connectivity of Graphs. *American Journal of Mathematics*, **54**, 150–168. 42, 44

[177] Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J.A., Lim, E., Sobsey, C.A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H.J. and Forsythe, I. (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, **37**, D603–610. 22

[178] Xia, J. and Wishart, D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, **38**, W71–W77. 82

[179] Yang, C., Chang, C., Yu, Y., Lin, T.E., Lee, S., Yen, C., Yang, J., Lai, J., Hong, Y., Tseng, T., Chao, K. and Huang, C.F. (2008) PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, **24**, i14–20. 18

[180] Yoon, K., Hwangbo, B., Kim, I., Park, S., Kim, H.S., Kee, H.J., Lee, J.E., Jang, Y.K., Park, J. and Lee, J.S. (2006) Novel polymorphisms in the SUV39H2 histone methyltransferase and the risk of lung cancer. *Carcinogenesis*, **27**, 2217–2222. 80

[181] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y. and Wang, S. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978. 54

[182] Yu, H., Paccanaro, A., Trifonov, V. and Gerstein, M. (2006) Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, **22**, 823–829. 43

[183] 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., Dricot, A., Vazquez, A., Murray, R.R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., Smet, A.d., Motyl, A., Hudson, M.E., Park, J., Xin, X., Cusick, M.E., Moore, T., Boone, C., Snyder, M., Roth, F.P., Barabasi, A., Tavernier, J., Hill, D.E. and Vidal, M. (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110. 39, 43, 61

[184] Yu, Y.P., Landsittel, D., Jing, L., Nelson, J., Ren, B., Liu, L., McDonald, C., Thomas, R., Dhir, R., Finkelstein, S., Michalopoulos, G., Becich, M. and Luo, J. (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *Journal of Clinical Oncology*, **22**, 2790–2799. 79, 81, 120, 121

[185] Yuan, H., Gong, A. and Young, C.Y.F. (2005) Involvement of transcription factor Sp1 in quercetin-mediated inhibitory effect on the androgen receptor in human prostate cancer cells. *Carcinogenesis*, **26**, 793–801. 80

Appendix

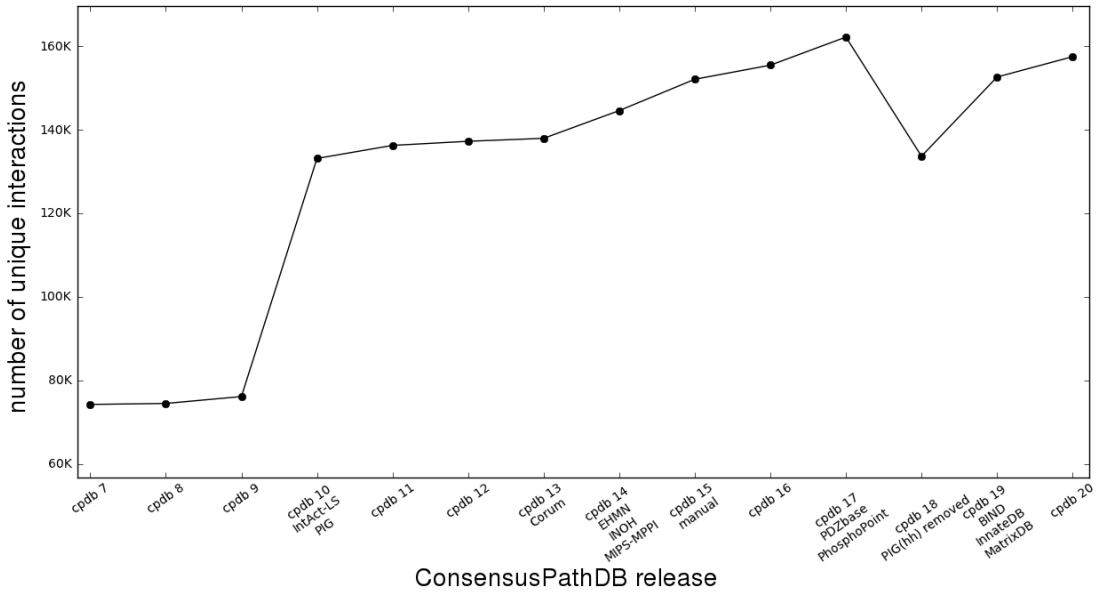

Figure A.1: Growth of ConsensusPathDB's unique interaction content since its initial publication. ConsensusPathDB is rebuilt automatically every three months with the newest versions of its source databases, and new databases are integrated at the rate of approximately one database per release. The plot shows the number of unique interactions in ConsensusPathDB for each release since its initial publication (89); newly integrated databases are listed for each release.

set centers	radius	set size	candidates contained	p-value	q-value	set sources
Leiomodin-1	1	21 (21)	10 (47.6%)	2.06e-13	6.3e-10	H R
Myosin regulatory light polypeptide 9	1	36 (36)	11 (30.6%)	4.31e-12	6.59e-09	P I H B S R
Vinculin	1	214 (180)	19 (10.6%)	4.96e-11	5.06e-08	N P M B B I M I I B P R S C D H
Sorbin and SH3 domain-containing protein 1	1	44 (44)	10 (22.7%)	1.06e-09	8.1e-07	N P B P M H B S I D R I
c-Jun	1	745 (704)	33 (4.7%)	1.35e-08	8.25e-06	N M I P M I P C D I B B I B R S H
JAK_Yp	1	15 (15)	6 (40.0%)	5.88e-08	3e-05	N P I P I S B
Integrin alpha-1	1	125 (116)	12 (10.3%)	2.48e-07	0.000108	N P B B I H I B P I S C D R I
Myosin light chain 6B	1	187 (185)	14 (7.6%)	1.18e-06	0.000451	P S D B M B R I
cytokine receptor	1	8 (8)	4 (50.0%)	3.79e-06	0.00129	I
Cyclin F	1	4 (4)	3 (75.0%)	1.47e-05	0.00448	S C B H B
Myosin-11	1	175 (174)	12 (6.9%)	1.78e-05	0.00495	I P M H B S D R I
TRAILR3	1	12 (12)	4 (33.3%)	2.55e-05	0.0065	P B I M H I B S D
Tropomyosin alpha-3 chain	1	185 (183)	12 (6.6%)	2.94e-05	0.00691	I P M H B S D R I
Integrin beta-5	1	187 (186)	12 (6.5%)	3.45e-05	0.00731	P B P M H I B S C R I
Paxillin	1	372 (351)	17 (4.8%)	3.58e-05	0.00731	N P P M H C D I B M B I I B S R
Catenin beta-1	1	707 (679)	25 (3.7%)	5.74e-05	0.011	N P P M M H C D I B M B I I B S I R
Vinexin	1	169 (169)	11 (6.5%)	6.76e-05	0.0122	P S M H B R I
Talin-1	1	877 (825)	28 (3.4%)	8.41e-05	0.0143	N P B B M H I B P S C I R I
Bone morphogenetic protein 2	1	17 (17)	4 (23.5%)	0.000115	0.0166	H
Myosin regulatory light chain 2-B, smooth muscle isoform	1	192 (180)	11 (6.1%)	0.000119	0.0166	D B H B R I
Mgp	1	7 (7)	3 (42.9%)	0.000124	0.0166	P H B
HOX B8	1	7 (7)	3 (42.9%)	0.000124	0.0166	M H B
CDK4I	1	74 (72)	7 (9.7%)	0.000125	0.0166	N B B P H I B P S R
Tropomyosin alpha-4 chain	1	168 (157)	10 (6.4%)	0.000173	0.0216	I S H I B R I
p53 protein	1	746 (687)	24 (3.5%)	0.000181	0.0216	N P M M I P C D I B M B I B R S H
HOXA9	1	19 (19)	4 (21.1%)	0.000183	0.0216	N M B I H I B S I
STAT	1	8 (8)	3 (37.5%)	0.000196	0.0221	S I
CDKN1A	1	56 (55)	6 (10.9%)	0.000203	0.0221	N H B
cGMP-dependent protein kinase 1, alpha isozyme	1	56 (56)	6 (10.7%)	0.000224	0.0235	N P P H B B I M H I B S D R
Orexin; HLA-DQB1	1	3 (2)	2 (100.0%)	0.00024	0.0235	B
Alpha-Actinin 4	1	2 (2)	2 (100.0%)	0.00024	0.0235	B
ATF-2	1	268 (263)	13 (4.9%)	0.000246	0.0235	N M P I M H P D I B B I I B S R
phospho-ERK-1	1	369 (345)	15 (4.3%)	0.000334	0.031	N P I P M H C D I B M B I B S I R
phospho-ERK-2	1	681 (643)	22 (3.4%)	0.000449	0.0404	N P I P M I C D I B B I B R S H
Tartrate-resistant acid phosphatase type 5	1	12 (11)	3 (27.3%)	0.000557	0.0487	I P M H B S
Cdc26	1	44 (44)	5 (11.4%)	0.000575	0.0489	P P H I B S D R I

Figure A.2: NESTs where cancer metastasis-associated genes are significantly over-represented. 36 NESTs where genes differentially expressed in metastatic prostate cancer (DE genes) are significantly over-represented are listed. For each NEST, the name of the central physical entity, the NEST radius (1 means that all entities in the NEST are direct neighbors of the center), the NEST size (number of physical entities in the NEST, followed by a corrected size according to the background), the number of DE genes contained, the p- and q-values of the hypergeometric test, as well as the source databases that contribute interactions to the NEST (see ConsensusPathDB web page for color key) are listed. Note that the majority of NESTs comprise data from several resources.

APPENDIX

Enriched interaction neighborhood based sets

set centers	set size	measured genes	p-value	q-value	set sources
Histone H3-K9 methyltransferase 2	43	24	8.34e-06	0.0483	11
Trap gamma	143	87	1.2e-05	0.0483	8
Myosin regulatory light polypeptide 9	34	30	1.82e-05	0.0483	6
PPA1	106	86	2.49e-05	0.0483	6
5S rRNA; 5.8S rRNA; 28S rRNA	120	80	2.62e-05	0.0483	1
np1l4_human	96	79	3.03e-05	0.0483	5
Hsc70	137	108	4.42e-05	0.0521	8
SET domain containing (lysine methyltransferase) 7 (EC:2.1.1.43)	40	20	4.77e-05	0.0521	9
40S ribosomal protein S4, Y isoform 2	81	59	5.13e-05	0.0521	1
PP2A-regulatory subunit B delta-2 isoform	217	143	5.45e-05	0.0521	13
peroxiredoxin 6 (EC:1.11.1.7 1.11.1.15)	211	145	6.5e-05	0.0531	6
Fructose-bisphosphate aldolase A	311	236	6.91e-05	0.0531	13
MAT1	229	146	8.17e-05	0.0531	11
MAP kinase p38 delta	115	95	8.21e-05	0.0531	11
lactate dehydrogenase B	174	130	8.85e-05	0.0531	9
Splicing factor 3B subunit 1	86	67	0.000101	0.0531	1
CHAF1A	59	47	0.000102	0.0531	10
SRP Receptor subunit alpha	110	74	0.000107	0.0531	7
SP1	290	210	0.00011	0.0531	13
eRF3	96	67	0.000115	0.0531	6
G-protein gamma 9 (GBGT2) subunit	364	202	0.000117	0.0531	7
Trap beta; SEC61 gamma	114	75	0.000122	0.0531	4
NudC	222	154	0.000138	0.0572	9
SET domain containing 1A (EC:2.1.1.43)	63	38	0.000145	0.0577	9
HSP90B1	263	195	0.000179	0.0639	8
T-cell receptor alpha chain	90	32	0.000195	0.0639	4
T-cell receptor alpha chain V region CTL-L17 precursor	91	32	0.000195	0.0639	2
SEC11C; SPCS1; SPCS3	105	70	0.000198	0.0639	1
7SL RNA (ENST00000410687); 7SL RNA (ENST00000410707)	110	74	0.000201	0.0639	1
hKNL1/CASC5	135	74	0.000206	0.0639	8
MARCKS	99	81	0.000207	0.0639	8
Nup85	458	137	0.000214	0.0639	9
nasp_human	124	102	0.000234	0.0679	7
SPCS2	110	72	0.000245	0.0689	7
Protein disulfide isomerase P5	190	145	0.000285	0.0759	6
HDAC1	495	351	0.000291	0.0759	18
Centromere protein P	122	65	0.000309	0.0759	2
Centromere protein N	123	65	0.000309	0.0759	3
Centromere protein Q	123	65	0.000309	0.0759	3
SRP19	117	80	0.000325	0.0776	6
40S small ribosomal protein 21	155	109	0.000334	0.078	7
G-protein beta2 (GBB2) subunit	391	213	0.000372	0.0848	9
Chk2	55	48	0.000387	0.085	11
SEC11A	112	75	0.000391	0.085	7
SET domain containing (lysine methyltransferase) 8 (EC:2.1.1.43)	33	16	0.000427	0.0881	8
60S ribosomal protein L40	173	123	0.00044	0.0881	4
SRP54	112	76	0.000443	0.0881	4
peroxiredoxin 4 (EC:1.11.1.15)	261	177	0.000452	0.0881	8
G-protein beta 3 (GBB3) subunit	362	192	0.000454	0.0881	7
ADPRT	431	308	0.00046	0.0881	12
60S ribosomal protein L34	155	105	0.000481	0.0896	7
CD151	17	13	0.000488	0.0896	6
TERT	97	75	0.000497	0.0896	8
Cyclin-H	228	149	0.000522	0.0925	13
SRP72	119	80	0.000549	0.0954	7

Table A.1: NESTs significantly associated with metastatic prostate cancer, based on data by Yu et al. The table lists all NESTs found to be significantly active (assessed through the Wilcoxon signed-rank test, FDR threshold=0.1) in metastatic prostate cancer compared to primary carcinoma (expression data from (184)).

Enriched pathway based sets

pathway name	set size	measured genes	p-value	q-value	pathway source
Ribosome - Homo sapiens (human)	88	62	1.34e-05	0.00957	KEGG
Eukaryotic Translation Termination	89	61	2.05e-05	0.00957	Reactome
Formation of a pool of free 40S subunits	98	71	2.64e-05	0.00957	Reactome
Peptide chain elongation	89	62	2.76e-05	0.00957	Reactome
Eukaryotic Translation Elongation	92	65	2.88e-05	0.00957	Reactome
Signaling by GPCR	893	298	5.83e-05	0.0137	Reactome
RNA Polymerase I Promoter Opening	30	16	6.1e-05	0.0137	Reactome
E2F transcription factor network	120	62	6.63e-05	0.0137	PID
insulin	42	68	0.000103	0.019	INOH
Insulin Synthesis and Processing	115	75	0.000122	0.0191	Reactome
Cell Cycle, Mitotic	276	183	0.000127	0.0191	Reactome
GPCR downstream signaling	831	263	0.000162	0.0209	Reactome
IL5	51	50	0.00017	0.0209	NetPath
insulin Mam	49	66	0.000191	0.0209	INOH
GPCR ligand binding	392	205	0.000191	0.0209	Reactome
Neuroactive ligand-receptor interaction - Homo sapiens (human)	272	158	0.000202	0.0209	KEGG
PDGFR-beta signaling pathway	58	49	0.000219	0.0214	PID
Muscle contraction	49	40	0.000317	0.0284	Reactome
Non-small cell lung cancer - Homo sapiens (human)	54	49	0.000325	0.0284	KEGG
Class A/1 (Rhodopsin-like receptors)	291	144	0.000413	0.0342	Reactome
vegf hypoxia and angiogenesis	37	26	0.000465	0.0364	BioCarta
Sema4D induced cell migration and growth-cone collapse	24	20	0.000483	0.0364	Reactome
Signalling to ERKs	35	31	0.000591	0.042	Reactome
JAK STAT pathway and regulation	121	218	0.000608	0.042	INOH
Signalling to RAS	26	24	0.00065	0.0431	Reactome
role of nicotinic acetylcholine receptors in the regulation of apoptosis	18	13	0.000732	0.0467	BioCarta
GTP hydrolysis and joining of the 60S ribosomal subunit	110	80	0.000767	0.0471	Reactome
BARD1 signaling events	31	25	0.00103	0.0588	PID
Translocation of ZAP-70 to Immunological synapse	29	17	0.00107	0.0588	Reactome
Validated targets of C-MYC transcriptional activation	153	67	0.00107	0.0588	PID
Insulin signaling pathway - Homo sapiens (human)	136	105	0.00112	0.0588	KEGG
Translation	125	92	0.00115	0.0588	Reactome
L13a-mediated translational silencing of Ceruloplasmin expression	110	79	0.00121	0.0588	Reactome
3, -UTR-mediated translational regulation	110	79	0.00121	0.0588	Reactome
Diabetes pathways	224	149	0.00134	0.0636	Reactome
superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass	61	28	0.00155	0.0698	HumanCyc
Phosphorylation of CD3 and TCR zeta chains	31	18	0.00158	0.0698	Reactome
Smooth Muscle Contraction	22	21	0.0016	0.0698	Reactome
DNA Replication	155	102	0.00174	0.074	Reactome
VEGF signaling pathway - Homo sapiens (human)	75	52	0.00193	0.0754	KEGG
Eukaryotic Translation Initiation	118	86	0.00195	0.0754	Reactome
Cap-dependent Translation Initiation	117	86	0.00195	0.0754	Reactome
human cytomegalovirus and map kinase pathways	17	15	0.00201	0.0754	BioCarta
ABC transporters - Homo sapiens (human)	44	25	0.00203	0.0754	KEGG
Generation of second messenger molecules	42	27	0.00205	0.0754	Reactome
NGF signalling via TRKA from the plasma membrane	134	114	0.00228	0.0823	Reactome
Chronic myeloid leukemia - Homo sapiens (human)	73	65	0.00273	0.0952	KEGG
Mitotic M-M/G1 phases	131	84	0.00279	0.0952	Reactome
Signaling events mediated by VEGFR1 and VEGFR2	70	57	0.00281	0.0952	PID

Table A.2: Pathways significantly associated with metastatic prostate cancer, based on data by Yu et al. The table lists all manually defined pathways from ConsensusPathDB found to be significantly active (assessed through the Wilcoxon signed-rank test, FDR threshold=0.1) in metastatic prostate cancer compared to primary carcinoma (expression data from (184)).

APPENDIX

Abbreviations

AP-MS	affinity purification coupled to mass spectrometry
AUC	area under the (ROC) curve
CAPPIC	cluster-based assessment of protein-protein interaction confidence
DE genes	differentially expressed genes
FDR	false discovery rate
GO	Gene Ontology
GOSemSim	Gene Ontology semantic similarity
IMPaLA	integrated molecular pathway-level analysis
NC genes	NEST center genes
NEST	neighborhood-based entity set
PCA	protein-fragment complementation assay
ROC	receiver operating characteristic
Y2H	yeast two-hybrid

Software availability

The interaction meta-database **ConsensusPathDB** described in Chapter 2 is freely accessible through a web interface at http://cpdb.molgen.mpg.de. NEST analysis described in Chapter 4 is implemented in ConsensusPathDB's web interface. Web service access to part of the functionality of the ConsensusPathDB web interface is available; it is documented on the ConsensusPathDB web site and the WSDL[1] file is available at http://cpdb.molgen.mpg.de/download/CPDB.wsdl.

The **ConsensusPathDB plugin for Cytoscape** described in Chapter 2 can be installed through Cytoscape's plugin manager, category "Network and Attribute I/O"

The **CAPPIC web-based tool** described in Chapter 3 is freely accessible at http://cpdb.molgen.mpg.de/cappic. The source code implementing CAPPIC is available on the web page.

The **IMPaLA web-based tool** for pathway-based analyses of large-scale gene expression and/or metabolomics data is available at http://impala.molgen.mpg.de. Web service access to IMPaLA is available; the WSDL[1] file is available at http://impala.molgen.mpg.de/download/IMPALA.wsdl and is documented on the IMPaLA web site.

[1]WSDL: web service definition language

CURRICULUM VITAE

Personal details have been omitted in the electronic version of the dissertation for data protection reasons.

Personal details have been omitted in the electronic version of the dissertation for data protection reasons.

Publications

Hegele, A.[1], **Kamburov, A.**[1], Grossmann, A., Sourlis, C., Wowro, S., Weimann, M., Will, C., Pena, V., Lührmann, R. and Stelzl, U. (2012) *Dynamic protein-protein interaction wiring of the human spliceosome.* Molecular Cell, 45: 567-580.

Kamburov, A.[1], Cavill, R.[1], Ebbels, T., Herwig, R. and Keun, H. (2011) *Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA.* Bioinformatics, 27: 2917-2918.

Dreher, F., **Kamburov, A.** and Herwig, R. (2011) *Construction of a pig physical interactome using sequence homology and a comprehensive reference human interactome.* Evolutionary Bioinformatics Online, 8: 119-126.

Yildirimman, R., Brolén, Vilardell, M., Eriksson, G., Synnergren, J., Gmuender, H, **Kamburov, A.**, Ingelman-Sundberg, M., Castell, J., Lahoz, A., Kleinjans, J., van Delft, J., Petter Björquist, P. and Herwig, R. (2011) *Human embryonic stem cell derived hepatocyte-like cells as a tool for in vitro hazard assessment of chemical carcinogenicity.* Toxicological Sciences, 124: 278-290.

Cavill, R., **Kamburov, A.**, Ellis, J., Athersuch, T., Blagrove, M., Herwig, R., Ebbels, T. and Keun, H. (2011) *Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells.* PLoS Computational Biology, 7:e1001113.

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H. and Herwig, R. (2011) *ConsensusPathDB: toward a more complete picture of cell biology.* Nucleic Acids Research, 39:D712-717.

Pentchev, K., Ono, K., Herwig, R., Ideker, T. and **Kamburov, A.** (2010) *Evidence mining and novelty assessment of protein-protein interactions with the ConsensusPathDB plugin for Cytoscape.* Bioinformatics, 26:2796-2797.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C.O., Forrest, A.R., Gough, J., Grimmond, S., Han, J.H., Hashimoto, T., Hide, W., Hofmann, O., **Kamburov, A.**, Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C.R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R.D., Tegnr, J., Lenhard, B., Teichmann, S.A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D.A., Ideker, T. and Hayashizaki, Y. (2010) *An atlas of combinatorial transcriptional regulation in mouse and man.* Cell, 140:744-752.

[1]Joint first authors.

Kamburov, A., Wierling, C., Lehrach, H. and Herwig, R. (2009) *ConsensusPathDB - a database for integrating human functional interaction networks.* Nucleic Acids Research, 37:D623-628.

Kamburov, A., Goldovsky, L., Freilich, S., Kapazoglou, A., Kunin, V., Enright, A., Tsaftaris, A. and Ouzounis, C. (2007) *Denoising inferred functional association networks obtained by gene fusion analysis.* BMC Genomics, 8:460.

Manuscripts currently submitted or in preparation:

Kamburov, A., Grossmann, A., Herwig, R. and Stelzl, U. *Cluster-based assessment of protein-protein interaction confidence.* Submitted.

Kamburov, A., Stelzl, U. and Herwig, R. *IntScore: a web tool for confidence assessment of biological interactions.* Submitted.

Li, J., Maschke-Dutz, E., Heeger, F., **Kamburov, A.**, Kühn, A., Hache, H., Lehrach, H., Herwig, R. and Wierling, C. *PyBioS: a tool for designing, modeling and simulating cellular systems.* In preparation.

Zusammenfassung

Die menschliche Zelle umfasst eine große Menge verschiedener Biomoleküle wie Nukleinsäuren, Proteine und Metabolite. Diese Biomoleküle erfüllen ihre Funktionen nicht isoliert, sondern durch ein komplexes Zusammenspiel untereinander. Erkenntnisse über die Gesamtheit der molekularen Wechselwirkungen, die in der Zelle stattfinden, ist unentbehrlich für das Verständnis zellulärer Prozesse auf der Systemebene. Zum Beispiel können molekulare Interaktionen oft erklären, wie Funktionsstörungen bestimmter Gene etwa durch Mutation zu einer bestimmten Krankheit führen. Gerade wegen diesem Aufklärungspotential molekularer Wechselwirkungen wurden zu ihrer Identifizierung unterschiedliche Techniken entwickelt. Viele molekulare Interaktionen in der menschlichen Zelle sind bereits entdeckt und veröffentlicht worden, wenngleich sie schätzungsweise nur einen kleinen Teil der wirklich existierenden Wechselwirkungen darstellen. Diverse Datenbanken sind entwickelt worden um Interaktionsdaten, die zum Beispiel über Datamining gewonnen werden, systematisch zu sammeln. Vorhandene Interaktionsnetzwerke werden bereits in verschiedenen Methoden eingesetzt, die zum Ziel haben, neue Erkenntnisse über krankheitsrelevante Gene, Stoffwechselwege und Signalwege zu gewinnen.

Ein tieferes Verständnis über normale und krankheitsbedingte zelluläre Prozesse auf der Systemebene ist allerdings durch zwei weitere Hauptfaktoren (neben der Unvollständigkeit vorhandener Interaktionsdaten) stark eingeschränkt. Zum einen sind solche Daten in der Regel fehlerhaft, das heißt, sie enthalten viele falsch positive Interaktionen. Diese entstehen meistens durch Fehler bei den experimentellen Messungen oder gegebenenfalls beim Datamining. Zum anderen sind vorhandene Daten in Hunderten von Datenbanken verstreut, wobei jede Datenbank Interaktionen nur einer oder weniger Arten enthält: manche Datenbanken enthalten ausschließlich Proteininteraktionen, während andere auf Genregulationen, metabolische Reaktionen oder Signalwege spezialisiert sind. In der Zelle wirken all diese Arten von Interaktionen zusammen um biologische Prozesse zu treiben. Interaktionsdatenbanken müssen also integriert werden, damit ein vollständigeres Modell der zellulären Biologie entsteht. Eine solche Integration ist

dadurch erschwert, dass die einzelnen Datenbanken sehr unterschiedliche Datenmodelle und -formate haben.

Diese Dissertation beschäftigt sich mit den Herausforderungen, dass vorhandene Interaktionsdaten zum einen fehlerhaft sind und zum anderen in vielen, wenig überlappenden Datenbanken zerstreut sind.

Zuerst wird eine neue Metadatenbank für molekulare Wechselwirkungen namens ConsensusPathDB vorgestellt. Hier werden unterschiedliche Arten von Interaktionen aus vielen öffentlichen Ressourcen integriert um ein vollständigeres Bild der molekularen Wechselwirkungen in der menschlichen Zelle zu erzielen. Zur Zeit sind Wechselwirkungen sowie Signal- und Stoffwechselwege aus sechsundzwanzig öffentlichen Ressourcen in der Metadatenbank integriert. Deshalb stellt das in der ConsensusPathDB vorhandene Interaktionsnetzwerk das umfangreichste Modell der Wechselwirkungen in der humanen Zelle dar. Der Mehrwert der Datenintergation wird anhand einiger Beispiele veranschaulicht. Die Webschnittstelle der Datenbank (http://cpdb.molgen.mpg.de) bietet zahlreiche Tools für Datensuche, Netzwerkanalyse und -visualiserung, sowie Interaktions- und Pathway-basierte Analysen von Genexpressionsdaten. Diese stellen wichtige Hilfsmittel für Biologen und Molekularmediziner dar.

Zweitens wird eine neue Methode vorgestellt, mir der Proteininteraktionen bezüglich ihrer Richtigkeit beurteilt werden. Die resultierenden Konfidenzwerte können benutzt werden um falsch positive Interaktionen zu detektieren, oder können als Interaktionsgewichte in netzwerkbasierten Methoden fungieren. Im Gegensatz zu vielen anderen Methoden werden hier keine Referenzdatensätze oder zusätzliche Informationen über die einzelnen Netzwerkelemente benötigt. Solche Daten sind oft nicht vorhanden, was vergleichbare Methoden zur Konfidenzwertbestimmung limitiert. Die vorgeschlagene Methode benutzt ausschließlich die Netzwerkstruktur, im Speziellen ihre Modularität, um die Konfidenzwerte zu berechnen.

Drittens wird ein zugleich vollständigeres und akkurateres Modell zellulärer Wechselwirkungen erstellt, indem die vorgestellte Konfidenzwert Methode auf die integrierten Daten aus ConsensusPathDB angewandt wird. Von dem resultierenden Netzwerk wird in einem neuen Verfahren zur Identifizierung von krankheitsrelevanten Genen und Subnetzwerken unter Berücksichtigung von Genexpressionsprofilen Gebrauch gemacht. Das integrative Verfahren wird auf Genexpressionsdaten aus Prostatakrebspatienten angewandt um sein Potential zu demonstrieren, Krebsgene richtig zu erkennen.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, 8. August 2011