FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Simulation in Metabolic Networks

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der technischen Wissenschaften

eingereicht von

Christian Siehs
Matrikelnummer 9203648

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Ao. Univ.-Prof. Dr. Rudolf Freund

Diese Dissertation haben begutachtet::

_____ _____

(Ao. Univ.-Prof. Dr. Rudolf Freund) (Univ.Doz. Dr. Bernd Mayer)

_____

Wien, 28.09.2015 (Christian Siehs)

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Simulation in Metabolic Networks

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der technischen Wissenschaften

by

Christian Siehs
Registration Number 9203648

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao. Univ.-Prof. Dr. Rudolf Freund

The dissertation has been reviewed by:

_____ _____

(Ao. Univ.-Prof. Dr. Rudolf Freund) (Univ.Doz. Dr. Bernd Mayer)

_____

Wien, 28.09.2015 (Christian Siehs)

Erklärung zur Verfassung der Arbeit

Christian Siehs
Keinergasse 20/1/12, 1030 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 28.09.2015	Christian Siehs

“The way to success has no short cuts“

Masahiko Tanaka

Acknowledgements

I would like to express my deep gratitude to Dr. Rudolf Freund and Dr. Bernd Mayer, my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this dissertation work. I would also like to thank all my scientific colleaques who have contributed to the parts of this work during the last years.

Finally, I wish to thank my parents for their support and encouragement throughout my studies.

Abstract

Molecular networks pose a challenge to biomedical research due to the complex interaction patterns and the comparably slow progress in identifying and characterizing the biological entities involved. Advances in high-throughput technologies for identifying and quantifying biological entities relevant for describing clinical phenotypes have narrowed the data gap over the last years, resulting in a large amount of heterogeneous biological data.

Analysis of this data landscape in a molecular mechanistic context, supported by data management and knowledge discovery technologies, has in the meantime provided network based approaches for discovery of novel diagnostics and therapeutics for addressing unmet clinical needs. This work presents and discusses extensively the advances in molecular network analysis specifically in the realm of Systems Medicine over the last years and gives an overview on network principles, data sources and representation standards and analytical methodologies.

Scientific contributions presented in this thesis address specifically (i) modeling of protein-protein interaction networks, (ii) tackling the false negative discovery rate of explorative omics experiments by proposing concepts of predicting relevant biological entities based on graph expansion algorithms and (iii) the practical analysis of omics data utilizing network-based methodologies.

The proposed prediction concepts for biological entities take into account our current knowledge of the topology of molecular networks, and are based on vertex neighborhood as well as minimum spanning trees. As a result, a novel algorithm was developed for efficiently calculating a connecting spanning tree over a subset of selected vertices of a graph with low but not necessarily minimal overall edge weight. These prediction concepts as well as other network based methodologies were then applied to the analysis of omics data characterizing ovarian cancer, mesothelial cells stress response, and angiogenesis in brain metastasis, providing additional insight regarding their underlying molecular mechanisms.

The complexity of network analysis remains challenging, nurturing research in life sciences as well as computer science. While the analysis of omics data with network methodologies sheds light on molecular mechanisms, the insights gained on natural network topologies and mechanisms further the design and analysis of networks in general. Analysis concepts and algorithms presented in this work serve as general approaches for omics data integration and analysis in the biomedical research area.

Kurzfassung

Molekulare Netzwerke stellen seit jeher aufgrund ihrer komplexen Interaktionsmuster und des vergleichbar langsamen Fortschritts bei der Identifizierung und der Charakterisierung von biologischen Entitäten für die biomedizinische Forschung eine Herausforderung dar. Fortschritte im Bereich der Hochdurchsatzverfahren zur Identifizierung und Quantifizierung von biologischen Entitäten mit Relevanz für die Beschreibung von klinischen Phenotypen haben die Datenkluft über die letzten Jahre verkleinert. Daraus resultieren große Mengen an heterogenen biologischen Daten.

Die Analyse dieser Datenlandschaft in einem molekularen mechanistischen Kontext, unterstützt durch Technologien des Datenmangements und der Wissensgewinnung, hat in der Zwischenzeit netzwerkbasierte Methoden für die Entdeckung von neuen Diagnostika und Therapeutika für den klinischen Bedarf hervorgebracht. In dieser Arbeit werden die Fortschritte der letzten Jahre im Bereich molekularer Netzwerke im Detail und speziell für die Systemmedizin präsentiert. Des weiteren gibt sie einen Überblick über Netzwerkprinzipien, Datenquellen sowie Datenstandards und analytische Methoden.

Die in dieser Arbeit vorgestellten wissenschaftlichen Beiträge addressieren spezifisch (i) Modellierung von Protein-Protein Interaktionsnetzwerken, (ii) Strategien für die falsch-negative Entdeckungsrate von explorativen Omics-Experimenten unter Vorstellung von Konzepten zur Vorhersage weiterer relevanter biologischer Entitäten, basierend auf Graphenexpansionsalgorithmen und (iii) die angewandte Analyse von Omics-Daten mittels netzwerkbasierten Methoden.

Die vorgestellten Konzepte zur Vorhersage von biologischen Entitäten berücksichtigen unser derzeitiges Wissen über die Topologie von molekularen Netzwerken und basieren auf Nachbarschaft von Knoten, sowie minimalen Spannbäumen. Daraus resultiert ein neuer Algorithmus zur effizienten Berechnung eines verbindenden Spannbaums über ein Subset von Knoten eines Graphen mit geringem, aber nicht notwendigerweise minimalem Kantengewicht. Diese Vorhersagekonzepte wurden gemeinsam mit weiteren netzwerkbasierten Methoden für die Analyse von Omics-Daten zur Charakterisierung von Ovarialkarzinom, Stressantwort in Mesothelzellen und Angiogenese in Hirnmetastasen angewandt, welche zusätzliche Erkenntnis zu deren zugrundeliegenden Mechanismen hervorbrachte.

Die Komplexität von Netzwerkanalysen ist nach wie vor eine Herausforderung und bietet ausgiebige Forschungsmöglichkeiten sowohl für die Lebenswissenschaften als auch die

Computerwissenschaften. Während die Analyse von Omics-Daten mittels Netzwerkmethoden Licht auf die molekularen Mechanismen wirft, erweitern die Erkenntnisse über Topologien und Mechanismen natürlicher Netzwerke das Design und die Analyse von Netzwerken im Allgemeinen. Die hier angewandten Analysekonzepte und Algorithmen stellen eine allgemeine Basis für die Omics-Datenintegration und die Analyse in der biomedizinischen Forschung dar.

Table of Contents

Introduction

The complexity of networks and systems has always been a fundamental challenge for scientists throughout many different academic fields. Over the last decade, and with the support of computer-aided technologies and the possibility to measure, administer and analyze huge sets of data, scientists have been rising to the challenge to investigate complex networks and systems [1], [2]. Among the many different areas, biological networks and systems, especially ecosystems and cellular systems, have been the most complex. With the advent of high-throughput technologies for measuring molecular biological entities over the last two decades, a huge amount of interconnected molecular data has been gathered. This influx on data, starting with the sequencing of the human genome, which was completed in 2003 [3], gave rise to the term genomics and many other '-omics' soon followed, like proteomics and metabolomics, each referring to the study of an aspect (or subsystem) in molecular biological systems. The rapid generation of biological data in any of the omics has led to a very heterogeneous data source landscape with different concepts and data formats. Over 500 data sources are available for network-related data alone [4]. This explosion of the amount of data has brought up many challenges on the level of integrating data, standardization and analysis [5] but harbors promising possibilities on the level of personalized health care and clinical pharmacology [6].

Huge amounts of data alone, though, will not solve problems and questions regarding complex networks and systems. The way we analyze them and infer our knowledge is fundamental and the first question arising is: "What makes a network and a system complex?" To answer the question we have to define the levels of difficulty of problems, an issue which has been addressed by scholars ever since Aristotle and Plato and the definitions vary [7]. Usually the definitions center on the relation between cause and effect. Simple problems have solutions that can be directly repeated and are not difficult to understand, like following a cooking recipe. Complicated problems are problems that are not easy to understand having complicated causes but once solved, the effect can be easily reproduced, like Dijkstra's algorithm for finding shortest paths within a graph [8]. Complex problems are difficult to solve and reproducing an effect is difficult as well, like the behavior of ecosystems. The difficulties in reproducing solutions of complex systems originate from the often non-linear relations between the elements and the topology of such a system's network, causing feedback loops and leading to a myriad of emergent properties. Those are a phenomenon already described by Aristotle who stated that the whole is more than the sum of its parts. The final group of problems are the chaotic ones which seem to behave completely random when changing starting parameters of a deterministic rule set by a small amount, like Lorenz attractors in weather models [9].

The major goals of the study of complex networks and systems, from a practical point of view, are to get them more predictable and to discern patterns of regularity which can be reproduced or even simulated. The sheer amount of possibilities how a complex system can behave still leads to the question of how to explore and analyze complex networks and systems deriving knowledge with practical implications. This question is directly linked to the philosophy of scientific methodology itself. Life sciences have been relying heavily on reductionism which aims at reducing and explaining complex phenomena in terms of their parts. For example, biological processes can be reduced to chemistry which itself can be explained by physics. As Stephen Weinberg did put it: "explanatory arrows always point downward" and ultimately lead to a final theory [10]. The reductionist approach can also be found in computer science and problem solving strategies. The 'divide and conquer' principle is an important algorithm design paradigm dividing a problem into sub-problems until the sub-problems become trivial to solve and build back together. An early algorithm based on this principle is Euclid's algorithm to compute the greatest common divisor and dates back to the 3rd century B.C. While reductionism and the hypothesis-driven knowledge discovery process have been proven very successful, the reductionist approach has its limitations, since it often fails to provide reliable explanations for an emergence of properties and complex system behavior on the way back upward from the 'basic building blocks'. Holism is the diametrically opposed philosophical point of view originating back to Aristotle, that the whole is more than its parts. Bertalanffy's System Theory [11], [12] is based on holistic principles which have been spreading ever since when investigating and analyzing complex networks and systems [13], [14]. So far, hypothesis-driven knowledge discovery based on reductionist principles has been dominating life sciences, and relies on testable hypothesis. Especially within complex systems those hypotheses are not testable most of the time. Inductive, computer-aided data-driven knowledge discovery addresses this issue and is also based on holistic principles and is discussed by Leonelli in the context of biological and biomedical sciences [15]. While we might not be able to formulate testable hypotheses for complex systems, we can observe patterns within our data pool without knowing their building blocks. Evaluating the patterns (for example with further experiments) and the data generated by the evaluation is added up to the data pool. Independent of the scientific approaches our knowledge about the mechanisms and concepts in life sciences, in general, and biological networks, in special, is far from complete and the unknown is a part of any topic which gets investigated. With observations and experiments we are creating data and derive models, test them and further our knowledge of life sciences.

In this work, an extensive summary of the state of the art of molecular networks are presented in chapter 1 beginning with an overview of network principles, followed by a description of different biological network types. Prominent data sources for biological networks are described next, focusing on the data concepts and presenting data models where available. Afterwards, modeling of biological networks is extensively discussed along data representation standards. Network analysis approaches and current network characteristics and knowledge follow thereafter. Chapter 1 concludes with a discussion of issues with the study of biological networks. Building on the presented state of the art from chapter 1, chapter 2 presents further scientific work from the author, which resulted in publications, as well as unpublished contributions. These contributuions are discussed in context of the state of the art as presented in chapter 1. This work concludes with a summary of the lessons learned and remaining challenges when studying biological networks.

CHAPTER 1
On networks

Introduction

Networks consist of entities forming relations with each other. For instance human relations can be represented as networks where human beings form the entities and the family status between them form the relations as like when forming a genealogical tree. The types of human relations are manifold. Considering social networks the relation between two people is defined by knowing each other and having linked themselves. Similarly personal phonebooks can be represented as network where each phone number represents a relation between the phonebooks owner and the phone number owner. This relation can be one sided as the individual phone number owner does not necessarily need to have the phone books owners phone number as well. Networks entities need not be limited to physical objects but also on human based concepts like street crossings or cities where the roads form the relations with the meaning of connecting two given cities or street crossings thus forming a street network. On a molecular scale we find proteins interacting with each other thus forming a network of protein-protein interactions. Networks exist in all aspects of human life and nature including economics, sociology, computer science, physics, biology, medicine and many more [1], [2].

Networks pose many challenges and problems that are associated with them, like finding the shortest route from one location to another for which algorithms, like the Bellman-Ford algorithm [16], [17] or Dijkstra's algorithm [8] have been developed. Calculating the minimal amount of phone lines required for connecting houses, formally creating a so called *minimal spanning tree*, is solved with the algorithms from Kruskal [18] and Prim [19]. Historically one typical network problem was the 'seven bridges of Königsberg' problem which was addressed by L. Euler [20]. He proved that there cannot be a route crossing each of the 7 bridge over the river Pregel while entering each of the 4 city sections only once (see Figure 1). The list goes on and for studying many of such network related problems graph theory, a discipline within mathematics, is used.

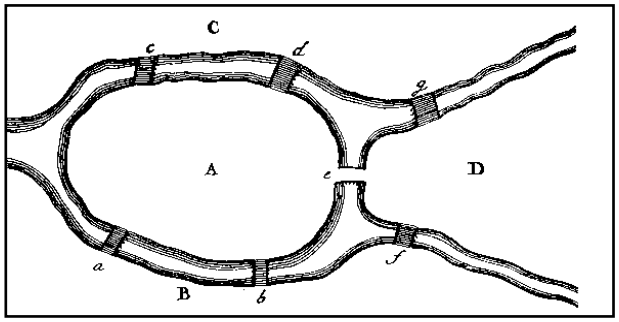


Figure 1: The seven bridges of Königsberg. Original drawing by L. Euler [Online]. http://www.matheprisma.uni-wuppertal.de/Module/Koenigsb/ [Accessed July 2014]

Formally describing networks with graph theory a network consists of nodes (or vertices) resembling the network entities which are connected by edges resembling the relations between the entities, together forming a graph. Thus a graph *G* is a pair of sets (*V*, *E*) where *V* is the set of nodes and *E* is the set of edges, formed by pairs of nodes. Nodes and edges have several core properties out of which many different graph types can be constructed and described (see Figure 2). Edges can be directed leading only in one direction, like a one-way street, or undirected leading in both directions. Edges and nodes can have one or more properties associated with them, like the names of cities and streets. Edges can have weights resembling different properties of the relations like the distance between two cities. Graphs can also evolve over time removing or adding nodes or edges like persons dropping out or entering social networks. Properties of edges, their weights and directions can change over time like the family status between persons.

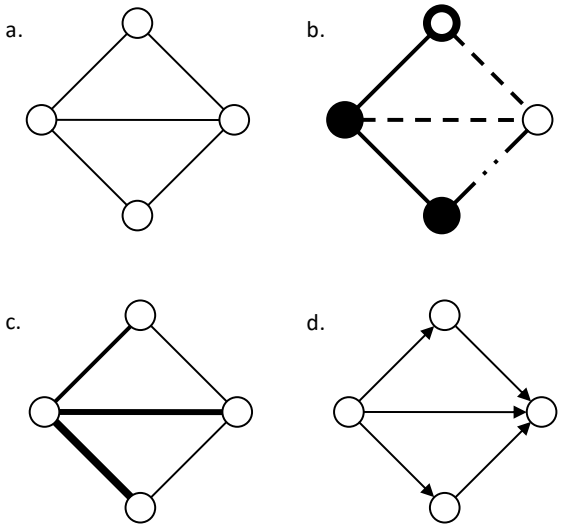


Figure 2: Graphs with different properties. (a.) Graph with single node type and single undirected edge type with a single edge weight. (b.) Graph with multiple node types and multiple undirected edge types and a single edge weight. (c.) Graph with a single node type and single undirected edge type and multiple different edge weights. (d.) Graph with a single node type and a single directed edge type and a single edge weight.

Depending on the arrangement of the nodes and edges, graphs can have many different characteristics like the graph diameter or degree distribution. The graph diameter is the length of the longest shortest path between any two nodes within a given graph. This would reflect for instance the longest travel distance between any two cities on the world assuming the shortest route is always chosen. The degree of a node is the amount of edges it has to other nodes and the degree distribution is the probability distribution of those degrees for the graph. Interestingly graph characteristics are often specific for certain types of networks like tumor networks [21]. The degree distributions in many real world networks, like the World Wide Web, very often do follow a power law [22].

Studying networks spans over many different academic fields and as such it is highly interdisciplinary. *Network science* as defined by the United States National Research Council is the study of network representations of physical, biological and social phenomena, leading to predictive models of these phenomena [23]. When studying networks we are interested in solving problems like finding the shortest path or characterizing and differentiating networks based on their graph properties, like characterizing tumor networks or disease networks [24]. Adding the dimension of time to networks and adding functions over time affecting network properties we can investigate the dynamics and evolvement of a network we are interested in, like the spreading of a disease during an epidemic or within a social network [25]. Networks changing over time describe systems and are the focus of *systems theory*, a term ascribed to Bertalanffy [11], [12], out of which the sub-discipline of *systems biology* emerged. The importance of studying the dynamics within the system of the human body became evident and resulted in the sub-discipline *systems medicine* [26]. The European Commission's funding for systems medicine related projects from 2004 to 2010 included more than 60 research projects with a total amount of ca. 400 million euro [27], like the CASyM project for developing an implementation strategy for Systems Medicine [28].

This chapter begins with a description of common molecular biological networks followed by an extensive list of molecular biological network data sources available and a brief description of the type of data, their origins, how to access them and technical concepts and frameworks which have been used for their implemention. Afterwards the modeling strategies of biological networks will be extensively presented as well as data standards commonly used with biological networks. The analysis of biological networks will list major conceptual aims when analyzing networks and present properties and characteristics of biological networks. Extensive examples are given of how this knowledge can be utilized especially in the context of human diseases. Finally this chapter concludes with a discussion of major issues within this academic field.

Molecular biological network types

Networks in biology consist of many different types of biological entities that can interact with each other. Proteins, nucleic acids and small-molecules form the basic biological entities and their direct interactions are based on two principles: chemical reactions and affinity docking. Chemical reactions are the molecular modification of usually covalent bonds between atoms changing substrate molecules to product molecules and vice versa with a reaction kinetic for each direction which is dependent on environmental factors, like temperature (see Figure 3).

$$A + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C + D$$

Figure 3: Basic scheme for chemical reactions. Substrates A and B are modified to products C and D with a certain reaction kinetic k_1 and vice versa with the reaction kinetic k_2.

Affinity docking between biological entities is based on matching 3-dimensional structures similar to a key-lock mechanism thus allowing weaker forces, hydrogen bonds, ionic charges and Van der Waals forces, to form stable non-covalent interactions like in protein-protein interactions. Besides networks based on physical interactions of biological entities, interactions within a biological network can also be based on other, often abstract, concepts. For example 'association' can be a concept in gene-disease networks where genes are connected to disease entities if known associations exist between them. As such depending on the level of observance many different networks can be identify within biology.

The following chapter describes common biological networks differing by the type of interactions and entities including:

- protein- protein interaction networks
- metabolic networks
- genetic/ transcription regulatory networks
- miRNA regulatory networks
- gene-gene interaction networks
- drug-target interaction networks
- disease / gene – disease networks

Application examples are given and additionally an extensive list of data sources available for biological networks data is provided for. Statistics on data size as well as core technical concepts, data models and computer technologies used are included for the data sources wherever the information was readily available.

Protein interaction networks

Protein interaction networks (PIN) or protein-protein interactions (PPI) describe the physical interactions between proteins where the proteins form the nodes and the interactions form the edges of the network graph. Examples of protein interactions include signal transduction where signals are propagated throughout the cell by various signaling molecules. Disturbances or mutations within such a signaling cascade like the MAPK signaling pathway (see Figure 4) are known causes for cancer development [29].

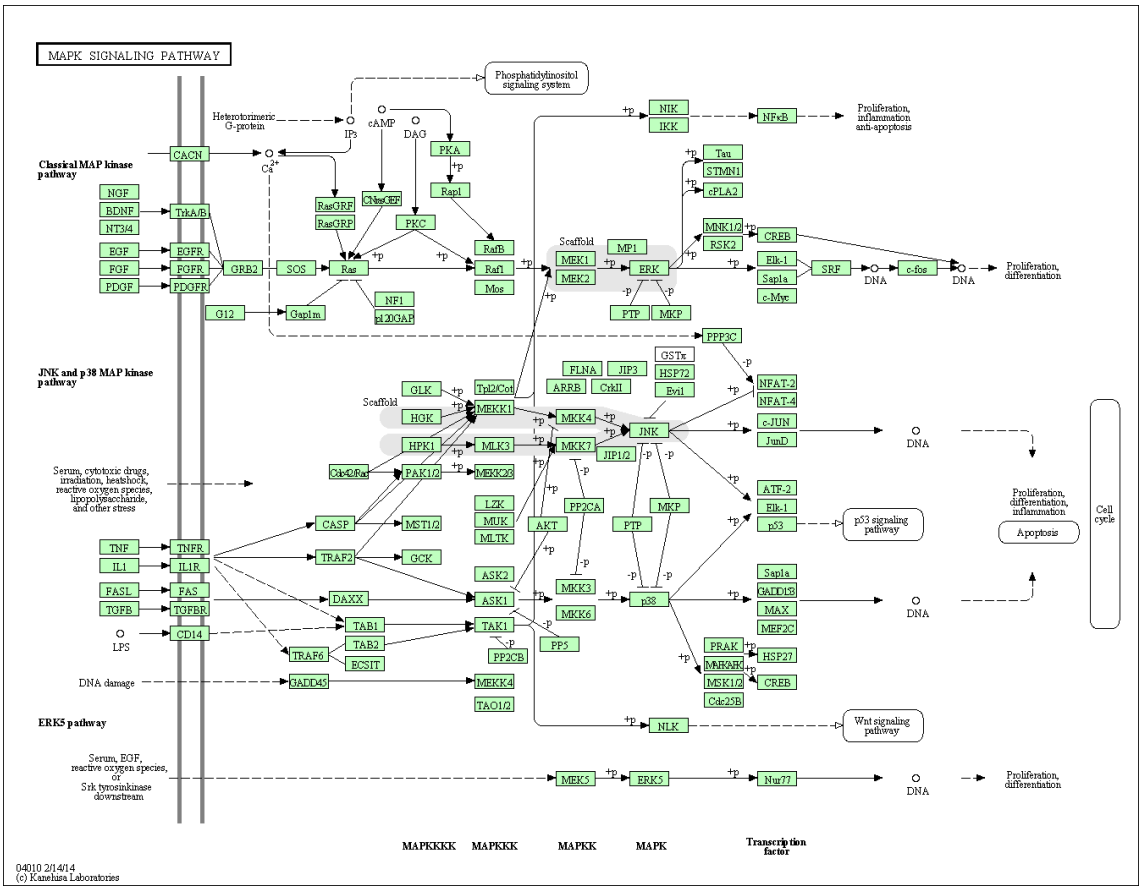


Figure 4: MAPK signaling pathway. [Online]. Available: http://www.genome.jp/kegg-bin/show_pathway?hsa04010. [Accessed August 2014].

Protein-protein interactions associate physically and lead to the formation of macromolecular structures and protein complexes like enzyme-inhibitor or antibody-antigen complexes. Proteins interact in pairs to form dimers like the reverse transcriptase or in sets of multiple proteins forming complex structures like the complement system. In most cases protein complex binding forces are non-covalent resulting from matching protein surfaces. Covalent bindings are known for example in posttranslational modifications of proteins like insulin where two polypeptide chains are connected by disulfide bonds. Protein-protein interactions are further regulated among others by protein concentration which is regulated by gene

expression and the degradation rates of proteins. Regulation is also strongly influenced by other proteins and protein types competitively docking with each other like the Bcl-2 family proteins for regulating apoptosis [30].

Due to the central role of proteins in the cellular mechanisms protein interaction networks serve as tool for investigating many aspects of cellular life. Proteins highly connected with other proteins seem to be fundamental for survival [31]. Protein interactions found in one species support the identification of interactions of similar or evolutionary linked proteins in another [32]. Protein interaction networks also support unraveling the molecular mechanisms of diseases [33], [34]. Jonsson and Bated [35] showed that cancer related genes had in average twice as many protein-protein interaction partners as non cancerous genes. Paik et al. [36] investigated protein-protein interactions for underlying co-occurrences of diseases and pathological conditions.

Methods for detecting protein-protein interactions

There are many methods for detecting protein-protein interactions ranging from experimental identification to prediction methods. Experimental methods include among others co-immunoprecipitation [37], the yeast two-hybrid screening [38], affinity purification coupled to mass spectrometry [39], protein microarrays [40], tandem affinity purification [41] and bimolecular fluorescence complementation [42]. Protein-protein interaction prediction can be based on homology transfer, where protein-protein interactions of one organism are used to predict interactions of homologous proteins in another [43]. Predictions based on test mining are another approach where protein-protein interaction prediction is based for example on co-occurrence of terms within PubMed abstracts [44] or natural language processing and rule based information extraction as well as machine learning approaches [45], [46].

Metabolic networks

Metabolic networks are a typically described by metabolic pathways like the citrate cycle (see Figure 5) forming a dense network of chemical compounds, enzymes and reactions defining the biochemical properties of a cell.

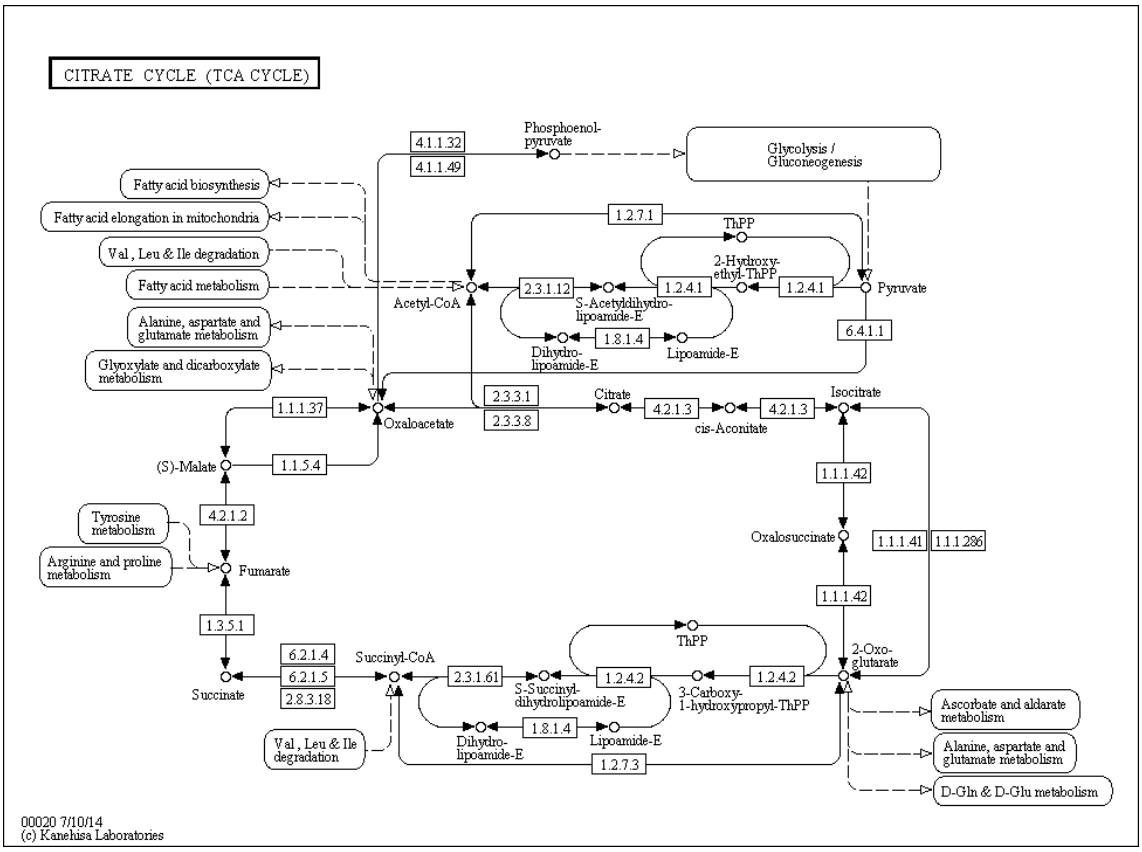


Figure 5: Metabolic pathway of the Citrate Cycle.[Online]. Available: http://www.genome.jp/kegg-bin/show_pathway?map00020. [Accessed August 2014].

Metabolism is governed by several biological entities, the chemical compounds known as metabolic substrates and products and enzymes catalyzing reactions. Processes within metabolism and the interactions are diverse as well. Biochemical reactions known as metabolic reactions producing the product from the substrate, often catalyzed by enzymes, docking of proteins and also regulation of metabolic steps are further different types of interactions. As such metabolic networks can be represented differently depending on the level of observance. Enzymes, metabolites and chemical reactions can be represented as nodes while the different interactions like mass flow, catalysis and regulation as edges. For visualization reasons these metabolic maps usually include multiple entries for a single chemical compound (see Figure 5, thiaminediphosphate: ThPP) which have to be modeled as a singular node within a graph representation of the metabolic network. Metabolic pathways have been studied, for example by Jeong et al. [47], Fell and Wagner [48] and Stelling et al. [49].

Genetic/Transcription regulatory networks

The expression of genes can be regulated by other proteins called transcription factors which can activate or inhibit a gene's expression. This regulation is fundamental for a cells function controlling the amount of proteins as required by the ongoing cellular processes. Depending on the graph representation of the involved entities and relations, proteins can form the nodes within genetic regulatory networks and directed edges can indicate if one protein influences another ones gene expression. Another representation includes two types of nodes, transcription factors and mRNA and two types of directed edges, regulation of transcription and translation. Initially transcription regulatory networks for *Saccharomyces cerevisiae* have been thoroughly investigated for example by Lee et al. [50], Farkas et al. [51] and Guelzim et al. [52].

Methods for constructing genetic regulatory networks

Construction of genetic regulatory networks is a difficult process, especially in eukaryotes with their complicated transcriptional regulation machinery [53]. The entities playing a role within regulation have to be identified, changes over time have to be monitored on the expression level and the impact on the phenotype needs to be detected. Experimental techniques supporting this process are chromatin immunoprecipitation (ChIP) combined with promoter DNA microarrays (known as ChIP on chip), gene expression profiling often combined with gene knock outs and genome-wide RNA interference (RNA1) screens. Computational approaches for predicting regulatory elements is challenging as well and most are based on clustering and supervised learning approaches [54]. Prediction of transcription factor binding sites was addressed for example by Perco et al. [55] who developed a genetic algorithm for detection of co-regulation including data from phylogenetic footprinting. Inferring networks was addressed by Beer and Tavazoie [56] who developed a method where they cluster genes from expression data into groups of regulons, identify over-expressed sequences and use Bayesian networks to deduce the relationships between expression profiles and sequence motifs. The difficulty of predicting genetic regulation is further emphasized by the DREAM initiative opening a challenge for best prediction algorithms in 2008 [57] with a best performer award for the algorithm from Gustafsson et al. [58] based on ordinary differential equations.

miRNA regulatory networks

MicroRNA (miRNA) are short RNA sequences (usually around 22 nucleotides) and seem to regulate gene expression on the level of translation by docking with mRNA and blocking translation as well as speeding up the degradation of the mRNA [59], [60]. As such the direct physical interaction for the regulation of a gene product takes place between the miRNA and mRNA. Usually when inferring regulatory networks, miRNA is linked to the target mRNA gene. miRNA has successfully been indentified for influencing several human diseases like skeletal growth [61] or cancers [62], [63].

Methods for constructing miRNA interactions

Experimental detection of miRNA-mRNA interaction is usually done on the expression level utilizing PCR or microarray technologies, often in combination with integrated mRNA and miRNA expression profiling approaches [64], [65]. Detection and prediction of transcription factor binding sites for miRNA genes are usually performed with chromatin

immunoprecipitation coupled with next-generation DNA sequencing (ChIP-Seq). A recent protocol termed cross-linking ligation and sequencing of hybrids (CLASH) allow high throughput screening of RNA-RNA based on UV-linked UV bait proteins. RNAs associated with the bait protein are linked, cDNA libraries prepared and after high throughput sequencing the resulting chimeric cDNAs identify RNA-RNA interactions [66]. Prediction of direct miRNA-mRNA interactions are usually sequence based combined with filtering based on structural knowledge, like binding energy, structural accessibility or nucleotide composition flanking in the binding sites [67]. Data mining technologies are also applied as demonstrated by McDermont et al. [68] who identified miRNA biomarkers for Luminal A-like Breast cancer utilizing neural networks.

Gene-gene interaction networks

Gene-gene (or genetic) interaction networks differs from gene regulatory networks in such as that a gene-gene interaction can be defined as logical interaction between two or more genes that affect the phenotype of an organism. Gene-gene interactions types are synthetic-interaction, epistatic interaction, suppressive interaction and additive interaction. Synthetic interactions between two genes exist if a mutation in either of the two does not have an effect on the phenotype while a mutation in both genes does have an effect on the phenotype. Usually both genes are on parallel pathways leading to the same relevant target gene for the phenotype (see Figure 6a). Two genes have an epistatic interaction if a mutation in one of the genes results in one of two different phenotypes while mutations in both results again in one of the two phenotypes. This usually occurs if one of the genes precedes the other in the same pathway (see Figure 6b.). Suppressive interactions between two genes exist if a mutation in one gene does not change the phenotype, a mutation in the other gene however does change the phenotype and a mutation in both genes does not change the phenotype. This is usually the case if one gene suppresses the other which itself suppresses a third phenotype relevant gene (see Figure 6c). Additive interaction between two genes exists if a mutation in one of both genes results in different phenotypes and a mutation in both results in another third phenotype. This can be the case if both genes are on separate pathways and leading to two separate phenotype relevant target genes (see Figure 6d).

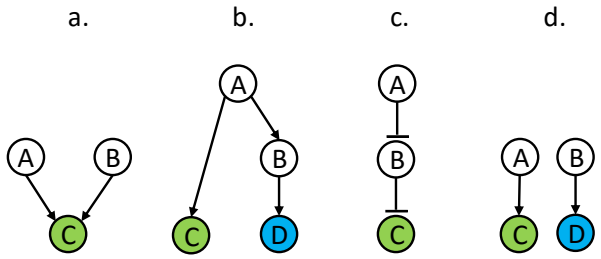


Figure 6: genetic interactions (a.) Synthetic interaction: both genes A and B activate a gene C for "green color". Knock out mutations in either A or B will not change the expression of C. (b.) Epistatic interaction: Gene C for "green color" and gene D for "blue color" both contribute in the wild type phenotype. Knock out mutations in gene B which activates gene D will result in a green color only phenotype. Knock out mutation in gene A (or gene A and gene B) which activates gene C as well as gene B will result in a phenotype without any color. (c.) Suppressive interaction: Wild-type phenotype expresses the "green color" gene C since it is not suppressed by gene B due to

gene A suppressing gene B. A knock out mutation in gene B or both genes A and gene B does not change the phenotype. A knockout mutation in gene A does activate gene B which then blocks gene C and changing the phenotype. (d.) Additive interaction: "green color" gene C and "blue color" gene D contribute both to the wild type phenotype. A knockout mutation in either gene A or gene B which are both activating gene C and gene D respectively results in an either green only or blue only phenotype. A knockout in both gene A and gene B results in a "no color" phenotype.

Genetic interaction networks play a role in many diseases since genes influence the disease phenotypes. Diseases based on mutations in genes can stem from mutations in single genes (monogenic) as in case of cystic fibrosis where mutations in the CFTR gene are the cause [69] or many (multigenic) as in case of atherosclerosis [70]. Synthetic interactions are the basis of the synthetic lethality concept as a targeted therapy against diseases, especially cancer [71]. Giving examples of applications, Söllner et al. link the mycophenolate mofetil mode of action with molecular disease and drug profiles [72]. Cramer-Morales induced synthetic lethality by targeting RAD52 in leukemia [73]. Mora et al. [74] investigated protein interaction data and multigenic inherited disorders.

Methods for constructing genetic networks

Detection of genetic interactions is usually done in genome-wide association studies and identifying single nucleotide polymorphisms (SNP) play a crucial role thereby [75]. High throughput detection of SNP are usually based on microarray technologies [76]. Prediction of genetic interaction is supported by machine learning approaches and statistical approaches. Koo et al. [77] investigate neural networks, support vector machines and random forests. A recent approach suggested by Zhang and Kim [78] implements a statistical framework and learning algorithm based on focusing on single nucleotide polymorphisms and perturbations of the network caused from expression quantitative trait loci. Wong et al. [79] integrate various sources to build a probabilistic decision trees to predict pairs of synthetic sick (or lethal) genetic interactions within *Saccharomyces cerevisiae* including data from localization, mRNA expression, physical interaction, protein function and characteristics of network topology.

Drug-target interaction networks

Drug-target interactions usually describe the physical interactions between biological entities, mostly proteins but also genes or miRNAs, and therapeutic drugs, mostly small molecules and also proteins. Identifying drug-target interactions is an essential step in the discovery of drugs for diseases [80]. The amount of drugs available increases on a rather slow rate while our knowledge about molecular mechanisms of diseases increase due to high throughput technologies and prediction techniques leading to a hypothesis rich environment of possible drug-targets. As of August 2014 DrugBank [81] holds 7,739 drug entries with 1,584 FDA-approved small molecule drugs. Therefore a strong emphasis exists for identifying possible new target for the relatively small set of available drugs.

Methods for detecting drug-target interactions

Experimental discovery of drug-target interactions include among others, thermal shift assays [82] and reverse pharmacological approaches [83], [84]. Basic approaches for prediction of drug-target interactions include docking simulations, literature mining, machine learning and

statistics. Cheng et al. [85] use a model-based approach solely on the crystal structure of the target binding site. Chou et al. [86] developed a probabilistic model for mining interactions from literature. Yamanishi et al. [87] characterize drug–target interaction networks and reveal correlations between drug structure similarity, target sequence similarity and the drug-target interaction network and propose a predictor based on supervised learning for bipartite graphs. Nagamine et al. [88] use a support vector machine approach for predicting protein-drug interactions based solely on protein sequence and chemical structure data. He et al. [89] propose a predictor based on nearest neighbor algorithms supported by a minimum-Redundancy-Maximum-Relevance algorithm for feature ranking.

Disease / Gene-disease networks

Disease networks focus on the representation and exploration of diseases. Therefore the representation of disease networks differs depending on the perspective. One perspective is to extend pathway categories in interaction networks by disease categories and include known physical interactions between the biological entities. KEGG [90] for example provides additional categories in the form of disease pathways similar to their protein-protein and metabolic pathways (for example see Figure 7 for the asthma pathway).

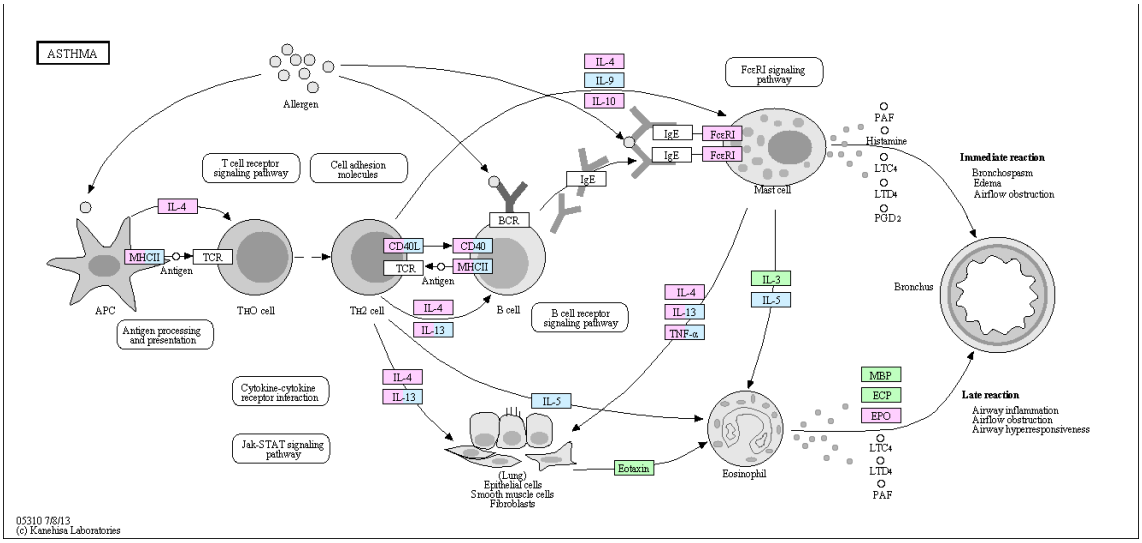


Figure 7: Asthma disease pathway from KEGG database. [Online]. http://www.genome.jp/kegg-bin/show_pathway?hsadd05310 [Accessed August 2014]

Genetic interaction networks represent networks of dependencies between genes resulting in different phenotype of which the majority studied are disease related phenotypes. Goh et al. [24] proposed new representations for disease networks. A human disease network (HDN) is a network of connected disease nodes where the weighted edges represent the amount of joint genes between both diseases. Their disease gene network consists of gene nodes which are connected by edges if both genes are associated with the same disease. Thirdly they created a *Diseasome* network by forming a bipartite graph consisting of two disjoint sets of nodes, one

for disease associated genes and one for diseases associated to genetic disorders and linking both sets by connecting a gene with a disease if known mutations exists within that gene which is associated to the disease (see Figure 8 for a schematic representation of all three network types). Data for constructing this disease network were obtained from the Online Mendelian Inheritance in Man (OMIM) database [91].

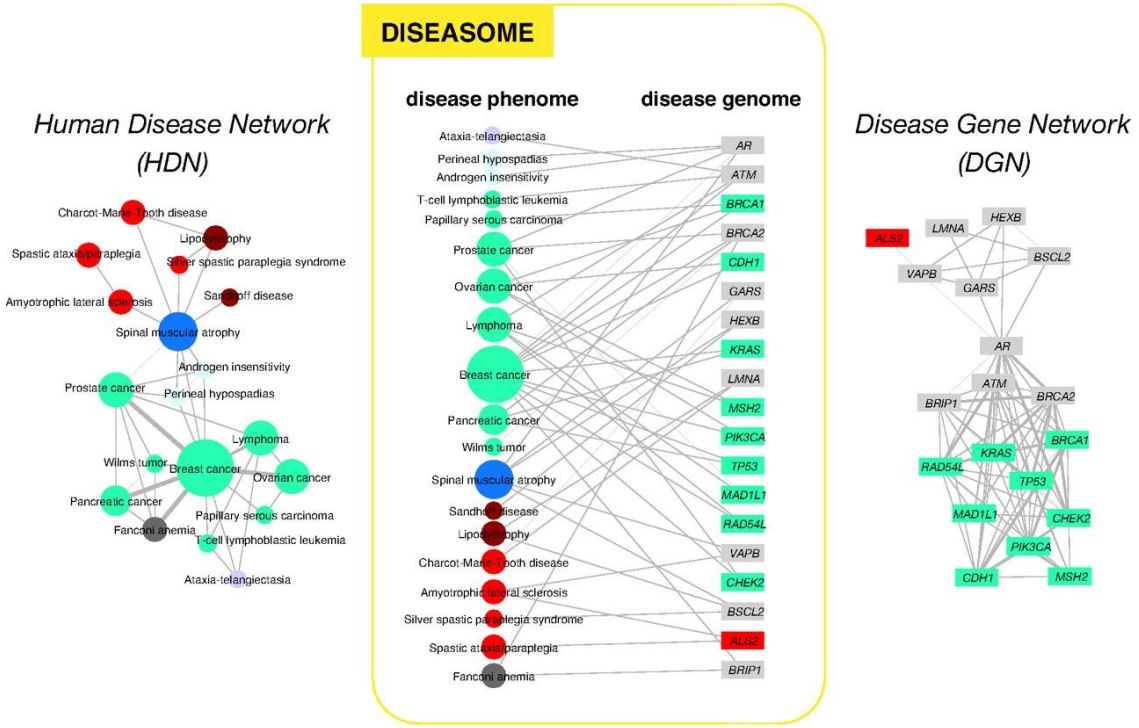


Figure 8 from Goh et al. 2007 [24]: Schematic representation of disease network representations. Circles represent diseases, rectangles represent genes. The size of a circle is proportional to the amount of genes associated to the disease. The color of the circles and rectangles indicate the disease class. Edge strength between disease nodes is proportional to the amount of genes both diseases share. (*Left*) Human Disease Network (HDN) where diseases are linked with each other if they share genes. (*Right*) Disease Gene Network (DGN) where genes are connected if they are part of the same disease. (*Middle*) Diseasome as bipartite graph where diseases are linked to genes if there are known mutations within that gene associated to that disease.

Data sources for biological networks

Data sources for interactions on the molecular scale are roughly categorized into primary databases which provide experimentally deduced interaction data, into meta-databases which provide experimental interactions from several other sources and into predicted databases which provide experimental as well as predicted interactions. The list of data sources presented here is extensive and while for sure not complete it covers most popular data sources discussed in literature and references. For a very extensive list of pathway related databases refer to the *Pathguide* resource list [4] which holds 547 pathway related data sources as of 14th August 2014. Some of the databases listed have not had an update within the last years or shortly after their introduction even though they can still be accessed and are often included in others. Data sources presented here are all publically available and free to use usually referring to academic or non-profit use of their data. Data sources are listed in alphabetical order.

APID

The *Agile Protein Interaction Data Analyzer* (APID) is a web-tool for analyzing and exploring known experimentally validated protein-protein interactions [92]. Data sources for APID are BIND [93], BioGRID [94], DIP [95], HPRD [96], IntAct [97] and MINT[98]. APID includes an edge weight for the reliability of given interactions including parameters like the connectivity, cluster coefficient, GO environment, GO environment enrichment, number of experimentally validated methods, GO overlapping and iPfam domain-domain interaction (see Figure 9 for a detailed schema of the building process). APID is freely accessible through the *Agile Protein Interaction Navigator* web-tool (APIN) on an exploration basis for proteins of interest. However the whole dataset cannot be downloaded. APID holds in total 56,460 proteins and 322,579 protein-protein interactions as of August 2014.

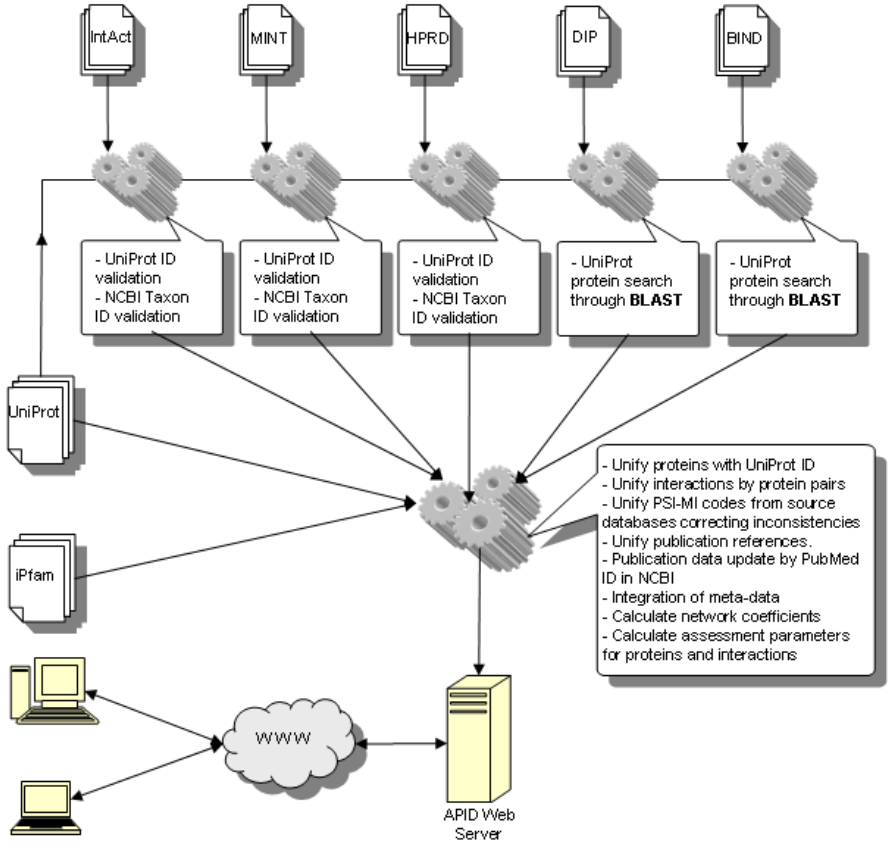


Figure 9: Agile Protein Interaction Network building schema. [Online]. http://bioinfow.dep.usal.es/apid/index.htm. [Accessed August 2014]

BiGG

The *Biochemical Genetic and Genomic* (BiGG) database is a freely accessible data source for knowledge of biochemically, genetically and genomically structured genome-scale metabolic network reconstructions [99]. BiGG integrates several genome-scale metabolic pathway sources and information like KEGG [100] and NCBI [101] by a model reconstruction process from literature, data sources and mathematical model building and is curated by experts. BiGG provides browsing of content and visualization of pathway data and cross references information with other sources. Pathway data is exportable in SBML format. As of 14th August 2014 BiGG contains about 100 metabolic pathways for *Homo sapiens* and other species.

BIND

The *Biomolecular Interaction Network Database* (BIND) is a freely available database of interactions, molecular complexes and pathways [93]. Objects for which interactions are stored include proteins, DNA, RNA, ligands or molecular complexes. Interaction data is built from peer-reviewed literature and direct submission. BIND follows an extensive ASN.1 data specification [102] as well as an XML data specification. Data access is supported by the NCBI

programming toolkit [103]. As of their latest release from December 2009 BIND holds in total 31,972 unique proteins and 58,266 unique interactions.

BioCarta

BioCarta (San Diego, CA) is a company providing among others freely accessible pathway related information which is frequently updated by experts following an "open source" approach [104], [105]. Pathways can be visually explored utilizing the web interface. As of August 2014 BioCarta holds information for over 120,000 genes from multiple species.

BioCyc/MetaCyc

The BioCyc/MetaCyc database is a freely accessible collection of over 3,000 species specific Pathway/Genome databases (as of October 2013) [106], [107]. MetaCyc contains over 2,100 experimentally determined metabolic pathways curated from over 37,000 publications including kinetic data. Data collections in BioCyc contain genomes for organisms and predicted metabolic networks and genomic information. The BioCyc web interface allows visual exploration of the data and offers a variety of tools for comparative analysis and modeling including among others pathway predictions, operon prediction, flux-balance analysis models, transcription factor based filtering, comparative overlay of pathways and omics data, metabolite path tracing, pathway enrichment analysis etc. BioCyc cross links the pathway information to 23 other databases. BioCyc provides data in various formats and access through the web interface and PERL, Java and Lisp APIs.

BioGRID

The *Biological General Repository for Interaction Datasets* (BioGRID) is a freely accessible database of physical and genetic interactions of several species [94], [108] and is freely downloadable as tab-delimited text files and PSI-MI XML [109]. Physical interactions are obtained by high throughput interaction studies (HTP). BioGRID holds for all organisms of the database 515,302 unique interactions and 44,528 unique proteins. For *Homo sapiens* 149,796 unique interactions and 18,757 unique proteins are stored (as of August 2014). For a detailed yearly statistics of the amount of interactions and proteins from *Homo sapiens* see Figure 10.

BioGRID's web interface is based on PHP [110] and hosted on an Apache web server [111]. The database management system is MySQL [112]. Their annotation compilation system is based on Java [113].

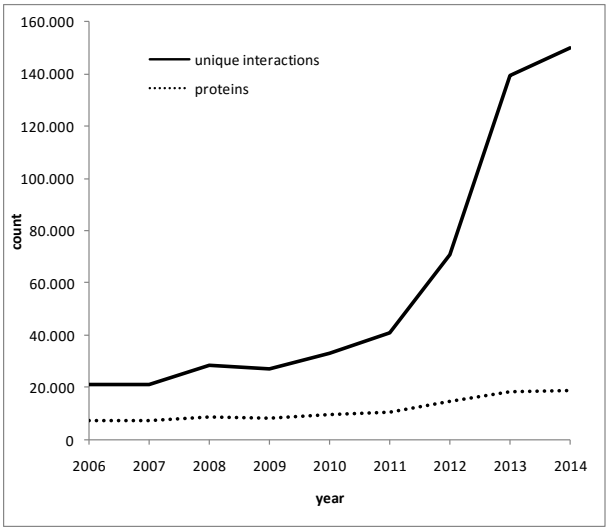


Figure 10: BioGRID's yearly statistics on the amount of non redundant interactions (solid line) and proteins for *Homo sapiens* (dotted line).

CPDB

The *ConsensusPathDB-human* (CPDB) is a freely accessible data source for interaction networks in *Homo sapiens* including protein-protein, genetic, metabolic, signaling, gene regulatory and drug-target interactions, as well as biochemical pathways [114], [115]. CPDB consolidates 32 data sources into a unified data source avoiding redundancy and complementing it with additional interactions curated from literature. As of its latest release version 29 from 27th June 2014 CPDB holds 154,537 unique physical entries, 416,872 unique interactions and 4,078 pathways. Data can be downloaded in tab-delimited or PSI-MI 2.5 format [116] and pathway results exported in BioPAX [117]. CPDB increases the confidence of interactions by analyzing text mining data for errors using an integrated approach exploiting network topology and annotation features. Data analysis is supported by a variety of options including the visualization of functional gene/metabolite sets as overlap graphs, gene set analysis based on protein complexes, induced network modules analysis with gene lists, graph visualization is based on Cytoscape.js (do not confuse with Cytoscape) [118]. A separate plug-in for Cytoscape [119] is available as well.

DIP

The *Database of Interacting Proteins* (DIP) catalogs experimentally determined protein-protein interactions obtained from peer-reviewed publications with manual and automated means [120], [95]. For DIP's data model see Figure 11. DIP is freely accessible and holds protein-protein interaction information for 693 organisms (as of August 2014). Protein-protein interaction data for *Homo sapiens* cover 4,283 proteins and 7,140 protein-protein interactions obtained from 3,122 experiments. DIP provides a plug-in (MiSink) for Cytoscape visualization software [119] for integration of its interaction data. MiSink can be freely downloaded [121].

DIP uses MySQL [112] as database management system.

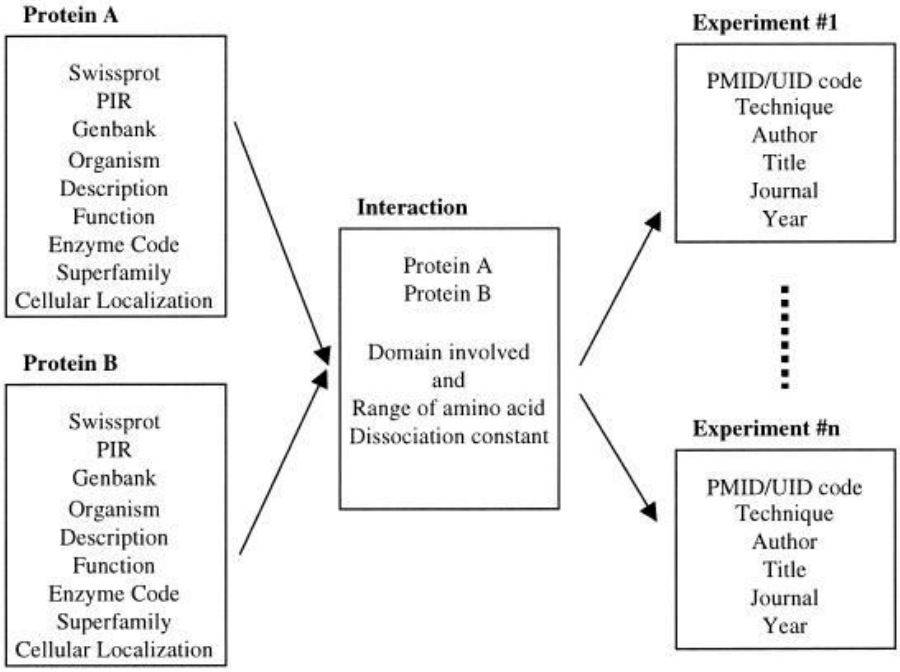


Figure 11 from Xenarios et al. 2000 [95]: Data model for the Database of Interacting Proteins (DIP)
.

DisGeNET

DisGeNET is a freely accessible database for human gene-disease associations integrating expert curated data sources and text mining approaches [122], [123]. Integration of the sources is performed by utilizing gene and disease vocabulary mapping and an association type ontology. A reliability score of the associations is computed depending on the supporting evidence and provided with the data (see

for the extensive data model). DisGeNET provides a Cytoscape [119] plug-in for visualization and analysis of its network data. Data is provided through a web interface, or by downloading a SQLite database dump. Additionally a Resource Description framework (RDF) is provided and RDF data access via a SPARQL endpoint. As of 14th August 2014 DisGeNET holds 381,056 associations between 16,666 genes and 13,172 diseases.

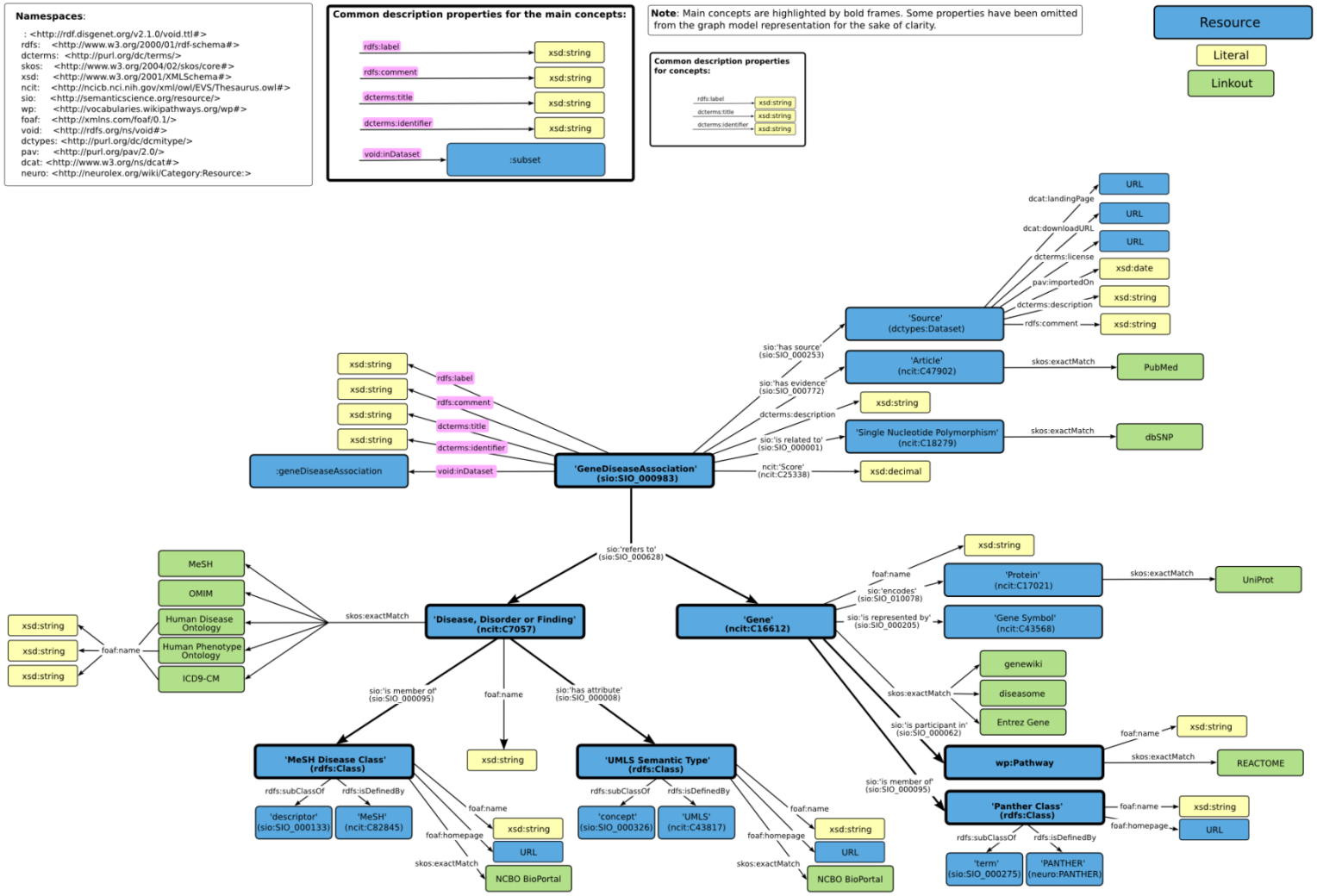


Figure 12: RDF data model for the DisGeNET gene-disease association database. [Online]. http://disgenet.org/web/DisGeNET/v2.1/rdf#schema [Accessed August 2014]

DrugBank

DrugBank is a freely accessible data source for drug and drug-target information [81], [124]. Drug and drug-target information is extracted manually from literature and curated by experts. DrugCards hold over 200 data fields for a drug and its target related information of relevance. As of August 2014 DrugBank holds 7,739 drug entries of which 1,584 are FDA approved small molecules, 156 FDA approved protein/peptide drugs, 89 nutraceuticals and over 6,000 experimental drugs. 4,283 non-redundant protein sequences are linked to these drug entries. Drug nomenclature proves especially challenging and is done manually.

HAPPI

The *Human Annotated and Predicted Protein Interaction* (HAPPI) database is a freely accessible data source for integrating protein interactions sources into a unified concept addressing semantic differences between various source concepts for protein-protein interactions [125]. Furthermore HAPPI uses a scoring system for the reliability of interactions. Data was integrated from HPRD [96], BIND [93], MINT [98], STRING [126], [127] and OPHID [128]. As of version 1.31 from 18th November, 2009 HAPPI holds 70,829 curated proteins of which 13,601 are denoted as interacting proteins and 601,757 protein-protein interactions and associations. Data can only be accessed via web interface.

HAPPI uses an Oracle [129] database management system and PERL [130] for implementations.

HitPredict

HitPredict is a database of high confidence protein-protein interactions [131]. HitPredict includes non-redundant interactions from IntAct [97], BioGRID [94] and HPRD [96] and calculates a confidence level of those interactions based on sequence, structure and functional annotation. As of its last update 1st May, 2012 HitPredict holds 49,071 proteins from 9 species with 239,584 interactions of which 168,458 are predicted to be of high confidence. Data sets are freely accessible except for HPRD data.

HMDD

The *Human microRNA Disease Database* (HMDD) is a freely available data source for expert curated miRNA to disease associations with experimental evidence [132]. As of the last update from 14th June 2014 HMDD holds 10,368 entries including 572 miRNA genes, 378 diseases and 3,511 publications and can be freely downloaded or browsed vie web interface.

HPRD

The *Human Protein Reference Database* (HPRD) is a freely accessible database of curated proteomic information pertaining to human proteins [96], [133]. Proteomic information includes protein-protein interaction, post translational modifications and tissue expression. HPRD holds 38,167 protein-protein interactions which have been detected based on yeast

two-hybrid, in vitro or in vivo. Proteomic data can be provided for by the scientific community via an annotation system called *Human Proteinpedia* [134], [135]. Pathway data can be downloaded in the data exchange formats BioPAX 2.0 [117], PSI-MI 2.5 [116] and SMBL 2.1 [136].

iHOP

The *Information Hyperlinked over Proteins* (iHOP) is a freely accessible service providing contextual clusters and hyperlinks of the biomedical literature based on genes and proteins providing a network of linked literature for exploration of genes and proteins of interest [137]. Strictly speaking iHOP does not provide a physical interaction network per se but a network of related information for a given gene or protein. iHOP search results distinguish between 'interaction information' between genes/proteins and 'defining information' between genes/proteins and biological terms. iHOP identifies genes and synonyms within biomedical text based on a dictionary approach for enabling cross-linking with other sources (see Figure 13 for the full concept). Data can be accessed via a web interface for a given gene or protein of interest. As of August, 2014 iHOP holds more than 2,700 organisms, 111,000 genes and over 28.4 million sentences.

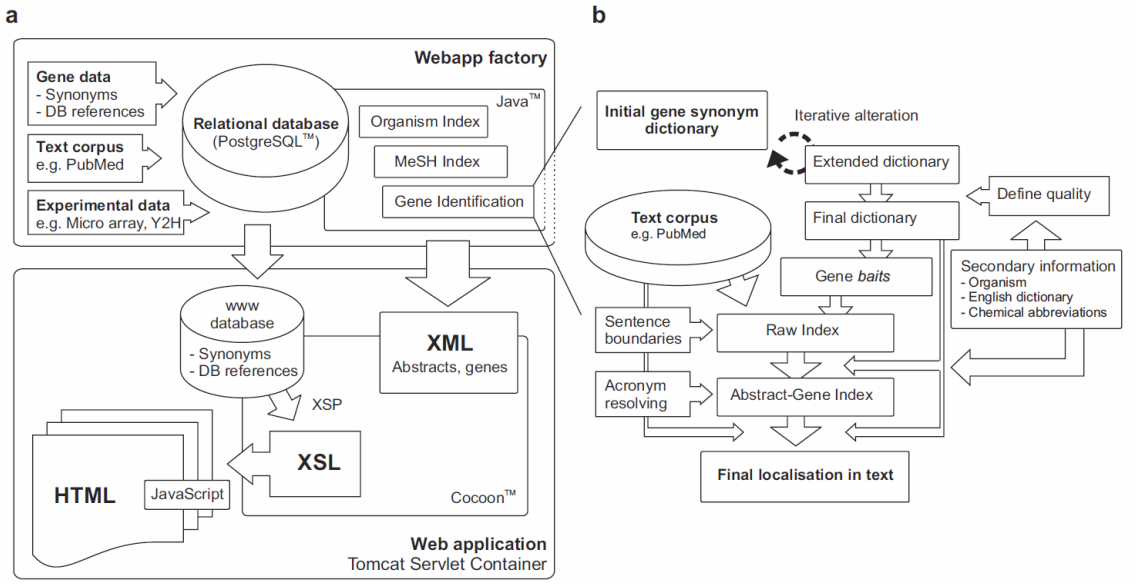


Figure 13 from Hoffmann and Valencia 2005 [137]: (a) iHOP System architecture. The iHOP system is divided into two separate parts: the web application factory and the web application itself. Production state data in the web application is based entirely on XML technology and extremely fast response times are obtained through avoidance of complex front-end database queries- For every gene, one static XML document was created. Dynamic effects were achieved though the HTML and JavaScript layer on the client side to minimize server load. (b) Gene synonym identification in biomedical abstracts. More than 12 million abstracts were examined for the occurrence of gene symbols, names and synonyms. At an average length of 200 words, the total number of examined terms reaches ~ 2 billion, of which each could be one of the 3.2 million gene synonyms in the dictionary. To accomplish this comparison it was crucial to subdivide the gene identification process into independent steps of increasing precision, going from a raw gene-article index to a stable index, and finally to an exact localization of gene synonyms in the text.

IMID

The *Integrated Molecular Interaction Database* (IMID) is a freely accessible data source for protein interaction including direct protein-protein interaction as well as protein to small molecules interactions [138]. IMID integrates molecular interaction data from literature by text mining utilizing a Bayesian network and from manually annotated data sources including Reactome [139], PID [140] and GO [141]. Furthermore interactions are linked to their biological context represented by biological terms allowing filtering of interactions based on biological terms. While the database is freely accessible web-based [142], download of all data is not possible. As of 10th August 2014 IMID holds in total 1,450,989 interactions.

IntAct

IntAct is a freely available open source database of molecular interactions derived from literature curation or direct user submissions [97], [143], [144]. Interactions between several biological entities called Interactors include binary interactions as well as multi-protein interactions. IntAct uses controlled vocabularies like the NCBI taxonomy database or Gene Ontology as well as own controlled vocabularies for annotating their data [141], [145]. IntAct holds 83,417 Interactors, 296,668 interactions from 33,742 experiments (as of August 2014). For yearly statistics on the amount of interactions and proteins see Figure 14 and for further statistics see [146].

IntAct uses an Oracle database management system [129] as well as a PostgreSQL system [147]. For the web interface the Struts framework [148] is used and the Tulip system for graph layouts [149]. IntAct is based on Java [113].

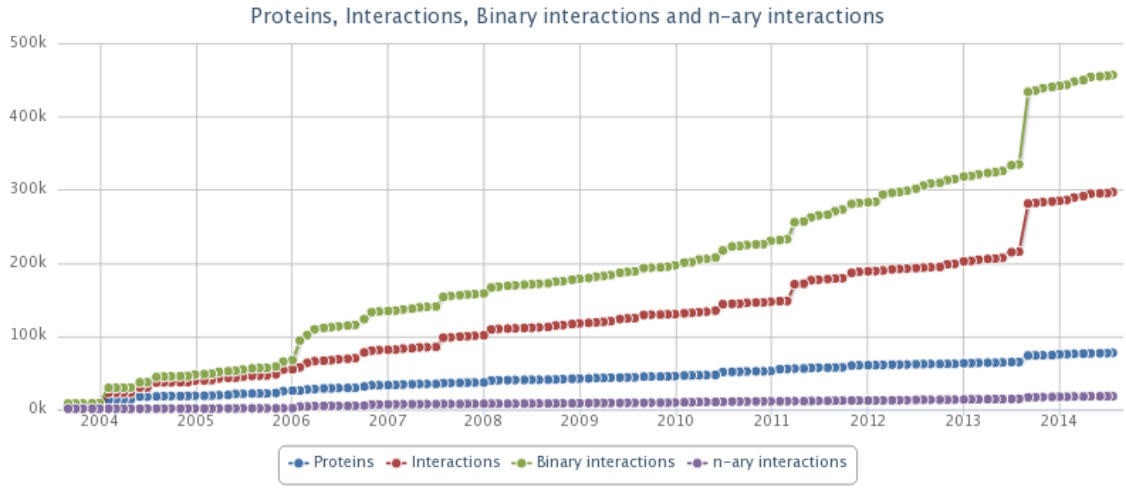


Figure 14: IntAct's statistic on the yearly amount of interactions and proteins stored at the database. Y-axis denotes the count in thousand. [Online].http://www.ebi.ac.uk/intact/pages/documentation/statistics.xhtml#tc01. [Accessed August 2014].

KEGG

The *Kyoto Encyclopedia of Genes and Genomes* (KEGG) is a freely available knowledge base linking genomic information with higher order functional information [90], [100], [150]. Genes are linked in a network of interacting molecules thus forming networks of direct or indirect protein-protein interactions. Annotation and pathway data is extracted from various sources with automated tools and supported manually building reference pathways. KEGG holds interactions within 463 pathways (as of August 2014). KEGG maps pathway information with experimental evidence manually to existing or new pathways and assigns genes from organisms to an ortholog group (KO) based on sequence similarity (see Figure 15). KEGG's SSDB is a graph which nodes are genes and the edges are weighted by sequence similarity scores. Assignment of annotation and modification of the genes within a KO group is highly computerized (Figure 15 left section, red color) and the identification of genes belonging to a KO group is manually done (Figure 15 right section, green color).

KEGG's database is based on an internal Oracle [129] database management system and a public PostgreSQL [147] system. Web services are based on REST and formerly SOAP/WSDL.

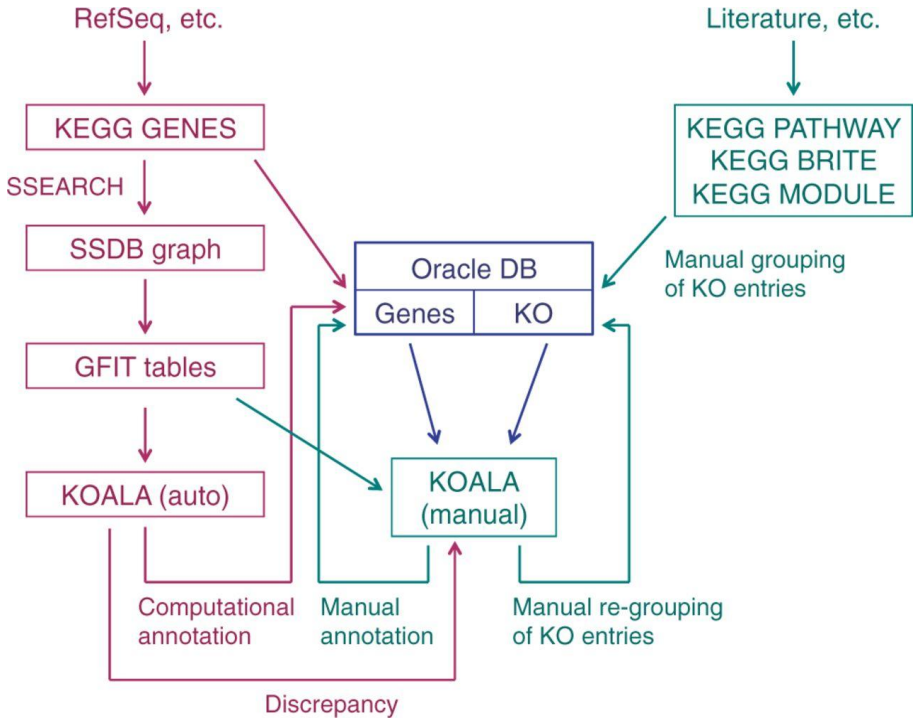


Figure 15: Schematic diagram of genome annotation within KEGG, obtained from Kanehisa et al. 2014 [90]. Steps relying on manual curation are colored green (right section) and steps relying on computerized automatization are colored red (left section).

MiMI

The *Michigan Molecular Interactions* (MiMI) database is a freely accessible data source providing knowledge from several protein interaction data sources [151]. MiMI merges the knowledge from all its sources into a single interaction concept addressing the issue of entities reported differently within separate databases thus generating and predicting a merged graph of joint interactions. The merging results are not curated and as such errors are possible. MiME supports users in identifying possible errors with natural language selection techniques applied on publications in which supposedly interactions had been reported. Additional protein information was integrated from GO [152], interPRO [153], IPI [154], miBLAST [155], OrganelleDB [156], OrthoMCL [157], PFam [158] and ProtoNet [159]. As of 2008 MiMI holds 3.7 million interactions, about 3.5 million genes, 19.2 million molecules and 1,288 pathways.

MINT

The *Molecular INTeraction* database (MINT) is a freely accessible database of experimentally verified and curated protein-protein interactions mined from scientific literature [98]. MINT is consistently curating all the issues of *FEBS Letters* (since January 2005) and *EMBO Journal* and *EMBO Reports* (since January 2006). MINT holds 241,458 interactions and 35,553 proteins (as of August 2014). For a detailed yearly statistics for the stored interactions and the number of curated articles see Figure 16 and [160]. Interaction data stored follows the PSI-MI 2.5 standard [116]. Moreover MINT uses the IntAct infrastructure since September 2013 joining efforts to reduce redundant efforts.

MINT's database is based on PostgreSQL [147] database management system, the object-relational mapping tool ObJectRelationalBridge OJB [161] from Apache. The web application is based on the Struts framework [148], a Tomcat servlet container [162] and an Apache server [111].

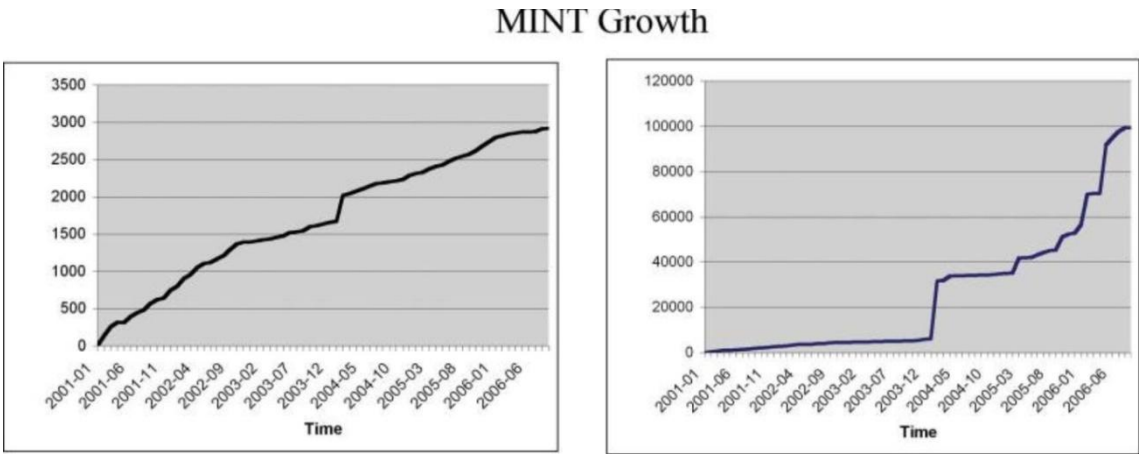


Figure 16: The Molecular INTeraction database's (MINT) statistics on the growth of the amount of stored interactions (left curve) and curated articles (right curve) as published by Chatr-aryamontri et al. 2007 [98].

MIPS

The *Munich Information Center for Protein Sequences* (MIPS) is a freely accessible database providing protein-protein interactions [163] with a specialized section for mammalians called MIPS MPPI [164]. Protein-protein interactions are extracted from literature and curated manually by experts including only data from individually performed experiments. MIPS holds over 900 proteins from 10 mammalian species with over 1,800 entries for protein-protein interactions (as of January 2005) available in the PSI-MI standard [116].

MIPS uses a MySQL [112] database management system and a web interface based on Perl [130] CGI scripts.

miRBase

The miRBase is a freely accessible database of published miRNA sequences and annotation and prediction of miRNA targets [165], [166]. Prediction of targets is based on sequence matching and on orthologous sequence similarities. Data are available through a web interface or downloadable in various file formats and database dump. As of its latest release version 21 from June 2014 miRBase holds 24,521 miRNA loci from 206 species and 30,424 mature miRNA products.

miRBase uses a MySQL [112] database management system.

MPIDB

The *microbial protein interaction database* (MPIDB) is a freely accessible data source for known microbial protein interactions. Interaction data is retrieved from literature and manually curated as well as imported from known interaction databases [167]. Data sources include IntAct [143], DIP [95], BIND [93] and MINT [98]. The latest release is from 18th November 2009 and holds 24,295 experimentally determined protein-protein interactions of 250 bacterial species or strains. In addition the interactions provided by MPIDB are further supported by 68,346 evidences based on interaction conversation, protein complex membership and 3D domain information from iPfam or 3did.

MPIDB uses three tier software architecture, separating data, logic and presentation. MySQL [112] database management system is used on the data tier, PHP [110] for the logic tier and CSS [168] for presentation. An Apache web-server [111] hosts the system.

NCI-PID

The *National Cancer Institute - Pathway Interaction Database* (NCI-PID, or PID) is a freely available data source for curated signaling pathways composed of human biomolecular interactions and cellular processes [140], [169]. As of its last release from September 2012 (no further updates are planned) PID holds 137 curated human pathways and 9,248 interactions as well as 322 human pathways with 7,575 interactions from BioCarta and Reactome. PID supports exploration of networks by creating interaction maps around molecules of interest. Data can be downloaded supporting BioPAX [117] and XML.

OMIM

The *Online Mendelian Inheritance in Man* (OMIM) is a freely accessible database for human genes and genetic disorders [91]. OMIM provides linkage information between genotypes and phenotypes of genetic disorders allowing the construction of gene-disease networks. Additionally OMIM provides extensive annotation and description of the genetic disorders and is manually created and curated from literature. As of 15th August 2014 OMIM holds a total of 14,660 gene descriptions and 4,174 phenotypes where the molecular basis is known. In total OMIM holds 22,480 entries. Access to OMIN is provided via a web interface integrated into the Entrez database collection [101] and by REST API for automated extraction.

OPHID

The *Online Predicted Human Interaction Database* (OPHID) is a freely available data source for predicted protein-protein interactions and is part of the Interologous Interaction Database (I2D) [128]. OPHID predicts interactions for humans if two putative human proteins have orthologous proteins identified by BLAST [170] in related species sharing a known protein interaction. Further co-occurrence of protein domains derived from BIND [93], DIP [95], HPRD [96] and MINT [98] were calculated by enrichment analysis. Co-expression analysis for proteins was calculated for interacting proteins utilizing Pearson correlation. A GO term similarity measure was determined by calculating a maximum semantic similarity of all GO term pairs of interacting proteins. Bootstrapping was applied to estimate statistically significant cutoffs for co-occurrence, gene co-expression and GO term similarity against randomized distributions. As of 12th August 2014 (last modification of the website was in April 2010) I2D holds 463,346 interactions (183,524 for *Homo sapiens*) from other sources and 460,948 predicted interactions (55,985 for *Homo sapiens*). The complete dataset for all 11 I2D versions can be downloaded in tab-delimited or in PSI-compliant XML format [109].

OPHID web interface and query engine is based on an IBM WebSphere system [171]. Additional software necessary was written in Java [113].

PANTHER-Pathway

PANTHER-Pathways is a freely accessible data source for over 176 (as of version 9, August 2014), primarily signaling, pathways and part of the *Protein ANalysis THrough Evolutionary Relationships* (PANTHER) database system [172], [173], [174], [175]. PANTHER is a curated data source focusing on inference of gene and protein function using phylogenetic trees to extrapolate from sequence. Pathways are drawn with CellDesigner™ [176] and can be exported in SBML [136] and SBGN [177].

PINA

The *Protein Interaction Network Analysis* (PINA) platform provides analytical and visualization tools for investigating protein-protein interaction and integrates data from 6 databases (IntAct, BioGRID, MINT, DIP, HPRD, MIPS) [178]. PINA provides in detail tools for network construction, collections of annotated interactome modules, network filters, network visualization and analytical tools including enriched GO term identification, topological feature selection,

identification of topologic important proteins and identification of common interacting proteins. PINA data are freely accessible and downloadable. As of the last update of 21st May, 2014 PINA holds for *Homo sapiens* 166,776 binary interactions and 5,211 complexes.

PINA uses the GPU accelerated AllegroMCODE [179] clustering plug in for Cytoscape [118]. RESTful web service was implemented in Java using the *jersey* library [180].

PIP

The *Human Protein-Protein Interaction Prediction* (PIP) database is a freely accessible data source for predicted human protein-protein interactions [181]. Protein-protein interactions are based on a naïve Bayesian classifier to calculate a score of interaction between all protein pairs. Of all calculated interaction scores calculated (ca. 17.6 million protein pairs) 37,606 (as of its last update from 12th September 2008) interactions with a score ≥ 1 indicating that the interaction is more likely to occur than not to occur. These predicted interactions do overlap with other databases only marginally. 34,215 out of the 37,606 are unique to PIP [182]. Several features are combined within the score's calculation including gene-co-expression, orthology, domain co-occurrence, co-localization, post translational modification and transitive topological network analysis.

PIP uses a MySQL [112] database management system, a Tomcat [162] web server and the front end is based on JSP [183].

POINT and POINeT

The *Prediction of Interactome* POINT database is a freely accessible data source for the prediction of protein-protein interactions based on the orthologous interactome [184]. Protein pairs have a predicted interaction if both have orthologous proteins (based on sequence similarity) interacting with each other. POINeT is a freely accessible web service for protein interaction network analysis and visualization [185]. Network analysis measurements include graph measures like closeness, degree, eccentricity, radiality and centroid centralities. Further analytical options include filtering using biological characteristics based on GO terms or using tissue specific expression profiles. POINeT further includes a filtering option for sub-network specificity scores determining the impact of a protein node on its sub-network. For a full system architecture design see Figure 17. POINeT additionally includes merged data from DIP [95], MINT [98], BIND [93], HPRD [96], MIPS [163], CYGD [186], BioGRID [94] and NCBI interactome [187] mapping all nodes onto NCBI Gene IDs. While POINeT is still available, the latest available version includes data from other sources up to 2008.

POINT and POINet web applications are based on Java [113], the Struts framework [148], JSTL [188] and AJAX [189].

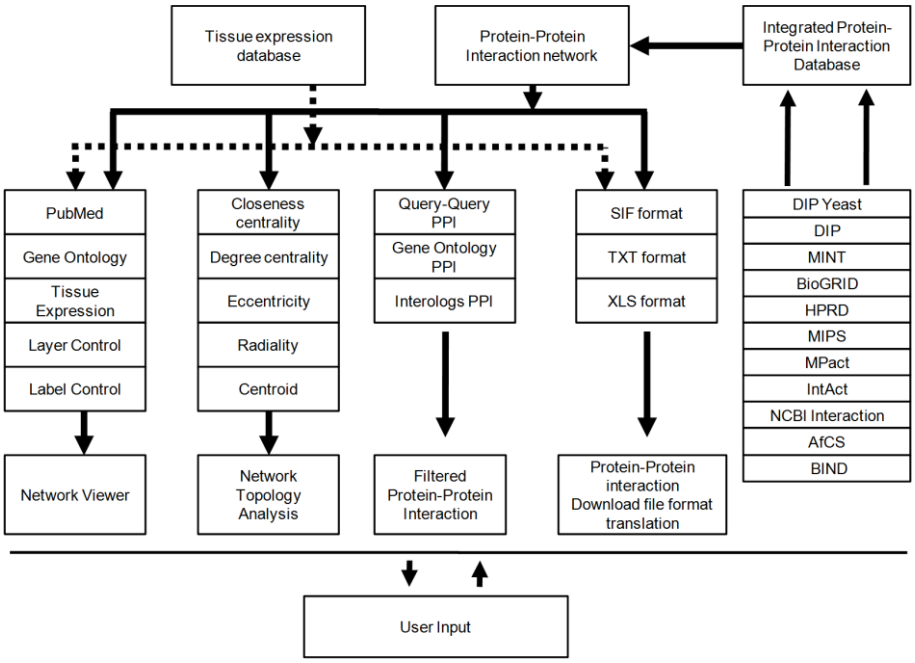


Figure 17: System architecture of POINeT obtained from Lee et al. 2009 [185]

Reactome

Reactome is a freely accessible curated and peer reviewed pathway database supporting visualization and data analysis of human pathways and reactions [190]. Reactome focuses its data model on biological reactions defined as events that change the state such as binding, activation, translocation, degradation or biochemical reactions etc. Information in Reactome is curated by expert and cross referenced to several other sources. As of version 46 from September 2013 Reactome holds annotations for 7,088 protein-encoding genes from the Ensembl human genome assembly (34% coverage), 15,107 literature references and 1,421 small molecules organized into 6,744 reactions collected in 1,481 pathways. Analysis of the resulting reaction network is supported by many different data annotations including post translational modifications, regulation by non–coding RNAs, functional annotation, disease association and many more allowing an extensive exploration of network related data via an extensive web based analysis tool. Reactome also supports several data representation standards.

Rhea

Rhea is a freely accessible data source of expert curated biochemical reactions providing a non-redundant set of chemical transformations for integration within construction of metabolic pathways or pathway inference [191]. Data is extensively cross referenced to other metabolic databases including KEGG [100], EcoCyc [192], UniPathway [193] and Reactome [139]. As of version 53 from 15th Jul 2014 Rhea holds 5,927 unique chemical compounds and 6,938 approved reactions. Data can be accessed via web interface or downloaded as files.

STITCH

The *Search Tool for Interactions of Chemicals* (STITCH) is a freely accessible data source for interactions between proteins and small molecules integrating data from metabolic pathways, crystal structure, binding experiments and drug-target interactions [194], [195]. Further relations between chemicals are predicted utilizing information from phenotypic effects, text mining and chemical structure similarity. Cross-species prediction of interactions is based on orthologous proteins. STITCH holds, as of version 4.0 from September 2013, 390,000 chemicals and 3.6 million proteins from 1,133 organisms with 367,000 interactions of high confidence within human. STITCH can be accessed with a web interface and downloadable files.

STRING

The *Search Tool for the Retrieval of Interacting Genes/Proteins* (STRING) is a freely accessible data source for known and predicted protein interactions [126], [127]. Interactions are derived from genomic context, high-throughput experiments, co-expression, literature mining and known sources. Interactions include direct physical interactions as well as functional, indirect interactions. Protein interactions are transferred between orthologous groups. Interactions predicted by genomic context are based on the observation that co-occurring genes often result in functionally associated proteins interacting with each other. The majority of predicted interactions result from literature text mining from OMIM [91] and PubMed. Data sources included within STRING are BIND [93], DIP [95], HPRD [96] and MINT [98], BioGRID [94], KEGG [100], Reactome [139], IntAct [143], EcoCyc [192], NCI-PID [196] and GO protein complexes [197]. As of version 9.1 STRING (2012) includes over 5 million proteins from 1,133 organisms. STRING's web interface supports searching and browsing of protein interaction data as well as inspecting underlying interaction evidence. A plug-in for Cytoscape [119] visualization tool implementing the PSICQUIC standard [198] is also provided for. HPRD [96], BIND [93], MINT [98]

STRING uses a PostgreSQL [147] database management system.

TRANSFAC

TRANSFAC is a data source for eukaryotic transcription factors, their experimentally proven binding sites, consensus binding sequences and regulated genes [199]. Public available data exist as of a snapshot from 2005, otherwise subscription is required.

TTD

The *Therapeutic Targets Database* (TTD) is a freely accessible data source for therapeutic protein and nucleic acids targets with their corresponding drug [200]. Disease and pathway information as well as annotation information of the targets are provided, including, function, 3D structure, ligand binding properties, enzyme nomenclature and drug structure, therapeutic class and clinical development. As of August 2012 TTD holds 2,025 targets and 17,816 drugs.

UNIHI7

The *Unified Human Interactome* (UNIHI7) database is a freely accessible data source for molecular interactions supporting the visualization, analysis and data retrieval of human molecular interaction networks [201]. As of 15th September 2013 UNHI7 holds about 350,000 molecular interactions for about 30,000 human proteins between genes, proteins and drugs including gene expression data and functional annotations. UNIHI7 includes several sources for physical protein interaction, functional association, regulatory transcriptional interaction, gene expression data and drug target interactions (see Table 1 for a full list of resources). UNIHI7 supports network based analysis by providing various filtering options and cross linking to known resources if available.

UNIHI7 is based on Java [113], a MySQL database management system [112], DAO [202] for database interaction, Hibernate [203] for object-relational mapping and a Tomcat web server [162].

Table 1: Data sources from the Unified Human Interactome database (UNIHI7) retrieved [Online]. http://www.unihi.org/. [Accessed August 2014]

Resource	Proteins	Interactions	Type of interaction	Methods	Reference
MDC-Y2H	1713	3340	Physical protein interaction	Y2H screen	Stelzl et al. 2005 Cell
CCSB	1549	2754	Physical protein interaction	Y2H screen	Rual et al. 2005 Nature
HPRD	12613	65227	Physical protein interaction	Literature curation	Prasad et al. 2009 NAR
BioGRID	14822	124035	Physical protein interaction	Literature curation	Chatr-Aryamontri, A et al. 2013 NAR
BIND	11524	19352	Physical protein interaction	Literature curation	Isserlin, R et al. Database(Oxford), 2011
DIP	3025	2925	Physical protein interaction	Literature curation	Salwínski et al. NAR 2004
IntAct	13611	37629	Physical protein interaction	Literature curation	Kerrien et al. NAR 2012
Reactome	5315	108867	Physical protein interaction	Pathway curation	Croft et al. NAR 2011
COCIT	3737	6580	Functional association	Computational predicton	Ramani et al. 2004 Genome Biology
ORTHO	6056	62863	Physical protein interaction	Computational predicton	Lehner et al. 2004 Genome Biology
HOMOMINT	6221	21863	Physical protein interaction	Computational prediction + Literature curation	Persico et al. 2005 BMC Bioinformatics
OPHID	7874	81677	Physical protein interaction	Computational predicton	Brown et al. 2005 Bioinformatics
Transfac	742	1554	Regulatory transcriptional interaction	Literature curation	Matys et al. 2006 NAR
miRTarBase	2234	3565	Regulatory transcriptional interaction	Experimentally validated	Hsu SD et al. 2011 NAR
HTRIdb	1634	2263	Regulatory transcriptional interaction	Literature curation	Bovolenta LA et al. 2012 BMC Genomics

Source	Data Type	Details	Type	Method	Reference
Symatlas	Gene expression	Expression data from 19 different tissue samples	Absolute expression	Affymetrix GeneChips	Data described in Su et al., 2004, PNAS and analysis described Russ and Futschik, 2010, BMC Genomics
DrugBank	Drug target	List of drugs and their targets	Drug->Target	Literature curation	Knox et al. 2011 NAR

UniPathway

UniPathway is a freely accessible data source of manually curated information for the representation and annotation of metabolic pathways. Data are cross linked to other metabolic sources including KEGG [100], MetaCyc [106] and [191]. UniPath integrates a hierarchical controlled vocabulary for building of larger pathways from smaller pathways in a controlled process which it also provides for UniProtKB [204]. For representing metabolic pathways UniProt developed a data model (see Figure 18). UniPathway data can be accessed and explored via web interface and downloaded in OBO format v1.2 [205].

UniPathway uses a PostgreSQL [147] database management system, a ZOPE [206] framework for the content management and the dojo toolkit [207] for web applications.

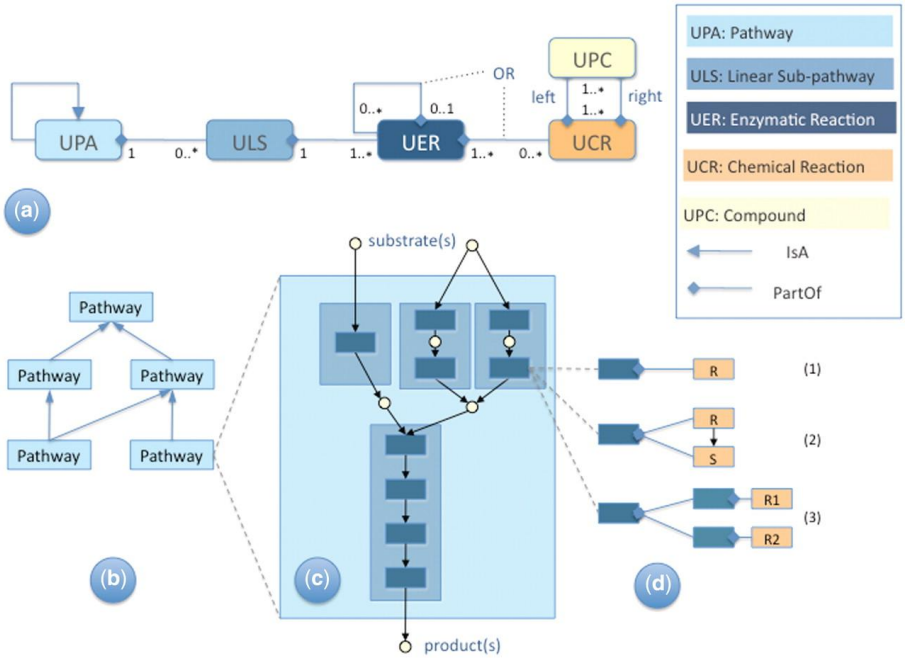


Figure 18 from Morgan et al. 2012 [193]: Overview of the UniPathway concepts. (a) Unified Modeling Language (UML)-like representation of the UniPathway classes and relationships. Legend is to the right of the main part of the figure. Multiplicity constraints read as: One UPA is composed of 0 or more ULS—One ULS is contained in exactly 1 UPA. One ULS is composed of 1 or more UER—One UER is contained in exactly 1 ULS. One UER is composed of 0 or more (alternate) UER—One UER is contained in 0 or at most 1 UER. One UER is composed of 0 or more UCR—One UCR is contained in 1 or more UER. One UCR is composed of 1 or more left UPC and 1 or more right UPC—One UPC is contained in 1 or more UCR. (b) Example of the IsA relationship defining the UniPathway controlled vocabulary hierarchy of pathway terms. A pathway instance may be a specific type of an abstract pathway entity. (c) Example of the PartOf relationship linking a pathway (UPA: light blue), its subpathways (ULS: blue) and individual enzymatic reactions that constitute the subpathway (UER: dark blue). (d) Three cases of the relationship between an UER and its chemical reaction components (UCR): (1) simple one-to-one relationship where R is catalyzed by a single enzyme; (2) R is catalyzed by an enzyme and S is a spontaneous reaction; (3) 'OR' relationship: the enzyme can catalyze two reactions differing by their co-substrates (e.g. NADH/NADPH).

VisANT

VisANT is an integrative network platform to connect genes, drugs, diseases and therapies [208]. VisANT predicts and consolidates interactions utilizing various sources into a metagraph with integrated disease and therapy hierarchy, disease-gene and therapy-drug association. Disease hierarchy is represented by ICD-10 and therapeutic hierarchy by ATC from the World Health Organization. Associations between genes and diseases were gathered from OMIM [91], KEGG [150], PharmGKB [209], GAD [210] and DrugBank [81]. Protein interaction data is predicted or imported from literature and existing data sources like BioGRID [94], MINT [98], BIND [93], MIPS [163], IntAct [97] and HPRD [96]. Filtering and enrichment analysis and topological calculations can be performed on 11 different types of networks supporting the

exploration of data (see Figure 19). As of August 2014 VisANT holds in total 1,316,571 interactions over 112 species. VisANT is freely accessible and runs as java applet.

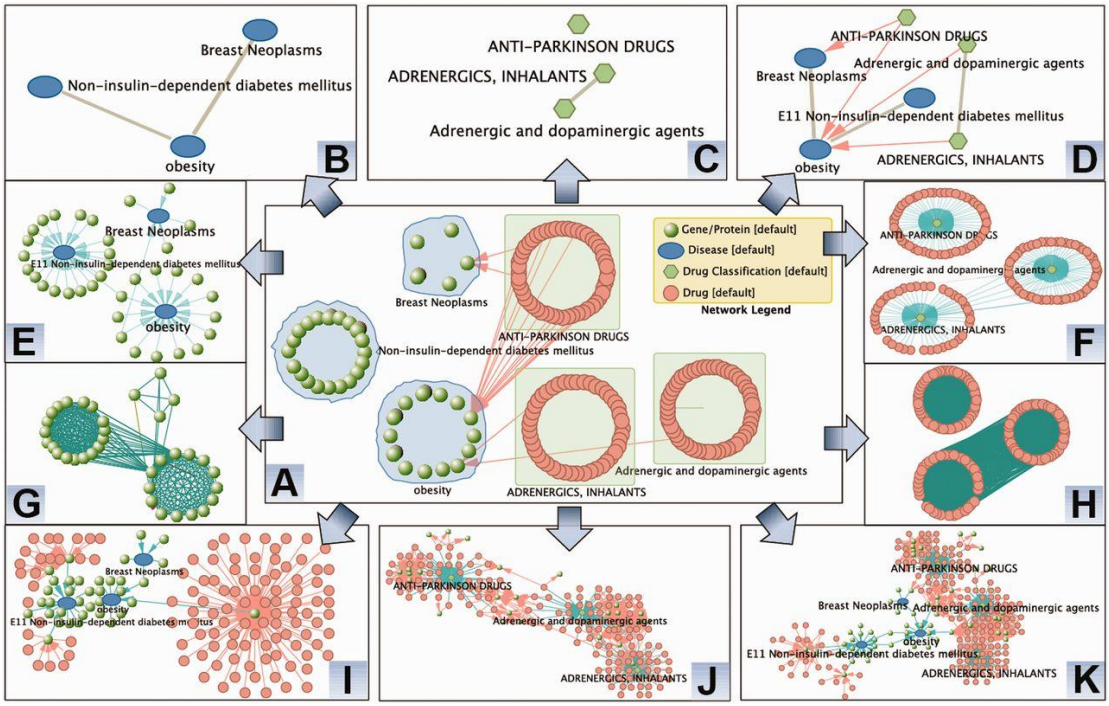


Figure 19 from Hu et al. 2013 [208]: Illustration of versatile network construction in VisANT 4.0 with integrated disease and therapy hierarchies, disease–gene and therapy–drug associations and drug–target interactions. Detailed explanation can be found in the session 'Network construction'. (A) Meta-network of diseases and therapies. Expanded metanodes of diseases are represented using convex polygons. Drugs are queried for their targets in three diseases (red lines). The rest of the networks (B–K) are all derived from this meta-network. (B) Disease network where grey edges indicate that there is shared genes associated with two diseases. (C) Therapy network similar to disease network. (D) Disease–therapy network. (E) Disease–gene network. (F) Therapy–drug network. (G) Co-disease gene network. (H) Co-therapy drug network. (I) Disease–gene–drug network. (J) Therapy–drug–gene network. (K) Disease–gene–drug–therapy network.

Modeling of biological networks

Ideally a network model represents our understanding of a system, its elements within the network, their state, their relationships and their dynamics defining the rates at which the changes of the states take place. When modeling a system of interest a first step is the mapping of our observations of the system to a causal network model explaining the observations and thus possibly highlighting gaps in our knowledge. By conducting further experiments, gathering more and more observations of the system, a network model can be extended. In a second step we can use our network models to simulate and predict effects on the system when manipulating a network's state. This is especially useful if experiments are difficult or impossible to conduct. Classical approaches for this kind of dynamic view on a network are models based on deterministic differential equations which are commonly applied in metabolic network analysis. The drawback though is that the elements, their connections and the dynamic rates have to be known for simulation and prediction of a system. Yet a major component, quantitative time resolved data, the dynamic rates (or kinetic rates), which is required for classical dynamically modeling of molecular networks is sparse or missing in most other biological data. Alternative strategies for inferring dynamic patterns are available if dynamic information can be deduced on a larger time scale, on the level of observable changes of phenotypes or on the sequence of process steps. Stochastic models for example approach the system dynamics on the level of observable phenotypes and probabilistic changes. Various concepts of stochastic models exist including master equation systems, cellular automata based approaches and agent based systems. Flux balance analysis in metabolic networks follows a constraint based approach where substance concentration and equilibrium assumptions are made for the network and the reaction kinetics do not need to be known. Boolean networks used for modeling gene regulatory networks and signaling networks rely on the qualitative information of the regulation sequence and simulating dynamic changes by abstract time steps where each state transition of a sequence node depends of the states of the prior nodes. The majority of biological network models existing rely on descriptive and conceptual time independent modeling approaches focusing on graph representations and topological features with various approaches of inferring a dynamic pattern of topology change out of molecular data. The choice of the modeling approach depends on the data available and the question of interest.

Depending on the level of detail of a model, the networks elements have to be defined. In case of physical molecular networks this usually includes cellular molecules, proteins, nucleic acids chemical compounds etc. Network elements can also be abstract concepts and information flow between those, like disease phenotypes associated with each other. When modeling a network on a higher level, information of the lower levels is usually hidden. In case of protein-protein interaction networks the level of observance is proteins as entities and their interactions. The internal organization of the protein's atoms is usually ignored. Similar, in a social interaction network between humans, we are usually not interested in the molecular composition of a human being. When modeling a biological network the data available usually dictate the possibilities for model selection, nevertheless it is as crucial to consider first the application goal of the model itself.

The following sections will shortly present and discuss network modeling approaches followed by a presentation of approaches for construction of networks from experimental data.

Afterwards random network models are presented. Data standards commonly applied in network biology finish this chapter.

Dynamic network models

Deterministic models with differential equations

Differential equation concepts usually used are ordinary differential equations where the concentration of a biological entity is a continuous function over time and not space, partial differential equations which include a spatial component like diffusion and stochastic differential equations, which are used for modeling irregular motion, variability or uncertainty due to time series. The latter two are computationally more demanding than ordinary differential equation systems. The core ingredient for such dynamic modeling is a time dependant variable, usually a dynamic rate like a reaction kinetic rate within chemical reaction equations (see Figure 3). Most dynamic models stem from metabolic networks which have been investigated for more than half a century and more kinetic data has been available for them than for other molecular systems. Chen et al. [211] applied stochastic differential equation models on transcriptional regulatory networks. Teusink et al. [212] investigated differential equation models of the yeast glycolysis. Lee et al. [213], [214] successfully modeled and analyzed the Wnt pathway with differential equation approaches. When considering time and space for modeling dynamical processes the differential equation approach can quickly become computationally very challenging. Additionally if we are able to observe a system's parameters and measure them, it is not necessarily possible to measure subparts of it composing the observable property. For example, bacterial growth depends simultaneously on the rate for generating new bacteria and the rate at which they die. Experimentally we measure the population change and not the rates with which they multiply and die. Consequently in such cases stochastic approaches where we do not need to know the subparts but the observable parts of a system are used or deterministic approaches where the parameters which need to be known are estimated if they are not known (which is the case most of the time). Parameter estimation approaches in dynamic models have been approached by several groups. Moles et al. [215] investigate global optimization approaches in biochemical pathways. Timmer et al. [216] performed parameter estimation with maximum likelihood and statistical testing in cellular signal transduction of the JAK-STAT signaling pathway. Huang et al. [217] used hierarchical Bayesian models for parameter estimation of their nonlinear differential equations for HIV models. Perrin et al. [218] combine deterministic models with parameter estimation based on dynamic Bayesian networks. Dai and Lai [219] use a simulated annealing variant for parameter estimation in biological networks.

Cellular automata and agent based models

Cellular automata and agent based models provide powerful frameworks where the global function is resolved by implementing the network entities and their relations discretely over time and space directly on an element by element basis (agents or objects), all together forming the global observable. In its core concept cellular automata and agent based models are defined by a discrete space, like a two dimensional square lattice for a surface, a definition of neighborhood between the discrete locations, the states available in each location and an update function which changes the states at a position during each update step in dependence of the states of the neighborhood and internal states. The difference between cellular

automata and agent based models lies in the synchronicity of the update steps. While in cellular automata the update of each object is done during each discrete time step quasi simultaneously, agent based models rely on an asynchronous update of an agents state often following a schedule or discrete event scheme. Additionally agent based systems usually do not rely on a grid based space. If the update function requires dynamic rates like in modeling chemical reaction networks [220], cellular automata suffer the differential equation modeling dependency on parameters that are mostly missing within biological interaction networks data. Siehs et al. discussed the use of cellular automata in context of chemical reaction networks [220] and for simulating apoptosis processes [221]. Chain et al. [222] investigate topological effects on the dynamics of feed forward motifs with cellular automata. Zhang et al. [223] simulated gene-protein interaction, cell phenotypes and multi-cellular patterns in brain cancer utilizing a 3D multi-scale agent based tumor model. Kleinstreuer et al. [224] describe a multi-cellular agent-based model of vasculogenesis for prediction of the disruption of blood vessel development.

Boolean networks

Boolean networks introduced in gene regulatory networks by Kauffman [225] are related to cellular automata and model a discrete sequence of steps where time and states are discrete without a discrete space. Applied on gene regulatory networks input molecules affecting the network, genes and their products are modeled as nodes and their causal link as directed edges. During each time step the state of the nodes are updated on the basis of Boolean functions of the nodes preceding them. Darabos et al. [226] recently addressed Boolean networks considering the recent advance in genomics and network knowledge.

Stochastic simulation based on master equations

Stochastic models rely on the fact that noise (stochasticity) is being frequently observed with consequence on the clonal population within transcription processes of eukaryotes [227], [228]. They describe that the true nature of chemical reactions and biological networks is not that of a continuous process but a discrete one since the amount of molecules cannot change by a fraction of molecules but only as integers. For large amounts of chemical entities the deterministic models seem adequate but in situations where the molecular populations are very small or, if the dynamical structure of the network makes it susceptible to noise amplification, the effects of stochasticity and discreteness can be important [229]. As such modeling a stochastic network considers reaction rates as probabilities of discrete state transitions accounting for noise based on master equations. The dynamics within a stochastic regulatory network are driven by a stochastic stimulation algorithm [230]. Thomas et al. [231] recently addressed phenotypic switching in gene regulatory networks on the basic of such a stochastic model.

Flux balance analysis

Flux balance analysis within a metabolic network does not rely on differential equation models [232], [233]. Two assumptions are made. The first being that the system is in equilibrium, a steady state and that the metabolite concentrations do not change anymore. This assumption is based on the observance from fermentation technologies that systems like bacteria or yeast enter a steady ‘production’ state once they pass the initial grow phase. The second assumption is optimality which states that a system has undergone an evolutionary optimization process. With those assumptions the system is reduced to a set of linear equations which is then solved

for identifying a flux distribution matching the constraints and maximizing the overall systems function from input substrates to the output products. Linear programming is used for the calculations which are not computationally intensive. A major advantage of flux balance analysis is that no reaction kinetics are required and that a calculated model can be easily tested (for microorganisms) in chemostats where the substrate concentrations can be kept constant. Lewis and Abdel-Haleem [234] addressed cancer metabolism with flux balance analysis. Lotz et al. [235] discuss flux balance analysis in context of plant metabolism.

Descriptive models

The sets of biological entities are by far from complete and as such the focus is still on identifying new biological entities and also their role and relation among each other. Deducing the identification and relations are primarily on a qualitative scale and not on a quantitative scale. With the advent of high-throughput technologies identification of biological entities and their relations sped up significantly especially in genomics and high throughput sequencing. Additionally it was possible to get a systemic snapshot of the amount of biological entities within a biological sample. For example gene expression arrays allow the quantification of all genes' mRNAs from a given sample. Within proteomics such systemic snapshots can be obtained from 2D gel electrophoresis approaches. The drawback with any present technologies for measuring the molecular systems is, that it is a quantification of the biological entities at a given time point and not within a time interval. As such when comparing two samples one can deduce their difference in states independent of a time scale. Several aspects influence this trend especially in human healthcare. Giving a few examples, patient samples over time are often not available especially with invasive sampling. Also molecular experimental measuring techniques usually destroy the probe material during the analysis. As such the next time step cannot be measured from the same probe but has to be taken from another probe at a later time from a hopefully similar sample. With the lack of measuring dynamic parameters modeling of complex interaction networks for dynamic simulation is next to impossible [236].

Descriptive time independent models dominate the field of interaction networks and are usually graph representation of the sum of the underlying biological observations. The nodes form the biological entities and the edges form the known interactions within the networks. Since those network representations are time independent the edges represent what has been observed at some timepoint in a given sample. This means that not all edges present within an interaction network need to be active at the same time or within the same cellular type and tissue. Uncovering these spatial and local differences in the interaction networks, several approaches emerged within the last years proposing other dynamical viewpoints and concepts mostly on a descriptive scale. Han et al. [237] investigated the yeast protein-protein interaction network for dynamically organized modularity. Based on co-expression data they identified 'party' hubs which are focusing on single functions all the time and 'date' hubs which connect between groups of proteins with varying functions and are active at different times. Taylor et al. [238] addressed temporal differential co-expression in hubs as predictive tool for breast cancer prognosis. Wallach et al. [239] explored the dynamics of circadian protein-protein interactions based on day-time dependent expression data and inferring descriptively the dynamical changes in interactions during the day. All these concepts do not describe dynamics as a continuous function over time but as chain like series of time points where the

phenotype of a time point *t* is dependent of the phenotype of the prior time point at *t-1*. Nevertheless identification of discrete differences in phenotypes based on underlying dynamical processes is possible by these approaches.

Constructing networks from experimental data

Experimental data from genomics, proteomics and metabolomics have been used to infer networks. Besides qualitatively describing networks, data stemming from omics experiments allow the construction of network models differentiating structure-only or structure-and-dynamics [240]. A basic approach is calculating a relation between entities based on co-expression or similar concepts resulting in a structural relation network. Such a relation can be Pearson correlation or biologically motivated [241]. This information is not necessarily causally linked. Correlation approaches can be extended with time series data to infer directionality and causality. Schmitt et al. used a time lagged correlation to infer a gene network from transcriptomic data [242]. The time series in such 'relevance networks' are mostly two time points implying that the later time point is dependent on the previous. Other approaches for inferring structural information rely on Gaussian graphical models [243], Bayesian networks [244] and including dynamics, Dynamic Bayesian networks [245]. Boolean networks [246] and probabilistic Boolean networks [247] are used for inferring structure and dynamics in discrete space. Markov models [248], state space models [249] and ordinary differential equations models [250] are used in continuous space. For a detailed comparison refer to Sima et al. [240]. Further approaches on the level of experimental data inference rely on perturbed experimental data as proposed by Markowetz et al. [251]

Random network models

Analysis of biological networks has also focused on descriptive networks finding patterns and differences between phenotypes and creating descriptive models of our observations. Analysis of biological networks on the topological scale revealed characteristic patterns which can be ascribed to real world networks and biological networks. These findings resulted in several models for generating randomized real world networks and biological networks. Graph properties of real world networks are described in *"Analysis of biological networks" later this chapter*

Erdős–Rényi_model

A basic model for generating random graphs was proposed and discussed by Gilbert, Erdős and Rényi [252], [253]. In their model (ER) edges are added randomly between any two nodes with a defined probability *P* independently of any prior added edges. While this model had been thoroughly studied it does not reflect the topology of many real world networks. The degree distribution in ER graphs follows a Poisson distribution while real world networks tend to follow a power law. Additionally ER graphs tend to have a low clustering coefficient since the generation of edges is independent from other edges.

Watts and Strogatz model

The Watts and Strogatz model (WS) for generating random graphs aims at generating graphs with small world properties, including short average shortest paths and high clustering coefficients [254]. WS graphs are produced by firstly creating a lattice ring and then rewiring the edges from each node with any other node with equal probability β such that no self-loops or link duplicates are generated (see Figure 20). WS models generate unrealistic degree distributions and as such do not fully satisfy real world network topologies. Additionally WS random graphs cannot be extended later.

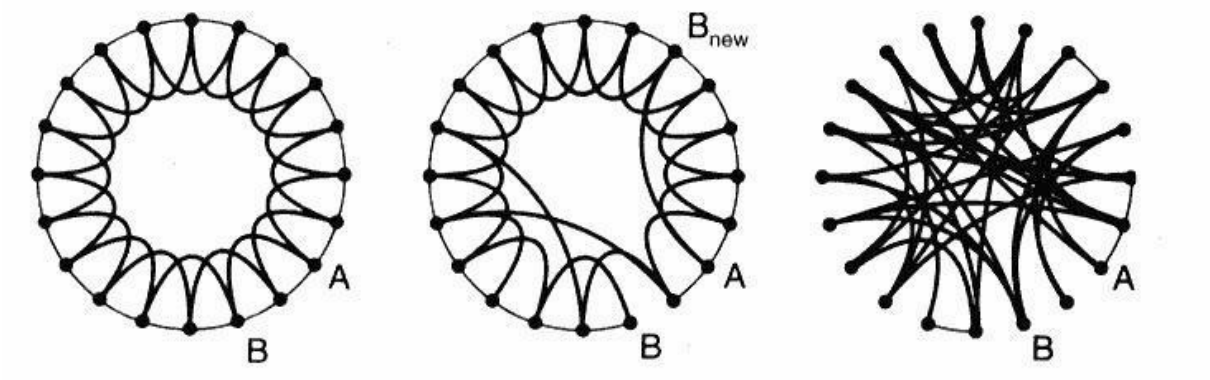


Figure 20 from Watts and Strogatz 1998 [254]: Generation of a Watts and Strogatz random graph. (Left) A lattice ring is created. (Middle) All the edges are rewired with the same probability β such that no self-loops or edge duplicates are created. Here the edge between node B and A is rewired from B to B_{new}. (Right) Final random graph.

Barabási–Albert model

The model for generating random graphs for generating scale free networks from Barabási and Alberts (BA) [22] uses a preferential attachment mechanism. Preferential attachment described by Yule [255] is the phenomenon that a new node (or edge) in many real world networks connects to an existing one and is proportional (i.e. higher) to the existing node's degree. The principle behind this phenomenon is that in an evolutionary system most of the time nothing is really created anew from scratch but a result of (slow) modification of existing objects and relations (by evolution). For example the species within genera evolve and form new species at some later time and most likely still belong to the same genera. Genera with more species are more likely to generate new species. Thus new species attach more likely to genera that already have a lot of species connected to them. Algorithmically a start set of connected nodes is defined and new nodes are added and connected with the existing nodes depending on their degree. BA models do not create high clustering as observed in real networks and as such are only partially realistic representations of real world networks.

Ravasz-Barabási model

The Ravasz-Barabási model (RB) is a random graph model based on hierarchical organization [256]. The scale free property and the high clustering coefficient, which is also independent from the network size, are characteristically for real world networks. Previously presented random network models fail to address both properties simultaneously. Hierarchical models try to overcome these shortcomings of the prior models by considering a fundamental principle within real world networks, their hierarchical topology. The hierarchical organization of networks describe the modular structure of real world networks formed by its entities based on common "interest" of the module groups like group of friends or co-workers. Algorithmically hierarchical networks are built iteratively by generating initial highly connected modules and then duplicating them and linking the peripheral nodes of modules to the center node of other modules (see Figure 21).

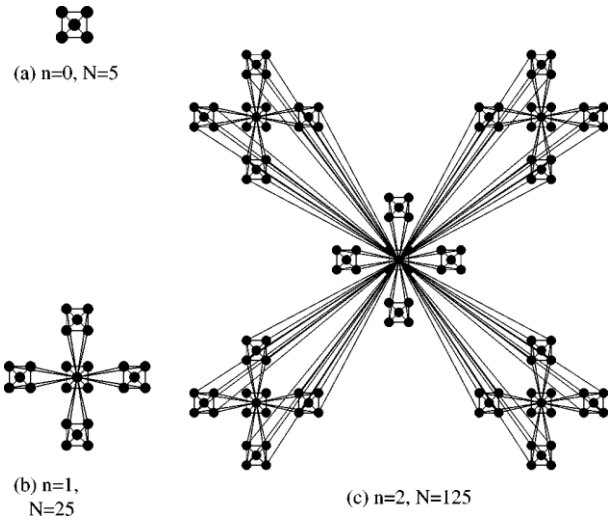


Figure 21 from Ravasz and Barabasi 2003 [256]: The iterative construction leading to a hierarchical network. Starting from a fully connected cluster of five nodes shown in (a) (note that the diagonal nodes are also connected –links not visible), we create four identical replicas, connecting the peripheral nodes of each cluster to the central node of the original cluster, obtaining a network of *N*=25 nodes (b). In the next step, we create four replicas of the obtained cluster, and connect the peripheral nodes again, as shown in (c), to the central node of the original module, obtaining an *N*= 125-node network. This process can be continued indefinitely.

Data representation standards

With the ever increasing amount of biological data several data representation and transfer standards, controlled vocabularies and guidelines for publishing molecular data have been developed. The following section gives an overview of the current state of the art.

ASN.1

The *Abstract Syntax Notation number One* (ASN.1) is a formal notion used for describing data transmitted [257], [102]. Key aspects are compressing data in binary format, speed, miniaturization and stability. The BIND database [93] uses ASN1 for describing their data, otherwise ASN.1 is not commonly used in biological network data description.

BIOPAX

"BioPAX is a standard language that aims to enable integration, exchange, visualization and analysis of biological pathway data. Specifically, BioPAX supports data exchange between pathway data groups and thus reduces the complexity of interchange between data formats by providing an accepted standard format for pathway data. It is an open and collaborative effort by the community of researchers, software developers, and institutions. BioPAX is defined in OWL DL and is represented in the RDF/XML format." (from [258], for a detailed description see [117]). Several databases support BioPAX format like Reactome [190] and BioCyc [106].

SBML

The *Systems Biology Markup Language* (SBML) is a XML based data format for exchange of models for biological processes [136], [259]. SBML does not aim at providing a universal language for representing quantitative models, rather seeking a common intermediate format that allows communication on the level of most essential aspects of the models. SMBL Level 3 provides a core set and optional packages which can be set atop of the core depending on the modeling needs. SBML libraries are available for several major programming languages.

SBGN

The *Systems Biology Graphical Notation* (SBGN) is a visual language for standardized graphically notating biological concepts [177]. SBGN consists of process diagrams, entity relationship diagrams and activity flow diagrams (see Figure 22 for examples). SBGN is supported by several platforms.

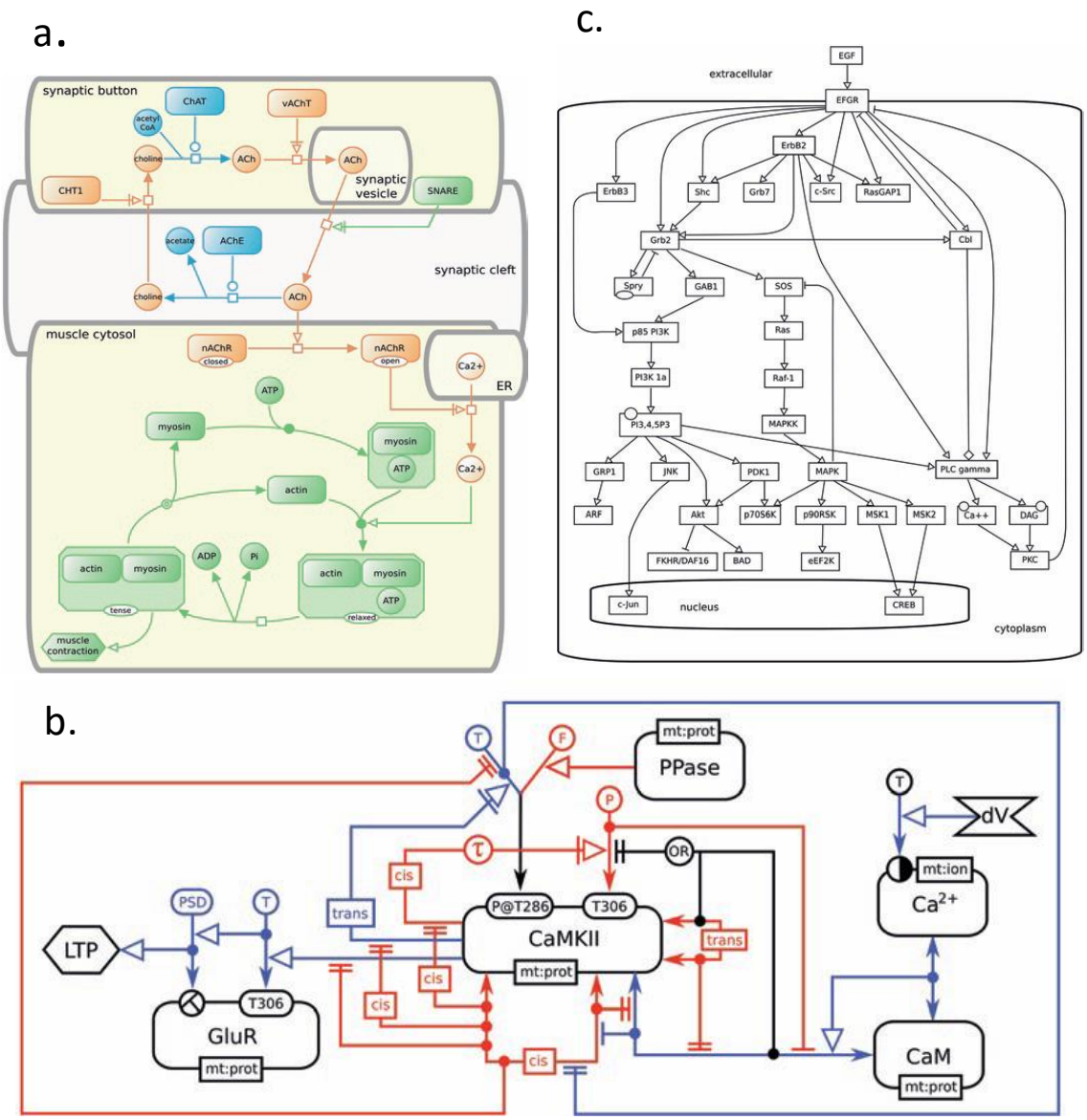


Figure 22 adapted from Le Novère, 2009 [177]: Example of complete SBGN diagrams. (a) Process diagram representing the synthesis of the neurotransmitter acetylcholine in the synaptic button of a nerve terminal, its release in the synaptic cleft, degradation in the synaptic cleft, the post-synaptic stimulation of its receptors and the subsequent effect on muscle contraction. Colors are used to enhance the biological semantics, blue representing catalytic reactions, orange for transport between compartments (including unrepresented ions, through channels) and green for the function of contractile proteins. However, it is important to note that those colors are not part of SBGN process diagram notation, and must not change the interpretation of the graph. (b) SBGN entity relationship diagram representing the transduction, by calcium/calmodulin kinase II, of the effect of voltage-induced increase of intracellular calcium onto the long-term potentiation (LTP) of the neuronal synapses, triggered by a translocation of glutamate receptors. The diagram describes the various relationships between the phosphorylations of the kinase monomers and their conformation. Colors highlight the direction of the relationships relative to the phenotype; blue relationships enhance LTP whereas red ones preclude this enhancement. (c) SBGN activity flow diagram representing the cascade of signals triggered by the epidermal growth factor, and going from the plasma membrane to the nucleus.

HUPO PSI standards and guidelines

The Proteomics Standard Initiative from the Human Proteome Organization (Santa Fe, New Mexico, US) produced several commonly used guidelines, data formats and controlled vocabularies for molecular interactions and others [116].

Among the guidelines MIMIx [260] advises on molecular interaction experiments documentation. The MIABE guideline [261] advices on reporting of bioactive entities. MIAPAR [262] advices on the minimum information of a protein affinity reagent.

Among data formats PSI-MI XML (and MI TAB) [109] proposes a molecular interaction exchange format and allows the use of open or closed external controlled vocabularies which have to be in OBO format. PSI-PAR is a data format for the exchange of protein affinity reagent data and is based on the PSI-MI XML2.5 schema with a controlled vocabulary for PSI-PAR.

Controlled vocabularies have been developed with EMBL-EBI and include PSI-MI CV for molecular interactions and PSI-PAR CV for protein affinity reagents.

The Proteomics Standard Initiative Common Query InterfaCe (PSICQUIC) [263] specification was developed for the access of molecular interaction data resources based on SOAP and REST based Web Services and a molecular interaction query languages (MIQL).

OBO 1.2

The Open Biological and Biomedical Ontologies (OBO) is collaboration for generating a suite of orthogonal interoperable reference ontologies in the biomedical domain [264]. A detailed specification can be found at [205]. AS of April 2013, 28 servers and over 150 million binary interactions are available.

Common graph representation formats

The Graph Modeling Language (GML) is a hierarchical ASCII based file format for representation of graphs [265]. GraphML is a XML based format with full compatibility with GML [266]. The eXtensible Graph Markup and Modeling Language (XGMML) [267] is another XML based format related to GML. Cytoscape software tool for visualization of biological networks [119] supports all of these formats presented.

Analysis of biological networks

Biological network analysis is motivated by several aims. On the one hand networks resemble operating biological systems and the interplay between all its entities constantly changes or is prone to disturbances often leading to changes in phenotypes such as diseases. As such a major aim of analyzing phenotypes on the network level is the identification or prediction of the phenotype (e.g. diseases) related processes and biological entities and their interplay hopefully identifying possible diagnostic or therapeutic approaches [268]. Another analytical aim is the characterization of phenotypes like diseases on the level of network topology or statistics and identifying patterns typical for a given phenotype and for biological networks in general. With the identification of common patterns the search for similar patterns within biological networks might be associated to similar phenotypes. Both analytical aims can be challenging even within networks where all real world interactions and entities are known. Biological network are far from complete. On the contrary the overlap even between major databases is rather poorly [269]. The differences have many reasons and consequently a further major goal is the analysis of biological network data on the level of reliability with the goal of integrating existing knowledge into a high quality human interaction network (also referred to as interactome). Analyzing network dynamics for generating reliable time dependent models is another goal but usually time resolved network data is rare or not available and thus dynamic interpretation of networks is rare or has to follow other concepts.

The following sections will first present graph properties and patterns typically found within biological networks followed by analytical approaches in context of these properties and patterns. Afterwards the identification and prediction of phenotype specific processes and biological entities is addressed followed by a section focusing on integration of data and methodologies.

Biological network graph properties and patterns

Networks represented as graphs can be characterized by specific graph properties. Additionally to graph properties patterns in structure and topology can be identified. Extensive studies [22], [270], [271] of biological networks characteristics revealed the following graph properties and patterns to be of common relevance.

Degree

The degree of a node *v*, *deg(v)* is the amount of edges connected to it. Within a directed graph the indegree $deg^-(v)$ is the amount of inbound edges and the outdegree $deg^+(v)$ is the amount of outbound edges.

Degree distribution

The degree distribution of a graph *G* can quantify the diversity of a network based on the amount of all degrees *deg(v)* from all nodes *v* from *G*.

The analysis of biological networks on the level of graph properties revealed that biological networks are not randomly organized. Within random networks based on the Erdős–Rényi model [253] nodes usually have a degree within the same range and follow a Poisson distribution The degree distribution of real world networks including biological networks seem

to follow a power law [22] and are called scale-free networks. The fraction *P(k)* of nodes within such a network having *k* connections follows approximately

$$P(k) \sim k^{\gamma} \tag{1}$$

The range of the parameter γ is usually in the range of 2 < γ < 3. Consequently scale-free networks have nodes with degrees high above the average also referred to as 'hubs'. Biological networks typically show a scale-free pattern and are extensively discussed by Albert [236].

Path

A path within a graph is a sequence of nodes where any two consecutive nodes are connected by an edge. The path length is the amount of edges within a path.

Distance / Shortest path

The shortest path sp_{ij} between two nodes *i,j* within a graph is the path with the shortest path length d_{ij} of all paths between *i* and *j* (also referred to as geodesic distance).

Diameter

The diameter *D* of a graph *G* is the longest shortest path within a graph.

$$D = \max \{d_{ij} | i, j \in G\} \tag{2}$$

Clique

A clique in an undirected graph *G* is a subset *C* of the node set *V* where *C* is complete (every two nodes are connected by an edge).

Clustering coefficient

The clustering coefficient (referring to the local clustering coefficient) of a node *i* within a graph *G* quantifies how closely its neighbors are to forming a clique. The neighborhood N_i of a node *i* is defined as the subset of nodes that are directly connected to node *i* by an edge. In an undirected graph the amount of all possible edges a_i between nodes of the neighborhood N_i of a node *i* is given as

$$a_i = \frac{\deg(i) * (\deg(i) - 1)}{2} \tag{3}$$

The (local) cluster coefficient c_i of a node *i* is defined as the ratio of existing edges e_i within the neighborhood N_i of a node *i* and the amount of all possible edges a_i between nodes of the neighborhood N_i of a node *i*:

$$c_i = \frac{e_i}{a_i} \tag{4}$$

Especially metabolic networks and protein interaction networks show a high clustering coefficient [272], [273].

Connectivity

Connectivity is a graph property that accounts for the minimum amount of edges (or nodes) that needs to be removed from a connected graph to disconnect him.

Betweenness centrality

Assuming that the biological processes favor the shortest path between two nodes the betweenness centrality (often called just betweenness) is a measure to estimate the importance of a node within a network by estimating the traffic load through the node *i* from a graph *G* and relates to the amount of shortest paths between all nodes *u* and *v* from *G* and the amount of shortest paths between nodes *u* and *v* going through node *i*. Let $asp_{uv}(i)$ be the amount of all shortest paths going through node *i* and asp_{uv} be the amount of all shortest paths between nodes *u* and *v* then the betweenness centrality b_i for node *i* is given as:

$$b_i = \sum_{uv} \frac{asp_{uv}(i)}{asp_{uv}} \; i \neq u \neq v \tag{5}$$

Yu et al. showed that betweenness centrality of nodes often correlate to essential functionality as well [274], especially in directed and regulatory networks. Scardoni et al. [275] investigated the essentiality of nodes based on graph measures with virtual Knock-Out Experiments within the leukocyte integrin activation network.

Small-world phenomenon

Biological networks, and real-world networks in general, exhibit a pattern called small-world phenomenon where most nodes can be reached over a few edges while also being poorly connected with all other nodes having rather small degrees [276]. While the path between any two nodes in the network is rather short and the degree of the nodes rather small, nodes tend to form clusters with higher clustering coefficients which are independent from the network size [22].

Motifs and modules

The local clustering of nodes within biological networks indicates relevant characteristics of the clustered subgraph. Motifs are subgraphs with usually few nodes and patterns that appear with a statistically significant higher frequency, like negative auto-regulation [277] or feed forward loops [278] within regulatory networks (see Figure 23 for examples of motifs with 3 nodes). Negative feed forward loops can give rise to adaptation and desensitization, while positive feedback loop can lead to emergent network properties such as ultrasensitivity and bistability [279], [280], [281].

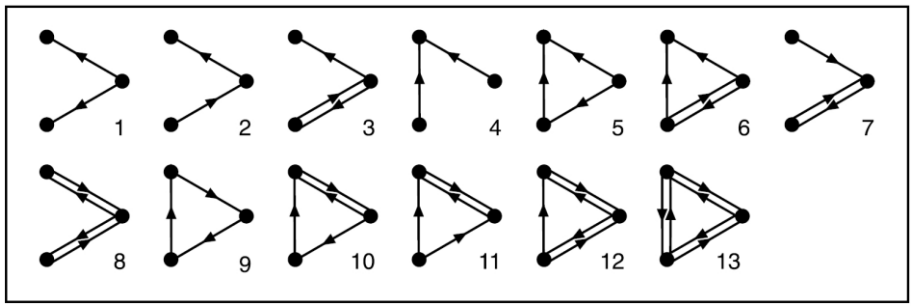


Figure 23 from Milo et al. 2002 [282]: 13 possible directed subgraph motifs with 3 nodes.

Modules within networks (also called communities) describe highly clustered entities that are physically or functionally linked like the proteins from the complement system, signaling pathways, transcriptional modules or disease modules. Modules are often found to be conserved between species as well, as Sharan illustrated between yeast, worm and fly protein networks [283].

Hierarchical organization

Real world networks show a hierarchical organization principle. Dorogovtsev et al. [284] found out that in deterministic scale-free graphs the clustering coefficient *C(k)* of a node with *k* links follows the scaling law

$$C(k) \sim k^{-1} \tag{6}$$

Ravasz and Barabási [256] investigated hierarchical organization principles in randomly generated and real world networks and found that the number and the size of modules of different cohesiveness is not random but follow a rather strict scaling law. Thus the *C(k)* curve of a network can be used to identify hierarchical organization within a network.

Analysis of network patterns and entities

The small world phenomenon and the scale-free pattern are typical within biological and many other real life networks. Interpreting hubs in biological networks revealed that the role of hubs and knockout of those is of high impact for a cell by correlating the phenotypic effect of knock-outs with the degree of a protein. A prominent example for this is p53, a cancer suppressor protein which is inactivated in 50% of human cancers [285]. Consequently the higher the degree of a node the more likely is its impact on the phenotype. Another observed consequence of scale-free networks is their overall robustness in case of network disturbances on an individual node scale as long as no hubs are involved. The whole architecture seems to stabilize networks in general and disease causing disturbances have to occur on a multi node level as has been reported for many diseases [286], [287]. Interestingly Goh et al. discovered that the majority of disease genes are non essential and only a few are related to hubs [24].

Motifs

Motifs are widely spread and seem to be conserved within many biological networks. Zhang et al. [288] showed that abundant motifs seem to form higher order network structures

associated to biological phenotypes. Motifs seem to represent evolutionary designed optimal control circuits [289]. The detection of motifs within networks is computationally very challenging since on the one hand the number of subgraphs increases exponentially with the motifs node count and on the other hand comparing subgraph topologies requires checking graph isomorphism which is computational highly intensive and one of the few NP problems and scales to $\exp(\sqrt{n}(\log n)^c$, addressed recently by Babai and Codenotti [290]. Additionally biological networks tend to be large and dense. Algorithmic approaches include 'subgraph sampling', 'exact enumeration', 'motif-centric approach', 'symmetry breaking' and 'mapping' [291]. Best practice approaches can at best find motifs with size of ~10 without utilizing heuristics [292]. Other prominent algorithms and tools include Mfinder [293], Pajek [294], MAVisto [295] and FANMOD [296]. An intriguing question related to biological network motifs is to what extend the effect of evolution has had an impact on the evolution of network topologies. One finding on the level of protein-protein interaction networks is that there seems to exist a preferential increase in the degree of highly connected proteins [297], [298].

Modules

Modules are associated with entities sharing or being part of the same phenotype (function). The rationale when analyzing modules follows that entities found within a module are more likely to play a role within the modules function [299]. Also finding topological module like subgraph patterns allows hypothesizing of novel functionality. Finding modules within networks is challenging since the amount of nodes is usually a magnitude larger then within motifs and the nodes often overlap with other modules [300]. Rives and Galitski [301] use hierarchical clustering for building modules. Girvan and Newman [302] elaborate on detection of communities within social and biological networks by focusing on edges with high betweenness. Due to the high overlap between nodes of modules Ahn et al. [303] propose a network based on linking modules. These techniques rely on graph properties of the modules. Fortunato [304] compared further module detection methods in great detail. Due to the high overlap of modules it has become evident that diseases are often associated with other diseases. When representing a disease network as disease based nodes linked together with edges representing co-morbidity, clustering of highly associated diseases can lead to further insight into the mechanisms of diseases [305]. Another approach is the identification of modules based on phenotypes using integrated approaches to formulate a module hypothesis combining topological, functional and experimental data. The steps usually include the preparation of relevant interactome data from data sources, identification of phenotype associated genes from linage analysis and genome wide association or other concepts, forming the seed of the modules within the interaction network and then identifying a highly connected subnetwork including these genes utilizing topological and functional module characteristics supported by clustering tools and others. Resulting significant subgraphs can be further investigated for their overlap with existing pathways possibly identifying known mechanisms which are disturbed. Depending on the data and tools, validation of the modules can be supported by gene expression data [306]. A magnitude of methods have been developed and are extensively discussed in Cho et al. [268]. Giving prominent examples, scoring techniques have been used by as Chuang et al. [307] for breast cancer metastasis classification. Correlation based approaches investigating for example co expression of genes in disease cases were applied by Mani et al. [308]. Shortest path based module detection have successfully been applied by Managbanag et al. [309] in context of longevity and by Shih and Parthasarathy [310]. Flow based approaches have been discussed by Kim et al. [311]. Liu et al.

[312] recently proposed an algorithm (DiME) for community detection within diseases utilizing heuristical approaches from social networks. Another recent algorithm from Leung et al. [313] (HyperModules) is a network search algorithm hat finds frequently mutated gene modules with significant clinical or phenotypic signatures from biomolecular interaction networks. Jiang and Singh propose a heuristic clustering algorithm SPICi [314] that builds clusters based on a greedy extension in confidence weighted interaction networks with guaranteed runtime of $O(V \log(V+E))$ (where V denote the size of the set of vertices and E the size of the set of edges within a given network graph). Rhrissorrakrai and Gunsalus presented the MINE algorithm [315] for module detection in networks which is based on an agglomerative clustering similar to the MCODE algorithm from Bader and Hogue [316]. Both agglomerate on the basis of vertex weights and heuristic expansions. Adamcsek et al. [317] proposed the CFinder algorithm which is based on overlapping clique searching. Enright et al. [318] propose MCL which is based on a Markov Clustering method.

Entities

Investigating network nodes and their properties in context of the whole network further our understanding of topological and structural characteristics of phenotypes. For example, Goh et al. [24] extensively studied the properties of nodes within the human disease network and discovered several principles. Diseases are rather highly connected where from 1,284 disorders 867 have at least a link to one other disorder (see Figure 24a) and that most diseases relate to a few disease genes and only a few relate to many disease genes (see Figure 24b). Among those diseases with high connections to other diseases, cancer related diseases with common repressor genes are very prominent. Genes associated with the same disease show increased interaction of their proteins, an increased co-expression in the same tissue, high co-expression levels, synchronized expression as a group and share GO terms.

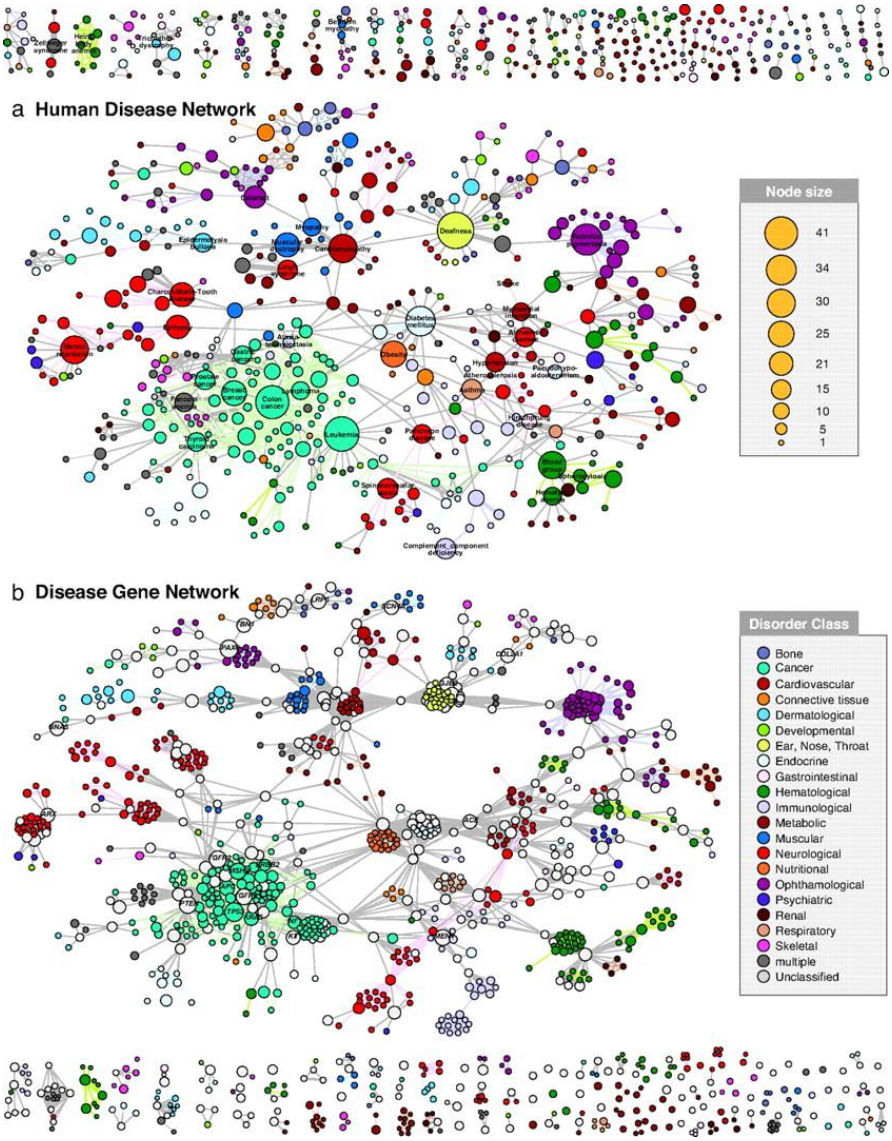


Figure 24 from Goh et al. 2007 [24] (caption modified): The HDN and the DGN. (*a*) In the HDN, each node corresponds to a distinct disorder, colored based on the disorder class to which it belongs, the name of the 22 disorder classes being shown on the right. A link between disorders in the same disorder class is colored with the corresponding dimmer color and links connecting different disorder classes are gray. The size of each node is proportional to the number of genes participating in the corresponding disorder (see key), and the link thickness is proportional to the number of genes shared by the disorders it connects. A common problem when investigating biological phenotypes on the molecular scale is that current state of the art explorative measurements do not provide complete set of relevant entities of the phenotypes molecular causes. Biological networks can further help in identifying entities relevant to phenotypes. (*b*) In the DGN, each node is a gene, with two genes being connected if they are implicated in the same disorder. The size of each node is proportional to the number of disorders in which the gene is implicated (see key). Nodes are light gray if the corresponding genes are associated with more than one disorder class. Only nodes with at least one link are shown.

Analysis and characterization of network entities and patterns has resulted in some basic hypotheses within biological networks and human network medicine [319].

- ***HUBS*:** Disease associated genes that are not essential avoid hubs while essential genes tend to associate with hubs.
- ***Local hypothesis*:** Proteins within the same phenotype tend to interact with other proteins from the same phenotype.
- ***Corollary of the local hypothesis*:** Mutations in interacting proteins tend to lead to similar disease phenotypes.
- ***Disease module hypothesis*:** Biological entities related to a disease phenotype tend to cluster in the same network neighborhood.
- ***Network parsimony principle*:** Causal molecular pathways tend to overlap with the shortest paths between disease phenotype associated biological entities.
- ***Shared component hypothesis*:** Diseases sharing the same biological entities show phenotypic similarity and co-morbidity.

These hypotheses are based on the analysis of current biological networks and as such are used for inferring, predicting and interpreting functional properties of newly investigated phenotypes.

Disease modules harbor another pitfall. When we observe disease phenotypes we have to be aware that the phenotypic observation can in truth be related to many different disease which happen to affect similar functions resulting into similar phenotypes. Fever is an example for a phenotype which is related to many different causes. Another prominent example is autism which is a highly complex genetic disorder [320]. While the causes may be different they can be nevertheless related to each other by targeting the same component in the cellular system [321].

Identification and prediction of phenotype specific processes and biological entities

Another area of analysis is the identification of phenotype associated processes and prediction of processes and nodes complementing noise level and error rates within biological experimental techniques and study setups. Identification of processes and pathways that are associated to a phenotype of interest is usually done by measuring a phenotype with experimental techniques, obtaining a list of biological entities as a candidate set that are relevant and performing an enrichment analysis of these candidates within given processes, pathways or other categorizations. An enrichment analysis calculates if the set of candidates results in a statistically over- or underrepresentation of given categories when compared to a random picking of candidates. Interpretation of categories found to be differentially represented between phenotypes follows that the function or category associated with the difference is relevant for the phenotypes. Methods for calculating the statistical significance can be divided into associative (or self-contained) and competitive (or enrichment) methods [322]. Methods include are the Fisher's product method [323], Truncated Product methods [324], the Tail Strength Measure [325], regression models [326] and more commonly used methods include Chi-square, Fisher's exact test, binomial probability and hypergeometric distribution, which are compared by Rivals et al. [327]. Another approach for defining

significance is called gene set enrichment analysis (GSEA), which avoids cut-offs considering biological relevance by small changes to concentration profiles and building a maximum enrichment score which is compared to randomly shuffled maximum enrichment score distributions [328]. Huang et al. [329] give an extensive overview on enrichment analysis and provide a list of 68 gene enrichment tools. Hung et al. [330] discuss recent advances in gene set enrichment.

Performing an enrichment analysis and identifying pathways, processes and categories is straightforward statistics but error rates and noise level in molecular experimental setups are actually a very serious issue with tremendous impact on the quality of our data making any analysis of the data and especially the interpretation very challenging [331]. The list of biological entities measured is usually not complete (false negatives) or include wrongly detected entities (false positives) and as such an enrichment analysis just gives us a hint on the underlying categories but is biased by those experimental errors. Addressing false negatives prediction of other biological entities of relevance within the given phenotype focuses on including network knowledge and hypotheses identified by the characterization of biological networks. For example following the local hypothesis any interaction neighbor of a candidate entity is more likely to be of relevance for the phenotype. Applying expansion on candidate lists based on such neighborhood and similar observed concepts will more likely identify false negative candidates. Predictions based on these approaches are very often done manually by network exploration or aided by utilizing expansion algorithms and merged or filtered with information from other perspectives resulting in a functional model that has to be validated experimentally. Chen et al. [332] identified additional Alzheimer related proteins by investigating protein interaction neighbors following a next neighbor approach. Hodges et al. [333] identified additional interactions in the ROS pathway from *Escherichia* coli utilizing a Bayesian networks expansion from gene microarray data. Expansion algorithms have been investigated extensively and discussing the impact on biases and enrichment and are presented in chapter 2 in the '*Comparative analysis of expansion methods on protein interaction networks*' project.

Integration of data

Comparisons of databases for the interactome revealed tremendous lack of overlap, even between primary databases as has been shown by Lehne and Schlitt [269] and has also been discussed by De Las Rivas and Fontanillo [334] (see Figure 25).

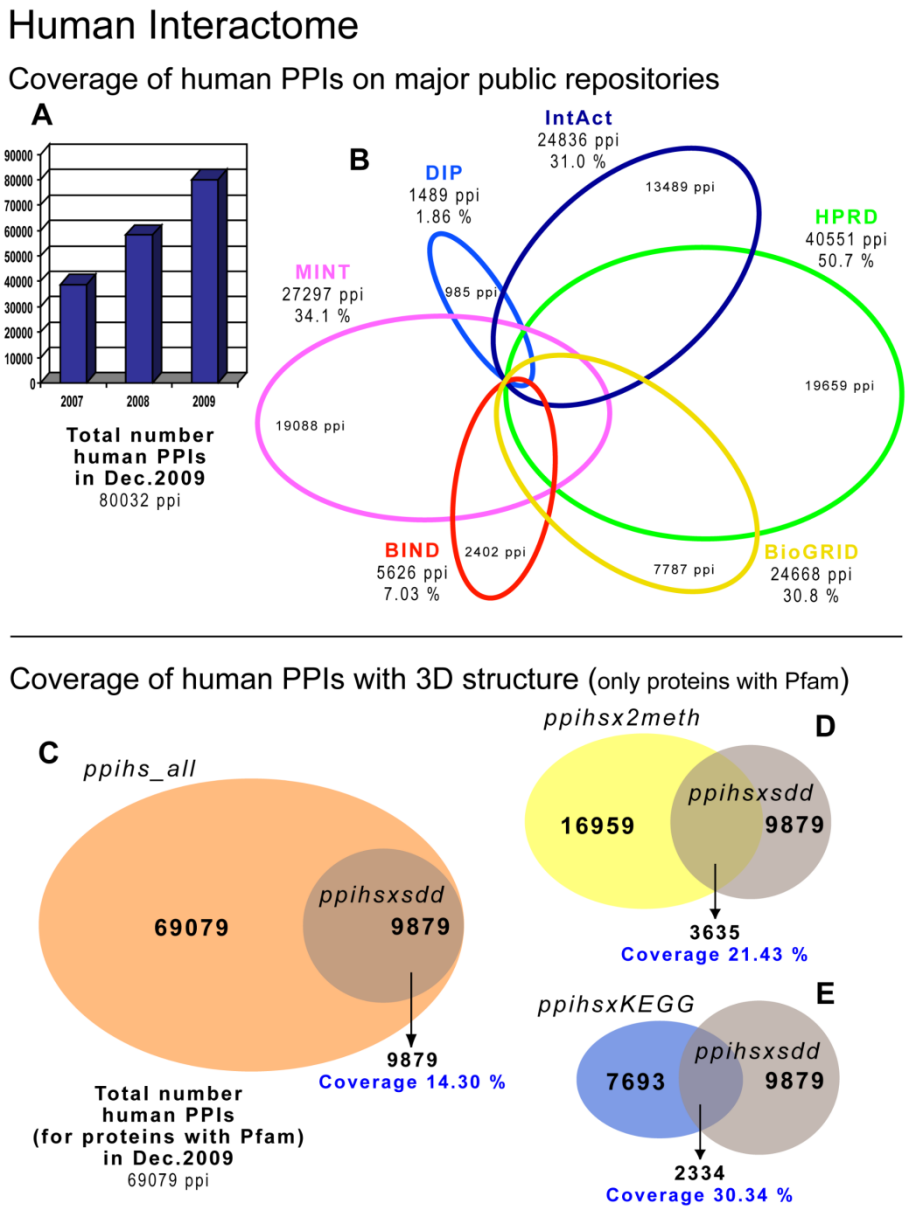


Figure 25 from Del Las Rivas and Fontanillo 2010 [334]: Human interactome: overlap of six databases and coverage of 3-D structural data. Analysis of human interactome PPI data showing the coverage of six major primary databases (BIND, BioGRID, DIP, HPRD, IntAct, and MINT), according to the integration provided by the meta-database APID. (A) Growth of the total number of human PPIs during the last 3 years. (B) Number of PPIs obtained from each primary repository showing the % (with respect to the total number of PPIs: 80,032 in December 2009) and the number of PPIs only reported by each database (shown inside the corresponding sector of the Venn diagram). Coverage and intersection of PPIs with 3-D structural information: (C) Intersection

between the PPIs of all human proteins that have at least one Pfam annotated (69,079 interactions, called ppihs_all) and the PPIs that include proteins with 3-D structural information (9,879 interactions, called ppihsxsdd); (D) intersection between the PPIs with 3-D structural information and a more stringent interactome constituted by PPIs proven at least by two experimental methods (16,959 interactions, called ppihsx2meth); (E) intersection between the PPIs with 3-D structural information and more stringent interactome constituted by interactions between proteins that are annotated to the same KEGG functional pathway (7,693 interactions, called ppihsxKEGG). doi:10.1371/journal.pcbi.1000807.g002

The ConsensusPathDB [115] provides an online statistics illustrating pair-wise overlaps between prominent data sources [335]. The best overlap of unique interaction could be found between BioGRID (version 3.2.112) and IntAct (version 2013-09-02+01:00) with just 25,847 joint interactions out of the 139,162 (18.6%) unique interactions in BioGRID and 47,585 (54.3%) unique interactions in IntAct. APID [92] also provides an online available statistics [336] of overlaps between IntAct [97], MINT [98], BioGRID [94], HPRD [96], BIND [93] and DIP [95] on the level of all combinations of overlaps between those based on interactions also available in APID with 63 interactions overlapping between all. The most likely cause for this small overlap between databases is the difference in expert curation of literature data as shown by Lehne and Schlitt [269] and the amount of experimental verifications of interactions. APID statistics show that from the total amount of the 322,579 interactions reported in APID 278,539 interactions were verified by 1 experiment only, 25,185 interactions by two experiments and 8,981 by three experiments. Now adding up data from prediction techniques, which heavily rely on the primary data sources, we can expect even further differences.

The scientific community approaches this issue by integrating data sources and adding a confidence score to the type of interactions and entities and by extending analytical approaches on a multi data scale. Thus a loss of information however likely is avoided as opposed to approaches which rely on overlapping information only (i.e. high confidence). While this concept sounds simple the approaches are manifold focusing mainly on data integration and the detection of significant or likely patterns for inferring a causal plausibility descriptive model for the phenotypes of interest. The following paragraph presents a few selected examples giving an overview on various integrative approaches.

Bernthaler et al. [337] integrate several sources into a protein dependency network 'omicsNET'. Approaches like Bayesian networks from Troyanska et al. [338] for protein function prediction illustrated the advantages of probabilistic models. Tu et al. [339] integrate genomic, gene expression, protein-protein interaction, protein phosphorylation and transcription factor binding information for causal gene identification and gene regulatory pathway inference. Albanese et al. [340] integrate imaging data in cancer treatment with multiple molecular information including a differential dependency network illustrating significant changes of topology within gene regulator networks. Neylan et al. [341] investigate posttraumatic stress disorder by integrating large-scale, high-dimensional molecular, physiological, clinical, and behavioral data for inferring perturbations within the molecular networks thus identifying possible biomarkers. Toubiana et al. [342] construct correlation based networks from time-resolved metabolomic data for studying plant metabolism. Pastrello et al. [343] discuss visual data mining within complex integrated data repositories utilizing their NAViGaTOR software [344]. Khurana et al.[345] integrate molecular data from various sources for generating a classifier for prediction of functionally essential loss of function

tolerant genes and additionally confirmed by correlation analysis that functionally essential genes tend to be more connected. Heinzel et al. [346] link clinical phenotypes and molecular functional unites derived from segmenting proteome interaction networks on a per patient level addressing biomarker and drug-target identification for specific patient strata. Mühlberger et al. [347] address molecular investigations of the cardiorenal syndrome by integrative bioinformatical approaches utilizing literature mining techniques and the integrated dependency graph 'omicsNET' from Bernthaler et al [337]. Kidd et al. [348] integrate various sources and methodologies for analyzing high-dimensional biological data in context of the immune system. For prediction of drug toxicity within systems pharmacology Bai and Abernethy [349] discuss the integration of biological network data focusing on structural ontological data and mathematical approaches. Mitra et al. [350] discuss integrative approaches for identification of modules in common biological functions. Leung et al. [351] elaborate on multi-target drug identification from herbs integrating 9 different platforms.

Discussion of limitations and issues with the study of biological networks

The study of biological networks has furthered the advances in the molecular life sciences significantly over the last decade, as has been exemplified in the previous sections. There are however many limitations and issues which have to be considered and kept in mind when working and analyzing biological network and omics data. This chapter will present a discussion of the following major limitations and issues for biological networks. The reliability of biological network data will be discussed on the level of heterogeneity and data integration. Since many interactions are derived from literature we will shortly address the issues of literature bias. Due to the heterogeneity of available data sources, a common problem throughout all omics is the matching of identifiers and the conceptual matching of data models, which makes data integrating a challenging task. On the level of technical issues the level of noise and the error rates resulting from measurements will be addressed. On the level of concepts we will address the difference in the actual effect, the phenotype and functions we are interested in and what is actually being measured, namely molecular features. A further conceptual issue addressed is the disparity of samples and tissues followed by a discussion of the time scale of measurements and network data and the interpretations derived from singular points in time. On the level of analysis issues can arise from analytical methods themselves due to many reasons. Issues from statistics permeate the whole field of omics data analysis, like the curse of dimensionality [352]. Statistical issues related to biological network reconstruction and hypothesis testing will be presented. For a detailed discussion of issues with omics related data in general see Prohaska and Stadler [331].

Reliability of biological network data

As has been outlined previously, the source of biological network data is highly heterogeneous and the overlaps between primary databases are rather small (see Figure 25 and [269], [334], [335]) due to differences in literature curation and experimental verifications of interactions [269]. Levy et al. [353] address the noise within protein-protein interaction networks and propose estimation approaches on the overall noise level. Kiemer and Cesareni [354] summarize that up to 30% of interaction artifacts stem from experimental issues. Extending on

the small overlap of primary data source, interaction data based on prediction methods should be carefully used when generating models and tested. Additionally we can be quite certain that our current knowledge of the interactome is far from complete and we have to be aware of missing and incorrect data. Integration of the data is currently an ongoing process which aims at unifying our current knowledge and usually includes solid interaction data with vague interaction data, thus allowing experts to consider their models carefully based on the evidence and their hypotheses. Mora et al. [74] for example showed that data integration is just beginning for many aspects. They constructed 497 multigenic disease groups from OMIM and tested the overlaps with interaction and pathway data. No single database in their survey contained all significant overlaps. Collaboration on a global scale is attempted by the IMEx consortium [355] which aims at curating major public interaction data sources from their members (as of August 2014 holding 14 members) and providing the interaction data in standards compliant download formats. Integration of network data is not trivial at all. The proteomics standards initiative provides a molecular interaction format which is widely used but the controlled vocabulary is often not used or used incorrectly [269]. Another problem permeating all aspects of the interactome is the issue of identifier matching, which will be discussed in the next section in more detail. Depending on the mapping approaches, mapping errors are not uncommon. Reliable mapping approaches like sequence similarity based mappings are advisable.

Literature bias

Text mining for extracting interactions from literature is commonly applied [356] and is prone to conceptual biases. Trends in research can lead to topics and biological entities of high interest. For example disease related genes tend to be more studied and as such more literature exists for them. Literature mining methodologies have to keep this bias in mind. For example in Ihop [137] the connectivity will be higher for disease related genes.

Biological identifier matching and data modeling

Almost all major molecular databases developed their own conventions when assigning identifiers for biological entities. Matching identifiers between different sources has been a major issue throughout omics research [357]. Errors in identifying biological entities thus needing corrections or the frequency of updates or databases or differences in the concepts of representation of the biological entities are common issues. For example literature text mining techniques often target HUGO gene symbols. Genes usually have an official symbol and a list of former aliases. A gene symbol found in especially older literature can easily be ambiguous and out of date. Tools like Excel can change gene symbols due to auto correction which can be overseen in large data sets [358]. Major biological databases aim at cross-referencing each other like Swissprot and NCBI. Additionally mapping tools exist like the SOURCE tool [359], MatchMiner [357] or Protein cross-reference tool PICR [360] which aims at mapping proteins based on sequence similarity.

Creating data models for integration of data is also challenging since concepts between the data sources often do not match. For example linking genes and proteins directly in one concept could be modeled as one-to-many relation while another concept could include transcripts in between them resulting in a one-to many-relation between genes and transcripts

and a one-to-one relation between transcripts and proteins. While this example can be resolved on the modeling level integrating the data into a joint new model can be challenging if the transcript information is missing in one. Since biological data is error prone even on the level of identifiers, data models and management have to be able to deal with that.

Technical issues caused by the measurement

Technical issues stem from the measurement technologies themselves. For example in microarray technology the sensitivity of oligonucleotide probes has been addressed by Binder et al. [361] and calibration by Binder and Preibisch [362]. Tandem affinity-purification/mass-spectrometry (TAP-MS) for measuring protein interactions is very error prone resulting in high rates of false positives and false negatives [363]. Proteomic explorative technologies like 2D gel electrophoresis cannot detect small peptides or variant types like membrane bound proteins for which a separate protocol needs to be used [364]. Technical issues can lead to upstream problems in the analysis of the data as will be demonstrated in the project 'Angiogenesis in Brain Metastasis' in chapter 2, where only 3 out of 4 housekeeping genes showed sufficient data. Giving another example, experimental detection methods for protein-protein interaction methods are affected by technical issues impacting the accuracy of those measurements [365], [366]. 'Nucleic acid'-protein measurements relying on antibodies depend on the sensitivity and specifity of the antibodies against the targeted protein component. Consequently we are confronted with a certain level of noise and false positive and false negative error rates within molecular measurements due to various technical issues.

Conceptual issues of measurements

"What do we actually measure?" is one important conceptual question one has to be aware of when interpreting and analyzing data. For example when we explore a disease utilizing genomic microarrays we get a list of genes but how do we interpret their role? What we did measure was the expression of genes, to be precise, the mRNA concentration of the sample. While there is a systematic link between gene expression and phenotype, it is not necessarily a causal link. The concentration of the resulting proteins which are actually responsible for the disease effects have not been measured in this case. Further, we measured quantities but not activities, while it might be safe to assume that more gene product will result in stronger effects (depending on what the protein does in its context) this does not have to be a linear effect nor is it guaranteed to have an effect at all (because for example another protein is masking its activity). We in fact measure molecular features but not the phenotypes (i.e. cellular functions). These fundamental issues have to be kept well in mind.

Sample and tissue specifity

When exploring phenotypes in search of causal explanations we map measured phenotype data on various omics level onto networks. The probe material we obtain comes from specific regions and tissues. Cells are highly differentiated and as such will of course show a different molecular pattern directly affecting the comparability and reproducibility. This issue is often avoided in vitro by cell culture where the material can be grown from the same cell line. While this is especially practical for conceptual homogenous probes or when there is not enough samples available, one has to be aware that cell cultures are based on immortalized cell lines

and as such are similar to cancerous cells. Comparison to in vivo tissue is questionable. We demonstrate in 'Mesothelial cell stress response and cytoprotection in peritoneal dialysis' in chapter 2, that the proteomic pattern of samples from different cellular splits within the same cell culture experiments could be clustered into different groups. The control samples and treatment samples on the other hand could not be separated into distinct groups. Tissue specifity on the level of male and female or grownups and children are seldomly investigated.

Time concept issues

As has been elaborated prior, dynamic data is rare within interaction networks. The interactome on a descriptive level, as it currently is, resembles a state of possibilities which have been gathered by our observations and also include errors. We have to be aware, as pointed out, that those interactions depend on time, on tissues, on developmental stages and other influences. As such they are possibilities of what can be, given the circumstances match.

Especially when we map explorative data from phenotypes to interaction networks we have to be aware that in most experimental setups data was collected at time points and as such we analyze time points like seeing single pictures from a movie, implying that the later time point is dependent on the previous. The more pictures we see the better our understanding of the story will be.

Computational and algorithmic issues

Computational issues arising within networks usually result from algorithmic complexity or amount of data. Graph theory knows many problems that are not solvable within realistic amount of time for practical applications like the *k*-coloring of a graph, the travelling salesman or Knapsack, etc. Applications in biological networks include the search for motifs which requires checking of graph isomorphism which is a NP problem or the search for the shortest path which can be calculated in reasonable time (linear-logarithmic). Searching for all shortest paths can get demanding though. Approaches for inferring networks from data can be very computational intensive [240]. An advantage with biological network data is that most calculations are not time critical (except for online services like BLAST) or need to be calculated only (like graph properties) once after every major update of a network. Nevertheless approaches for increasing calculation speed can be addressed by parallelization. Salwinski and Eisenberg [367] for example use a highly parallel architecture based on Field Programmable Gate Arrays for in-silico simulation of biological network dynamics. Zhou et al. [368] perform parallel programs on graphics processing units for dynamic network simulation. Another classical strategy for addressing computationally intensive algorithms is approximations thus accepting suboptimal solutions.

Another challenge with biological networks is the design of algorithms mirroring or searching for biological patterns. For example the generation of random graphs matching real world properties has led to several random graph models with none being truly ideal. In 'Comparative analysis of expansion methods on protein interaction networks ' in chapter 2 an algorithm for approximating a *k-minimum spanning tree* that spans over a defined set of *k* nodes is presented. While it is not optimal in all cases, it is sufficient from a biological point of view.

Modules that have been identified by prediction approaches will be noisy and contain errors. Barabási and Oltvai [270] showed that the predictions of modules are often dependent on the methods and parameters used in the initial data partitioning and that inaccurate or missing data of the interaction networks can lead to biased predictions. Consequently no preference for any detection method can be given. Additionally detected modules do not necessarily provide a mechanistic model and are rather a collection of entities which serve most likely a common function. Nevertheless this can be advantageous for identifying functions of biological entities within a module since their function will be closely related to the modules function.

Statistical issues

Construction of networks from genomic or proteomic data is supported by various statistical approaches like clustering techniques, Bayesian network or supervised learning approaches [54]. Clustering approaches from expression rely on the assumption that co-expression is related to function [369] yet recent studies indicate that this is rather the exception then the rule [370]. As such approaches focusing on conditional independence, like Gaussian graphical models were applied [243] but capture undirected relationships and are thus not suited for regulatory networks. Bayesian networks address this issue but can only infer acyclic directed networks [244]. Additionally Dynamic Bayesian Networks are computationally very demanding [240]. Including perturbations into experimental data for inferring network structure were proposed by Markowetz et al. [251], [371] with nested effects models which tend to be computationally challenging [372], [373].

Gene set enrichment is a prominent analytical approach that provides a system level point of view on the changes in molecular systems [330]. Emmert-Streib and Glazko discussed that the gene-gene relations have an influence on the power of the tests [374]. A network based gene set analysis was proposed by Shojaie and Michailidis [375].

CHAPTER 2
Scientific Contributions

Overview

This chapter presents scientific projects the author had undertaken and contributed to during the course of his dissertation and are based on the extensive summary on the state of the art of molecular networks, as presented in chapter 1. From the presented network types, protein-protein interaction networks were mainly investigated and modeled. Available data required descriptive models. Several data sources were required for building reference networks and providing annotation and additional information for the analysis within the contributions. Various analytical methods were applied and new ones developed based on toplogical features and known patterns.

The first project "*Basic network analysis software suite and KEGG interaction network modeling"* aimed at implementing a basic graph analysis platform, which served as network-algorithm tool for the other projects presented in this chapter. Additionally a data extraction routine and a data model was developed for a KEGG protein-protein interaction network, which was used in the investigation of synthetic lethality for linking the mycophenolate mofetil mode of action with molecular disease and drug profiles [72] and during the "comparative analysis of expansion methods on protein interaction networks" presented later in this chapter.

The second project *"Ovarian cancer analysis"* was a meta-analysis of ovarian cancer based on gene expression and autoimmune data. The project was a cooperation between the Institute for Theoretical Chemistry, University of Vienna, Austria; emergentec biodevelopment GmbH, Vienna, Austria. Contributions presented focused on data preparation and integration, a subcellular location analysis, supporting a co-regulation analysis and performing a pathway enrichment analysis. Additionally a conceptual workflow had been discussed for integrating and analyzing omics data. Results of these contributions have been published in [376], [377] and were presented at [378], [379], [380] and [381].

The third project *"Mesothelial cell stress response and cytoprotection in peritoneal dialysis"* was an explorative proteomics analysis of cellular stress of mesothelial cells during peritoneal

dialysis. The project was a cooperation between the Department of Pediatrics and Adolescent Medicine, Medical University of Vienna, Austria, Institute of Analytical Chemistry and Food Chemistry, University of Vienna, Austria, emergentec biodevelopment GmbH, Vienna, Austria, and Institute of Analytical Chemistry, Academy of Sciences of the Czech Republic, Brno, Czech Republic. Contributions presented focused on the analytical preprocessing and data integration of proteomics data and the analysis of proteomic profiles between several setups as well as the identification of putative false negative proteins by means of expansion algorithms. Results of these contributions have been published in [382], [383] and were presented at [380]

The fourth project *"Comparative analysis of expansion methods on protein interaction networks "* was a research question followed by the author and was a cooperation between the author and emergentec biodevelopment GmbH, Vienna, Austria. The hypothesis addressed and investigated in detail focused on the comparison of expansion algorithms for identifying false negative proteins from omics measurements on the level of topology and pathway enrichment within reference protein-protein interaction networks.

The fifth project *"Angiogenesis in brain metastasis"* was an explorative gene expression analysis on the level of RT-PCR assays of angiogenesis related cancer types from patient samples. The project was a cooperation between the author and Dr. Matthias Preusser and Dr. Aysegül Ilhan-Mutlu of the Clinical Division of Oncology, Department of Medicine I, Medical University of Vienna, Austria. Contributions presented focused on the data preprocessing and data integration of RT-PCR data and a network based expansion of the candidate set. Results of these contributions were published in [384].

For further scientific contributions not related to biological networks refer to the curriculum vitae.

Basic network analysis software suite and KEGG interaction network modeling

Introduction

For the investigation of biological networks a graph software suite was required and developed in Visual C# 2005 since no public available library was found at the start of the dissertation work. Graph libraries already existed for many programming languages like the Boost Graph Library in C++ [385] and implementing one in C# is straightforward. The graph software suite contains core graph functionality to represent any type of graph. Algorithms for analyses were implemented on a per need basis during the different projects addressed in this work and will be presented there. Besides construction of a C# library, biological network data had to be retrieved and the models mapped on the graph structure. Depending on the available types of data retrieval options, import routines were specifically implemented for a data source where necessary. Additionally the data models were adapted where necessary to better match biological concepts and graph concepts. KEGG needed such a conceptual remodeling for adapting the information available in KEGG to be comparable with the data sources and will be discussed extensively in the results.

Material and Methods

Graph data structure

Biological entities are forming the nodes of a graph and the relations between them form the edges. For graph representation adjacency lists are chosen, storing for each graph node the list of neighbors with the weight of the edges. The first level of the adjacency list representing the nodes is realized in form of a hash table allowing for a quicker average search of nodes when compared to list-based data structures. The second level of the adjacency list representing the neighbors of a given node is realized as double linked list for quicker walking though all neighbors, storing neighbor node identifier as well as the edge weight (see Figure 26b).

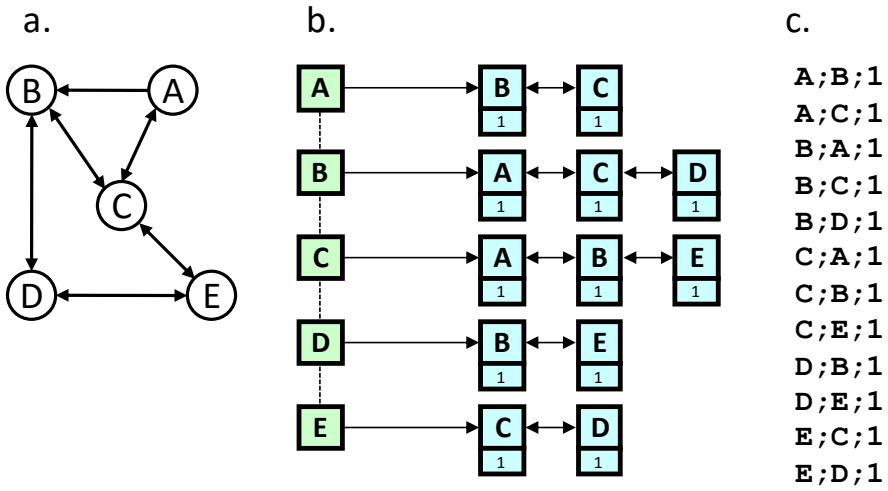


Figure 26: Graph representation as adjacency list. (a.) depicts an undirected, unweighted graph. (b.) represents a scheme of the corresponding adjacency list. The left green nodes represent the first level implemented as Hashtable. Each node has a reference to its neighbor nodes (blue list on the right) in form of a double linked list and representing the second level. The reference to the neighbors additionally stores the edge weight for each neighbor edge (here 1 for representing uniformly an unweighted edge). (c.) represents a flat file representation with semicolon separator. Each row denotes an edge from the first node entry towards the second node entry with the edge weight as third entry. Since the graph here is undirected, both directions are provided.

Biological network data sources

Example datasets were included from OPHID [128], BIOGRID [94] and KEGG [90]. OHID and BIOGRID provide data as downloadable flat file formats in a simple graph format. KEGG provides data as downloadable XML files and as web services using SOAP/WSDL and REST APIs.

Network construction

The information needed for constructing graphs from network data includes the nodes and edges. A simple file representation consists of a delimited flat file format where each row represents the networks edges by listing both nodes of the edge. In case of weighted graphs a third entry could denote the edge weight. In case of directed graphs the order of the nodes within each row defines the edges direction. This is the minimal information required for building a graph from network data (see Figure 26c). Several databases like OPHID and BIOGRID provide such a network representation and the construction of a graph is straightforward. KEGG does not provide network data in a simple tabular format.

KEGG Pathway data

KEGG provides among others molecular pathway and protein-protein interaction related information representing our knowledge on molecular interaction and reaction networks. KEGG developed an xml based format, the KEGG Markup Language (KGML) [386] which provides an exchange format of the KEGG graph objects supporting the manually drawn pathway map, computational analysis and modeling of protein and chemical networks. For the KEGG pathway data model see Figure 27. The model is centered on separate pathways all

having their own abstract identifiers and was adapted for merging all pathway centered interaction data into one protein-protein interaction network. Access to the data is provided for via ftp download and web services based on SOAP/WSDL and REST. The REST based API service succeeded the former SOAP based service on July 1, 2012, the later being shut down on 31.12.2012.

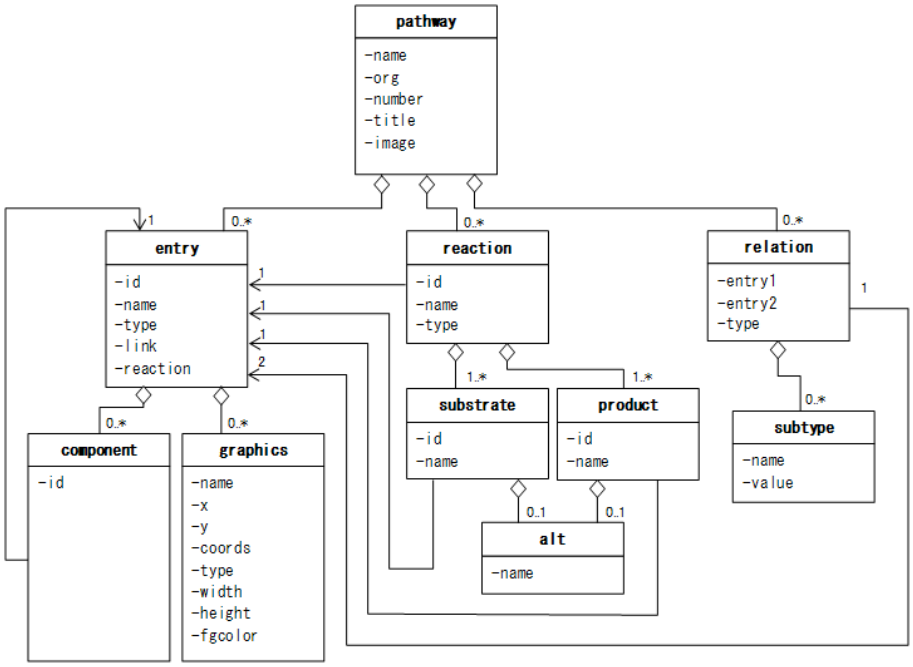


Figure 27: KGML entity relationship model. Source: M. Kanehisa, "KEGG Markup Language," Kanehisa Laboratories, [Online]. Available: http://www.kegg.jp/kegg/xml/docs/. [Accessed August 2014].

Network construction from KGML files

After downloading the network data files (KGML version 0.6.1), construction of the protein interaction network graph was performed identifying the biological components relevant for protein network construction and writing a program in PERL for parsing the xml files.

Network construction from SOAP/WSDL API

KEGG provided a SOAP/WSDL based web service for accessing the database [387] allowing an integrated extraction and processing of network data without the need for writing a parser. The SOAP/WSDL service was included by linking to the wsdl file at http://soap.genome.jp/KEGG.wsdl. Extraction of protein interaction data for human was done firstly by extracting all human pathways with the *list_pathways("hsa")* function and extracting the KEGG pathways identifier for each pathway. Secondly entries from the pathways according to the KGML model are extracted with the *get_elements_by_pathway(CurrentPath)* function. Pathway identifiers, entry identifiers, entry names referencing to biological entities and entry types were extracted and saved to file. In case of entry types ‘group’ and ‘ortholog’, the set of

biological entities had to be extracted additionally for each type from the functions resulting entry object. The biological entities resulting from this set were merged with a separator and saved as entry name. After extracting the entries, the relations were extracted for each pathway by utilizing the *get_element_relations_by_pathway(CurrentPath)* function resulting in a list of pathway entry pairs. Pathway identifier, entry pairs and relation type were extracted and saved to file. Interactions between the proteins and complexes were extracted according to the model with added functionality, since KEGG SOAP service does not provide this information directly.

Results

Each KEGG pathway stores its biological elements via an intermediate, abstract entry element which is unique for a given pathway only and can consist of one of 7 different biological entities like genes, chemical compounds and protein complexes or even map links. Interactions between two entries are defined via the relation type which can consist of one of 5 different relation types like enzyme-enzyme relation or protein-protein interaction. Extracting from the data thus requires identifying for each pathway the entries (nodes), filtering for protein or gene related types, extracting the relations (edges) and filtering for protein interaction types.

Interactions provided by KEGG are not only direct protein-protein interactions but also interactions between protein complexes or protein complexes and singular proteins. Thus the construction of a protein interaction network graph has to either dissolve the biological complexes and connect the proteins with each other or include the complexes as additional node types. Preserving the protein complex interaction information the second approach was chosen and the graph constructed consists of protein nodes and protein complex nodes. Identifier for the nodes is the KEGG gene identifier in case of singular protein nodes and all gene identifiers merged with a separator in case of protein complexes. While the later can generate long identifier for a protein complex node, this approach was nevertheless chosen since the elements of the node can easily be split avoiding an abstract additional identifier for protein complexes. Additionally each singular protein from a protein complex was included as singular node within the graph and linked to all other proteins from the same complex, as well as to the protein complex resembling the protein-protein interactions required to form a protein (see Figure 28).

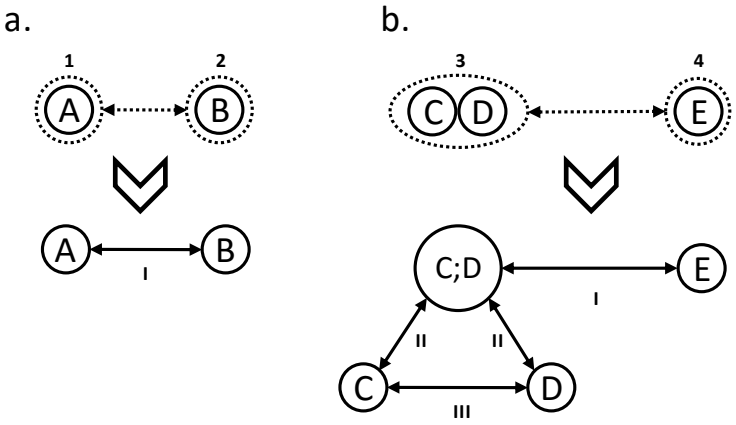


Figure 28: KEGG pathway data model adaption. In each KEGG pathway the entries are the abstract nodes (nodes with dotted circle and number) having a relation with other entries (dotted arrows). The adapted model has nodes representing directly proteins (solid circles) and direct interactions (solid arrows). (a.) depicts the adaption of two singular entry type nodes from KEGG into the adapted model. (b.) depicts the adaption of a grouped entry type node from KEGG with a singular entry type node. The grouped node gets a merged label from its protein elements (*C* and *D*) and both elements are additionally linked as new nodes with each other and the node protein complex node *C;D.* Edge types are marked separately. 'I' denotes edges as reported from KEGG, 'II' denotes edges that link a new protein node from a group to the complex protein node and 'III' denotes edges that connect each new protein node from a group node.

Discussion of the adapted KEGG model

The adapted model presented here dissolves the abstract data representation of the pathway centric entries in the KEGG KGML data model and merges the entries on a more realistic basis. The concept of groups and protein complexes like the complement system interacting with other proteins is addressed hereby. Singular proteins constituting to a complex have to interact with each other to form the complex. We can differentiate two types of interactions in this context. On the one hand the interactions between the singular proteins with each other and on the other hand the interaction of each singular protein with the complex itself. The complement system for example forms a ring like structure. Two critics to this model are thereby possible. On the one hand side creating a full clique out of the proteins from a complex can add false edges. For example the proteins within the complement system are not all interacting with each other directly to form the ring like structure. The second critic is if the interactions of these singular proteins with the complex are real protein interaction or conceptually a 'is part of' relation. Addressing the second critic the edges were labeled accordingly using the graph data structure, which provides such multi edge type graphs, thus differentiating between initial edges as reported by KEGG, protein interactions between the members of a complex and thirdly interactions between the members of a complex and the complex node (see Figure 28). Zhang and Wieman [388] proposed a R package KEGGgraph which addresses the same issue and provide separate topological arrangements of the KEGG data.

Ovarian cancer analysis

Introduction

It has been observed that cancerous tissue can lead to cellular and humoral immune responses [389]. It has been unclear though why proteins become auto-antigens in a humoral immune response due to cancer development. Studies indicate that proteins are presented in cancer cells that would normally not be presented [390]. Other findings indicate that intracellular proteins are released by cancer cells into the cellular environment and that abnormal splice variants are expressed [391], [392]. Another interestingly effect that might explain auto-antigenicity is that an increase of protein concentration can lead to a humoral response as has been found for *p53* mutations, which often increase the stability of the protein [393]. Following these findings the core hypothesis during this project is that the abundance of a protein in cancerous tissue is related to the probability that the protein will induce a humoral response.

For investigating this hypothesis ovarian carcinoma was chosen as cancer case study since the data available was adequate. Gene expression data was collected in a meta-study from 25 publications. Auto-antigenicity data was retrieved from the SEREX data base [394]. The investigations during this contribution addressed the hypothesis by directly comparing transcriptional up-regulated gene lists form the meta-study with the auto-antigenicity gene lists. Further analytical steps included the identification of indirect relations via subcellular location analysis and network analysis based on co-regulation and protein-protein interactions.

Material and Methods

Data sets

Data for investigation consisted of two major datasets related to ovarian cancer. The first dataset is based on differential gene expression and was derived from a meta-analysis of 20 publications from 1999-2005 that compared cancerous and healthy tissue [377]. The second data set covering auto-antigens was derived from serological expression cloning analysis (SEREX) or protein arrays publically available at a web database [394].

Meta-analysis data preparation

Publications were screened for reported genes being either up or down regulated significantly between healthy and cancerous ovarian tissue. Gene identifiers were matched utilizing Stanford microarray database SOURCE tool [359] and genes were filtered depending on the difference between reported up- or down-regulation.

Subcellular location analysis

Subcellular location for given gene sets were extracted utilizing the Stanford microarray database SOURCE website [395] and genes marked if their subcellular location description could be associated to membrane bound or secreted. Missing information was computed utilizing PSORT algorithm [396] and added if the subcellular location categories from PSORT

results indicated at least a 20% prediction for extracellular/including cell wall. Using PSORT requires as input sequence information in FASTA format. Data was extracted from the National Center for Biotechnology Information (NCBI) [397] utilizing prior a PERL script for mapping Entrez GeneID to RefSeq protein identifiers using batched http requests and parsing the returned websites.

Supported co-regulation analysis

A co-regulation analysis based on GAnalyzer [55] aimed at identifying additional genes having similar transcription factor binding sites (TFBS) to relevant genes from the meta-analysis gene expression sets. Those genes were additionally considered as relevant and compared to the auto-antigene set from SEREX. Co-regulation analysis was supported by transforming data inputs between Confac analysis data [398] and GAnalyzer as well as filtering multiple redundant entries for genes having more than one mRNA sequence.

Pathway analysis

Databases providing pathway related information included Kyoto Encyclopedia of Genes and Genomes (KEGG) [100], [399], [90], BioCarta [104], [105] and Panther [172], [173], [400, p. 6], [174]. Pathway lists for KEGG and BioCarta including genes in form of HUGO gene symbols [401] were provided by Tomfohr et al. [402]. KEGG dataset holds 118 unique pathways and 1,997 unique genes and BioCarta dataset holds 263 unique pathways and 1,264 unique genes. Panther data was accessed via the Panther classification website [175] and provided pathways including genes in form of Entrez Gene identifier [403]. Pathway analysis was performed by matching genes to pathway data sets and extracting associated pathways for each set. Amount of pathways and found genes were counted. Merging the resulting pathway and gene list between gene expression sets and auto-antigen sets the joint pathways and genes resulting from those sets were identified (calculated for each pathway data set). Finally considering only pathways that contained at least one gene from a given gene expression sets as well as from a given auto-antigen set, the amount of such joint pathways was counted and compared to a randomized distribution. Calculation of the randomized distribution was developed utilizing PERL and included the generation of 1,000 gene set pairs for each randomization. Gene set pairs consisted each of randomized gene set (from a given pathway data set) of same size as the gene expression gene sets and auto-antigen sets of interest.

Software Tools

EXCEL 2003 (Microsoft Corporation, Redmond U.S.) was used for data handling and graphics. SPSS 13.0 (IBM, New York, U.S.) was used for statistics and graphics. Network calculations were done using the Basic network analysis software suite. Perl [130] was used for programming of data handling and analysis routines.

Results

Firstly the publication datasets were consolidated. The meta-analysis for the gene expression data focused on extracting all genes with reported significant difference between healthy and cancerous ovarian tissue. Reports needed to be filtered depending on the differences of reported up- and down-regulation. Afterwards genes were filtered having secreted protein products. SEREX data was provided and overlaps between both data sets determined. Further

investigations included supporting a co-regulation analysis for identifying further genes being possibly relevant in gene expression and a pathway analysis comparing similarities on the level of pathways between gene expression data and auto-antigens.

Data preparation

Consolidating the data for genes related to ovarian cancer in the set of 20 publications resulted in a list of 2,200 partially redundant genes with various annotations and if they showed an up or down regulation. Unique identifiers available covered Genebank Accession number [404], Entrez Gene identifier [403], HUGO Gene Symbol [401], UniGene identifier [405] and Reference Sequence protein identifier [406]. A PERL script was written for counting the amount of publications reporting up or down-regulations for each gene. Genes were found as having a different amount of publications reporting them as either up- or down-regulated. Thus we defined a gene as overall up- or down-regulated if it was overall reported more often as up- or down-regulated resulting in 786 genes being up-regulated and 871 being down-regulated with 53 having equal amounts of publications reporting them as up-or down-regulated. Considering genes which had been reported at least in two publications and omitting genes with no difference between reported up- and down regulations resulted in a unique list of 192 genes (see Figure 29). Maximum amount of publications referring to a given gene was 8.

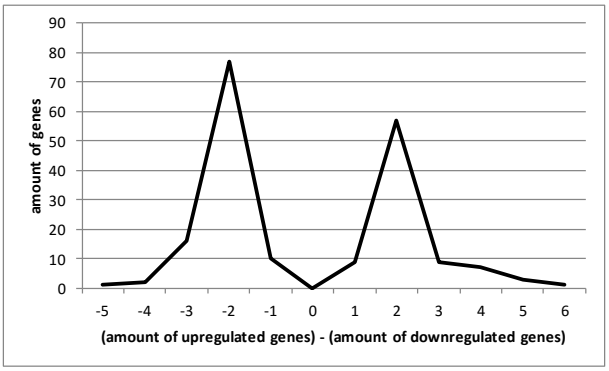


Figure 29: amount of genes (y-axis) having the same difference (x-axis) between up-regulated and down-regulated genes from the set of 192 genes being reported in more than one publication from the meta-analysis.

Auto-antigenes from SEREX were provided for consisted in a first setup of 118 genes and in a second refined setup of 81 auto-antigens.

Data analysis

Subcellular location analysis

After extraction of subcellular locations with the SOURCE tool 123 genes from the 192 genes set could be assigned to a subcellular location and 29 genes from the 118 SEREX set. Genes with membrane or secreted association were derived for 60 genes from the 192 gene set and for 7 genes from the 118 SEREX set. Missing information was calculated utilizing PSORT

subcellular location prediction. Extraction of FASTA sequences for the set of 192 genes set resulted in 186 sequences for subcellular location prediction and 112 out of 118 sequences for the SEREX gene set. Distributions of the resulting PSORT probability scores for each subcellular location category (see Figure 30) reveals that the largest fraction stems from nuclear associated subcellular locations for the given genes in both sets. Nevertheless additional genes could be identified for the 192 gene expression set resulting in a merged list of 86 genes from the 192 gene expression set as being associated with cell walls or secreted (further called Meta-UP set).

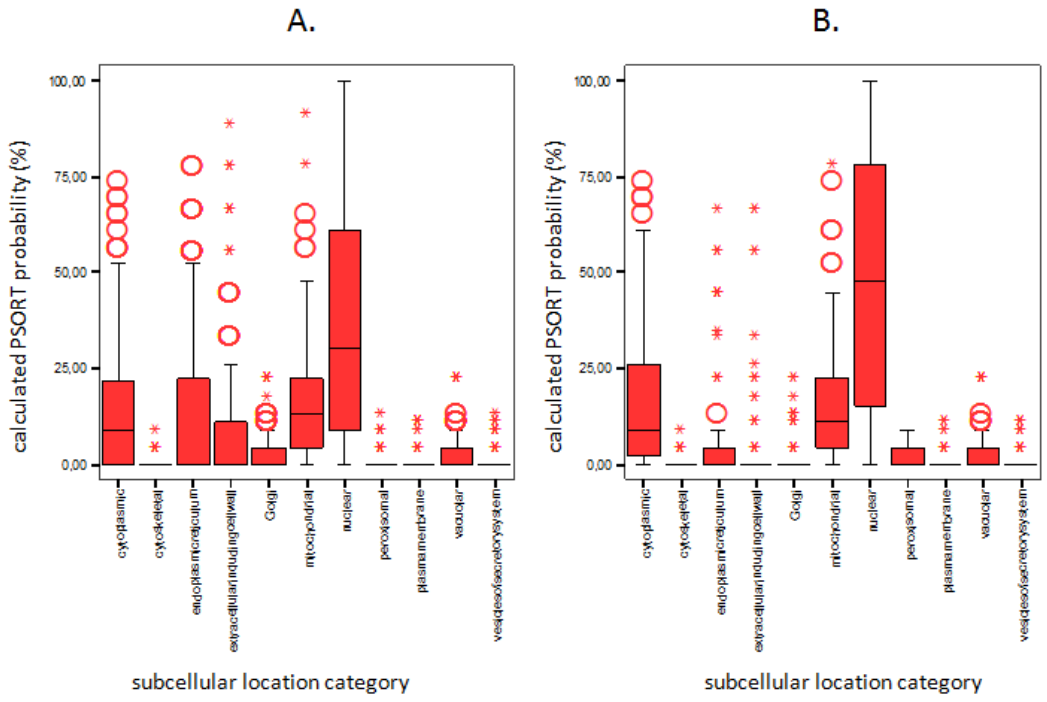


Figure 30: Box plots for distribution of calculated PSORT probabilities of the 192 genes from the gene expression meta-analysis (A.) and the 118 auto-antigenes set from SEREX (B.).

Co regulation analysis support

Additionally to the reported 192 genes from the meta-analysis we performed a Transcription factor analysis for identification of possible co-regulated genes utilizing GAnalyzer [55]. For supporting this analysis several data processing steps were included. Provided result files from Confac analysis [398] consisting of the consolidated transcription factor binding site (TFBS) matrix and Wilcox statistical test for significantly represented TFBS were transformed for input into GAnalyzer software. The consolidated TFBS matrix included redundancy regarding multiple similar entries of a gene due to multiple mRNA versions which were merged into a single gene entry and delivered for further analysis. Finally a list of 29 genes was reported back after co-regulation analysis.

Overlap analysis

The datasets from gene expression consisting of the full set, the refined set of 192 genes, the set of 86 Meta-UP secreted and up-regulated genes as well as the set of 29 genes from co-regulation analysis were compared for overlaps between the SEREX data set consisting of 118 genes. 15 genes matched between SEREX gene set and the full gene expression set, 4 between SEREX gene set and 192 refined gene set out of which 3 were also members of the 86 Meta-UP set and no genes were found from the co-regulation analysis within the SEREX gene set (see Table 2).

Table 2: Overlapping genes between the full gene expressions set, the refined set of 192 genes and the 86 secreted Meta-UP set from the meta-analysis with the auto-antigen set from SEREX. Genes are denoted by the identifiers Hugo Gene Symbol, Unigene identifier, Entrez Gene identifier. 'up' denotes the amount of publications in which this gene had been reported as up-regulated, 'down' for the down-regulated publications and the calculated 'difference' between both.

	Gene Symbol	UniGene	GeneID	up	down	difference	
	DNCH2	Hs.503721	79659	0	1	-1	
	RBM25	Hs.531106	58517	0	1	-1	
	BRAP	Hs.530940	8315	0	1	-1	
	TNNT1	Hs.534085	7138	1	1	0	
	HMMR	Hs.72550	3161	1	1	0	
	GPX1	Hs.76686	2876	1	0	1	
	RHOA	Hs.247077	387	1	0	1	
	MSLN	Hs.408488	10232	2	0	2	
86 Meta-Up set	BARD1	Hs.54089	580	2	0	2	***192 set***
	KRT8	Hs.533782	3856	2	0	2	
	PDGFRA	Hs.74615	5156	0	5	-5	

Pathway analysis

Comparative analysis of pathways between the 192 set, the 86-Meta-UP set and the 118 SEREX set was utilized with KEGG, BioCarta and PANTHER pathway data. For the 192 set within BioCarta pathway data 29 unique genes (15% of total set size) were found being spread multiple times over 77 unique pathways, within KEGG pathway data 55 unique genes (19% of total set size) were found being spread over 59 unique pathways and in PANTHER pathway data 46 unique genes (24% of total set size) were found being spread multiple times over 42 unique pathways. For the 86 meta-UP set 11 unique genes (13% of total set size) were found being spread multiple times over 24 unique pathways within BioCarta pathway data, within KEGG pathway data 23 unique genes (27% of total set size) were found being spread over 21 unique pathways and in PANTHER pathway data 17 unique genes (20% of total set size) were found being spread multiple times over 24 unique pathways. For the 118 SEREX set 7 unique

genes (6% of total set size) were found being spread multiple times over 22 unique pathways within BioCarta pathway data, within KEGG pathway data 17 unique genes (14% of total set size) were found being spread over 25 unique pathways and in PANTHER pathway data 18 unique genes (15% of total set size) were found being spread multiple times over 18 unique pathways.

Merging pathways from BioCarta between the 86 meta-UP and the 118 SEREX set resulted in 46 distinct pathways holding 36 distinct genes from both sets (see Table 3). Merging pathways from KEGG between the 86 meta-UP and the 118 SEREX set resulted in 21 distinct pathways holding 23 distinct genes from both sets (see Table 4). Merging pathways from PANTHER between the 86 meta-UP and the 118 SEREX set resulted in 10 distinct pathways holding 48 distinct genes from both sets (see Table 5).

Table 3: Pathways and genes covered within BioCarta pathway data sets merged between the 86 meta-UP gene expression data set and the auto-antigen set 118 from SEREX.

BioCarta Pathway	genes
Actions of Nitric Oxide in the Heart	FLT1, HSPCA, VEGF
Adhesion Molecules on Lymphocyte	CD44
Ahr Signal Transduction Pathway	HSPCA
AKT Signaling Pathway	HSPCA
ALK in cardiac myocytes	BMP7
Apoptotic DNA fragmentation and tissue homeostasis	HMGB1
Apoptotic Signaling in Response to DNA Damage	BCL2L1
Basic Mechanisms of SUMOylation	UBE2I
Cardiac Protection Against ROS	GPX1
receptors	PDGFRA
CCR3 signaling in Eosinophils	RHOA
Corticosteroids and cardioprotection	HSPCA
Endocytotic role of NDK, Phosphins and Dynamin	PICALM
Erk and PI-3 Kinase Are Necessary for Collagen Binding in Corneal Epithelia	RHOA
Erk1/Erk2 Mapk Signaling pathway	PDGFRA
Free Radical Induced Apoptosis	GPX1
Gamma-aminobutyric Acid Receptor Life Cycle	NSF
Hypoxia and p53 in the Cardiovascular system	HSPCA
Hypoxia-Inducible Factor in the Cardiovascular System	HSPCA, VEGF
IL-2 Receptor Beta Chain in T cell Activation	BCL2L1
Influence of Ras and Rho proteins on G1 to S Transition	RHOA
Inhibition of Matrix Metalloproteinases	MMP14
Integrin Signaling Pathway	RHOA
Keratinocyte Differentiation	PRKCH
Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa(alpha)	HSPCA
Monocyte and its Surface Molecules	CD44
Neutrophil and Its Surface Molecules	CD44
Opposing roles of AIF in Apoptosis and Cell Survival	BCL2L1
PDGF Signaling Pathway	PDGFRA
Phospholipids as signalling intermediaries	PDGFRA, RHOA
Control	SLPI
Protein Kinase A at the Centrosome	RHOA
Rac 1 cell motility signaling pathway	PDGFRA, RALBP1
Ras Signaling Pathway	BCL2L1
Ras Signaling Pathway	RALBP1, RHOA
Regulation of BAD phosphorylation	BCL2L1
Progression	CKS1B
Rho cell motility signaling pathway	RHOA
Rho-Selective Guanine Exchange Factor AKAP13 Mediates Stress Fiber Formation	RHOA
Hypertrophy	RHOA
Role of Mitochondria in Apoptotic Signaling	BCL2L1
Role of PI3K subunit p85 in regulation of Actin Organization and Cell Migration	PDGFRA, RHOA
Telomeres, Telomerase, Cellular Aging, and Immortality	HSPCA
Thrombin signaling and protease-activated receptors	RHOA
uCalpain and friends in Cell spread	RHOA
VEGF, Hypoxia, and Angiogenesis	FLT1, VEGF

Table 4: Pathways and genes covered within KEGG pathway data sets merged between the 86 meta-UP gene expression data set and the auto-antigen set 118 from SEREX.

KEGG pathways	gene
Adherens junction	RHOA, MLLT4, PTPRF
Alzheimer's disease	APOE
Amyotrophic lateral sclerosis (ALS)	GPX1, BCL2L1
Apoptosis	BCL2L1
Arginine and proline metabolism	CKB
Benzoate degradation via CoA ligation	CDC2L5
Calcium signaling pathway	PDGFRA
Cytokine-cytokine receptor interaction	ACVR2B, PDGFRA, BMP7, FLT1, VEGF
ECM-receptor interaction	LAMB1, HMMR, COL5A1, CD47, SDC4, ITGB8, CD44
Focal adhesion	RHOA, PDGFRA, ITGB8, FLT1, VEGF
Glutathione metabolism	GPX1, GPX3
Glycine, serine and threonine metabolism	AMT
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	PIGQ
Hedgehog signaling pathway	BMP7
Huntington's disease	DCTN1
Inositol phosphate metabolism	CDC2L5
Jak-STAT signaling pathway	BCL2L1
MAPK signaling pathway	PDGFRA
Neuroactive ligand-receptor interaction	SSTR5, NMU
Neurodegenerative Disorders	APOE, BCL2L1
N-Glycan biosynthesis	DPM1
Nicotinate and nicotinamide metabolism	CDC2L5
Nitrogen metabolism	AMT
Notch signaling pathway	JAG2
One carbon pool by folate	AMT
Phosphatidylinositol signaling system	CDC2L5
Porphyrin and chlorophyll metabolism	CP
Prion disease	LAMB1
Pyruvate metabolism	HAGH, HAGHL
Regulation of actin cytoskeleton	PFN2, MYH9, RHOA, PDGFRA, ITGB8
TGF-beta signaling pathway	ACVR2B, RHOA, BMP7
Tight junction	MYH9, RHOA, MLLT4, PRKCH, CLDN4, CLDN3, PRKCI
Tryptophan metabolism	UBE3A
Ubiquitin mediated proteolysis	UBE2C, EDD
Urea cycle and metabolism of amino groups	CKB
Wnt signaling pathway	RHOA

Table 5: Pathways and genes covered within PANTHER pathway data sets merged between the 86 meta-UP gene expression data set and the auto-antigen set 118 from SEREX.

PANTHER pathway	genes
Alzheimer disease-presenilin pathway	CD44, MMP14
Angiogenesis	JAG2, PRKCH, PRKCI, VEGF
Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway	PRKCH, PRKCI
signaling pathway	PRKCH, PRKCI
Integrin signalling pathway	COL5A1, COL9A2, ITGB8
Notch signaling pathway	JAG2
Parkinson disease	SFN
PDGF signaling pathway	ELF3, PRKCH, PRKCI
TGF-beta signaling pathway	BMP7
Ubiquitin proteasome pathway	UBE2C, EDD

Investigating the significance of pathways (focusing on KEGG pathway data) overlapping between the meta-UP set and 118 SEREX set where each pathway holds at least one gene from both sets we identified 9 joint pathways (see Table 6).

Table 6: Pathways from KEGG pathway data holding at least one gene from the 86 meta-UP set from gene expression data as well as one gene from 118 SEREX set from auto-antigen data.

KEGG pathways
Cell Communication
Cytokine-cytokine receptor interaction
TGF-beta signaling pathway
Focal adhesion
ECM-receptor interaction
Adherens junction
Tight junction
Leukocyte transendothelial migration
Regulation of actin cytoskeleton

For estimating the significance of this finding we calculated the average expected amount of pathways holding at least one gene from two sets by creating 1,000 randomized gene set pairs from KEGG pathway one of same size as the 86 meta-UP set and the other of same size as the 118 SEREX set. Pathways holding at least one gene from both sets were counted. No significance was found since the average amount of joint pathways over all randomizations was 6.7 with standard deviation of 3.4 (see Figure 31).

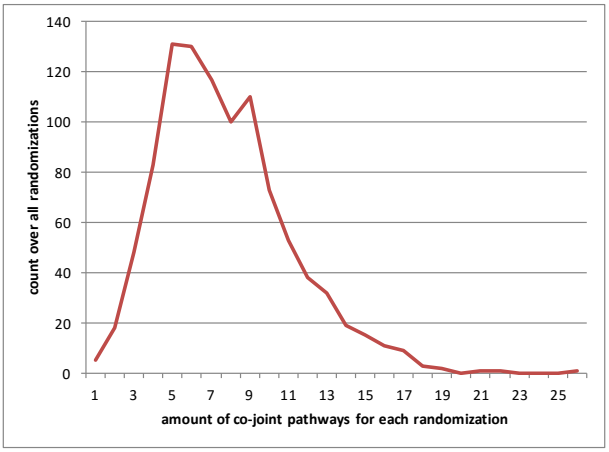


Figure 31: Distribution of the amount of joint KEGG pathways between two randomized gene sets of same size as 86 meta-UP gene expression set and 118 SEREX auto-antigen. Joint pathways are holding at least one gene from both sets (randomizations n=1,000).

Discussion

The preparation of the data needed extensive integration on various levels before a comparative analysis was possible. Major issue was the difference in used identifiers throughout the publications. While matching tools were available like the SOURCE tool [359], [395], caution was advisable and we had to correct several ambiguous results manually. Identifier matching between biological database identifiers is still a common issue in bioinformatics. Platforms, which provide a conversion tool, are for example the Database for Annotation, Visualization and Integrated Discovery (DAVID) [407] and the ID Mapping service at UniProt [408].

Another issue with possible impact is the difference in reported up or down regulation of relevant genes throughout the publications. In worst we found genes as being reported up-regulated, as well as down-regulated. This pattern can be related to mistakes on the experimental technique or is correct and implying biological differences between the different studies on a cellular level. Either explanation is possible and a common issue within omics data analysis of high-throughput data [331]. Our solution was to investigate only those genes where there was a clear trend into either up or down regulated. This approach can be flawed considering the hypothesis. If we are looking for genes with up-regulation within ovarian cancer and if the expression levels of genes differ due to different cancer types, then we can expect to find auto-antibodies for genes which might be up-regulated as well as down-regulated, depending on the cancer stage. In that case we would produce false negatives by our filtering.

Another issue related to data integration on the level of pathways is the heterogeneous results between BioCarta, Panther and KEGG for involved pathways after overlapping SEREX genes with the significant gene set of the meta-analysis. The heterogeneity between pathway data is still high as outlined in the first chapter. Currently databases like APID [92] provide a good level of integration of network data publically available. For terms of pathway categories KEGG are

accepted as high quality pathway data but might not be applicable to other databases if the overlap to KEGG is small. A classification that is independent of the pathway data source is provided by the GO classification [141].

Nevertheless the identified small set of 3 genes (BARD1, KRT8, MSLN) with reported up-regulation and auto-antigenicity was statistically significant when compared to the number of conjoint members of randomly generated datasets thus supporting the hypothesis.

The analysis workflow of this project, as a whole, is not a unique workflow but can be adapted to similar omics explorative investigations. Consequently a proposed workflow might consist of the following steps. Experimental high-throughput omics data are collected and raw data preprocessed and integrated, supported by automatic annotation. For in house data a data management system has to be established ideally following standardized guidelines for data representation. Preprocessing of the data can include several automated steps including a missing value analysis, identification of outliers etc. Normalization routines have to be chosen with care, depending on the experimental setup but can be automat zed once similar workflows have been defined. Additional filtering techniques can be defined. Following the data preprocessing a first line analysis can be undertaken. Approaches can include unsupervised analysis like hierarchical clustering or supervised approaches including statistical tests for identifying candidates with a significant change of pattern. Further analytical steps on the level of network knowledge can be the identification of co-expression patterns and category enrichment analysis. Realizing such a workflow with computer aided support or even automatized steps is feasible and requires integration aspects not only on the data level but also on the workflow and process level. discoveryBase is such an analysis platform as proposed in [376].

Mesothelial cell stress response and cytoprotection in peritoneal dialysis

Introduction

Peritoneal dialysis is an alternate therapy to hemodialysis for patients suffering under severe kidney diseases. The advantage lies in its relative safe and cost-effective mode. The drawback to this treatment method is the bio-incompatibility of peritoneal dialysis fluid (PDF) which damages the peritoneal membrane function in the long run and leads to peritoneal inflammation [409]. Investigating mesothelial tissue on a molecular scale is usually not possible in-vivo hence studies are usually performed with in-vitro cell culture models [410]. It could be shown in prior studies that the exposure to such cellular stress inducing PDF also leads to the increase of cytoprotective responses by over-expression of heat-shock proteins (HSP) [411]. Other cytoprotective mechanisms could be observed in renal medullary derived cells including apoptosis, mitotic arrest and cellular stress response. Different external stress stimuli including for example, hypertonicity, acidosis, cytotoxicity and temperature will require appropriate specific molecular protective mechanisms leading to a complex cytoprotective system within cells.

Goal of this project was to identify the cellular mechanisms and proteins of relevance on a systems and biological network level and propose a mode of cytoprotection against the effects of PDF. Two strategies were pursued during this project for increasing the cytoprotection of mesothelial cells. The first strategy aimed at prevention of injury in mesothelial cells upon PDF exposure and the second strategy aimed at increasing cellular repair mechanisms.

A combined proteomics and bioinformatics analysis for exploring of the molecular mechanisms of PDF exposure was performed. Proteomics was based on two-dimensional gel electrophoresis measuring differential protein abundance in cell culture between samples being exposed to PDF and samples without treatment. Two-dimensional gel electrophoresis is a popular methodology for explorative analysis of the proteome but has its limitations as membrane-bound proteins or proteins with low abundance are less likely detected [412]. For investigating the effects of peritoneal dialysis fluid (PDF) on human immortalized mesothelial cells, two major setups were defined. Firstly, cell culture was exposed to PDF once (single exposure) and secondly, cell culture was exposed to PDF repeatedly (repeat exposure). For each setup the general workflow consisted of several steps beginning with the cultivation of the cell culture, splitting it into separate control (CO) and exposure (D2) groups on which PDF was applied depending on the setup. Afterwards proteomic measurement was performed by two-dimensional gel electrophoresis including imaging and spot matching between the samples.

The project was supported by several contributions including data preprocessing and management, descriptive statistics, identifying protein candidates with differential expression patterns and enrichment analysis of biological processes and pathways. Due to the limitations of two dimensional gel electrophoresis the false negative rate was further addressed by applying expansion methods on a protein-protein interaction network level for predicting putative protein candidates which are additionally involved in the molecular mechanisms.

Material and Methods

2D gel electrophoresis spot data preparation by Delta2D software

Data for protein spot intensities was provided for further statistics and bioinformatics analysis after preparation of raw data gel images with Delta2D software (Decodon GmbH, Greifswald, Germany) [413]. Gel images were fused with the group warping strategy followed by a spot detection and normalization of intensities on each gel with the average intensity of all spots on the gel as normalization basis [414].

Statistics of prepared 2D gel electrophoresis spot data

Prepared spot data was described by (i) distribution of spot intensities (as % of total spot-volume) of the groups CO and D2, (ii) distribution of difference in variance between CO and D2 and (iii) by the distribution of difference in spot intensity ratio between D2 and CO.

Significant differences in spot intensities between CO and D2 were calculated by applying an unpaired t-test with equal variance and a significance level set to 0.05. Correction for multiple testing was done by applying the Bonferroni method [415], [416].

Identification of proteins utilizing MASCOT and MS-Tag software

After delivering spot test statistics to the identification group proteins from significant spots were analyzed with mass spectrometry was utilizing Mascot software (Matrixscience Inc, Boston, USA) and MS-Fit, MS-Tag software from the ProteinProspector bundle (UCLA, San Francisco, USA) [417], [418]. Results were provided in the form of lists of SwissProt identifier [419], [408]. Additional mapping between SwissProt identifier and entrez gene identifier had to be done separately. All human SwissProt identifier were extracted from SwissProt database and protein annotation for each entry parsed for corresponding entrez gene identifier using http requests.

Hierarchical Clustering of 2D gel electrophoresis spot data

Prepared data was hierarchically clustered using Pearson correlation as distance metric with average linkage rule [420]. MultiExperiment Viewer Software (MeV) was used for clustering [421], [422].

Principal component analysis of 2D gel electrophoresis spot data

In case of biological noise within samples Alter et al. suggested a principal component analysis (or single vector decomposition) to reduce the effect on noise of genome wide expression data [423]. Applying this approach on 2D gel electrophoresis is done identically. After calculating the relevant matrices noise reduction is accomplished by normalization of spot data by filtering out those eigenspots (and eigensamples) representing the biological noise by substituting zero for the eigenintensity level of the diagonal eigenmatrix where the noise can be associated to a biological interpretation. Identification of the corresponding eigeninstensity level is done by plotting and interpreting the relative intensity values of each eigenintensity level. Afterwards the calculation is reversed with the modified eigenmatrix obtaining the noise-filtered spot intensity matrix. Calculations were done using R and the svd package [424], [425].

Analysis of enriched pathways and biological processes

For investigating over and underrepresented pathways and biological processes the PANTHER classification system was used [172], [173], [400], [174, p. 20]. PANTHER identifies such proteins by computing a binomial test (with optional Bonferroni correction) comparing the set of significant proteins measured to the sets of proteins belonging to pathways or biological processes as given by the PANTHER classification.

Protein interaction network connections analysis

Proteins that are located topologically close within a protein interaction network are expected to have a higher degree of direct interactions when compared to random sets of same size. The amount of direct connections between two proteins are counted within a given experimental protein set and compared to the average amount from direct connections within 100 randomly picked protein sets of same size.

Protein interaction neighbors and neighbor expansion using OPHID protein-protein interaction network

Sets of significant proteins were investigated in context of their direct protein interactions within given protein-protein interaction networks (PIN). Proteins that are located topologically close within a protein interaction network are expected to have a higher degree of direct interactions when compared to random sets of same size. We therefore compared the amount of direct interactions in a given experimental protein sets with the average amount of connections in 100 randomized protein sets of same size. Furthermore it can be assumed that a direct interaction between two given proteins is related to same biological processes and pathways [35], [426]. Thus we additionally defined an expanded protein as being connected to at least two proteins from a given set of significant proteins within a given PIN expecting a higher amount of expanded proteins if the significant proteins are topologically in the same vicinity of the PIN. We used the Online Predicted Human Interaction Database (OPHID) Version 2007 [128] as reference PIN for identifying interactions between significant proteins and for calculating expanded neighbors. Constructing the network graph structure for OPHID PIN resulted in 7,266 distinct nodes and 24,701 distinct edges.

Distribution of shortest path length of significant protein sets

Causal molecular pathways appear close to the shortest path between given protein sets [427], [428]. Significant protein sets were investigating for their distribution of all shortest path lengths of all protein pairs from the given sets within OPHID PIN and compared to the distribution of 100 randomized protein sets of same size. Additionally PANTHER classification holds a unique set of 200 known stress associated genes for which the distribution of shortest path lengths is calculated as well and compared to the distributions of the significant sets.

Software Tools

EXCEL 2007 (Microsoft Corporation, Redmond U.S.) was used for data handling and graphics. SPSS 13.0 (IBM, New York, U.S.) and R [424] were used for statistics and graphics. MeV was used for hierarchical clustering [422]. Network calculations were done using the Basic network analysis software suite. Perl [130] was used for programming of data handling and analysis routines. Cytoscape [119] was used for network graphics.

Results

Beginning with single exposure experiments the first experimental setup was conducted for establishing the experimental environment and the data analysis workflows. Results of interest were by how much the control group and exposed group differ in their overall spot intensities, their variance and ratio. Extracted protein amount from cell culture cultivation resulting in 6 samples (3 CO, 3 D2). Having improved the experimental workflow establishing the workflows a second experimental setup consisting of 10 samples (5 CO, 5 D2) was conducted for single exposure. Results of interest concerning the first setup were if the spot intensity patterns as investigated in the first setup differed in the second setup. Having finalized the experimental workflow, the second setup was further investigated for significant proteins between the groups followed by a bioinformatics and network analysis. A third setup for single exposure was conducted at a later time addressing the issue with the low sample size of setup 2 consisting of three biological replicates with the first two having 10 samples (5 CO and 5 D2). The third replicate consisted of 12 samples in total (6 CO and 6 D2). Cell culture for the third replicate was additionally run as duplicate with one half of the samples serving for the repeat exposure experiment.

Single exposure first setup

Spot intensity distribution investigations

The amount of protein spots within the first setup were 744 distinct spots within the CO group and 743 matching protein spots within the D2 group.

For distribution of spot intensities (as % of total spot-volume) of the groups CO and D2 see Figure 32 and Figure 33 respectively. Distribution of the spot intensities reveals that ca. 70% of all spots contribute with lower than average intensities while 30% contribute with higher than average spot intensities.

For distribution of difference in variance between CO and D2 see Figure 34. The distribution has an approximated Gaussian form with the mean being at 0.312 indicating that the variance within the control group is in average about 2 times higher than in the exposed group.

For the distribution of difference in spot intensity ratio between D2 and CO see Figure 35 for all spots and Figure 36 for significant spots only. The distribution for all spots has an approximated Gaussian form with the mean at -0.13 with a standard deviation of 0.238 indicating that most spots between D2 and CO differ in their intensities within the range of +/- 2 times.

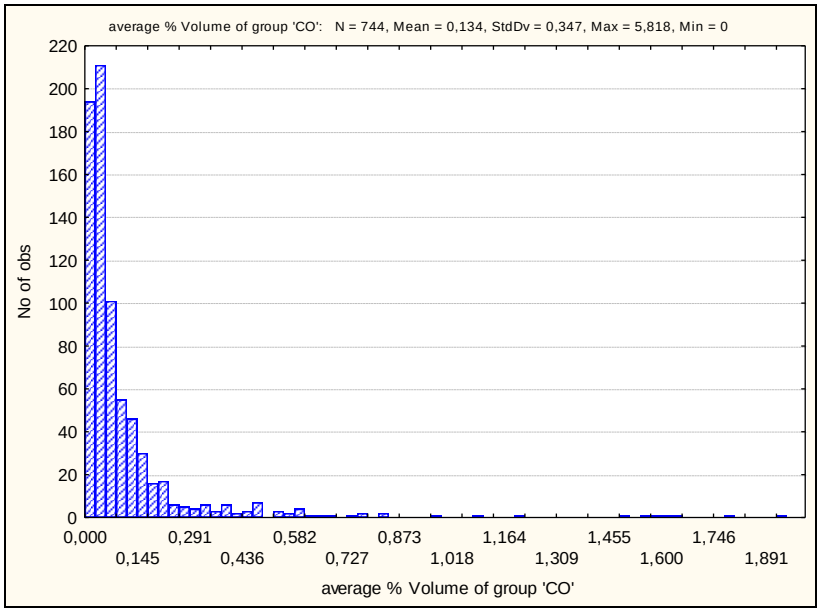


Figure 32: Distribution of average spot intensities of the control group (CO) as fraction of total intensity volume over all spots (n =744, x-axis) plotted against the amount of spots ('No of obs', y-axis).

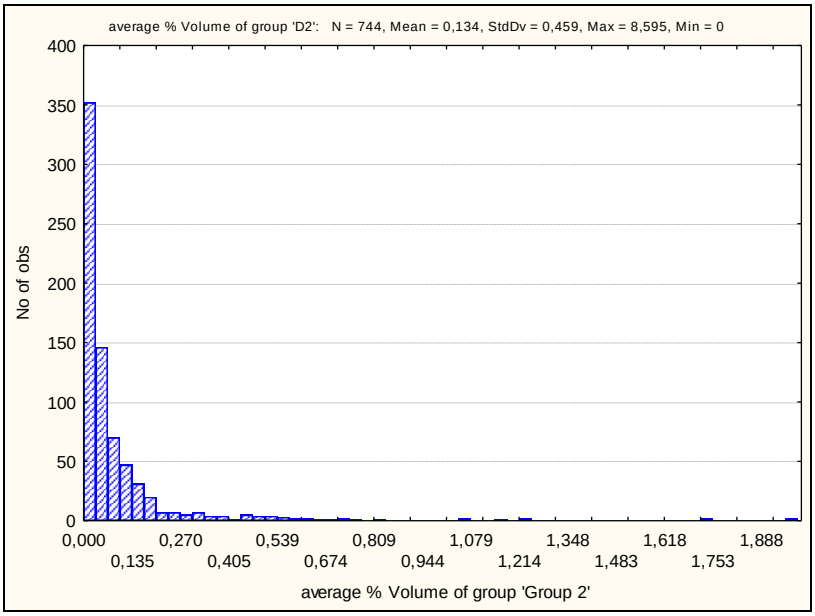


Figure 33: Distribution of average spot intensities of the exposed group (D2) as fraction of total intensity volume over all spots (n =743, x-axis) plotted against the amount of spots ('No of obs', y-axis).

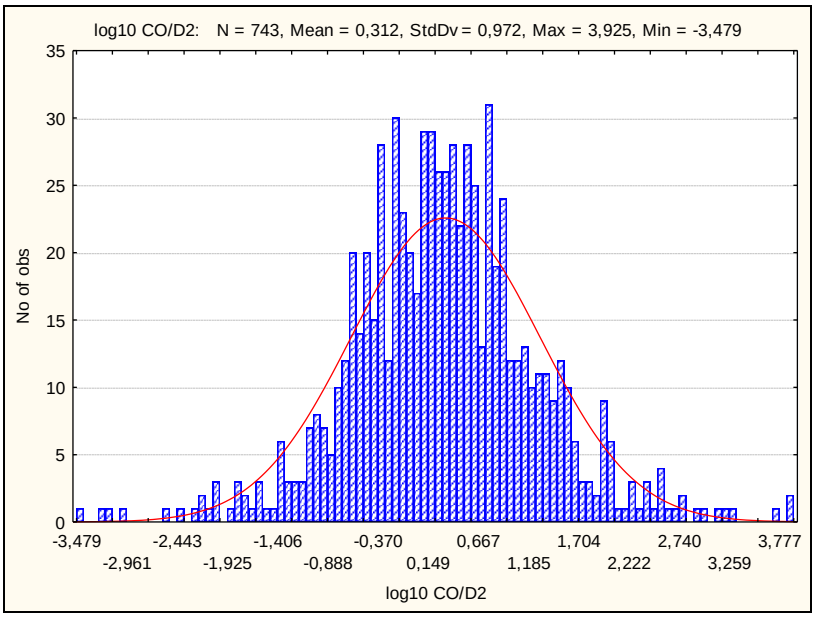


Figure 34: Distribution of variance between control group (CO) and exposed group (D2) as log10 CO/D2 (x-axis) plotted against the amount of spots ('No of obs', y-axis).

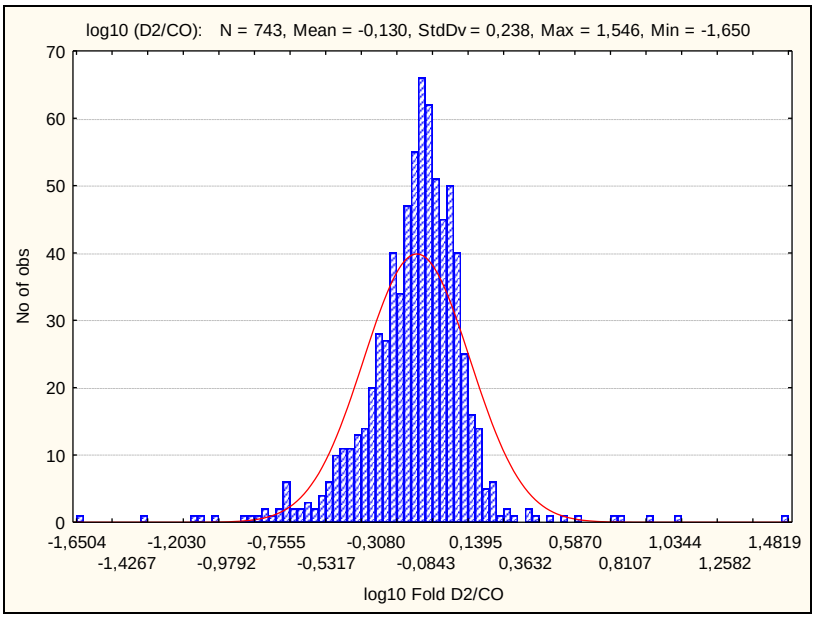


Figure 35: Distribution of difference in spot intensity ratio (Fold) between all spots of the exposed group (D2) and the control group (CO) as log 10 D2/CO (x-axis) plotted against the amount of spots ('No of obs', y-axis).

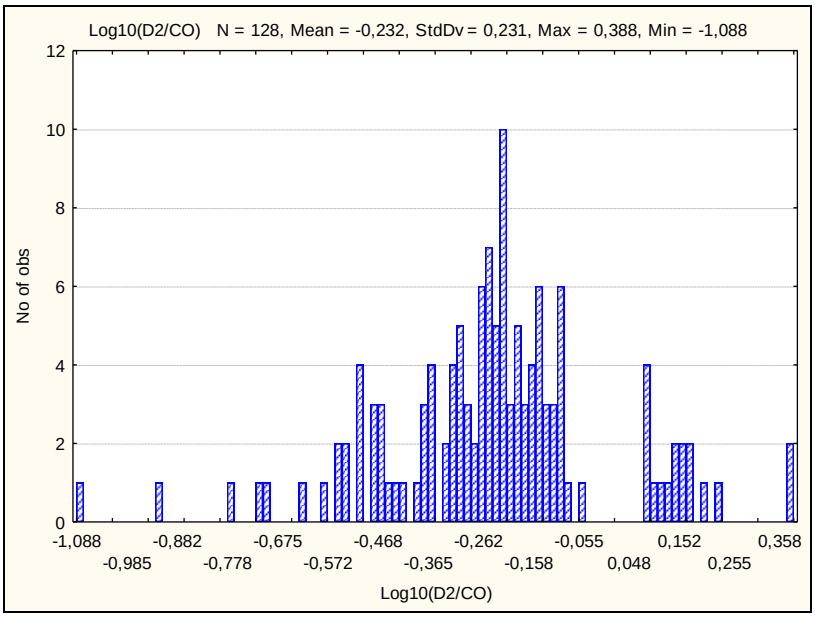


Figure 36: Distribution of difference in spot intensity ratio (Fold) between significant spots of the exposed group (D2) and the control group (CO) as log 10 D2/CO (x-axis) plotted against the amount of spots ('No of obs', y-axis).

Test statistics and protein identification

Calculating the t-test without correction for multiple testing revealed 17 spots within the exposed group having a significant increase and 111 spots have a significant decrease in abundance relative to the control group. Including correction for multiple testing, no spots with significant differences could be identified.

For establishing the identification workflow of the proteins with mass spectrometry spot analysis was performed on selected spots focusing on the location of known housekeeping proteins and proteins associated with stress reactions. In total 43 spots were identified of which 13 belonged to the set of significant spots (7 with increased and 6 with decreased abundance).

Hierarchical clustering

Hierarchical clustering of the spot data using Pearson correlation as distance metric with average linkage rule revealed a separation of the control group (CO) and the exposed group (D2) (see Figure 37).

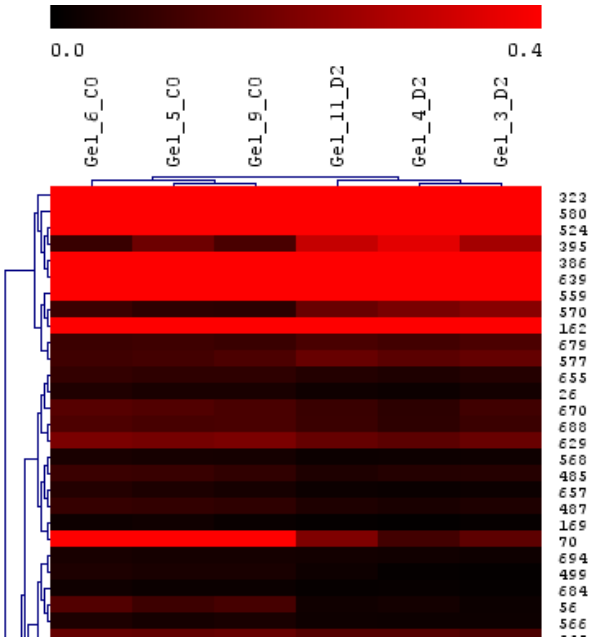


Figure 37: Hierarchical clustering of spot data using Person correlation as distance metric with average linkage rule. Column 1-3 from left denote the control group (CO), column 4-6 the exposed group (D2). Rows have been truncated (in total n = 743)

Single exposure second setup

Spot intensity distribution investigations

The amount of protein spots within the second setup were 1197 distinct spots that could be matched between the CO group and the D2 group, an increase of 61% from the first setup.

For distribution of spot intensities (as % of total spot-volume) of the groups CO and D2 see Figure 38 and Figure 39 respectively. Distribution of the spot intensities reveals a similar result to the first setup with ca. 80% of all spots contribute with lower than mean intensities while 20% contribute with higher than mean intensities.

For distribution of difference in variance between CO and D2 see Figure 40. The distribution has an approximated Gaussian form with the mean being at 0.013 indicating that the variance between both groups differs only by 3%.

For the distribution of difference in spot intensity ratio between D2 and CO see Figure 41 for all spots and Figure 42 for significant spots only. The distribution for all spots has an approximated Gaussian form with the mean at 0.01 with a standard deviation of 0.13 indicating that most spots between D2 and CO differ in their intensities within the range of +/- 30%.

Investigating the various spot distributions between the first and second setup the overall shape of the distributions is similar. Differences in the variance and the ratio between control and exposure group during both setups can be most likely explained by the overall

improvement of the experimental setup during the establishing of the workflow and the resulting increase of detected spots.

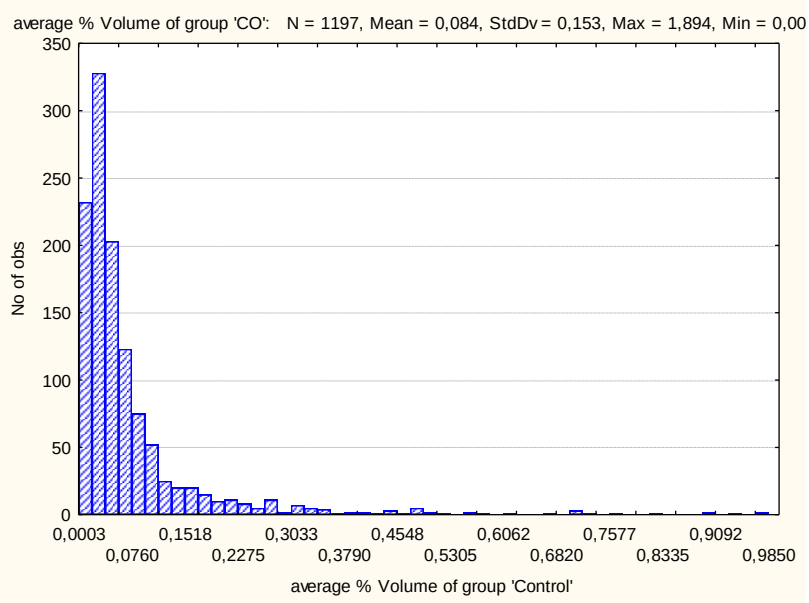


Figure 38: Distribution of average spot intensities of the control group (CO) as fraction of total intensity volume over all spots (n =1197, x-axis) plotted against the amount of spots ('No of obs', y-axis).

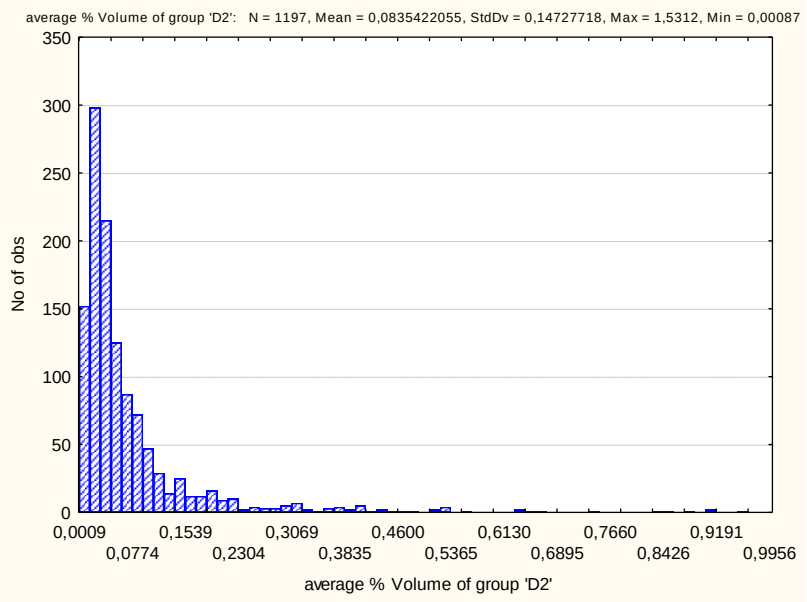


Figure 39: Distribution of average spot intensities of the exposed group (D2) as fraction of total intensity volume over all spots (n =1197, x-axis) plotted against the amount of spots ('No of obs', y-axis).

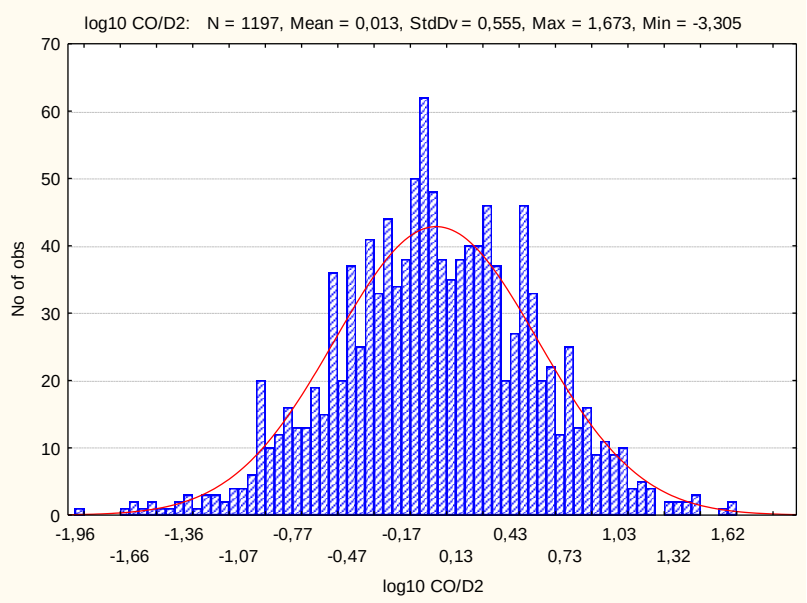


Figure 40: Distribution of variance between control group (CO) and exposed group (D2) as log10 CO/D2 (x-axis) plotted against the amount of spots ('No of obs', y-axis).

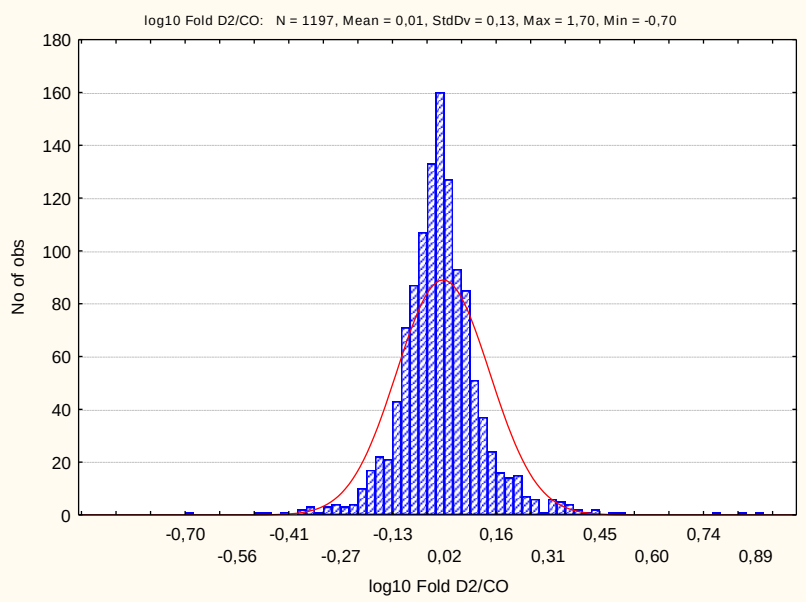


Figure 41: Distribution of difference in spot intensity ratio (Fold) between all spots of the exposed group (D2) and the control group (CO) as log 10 D2/CO (x-axis) plotted against the amount of spots ('No of obs', y-axis).

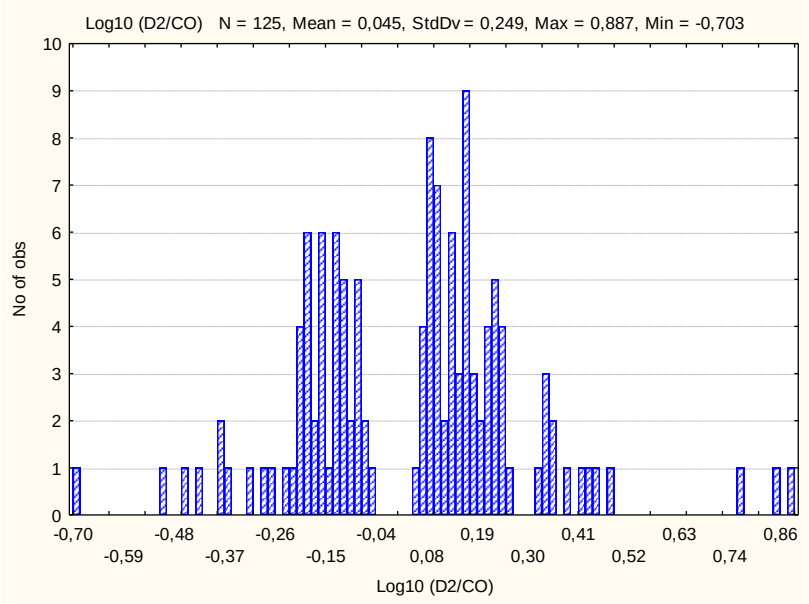


Figure 42: Distribution of difference in spot intensity ratio (Fold) between significant spots of the exposed group (D2) and the control group (CO) as log 10 D2/CO (x-axis) plotted against the amount of spots ('No of obs', y-axis).

Test statistics and protein identification

Calculating the t-test without correction for multiple testing revealed 73 spots within the exposed group having a significant increase and 52 spots have a significant decrease in abundance relative to the control group. Including correction for multiple testing, 4 spots with significant differences could be identified (2 with an increase, 2 with a decrease in abundance).

Identification of significant spots via mass spectrometry provided 47 proteins. Not all spots could be successfully identified, especially those with low abundance. Of those 47 proteins 4 were also present in the first setup.

Hierarchical clustering

Hierarchical clustering of the spot data using Pearson correlation as distance metric with average linkage rule revealed no separation of the control group (CO) and the exposed group (D2) (see Figure 43A). While the first setup revealed a clear separation of the groups with hierarchical clustering, the second setup could not reproduce this separation. It was assumed that the much higher amount of spots and the overall low abundance could have an effect on the clustering. Therefore low abundant proteins (intensities < 0.4) were pruned and clustering was repeated (see Figure 43B. The resulting dendrogram was almost identical to the unpruned clustering indicating that the low abundant proteins did not have the suspected influence.

Sample size calculations

Calculating a required sample size using the spot intensity data from setup 2 as a reference results in 72 required samples (i.e. gels) 36 for each group (sigma CO= 0.007568, sigma D2=0.008305, two-tailed alpha = 0.05, true difference of means = 0.005319, power 0.8).

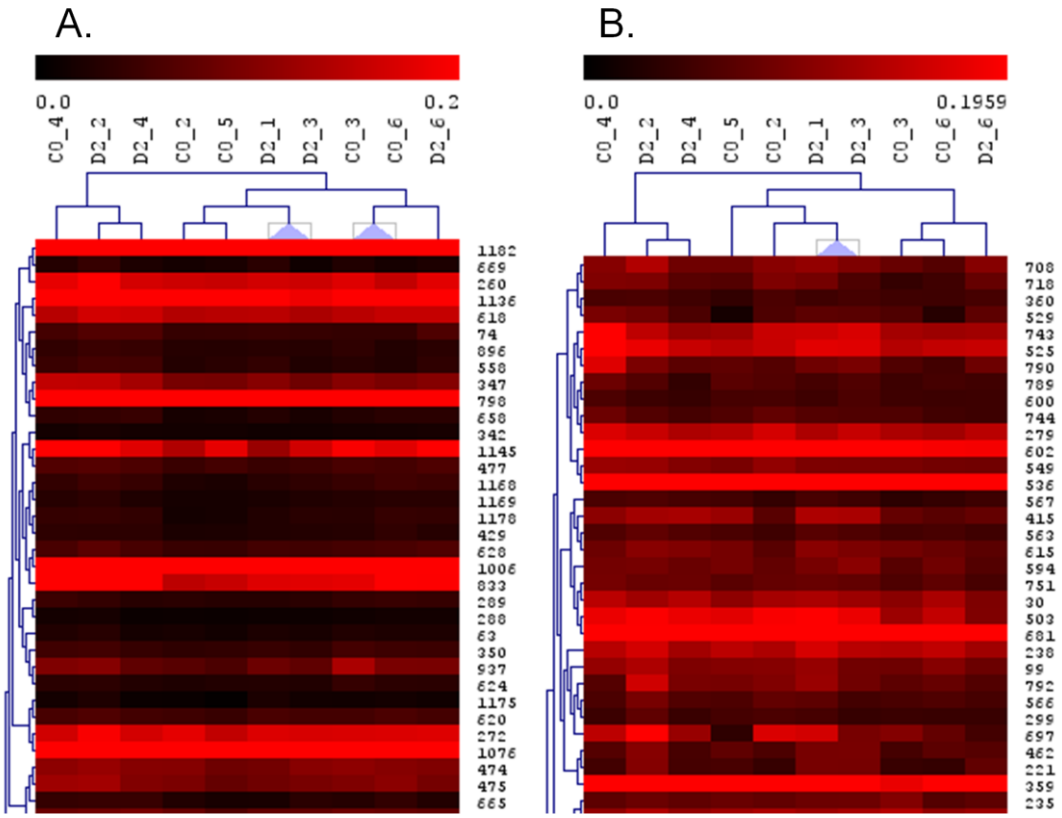


Figure 43: Hierarchical clustering of spot data using Person correlation as distance metric with average linkage rule. (A.) shows results for full set spot data (n=1197); (B.) shows the results for the set of spots with abundance >0.4 (n~500). Rows have been truncated.

Shortest path distribution of the significant protein sets

PHANTER classification provides a set of 200 proteins associated with stress response. A shortest path distribution of the pair wise path lengths of all protein pairs was performed for investigating the topological distance of the set of 200 proteins. If the set is topologically close, we expected in finding in average shorter distances as opposed to random sets. Comparing the set of 200 stress proteins from PANTHER classification with a distribution of 100 randomized sets of same size using OPHID PIN revealed indeed that the proteins from the stress response set are topologically in close neighborhood compared to a randomized set (see Figure 44). Performing this shortest path analysis on the set of significant proteins from the single exposure setup 2 and a randomized set of same size revealed no significant difference (see Figure 45).

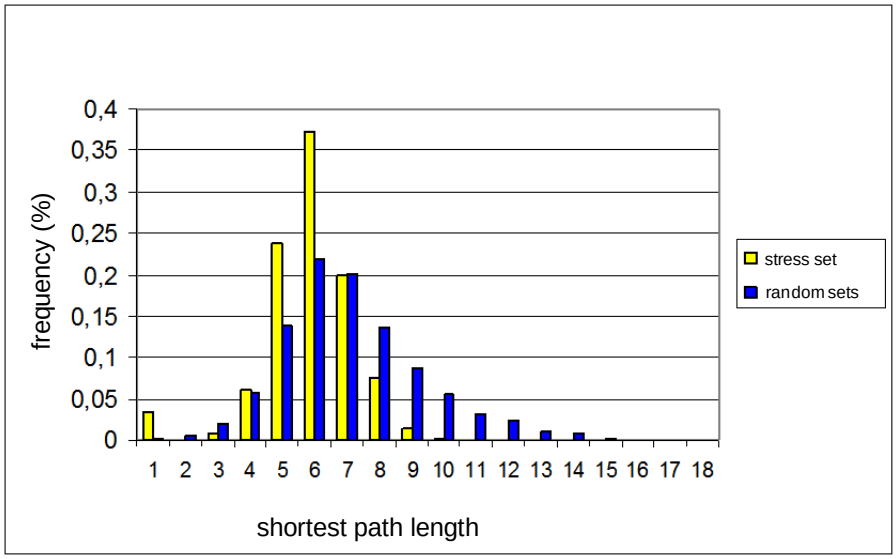


Figure 44: Shortest path length distribution utilizing OPHID protein interaction network of PANTHER classification stress response set (yellow bars) vs. 100 randomized sets of same size (blue bars).

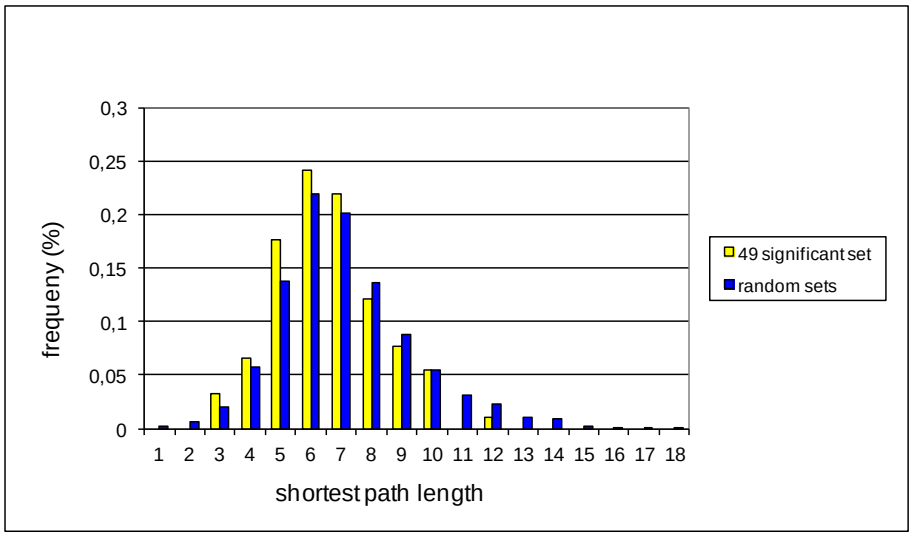


Figure 45: Shortest path length distribution utilizing OPHID protein interaction network of the single exposure setup 2 set of proteins with significant difference in their abundance between control group and exposure group (yellow bars) vs. 100 randomized sets of same size (blue bars).

Classification of enriched pathways, biological processes and molecular functions.

Utilizing PANTHER classification software the set of 49 significant proteins was mapped on the PANTHER database and was investigated for significantly enriched pathways, biological processes and molecular functions (including Bonferroni correction). No pathways could be identified with significant enrichment. Four biological processes as well as metabolic functions

were successfully identified as being overrepresented (see Table 7) including 23 proteins of which 15 showed significantly increase of abundance and 8 significantly decrease of abundance.

Table 7: Biological processes and molecular functions found to be enriched for the set of significant proteins from the second single exposure setup, utilizing PANTHER classification software with Bonferroni correction. ('ref hsa #' denotes the amount of human proteins representing the given category, 'found #' denotes the amount of proteins from the measurement and 'expected #' denotes the amount of proteins one would find by randomly picking protein sets of same size)

biological process	ref hsa #	found #	expected #	+/-	p-value
protein folding	186	10	0.41	+	1.66E-09
protein complex assembly	68	5	0.15	+	6.76E-05
protein metabolism and modification	3040	18	6.70	+	1.76E-03
stress response	200	5	0.44	+	1.19E-02

molecular function	ref hsa #	found #	expected #	+/-	p-value
Chaperone	176	8	0.39	+	1.61E-07
Hsp 70 family chaperone	15	4	0.03	+	6.98E-06
Select calcium binding protein	274	6	0.60	+	9.28E-04
Isomerase	178	4	0.39	+	1.91E-02

Protein interaction network analysis

Proteins with a significant difference in their abundance between control and exposure group were mapped to OPHID protein interaction network. 36 of the 49 identified proteins were successfully matched to entries within OPHID resulting in a subgraph with two proteins having a direct protein-protein interaction (see Figure 46) which is similar to the average amount of interactions in a randomly picked protein set of size 36 within OPHID.

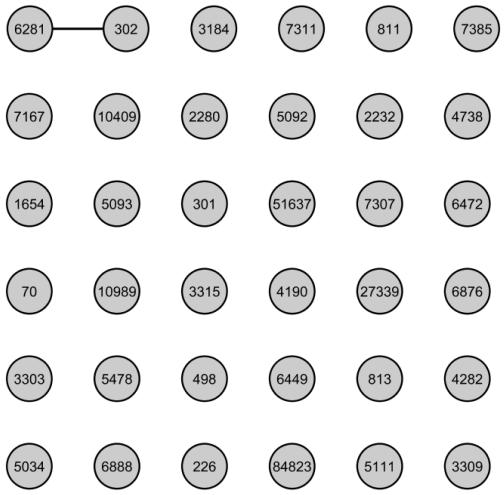


Figure 46: Subgraph of proteins with significant difference in abundance between control and exposure group from single exposure setup 2 mapped onto OPHID protein interaction network. Node labels denote corresponding entrez gene identifier.

Following up we investigated the connections between the set of 36 mapped significant proteins and the set of 200 stress response related proteins from PANTHER classification of which 123 proteins could be matched to OPHID PIN expecting some additional connections. Three proteins from the set of 36 mapped significant proteins were also part of the stress response protein subset. An additional 6 proteins from the set of 36 mapped significant proteins were also connected to the set of 200 stress response associated proteins (see Figure 47).

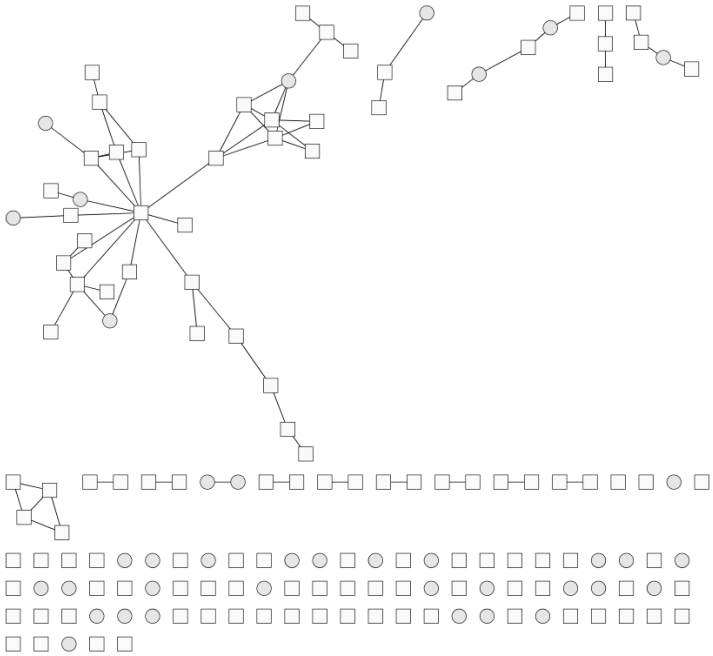


Figure 47: Subgraph of stress response associated proteins from the PANTHER classification system mapped to OPHID protein interaction network (white squares) connected to proteins with significant difference in abundance between control and exposure group from single exposure setup 2 (light grey circles).

Expanding the set of 36 successfully mapped significantly enriched proteins by proteins having a direct connection to at least 2 proteins from the set of 36 proteins within OPHID PIN we identify additionally 55 proteins revealing a large subgraph of 75 connected proteins (see Figure 48) indicating that the proteins with significant difference are indeed in close topological vicinity within the OPHID PIN.

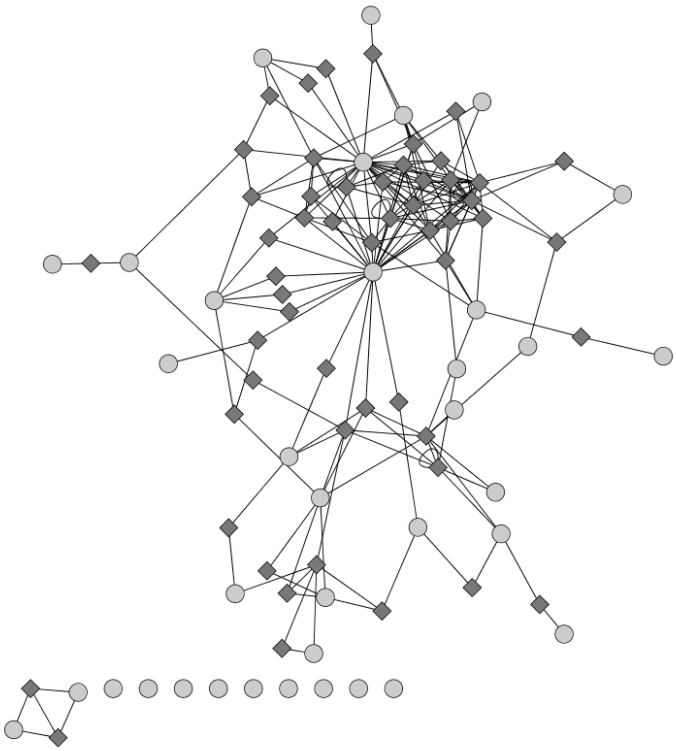


Figure 48: OPHID protein interaction network subgraph of proteins with significant difference in abundance between control and exposure group (light grey circles) from single exposure setup 2 and their expanded additional proteins (dark grey diamonds) connecting at least two proteins from the set of significant proteins.

Joining the merged set of the 123 mapped stress response associated proteins from PANTHER classification and 36 mapped proteins with significant abundance with their expanded set of 55 neighbors (of which one protein was already part of the stress response set) we identify a merged subgraph with 210 protein nodes of which 140 (66,6%) form a closely connected subgraph (see Figure 49).

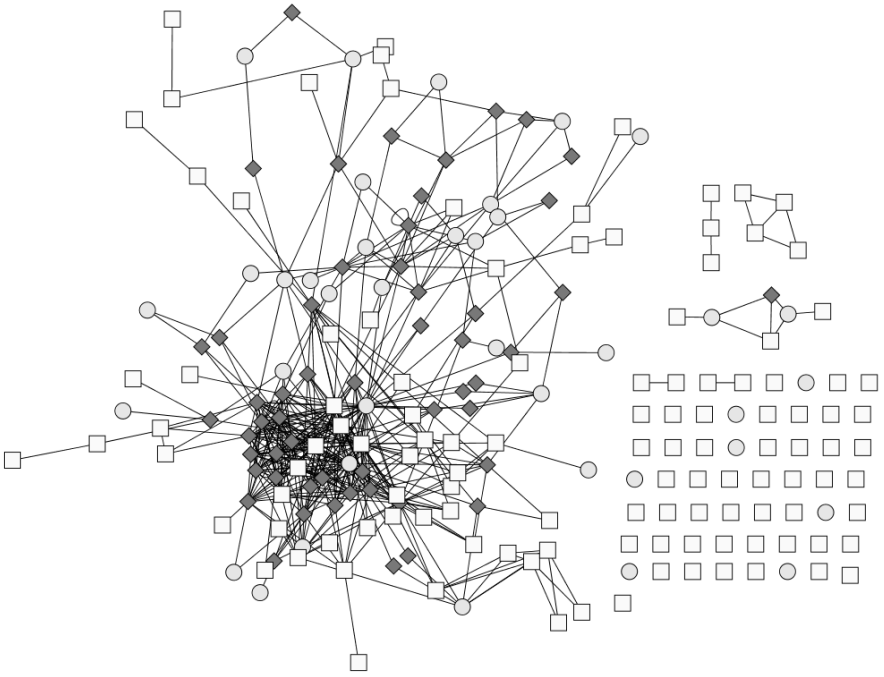


Figure 49: Subgraph of stress response associated proteins from the PANTHER classification system mapped to OPHID protein interaction network (white squares) connected to proteins with significant difference in abundance between control and exposure group from single exposure setup 2 (light grey circles) and their expanded protein interaction neighbors connecting to at least two proteins (dark gray diamonds).

Shortest path distribution of the expanded protein set

Investigating the distribution of pair wise shortest path lengths of all protein pairs within the neighborhood expanded protein set of significant proteins we observe an increase of shorter path lengths and a decrease of longer path lengths when compared to a distribution of 100 randomized sets of same size (see Figure 50). This pattern is similar to the distributions of the stress response set from PANTHER classification vs. randomized sets (see Figure 44).

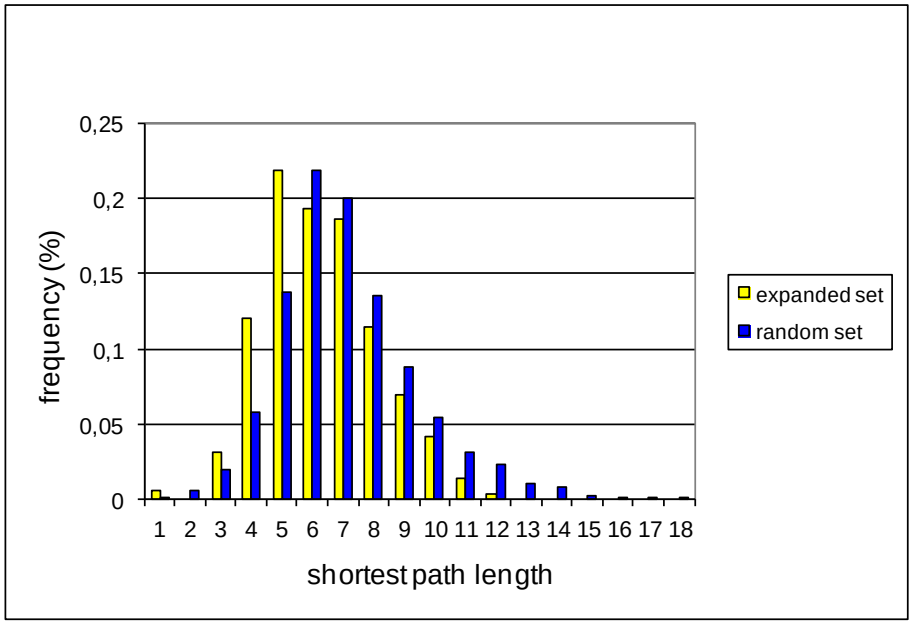


Figure 50: Shortest path length distribution utilizing OPHID protein interaction network of the single exposure setup 2 set of proteins with significant difference in their abundance between control group and exposure group and expanded with protein interaction neighbors connecting at least two proteins from the set (yellow bars) vs. 100 randomized sets of same size (blue bars).

Classification of enriched pathways, biological processes and molecular functions of the expanded protein set

After expansion of the set of 49 significant proteins by their 55 neighbors as described, analysis of the enrichment of pathways, biological processes and molecular functions was repeated. Three pathways were found enriched as well as 11 biological processes of which 3 were also enriched without neighbor expansion and 4 molecular functions of which 3 were also enriched without neighbor expansion (see Table 8).

Table 8: Enrichment of pathways, biological processes and molecular functions of the set of proteins with significant difference in abundance between single exposure setup 2 control group and exposure group utilizing PANTHER classification system. Enrichment without neighbor expansion is compared with enrichment resulting from expanding the protein set with protein interaction neighbors connecting at least two proteins from the set utilizing OPHID protein interaction network. Categories having a p-value > 0.05 at a given setup were left blank.

		without neighbor expansion			with neighbor expansion		
	ref hsa #	found #	+/-	p-value	found #	+/-	p-value
pathway							
Apoptosis signaling pathway	131				15	+	5.91E-16
Toll receptor signaling pathway	71				9	+	1.50E-09
B cell activation	86				4	+	4.75E-02
biological process							
Protein folding	186	8	+	2.90E-07	8	+	8.89E-05
Stress response	200	5	+	5.08E-03	6	+	1.74E-02
Protein metabolism and modification	3040	15	+	7.83E-03	29	+	4.74E-05
Protein complex assembly	68	3	+	4.12E-02			
NF-kappaB cascade	71				10	+	5.17E-11
Immunity and defense	1318				20	+	2.85E-06
Apoptosis	531				13	+	3.56E-06
Induction of apoptosis	165				7	+	5.05E-04
Intracellular signaling cascade	871				14	+	7.52E-04
JNK cascade	61				4	+	1.83E-02
Signal transduction	3406				24	+	5.02E-02
B-cell- and antibody-mediated immunity	97				4	+	7.71E-02
molecular function							
Hsp 70 family chaperone	15	4	+	3.41E-06	4	+	6.20E-05
Chaperone	176	6	+	2.68E-05	7	+	1.53E-04
Select calcium binding protein	274	5	+	4.43E-03	8	+	3.01E-04
Isomerase	178	4	+	9.76E-03			
Cytoskeletal protein					10	+	5.26E-02

Single exposure third setup

Addressing the issue with the low sample size from setup 2, the third setup consisted of 32 samples in total 16 for CO and 16 for D2 distributed over 3 biological replicates. A total of 947 distinct spot were matched over all samples.

Hierarchical clustering and test statistics

All biological replicates were separately clustered hierarchically using Pearson correlation with average linkage rule (see Figure 51). Similar two the second single exposure setup no distinct grouping can be identified between control and exposure samples in any of the replicate setups.

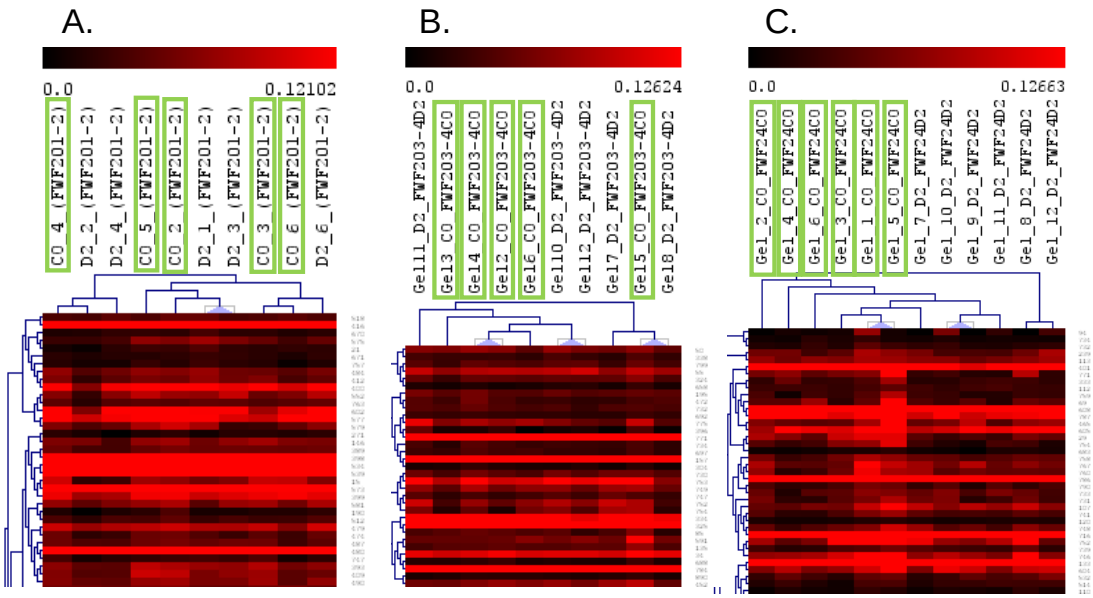


Figure 51: Hierarchical clustering of spot data from single exposure setup 3 for each biological replicate (A. replicate 1, B. replicate 2, C. replicate 3.) using Pearson correlation as distance metric with average linkage rule. Columns denote the sample gels labeled with the internal reference for each gel. Green boxes highlight sample gels from the control group CO and samples gels not highlighted belong to the exposed group D2. Rows have been truncated (total spot amount n = 947).

Calculating t-test without correction for multiple testing was done for the CO and D2 of each replicate group and for the combined CO and D2 of all 32 samples. The amount of significant spots found within the first replicate group was a total of 90 spots (overlapping with 18 spots from the second setup), 147 spots (overlapping with 8 spots from the second setup) for the second replicate group and 127 spots (overlapping with 9 spots in from the second setup) within the third replicate group. Significant proteins overlapping between all three replicates were 5 of which 2 were also significant in the second setup (see Figure 52 for a detailed VENN diagram of all overlaps).

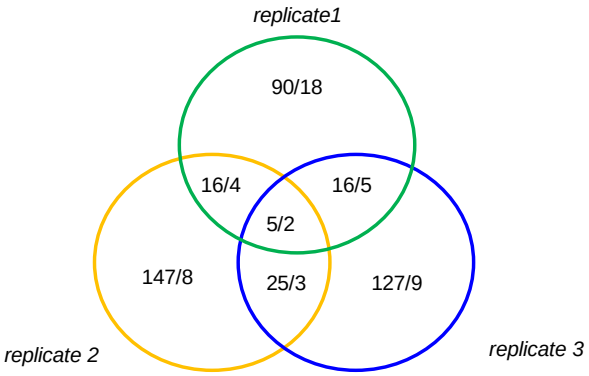


Figure 52: VENN diagram for the amounts of overlapping protein spots with significant difference in their abundance between the replicate groups from single exposure setup 3 (first number) and their overlap with significant protein spots from single exposure setup 2 (second number).

Due to the small overlap of significant spots a pair-wise correlation matrix was computed for the complete single exposure sample gels set using Pearson correlation coefficient for estimating the overall similarity between the sample gels (see Figure 53). The distribution reveals a very strong correlation (>= +0.84) of all sample pairs indicating that even though the overlap of significant spots was low, the overall pattern between the sample gels is strongly correlated.

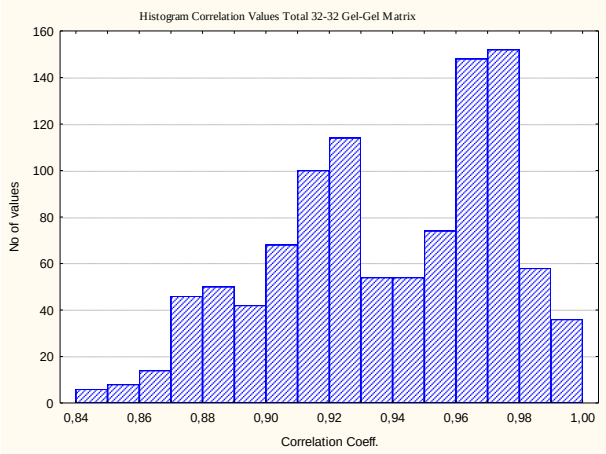


Figure 53: Histogram for pair-wise correlation coefficient distribution of all sample gel pairs of the single exposure third setup (n = 32).

Therefore all replicate setups were merged resulting in 16 samples for the CO group and 16 samples for the D2 group. Normalized spot intensities were plotted for each sample gel (see Figure 54).

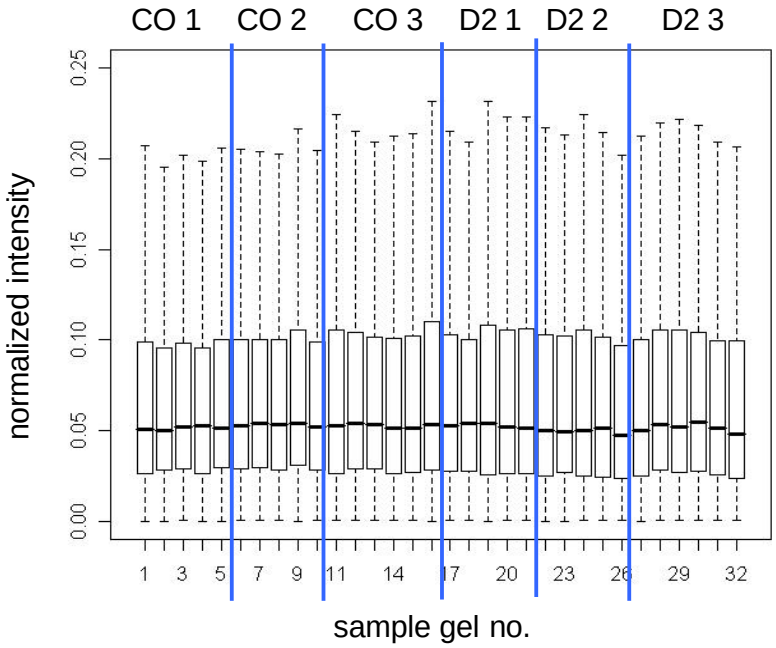


Figure 54: Box plots for the normalized spot intensity distributions of each sample gel grouped by control group and replicate setup (CO 1-3 and D2 1-3).

Hierarchical clustering was calculated for the spooled groups (see Figure 55) again resulting in no distinct separation between CO and D2 groups. Test statistics were calculated resulting in a total list of 90 protein spots showing significant differences in abundance between CO and D2. Interestingly though replicate setup 1 and replicate setup 2 (see Figure 55 blue dotted box) formed a cluster separating those sample gels from those of replicate setup 3. Investigating the reasons for this surprising difference on the level of cell culture we identified that the cells being grown for the replicate setups originated from different cell splits. Replicate setup 1 and 2 originated from the same split while replicate setup 3 originated from a few splits later.

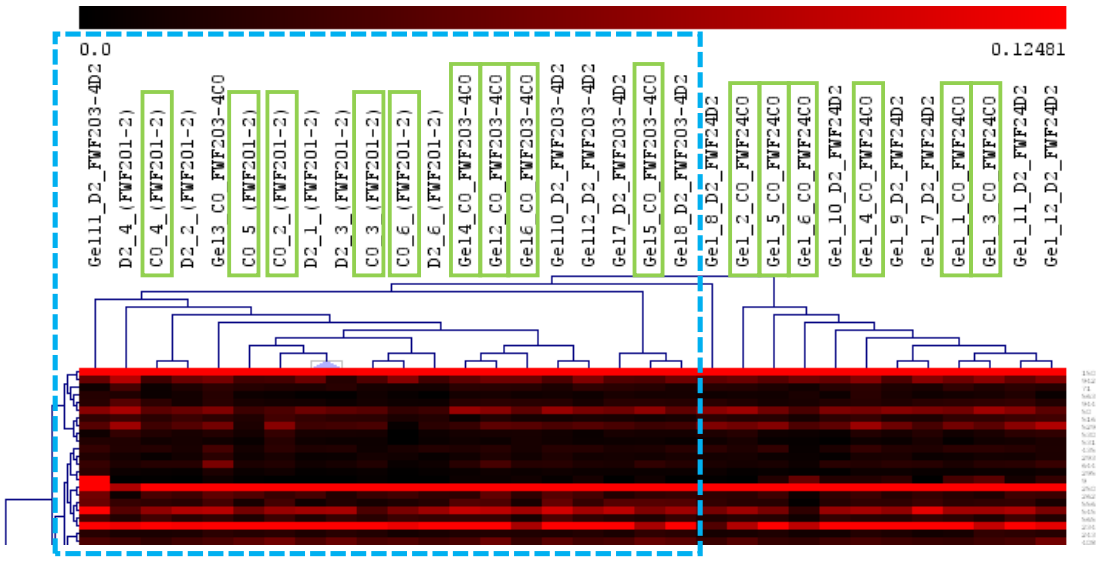


Figure 55: Hierarchical clustering of spot data from single exposure setup 3 for the spooled biological replicates using Pearson correlation as distance metric with average linkage rule. Columns denote the sample gels labeled with the internal reference for each gel. Green boxes highlight sample gels from the control group CO and samples gels not highlighted belong to the exposed group D. The blue dotted box denotes the joint group of replicate group 1 and 2. Rows have been truncated (total spot amount n =947).

Principal Component Analysis (PCA)

For further analysis we decided to perform a PCA and to investigate the PCA's eigenvalue levels (components) for identifying possible noise that can be associated to single exposure setup 3 replicate 3, thus correcting the influence of the cell split by normalizing the data according to Alter et.al. [82]. Calculation values for the components can be seen in Table 9.

Table 9: PCA calculation of the components values and relative share for the single exposure setup 3 (32 samples, 947 spot intensities).

component	value	%	component	value	%
1	36,90	0,53	17	0,69	0,01
2	5,72	0,08	18	0,67	0,01
3	2,67	0,04	19	0,64	0,01
4	2,03	0,03	20	0,63	0,01
5	1,65	0,02	21	0,60	0,01
6	1,58	0,02	22	0,59	0,01
7	1,40	0,02	23	0,57	0,01
8	1,33	0,02	24	0,54	0,01
9	1,22	0,02	25	0,53	0,01
10	1,17	0,02	26	0,50	0,01
11	1,11	0,02	27	0,48	0,01
12	0,96	0,01	28	0,46	0,01
13	0,94	0,01	29	0,44	0,01
14	0,88	0,01	30	0,40	0,01
15	0,83	0,01	31	0,39	0,01
16	0,81	0,01	32	0,37	0,01

Searching for the component of relevance we investigated the relative expression values for each component and could successfully identify component 2 as having a distinct difference between replicate setup 1,2 and setup 3 (see Figure 56). The second components value was set to zero and the intensity matrix recalculated.

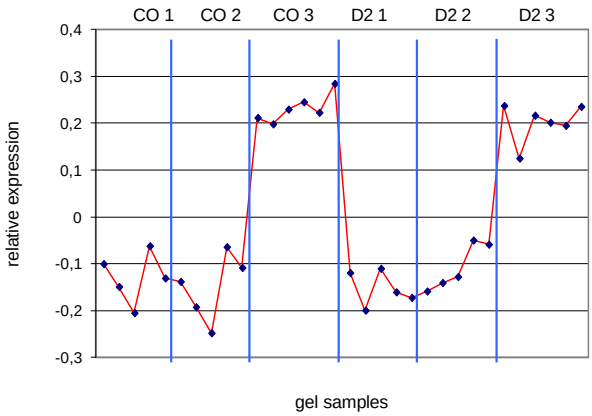


Figure 56: relative expressions of the gel samples of the second component of the PCA calculation for the single exposure setup 3. Samples are grouped by control group (CO) and exposure group (D2) and biological replicate (1-3).

Recalculation of test statistics after PCA correction and protein identification

Repeating the test statistic on the corrected spot intensity data from single exposure setup 3 revealed a total of 140 spots having a significant difference in their abundance between control and exposure group of which 13 were also found in single exposure setup 2. 82 spot were found having an increased and 58 having a decreased abundance. Comparing the significant spots with and without PCA correction, all 90 spots that have been detected without PCA correction were also found to be significant after the correction.

Identification of all the significantly changed protein spots resulted in a list of 104 unique SwissProt identifiers.

Classification of enriched pathways, biological processes and molecular functions.

Utilizing PANTHER classification software the set of 104 significant proteins was mapped on the PANTHER database and was investigated for significantly enriched pathways, biological processes and molecular functions (including Bonferroni correction). 4 pathways and 11 biological processes were found significantly enriched as opposed to 0 pathways and 4 biological processes from the second setup (molecular functions were omitted) of which all 4 were also present in the third setup (see Table 10).

Protein interaction network analysis

Proteins with a significant difference in their abundance between control and exposure group were mapped to OPHID protein interaction network. 81 of the 104 identified proteins were successfully matched to entries within OPHID resulting in a subgraph with 16 direct protein-protein interactions (see Figure 57) and the largest subgraph having a node count of 7 which is again not significantly different from randomly picked subgraphs of same size.

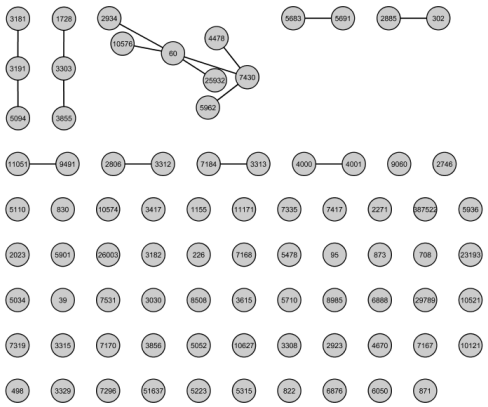


Figure 57: Subgraph of proteins with significant difference in abundance between control and exposure group from single exposure setup 3 mapped onto OPHID protein interaction network. Node labels denote corresponding entrez gene identifier.

Expanding the set of 81 proteins found in OPHID PIN by their neighbors according to our method, we identified an additional 132 proteins totaling in a subgraph with 213 proteins and 730 interactions. Interestingly this resulting subgraph held a connected subgraph of 197 proteins with 728 interactions which is 92% of all the nodes and 99% of all the edges (see Figure 58). For verifying that this finding is likely not random, we calculated the average size of the largest subgraphs when generating 1000 sets of randomly picked proteins within OPHID and expanding them with our neighbor expansion method resulting in an average size of 90 nodes.

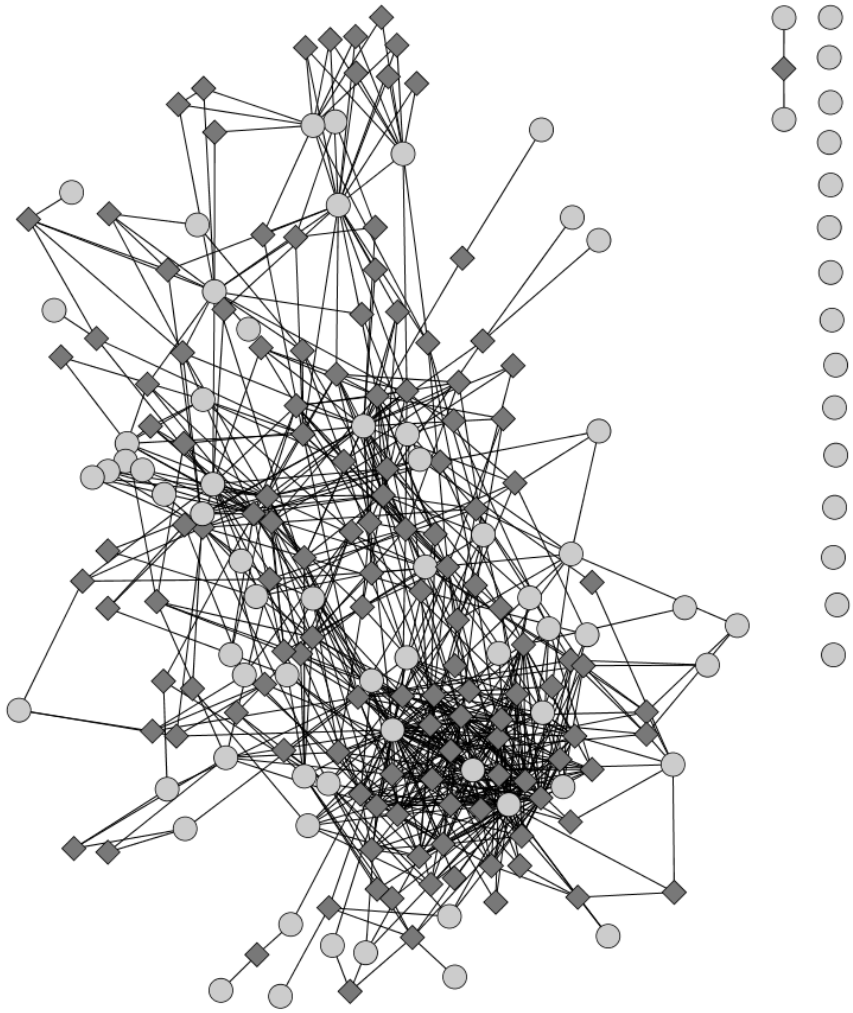


Figure 58: OPHID protein interaction network subgraph of proteins with significant difference in abundance between control and exposure group (light grey circles) from single exposure setup 3 and their expanded additional proteins (dark grey diamonds) connecting at least two proteins from the set of significant proteins.

Classifications of enriched pathways, biological processes and molecular functions of the expanded protein set

After expansion of the set of 104 significant proteins OPHID by their 132 neighbors in OPHID as described, analysis of the enrichment of pathways and, biological processes was repeated. 19 pathways were found enriched including the 4 significantly enriched pathways without expansion as well as 22 biological processes of which 8 were also enriched without neighbor expansion (see Table 10).

Table 10: Enrichment of pathways and biological processes of the set of proteins with significant difference in abundance between single exposure setup 3 control group and exposure group utilizing PANTHER classification system. Enrichment without neighbor expansion is compared with enrichment resulting from expanding the protein set with protein interaction neighbors connecting at least two proteins from the set utilizing OPHID protein interaction network. Categories having a p-value > 0.05 at a given setup are labeled n.s.

	single exposure set		with neighbor expansion	
	found #	p-value	found #	p-value
pathways				
Apoptosis signaling pathway	4	n.s.	25	1,19E-22
Toll receptor signaling pathway	2	n.s.	14	1,98E-12
Parkinson disease	7	5,41E-05	14	4,06E-10
EGF receptor signaling pathway	2	n.s.	13	3,93E-07
B cell activation	1	n.s.	10	1,90E-06
Integrin signalling pathway	2	n.s.	14	6,50E-06
Angiogenesis	2	n.s.	13	5,08E-05
Inflammation mediated by chemokine and cytokine signaling pathway	2	n.s.	15	5,70E-05
T cell activation	1	n.s.	9	2,03E-04
FAS signaling pathway	4	2,74E-03	6	2,22E-04
FGF signaling pathway	2	n.s.	10	2,63E-04
Glycolysis	5	1,89E-05	5	1,05E-03
VEGF signaling pathway	0	n.s.	7	1,94E-03
PI3 kinase pathway	2	n.s.	8	3,43E-03
Ubiquitin proteasome pathway	4	n.s.	7	3,82E-03
Interleukin signaling pathway	1	n.s.	9	1,60E-02
p38 MAPK pathway	1	n.s.	5	3,79E-02
p53 pathway	1	n.s.	7	5,16E-02
De novo purine biosynthesis	4	3,74E-03	4	8,58E-02

	single exposure set		with neighbor expansion	
	found #	p-value	found #	p-value
biological process				
Protein metabolism and modification	33	2,57E-06	76	5,18E-15
Protein folding	14	7,81E-12	21	2,20E-14
NF-kappaB cascade	0	n.s.	13	5,32E-11
Immunity and defense	12	n.s.	40	1,25E-09
Apoptosis	0	n.s.	25	1,42E-09
Protein modification	10	n.s.	38	2,03E-09
Intracellular signaling cascade	1	n.s.	32	6,27E-09
Cell structure and motility	19	6,27E-06	33	2,91E-07
Protein complex assembly	7	2,40E-06	9	3,04E-06
Protein phosphorylation	0	n.s.	24	3,32E-06
Cell structure	16	3,24E-06	24	5,12E-06
Induction of apoptosis	0	n.s.	12	9,46E-06
Stress response	6	2,70E-02	13	9,65E-06
Cell proliferation and differentiation	1	n.s.	28	1,25E-05
Signal transduction	5	n.s.	60	1,52E-05
Cell cycle	8	n.s.	27	2,87E-05
Glycolysis	6	6,62E-06	6	7,82E-04
Cell cycle control	3	n.s.	15	1,51E-03
B-cell- and antibody-mediated immunity	0	n.s.	7	5,73E-03
Oncogenesis	0	n.s.	13	1,62E-02
Intracellular protein traffic	8	n.s.	20	3,80E-02
Carbohydrate metabolism	11	1,03E-03	14	4,28E-02
Purine metabolism	5	1,18E-03	5	n.s.
mRNA transcription	0	4,24E-02	13	n.s.
Muscle contraction	5	4,31E-02	5	n.s.

Repeat exposure setup

Data for the repeat exposure was provided in the form of already preprocessed and prepared lists of proteins with difference in abundance between control and exposure groups. 44 protein spots had been reported having an increased abundance, 46 a decreased abundance resulting in a final list of identified unique SwissProt entries of 58.

Classification of enriched pathways, biological processes and molecular functions.

Utilizing PANTHER classification software the set of 58 significant proteins was mapped on the PANTHER database and was investigated for significantly enriched pathways and biological processes (including Bonferroni correction). No pathway and one biological process was found significantly enriched (Carbohydrate metabolism) which was not found enriched in the single exposure setup 3 (see Table 11)

Protein interaction network analysis

Proteins with a significant difference in their abundance between control and exposure group were mapped to OPHID protein interaction network. 47 of the 58 identified proteins were successfully matched to entries within OPHID resulting in a subgraph with 6 direct protein-protein interactions (see Figure 59) and the largest subgraph having a node count of 3 which is again not significantly different from randomly picked subgraphs of same size. 3 proteins were also found in the expanded neighbor set from single exposure setup 3.

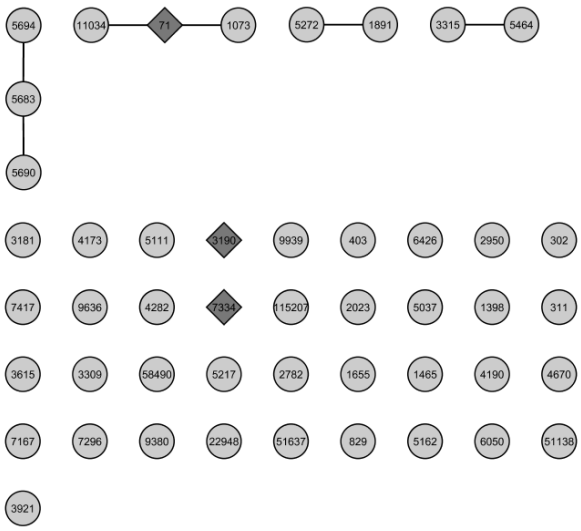


Figure 59: Subgraph of proteins with significant difference in abundance between control and exposure group from repeat exposure setup mapped onto OPHID protein interaction network. Node labels denote corresponding entrez gene identifier and dark gray diamond shaped proteins were also found in the expanded protein set from single exposure setup 3.

Expanding the set of 47 proteins found in OPHID PIN by their neighbors according to our method, we identified an additional 51 proteins totaling in a subgraph with 98 proteins and 213 interactions. Interestingly this resulting subgraph held a connected subgraph of 87 proteins with 209 interactions which is 88% of all the nodes and 98% of all the edges (see Figure 60). For verifying that this finding is indeed less likely to be random, we calculated the average size of the largest subgraphs when generating 1000 sets of randomly picked proteins within OPHID and expanding them with our neighbor expansion method resulting in an average size of 34 nodes.

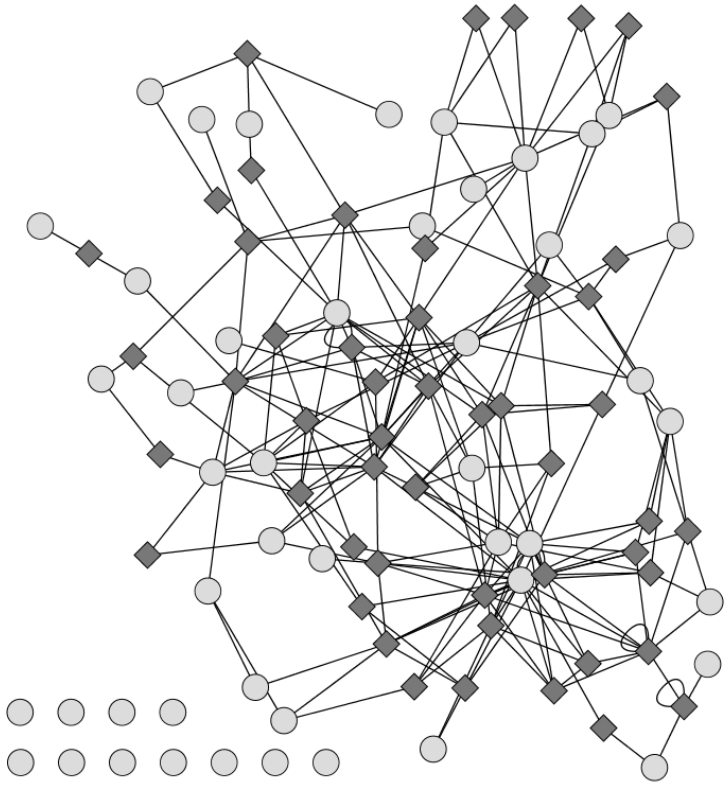


Figure 60: OPHID protein interaction network subgraph of proteins with significant difference in abundance between control and exposure group (light grey circles) from repeat exposure setup and their expanded additional proteins (dark grey diamonds) connecting at least two proteins from the set of significant proteins.

Classification of enriched pathways, biological processes and molecular functions of the expanded protein set

After expansion of the set of 58 significant proteins by their 51 neighbors in OPHID PIN as described, analysis of the enrichment of pathways and, biological processes was repeated. 6 pathways were found enriched as well as 6 biological processes of which none were also enriched without neighbor expansion (see Table 11).

Table 11: Enrichment of pathways and biological processes of the set of proteins with significant difference in abundance between repeat exposure setup control group and exposure group utilizing PANTHER classification system. Enrichment without neighbor expansion is compared with enrichment resulting from expanding the protein set with protein interaction neighbors connecting at least two proteins from the set utilizing OPHID protein interaction network. Categories having a p-value > 0.05 at a given setup are labeled n.s.

	repeat exposure set		with neighbor expansion	
	found #	p-value	found #	p-value
pathways				
Toll receptor signaling pathway	1	n.s.	9	5,67E-09
Apoptosis signaling pathway	1	n.s.	10	5,85E-08
Integrin signalling pathway	0	n.s.	9	1,14E-04
B cell activation	0	n.s.	6	3,73E-04
Cytoskeletal regulation by Rho GTPase	3	n.s.	5	2,11E-02
T cell activation	0	n.s.	5	2,11E-02
biological process				
Carbohydrate metabolism	7	1,22E-02	8	n.s.
NF-kappaB cascade	0	n.s.	6	1,48E-04
Immunity and defense	6	n.s.	18	3,99E-04
Nucleoside, nucleotide and nucleic acid m	12	n.s.	29	4,11E-03
Apoptosis	2	n.s.	9	1,53E-02
Pre-mRNA processing	5	n.s.	7	4,01E-02
Cell structure and motility	5	n.s.	13	4,05E-02

Comparison single exposure setup 3 and repeat exposure

For investigating the physiological change between single exposure and repeat exposure similarities and differences between both setups are searched for.

Single exposure setup 3 provided a total of 104 identified proteins including additionally 132 expanded neighbors. Repeat exposure provided 58 identified proteins including additionally 51 expanded neighbors. 13 proteins overlapped between both sets of significantly changed proteins and an additional 35 proteins between both expanded sets.

Comparison of classification of enriched pathways, biological processes and molecular functions.

Comparing single exposure setup 3 and repeat exposure enriched pathways and processes no significant pathways and one biological process were found (Carbohydrate metabolism). Comparing the expanded protein sets from both setups 5 overlapping pathways and 4 overlapping biological processes can be identified as being enriched (see Table 12).

Table 12: List of significantly enriched pathways and biological sets within single exposure setup 3 and repeat exposure setup including expanded neighbors.

pathways	biological process
Toll receptor signaling pathway	NF-kappaB cascade
Apoptosis signaling pathway	Immunity and defense
Integrin signalling pathway	Apoptosis
B cell activation	Cell structure and motility
T cell activation	

Protein interaction network analysis

Combining the OPHID PIN subgraphs from single exposure setup 3 and repeat exposure resulted in a joint subgraph of 116 nodes and 38 interactions with 11 overlapping proteins out of which a rather large subgraph with 15 proteins and 38 interactions can be identified (see Figure 61).

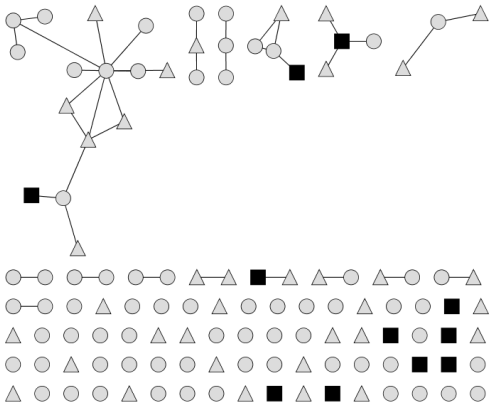


Figure 61: Subgraph of proteins with significant difference in abundance between control and exposure group from repeat exposure setup (triangle shaped nodes) and single exposure setup 3 (circle shaped nodes) mapped onto OPHID protein interaction network. Proteins being present in both sets are presented in black rectangle shaped nodes.

Combining the OPHID PIN subgraphs from single exposure setup 3 and repeat exposure including their expanded neighbors resulted in a joint subgraph of 264 proteins and 921 interactions with 48 overlapping proteins. Out of which a rather large subgraph with 239 proteins which is 91% of all nodes and 919 interactions which is 99% of all edges can be identified (see Figure 62).

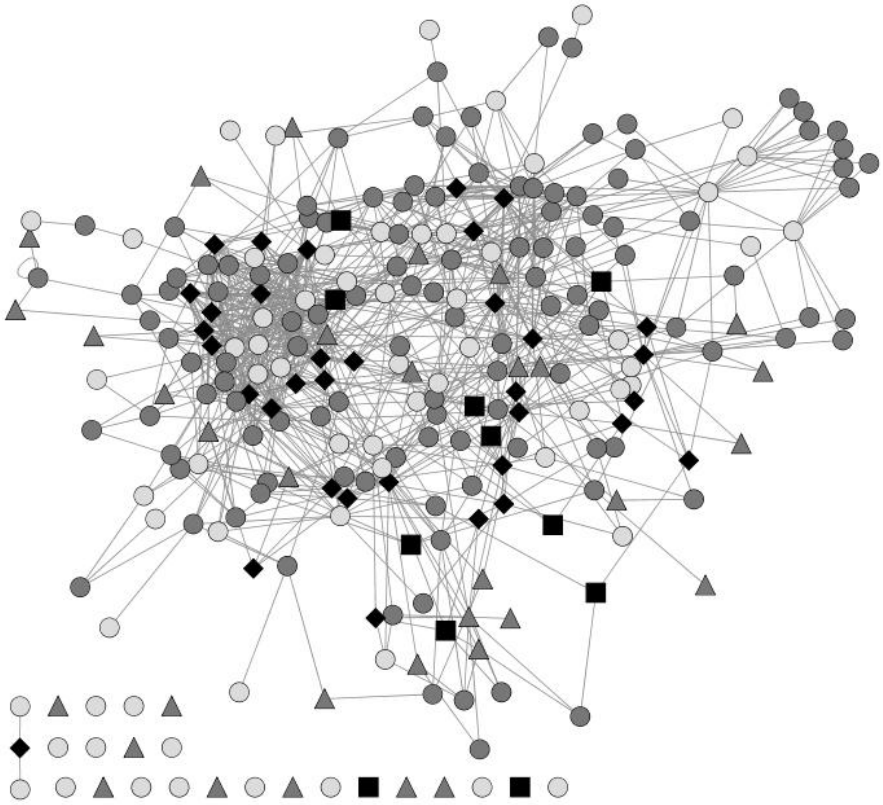


Figure 62: Subgraph of proteins with significant difference in abundance between control and exposure group from repeat exposure setup (triangle shaped nodes, light grey color) and single exposure setup 3 (circle shaped nodes, light grey color) mapped onto OPHID protein interaction network. Expanded neighbors connecting at least two proteins from each set are colored dark grey. Proteins being present in both sets are presented as black rectangle shaped nodes and proteins being present in both expanded neighbor sets as black diamond shaped nodes.

Discussion

The first setup was for establishing a workflow and getting a first hand on experimental data and preliminary patterns. It was revealed that the majority of spot intensities are rather low abundant (regarding the resolution of the dye) while only 30 % of the spots had intensities above average. The low abundance of protein spots also influenced the protein identification process since the manual excision of almost not visible spots often led to too little protein amounts after excision for identification with mass spectrometry. This issue could be solved with a robotic excision of spots defined by location coordinates. Variance of spot intensities

between the samples between the control group and the exposed group revealed that the control group showed higher variance. A possible interpretation could be that the pattern due to cellular stress resulted in a more controlled and uniform reaction in all three exposed samples resulting also in a smaller variance, while the control group samples differed more. While this is a 'could be' explanation based on the small sample sizes this is rather hypothetical and the later setups did not show this difference any more. The sample size of the groups was small, three for each. The resulting sample sizes for the further setups were increased to 5-6 per group. The sample size calculation performed with the statistical data from single exposure setup 2 resulted in a total of 72 gels, 36 for each group. This large amount of samples exceeded the possibilities of the laboratory since 36 samples could not be run in parallel nor could the biomass be grown in the required amounts simultaneously. With this small amount of samples and the large amounts of protein spots (>700) all the issues associated to the curse of dimensionality [429] have to be kept in mind.

Unsupervised clustering interestingly revealed a clear grouping and differentiation between control group and exposure group, a pattern that could not be repeated for any later setup. All further clusterings show a heterogeneous mixture of reference and exposure samples. A possible explanation could be that many proteins of low abundance were missing due to the resolution of the measurement including especially the signaling proteins and transcription factors, which play a pivotal role within the regulation within cellular stress. In that case the measured proteins could cover other typical housekeeping proteins. We would not expect so much of a difference then between exposed samples and control samples and the correlation for the clustering would be rather non specific as can be seen from the results. Far more astonishing was the observation that the difference between biological replicate groups within single exposure setup 3 showed a clear separation between the first two replicate groups and the third replicate group (see Figure 55). The difference was ascribed to the later cell split for generating the replicate group 3. The interpretation on a molecular scale leads us to the nature of immortalized cell lines and raises the question if immortalized cell lines, i.e. cancer cells, and their high variability due to genetic and epigenetic instability [430] are an adequate replacement for primary cell. But then again, primary cells are not always available, as in this study project and underlines the necessity to validate any findings in vivo. Comparing the overlapping proteins with significant difference in their abundance between all three replicate groups also revealed very small overlaps (see Figure 52) but interestingly the overall pair-wise correlation coefficient distribution of all sample gel pairs of the single exposure setup 3 revealed a strong positive correlation. This could be another hint that the majority of proteins that were detectable on all sample gels are not strongly related to the cellular stress response but rather to basic cellular processes.

Investigations of the topological characteristics within OPHID protein-protein interaction network were done by (i) measuring the shortest path distributions of the significant protein sets and comparing them with randomized protein sets of same size and (ii) by expansion of the sets with their neighbors as described. The shortest path distributions of all setup sets did reveal an average shorter distance, as one would expect if the proteins were related and in topological close vicinity but the change is small when compared to random sets. While the size of the subgraphs matched randomly generated subgraphs, the expansions revealed in all setups that most of the proteins identified seem to belong into a network module. Topologically speaking they are all in a close neighborhood. This suggests a close functional relation among them following the network parsimony principle [319]. Comparisons between

the reported set of stress response related proteins from PANTHER also revealed high overlaps on the topological scale. On a side note we noticed that the stress response set from PANTHER missed one of the HSP70 proteins, which was found within our sets as being significant. Comparing the single exposure setups and the repeat exposure setup we also found high topological overlaps on the network level. This seems very likely since the underlying stress related processes are the same. The level of cytoprotection induced by firstly stimulating the cells will be hidden most likely within that small difference. Given the high variability of the biological data, as discussed above, this finding underlines that it might make more sense to consider a disease (or phenotype in general) on such a module level taking into account the proteins measured and the proteins in the same modular vicinity for finding causal models of explanation of phenotypes.

On the level of enriched biological processes and pathways analysis, bioinformatics analysis showed simultaneous activation of apoptosis and inflammatory pathways. The data indicates that both intrinsic and extrinsic pathways of cellular apoptotic mechanisms are activated following PDF exposure. Several studies in peritoneal dialysis have described apoptotic changes involving several proteins [431], [432], [433]. Among the proteins associated with the found processes and pathways we find several, especially chaperons which are essential in protein folding and stress response mechanisms like heat shock (HSPA1B, HSPB1, HSPA1A), glucose regulation (HSPA5) or hypoxia regulation (HYOU1) typically associated with cellular stress [434], [435], [436]. We have to be aware of a bias generated by the expansion method since the neighbors will most likely belong to the same processes and pathways thus enrichment of the same is more likely. A correction for this bias is proposed in the study “Comparative analysis of expansion methods on protein interaction networks” later in this chapter. The results from the contributions to this project and further findings have been discussed extensively in [382] and [383].

From the network point of view it could be demonstrated that peritoneal dialysis induces system-wide effects on a molecular scale and the measuring perspective on a proteomic level is impaired by issues based on the measuring technique thus missing proteins also involved. Combining proteomic investigations with other omics fields, for example including gene expression analysis would provide a powerful additional perspective on the stress response mechanisms, especially when inducing non-lethal PDF stimuli for activating cellular stress responses without damaging the cells. On the computational side, already existing data on networks proved helpful in asserting that the found proteins are indeed part of a close subnetwork. Identifying additional proteins, be it by the applied neighbor expansion method or by applying or inventing module detection algorithms are a powerful approach for utilizing this existing network knowledge for compensating the technical issues from two-dimensional gel electrophoresis. Further expansion methods will be proposed and discussed in great detail in the study “Comparative analysis of expansion methods on protein interaction networks” later in this chapter.

Comparative analysis of expansion methods on protein interaction networks

Introduction

Omics profiling has opened up the opportunity of explorative analysis, in translational clinical research specifically focusing on identifying molecular signatures being characteristic for clinical phenotypes. Tailored platforms have been established for high-throughput assessments spanning from the genome to the metabolomic level, and deriving omics profiles has to some extent become a standard laboratory technique. The challenge has shifted towards data interpretation, essentially for traversing descriptive lists of molecular features being associated with a given phenotype into a molecular model of feature dependencies [352]. Such models promise improved understanding of omics results, specifically for dissecting causality and association of molecular processes in the realm of clinical phenotypes under study.

Analysis on the level of interaction networks has become a standard approach in omics results interpretation, with protein interaction networks being the most prominent representative. On the level of human protein coding genes prominent interaction data repositories mainly encoding direct (physical) interactions include BioGRID [94], [108], IntAct, [144], [437], and Reactome [190], [438], [439], whereas more complex interplay is provided by KEGG [90], [100], [399], and PANTHER [172], [173], [174], [400, p. 6]. Next to the various types of interactions as defined in PSI MI format [109] the coverage in terms of protein coding genes and number of interactions represented varies substantially. As example, Reactome holds about 7,088 human proteins and 144,449 unique interactions as compared to BIOGRID with about 16,469 proteins and 104,832 interactions, whereas pathway resources as KEGG encode on the level of human about 4,823 proteins and 38,129 interactions. Consequently, analysis of omics profiles on such networks is significantly impacted by the specific network used, with the main factors being coverage of omics features on a network as well as topology of the network as such [440]. On this background hybrid networks were introduced integrating heterogeneous sources aiming at improving feature coverage and providing a more comprehensive representation of interactions [126], [347], [441].

Comprehensiveness and quality of such interaction networks has become a core issue in translational medicine, specifically in Systems/Network Medicine [36] aiming at linking molecular processes and disease phenotypes, since diseases tend to be a result from disturbances within such networks and less the source of single effectors gene products [33], [319]. Populating hybrid interaction networks with phenotype-specific omics profiles, followed by network segmentation for deciphering phenotype-specific molecular processes was recently discussed by Heinzel et al. [346], exemplified for diabetic nephropathy in [442]. Identification of such disease specific subgraphs, however, shows major hurdles as recently reviewed by Barabasi [443]. Biological networks are scale free [444], i.e. most nodes show a low degree of connectivity while several nodes form hubs of high connectivity. At least when analyzing given interaction networks the prevalence of topological modules of highly interlinked local regions within the network appears sparse but are identified [282]. In context of diseases, proteins and cellular components involved in the same disease phenotype tend to interact with each other (local hypothesis) [21], [35], [426] and show a tendency to cluster in

the same network neighborhood (disease module hypothesis) [445]. Causal molecular pathways appear further close to the shortest paths between disease associated components (network parsimony principle) [88], [427], [428].

A start point for any such analyses is population of a selected network with phenotype-associated molecular features, be it from mining of literature [446], cross-omics data integration [447], or utilizing results from individual omics profiling [442]. A frequently applied method is analyzing the frequency of phenotype-specific features in context of molecular pathways if as such provided by the network used. Such gene set enrichment analysis [448] is regularly applied on networks as KEGG or PANTHER. Alternative methods as e.g. introduced by Draghici et al. [449] follow an impact analysis taking pathway specific factors into account. The assumption of such approaches is that features being identified as relevant in statistical analysis of omics profiles mirror changes in specific molecular processes, in turn being identified via enrichment analysis. However, comprehensive identification of all features associated with a phenotype is usually not the case in omics profiling. One reason is the study design of omics experiments, and the intrinsic curse of dimensionality [429] In contrast to clinical studies the number of features tested exceeds by far the number of samples, and although utilizing correction methods for multiple testing a fraction of features identified as relevant are false positive as well as false negative. This limitation in identifying complete feature sets is complemented by shortcomings of multiplexed assays used in omics. Whereas modern transcriptomics arrays cover most protein coding genes completeness is not reached in proteomics and metabolomics, being a combination of analytical resolution together with incompletes of catalogs as such [331]. As a consequence, reconstruction of functional dependencies on a network level has to deal with missing features (being either not identified due to experimental constraints, or having failed in showing statistical significance in the light of the sample size used).

With this background concepts have been developed for imputing such missing but still relevant features. Based upon the local hypothesis one approach for predicting additional such genes or proteins is by investigating protein interaction partners within a given protein interaction network (PIN). Given a protein that shows significant differential abundance between biological samples, its direct interaction partners may play a relevant role as well [35], [426]. Following this assumption e.g. Chen et al. identified additional Alzheimer related proteins by investigating protein interaction neighbors (applying next/nearest neighbor expansion algorithms) of a set of known Alzheimer related proteins using the Online Predicated Human Interaction Database (OPHID) database as interaction data basis [332]. Such methods in the meantime saw implementation in tools like Cytoscape [119]. The next neighbor expansion increases a given set of features by all direct interaction partners from the set as identified in an underlying PIN. Such expansion (depending on the PIN used) usually leads to a significant increase of the number of proteins included in analysis. More stringent neighbor expansion methods require that additionally predicted proteins connect to at least two proteins from the measured set [382]. Another approach introduced by Zhou [450] proposes a maximum clique search in the neighborhood of a seed protein until all significant proteins are aggregated within the clique. Using graph search algorithms and taking into account biological network characteristics additional algorithms can be defined for expanding a given measurement set within a PIN [442].

We in this work systematically analyze four expansion methods utilizing two different PINS for evaluating the impact of expansion method as well as underlying PIN for the interpretation of transcriptomics profiles. As expansion concepts we define and present next neighbor expansion, inter-neighbor expansion of degree 2, inter-neighbor expansion of degree 3 and minimum spanning tree based expansion. As PINS we use KEGG and BIOGRID, and analyze the impact of expansion on resulting feature sets and their interpretation.

Material and Methods

Reference interaction networks and transcriptomics data sets

A protein interaction network is defined as a graph with proteins forming the set of nodes and their interactions forming the set of edges between the nodes. Reference human protein interaction networks were generated using (i) the Kyoto Encyclopedia of Genes and Genomes database (KEGG, database status free download version as of Feb. 2013) and (ii) BIOGRID (database version 3.2.96). KEGG provides information on genes and proteins and their relations in the form of molecular pathways, and protein-protein interactions were derived for all 262 human pathways as provided via the KEGG REST API (http://rest.kegg.jp/list/pathway/hsa, status Feb. 2013). Network reconstruction for KEGG was done by utilizing the R/Bioconductor package: KEGGgraph [388]. Extracting interactions from KEGG resulted in 4,823 unique nodes (protein coding genes) and 38,129 edges, respective numbers for BIOGRID were 16,469 nodes and 104,832 edges.

These two PINs served as reference for analyzing expansion algorithms for example transcriptomics feature sets. Here we selected two specific sets of proteins originating from transcriptomics profiling on diabetic nephropathy. The data set of Baelde et al. [451] held 171 significantly differentially regulated genes from kidney biopsies of patients in the early stages of diabetic nephropathy as compared to healthy control tissues samples. As second set a study of Cohen et al. [452] was included holding 70 significantly differentially expressed genes comparing diabetic nephropathy to control samples. Mapping the protein identifiers of the respective genes to KEGG provided 66 and 45 nodes being effectively covered for Baelde et al. and Cohen et al., for BIOGRID (covering significantly more protein coding genes) positive mapping was obtained for 77 and 45 nodes of the respective input feature sets, respectively.

Graph expansion algorithms and analysis steps

Following the local hypothesis and the network parsimony principle (and biological assumption that closely connected proteins in a PIN are also functionally dependent), the expansion of a given set of proteins by closely connected proteins should include further features of relevance. Following this hypothesis we studied expansion utilizing next neighbors, inter-neighbors of degree 2 and degree 3 and minimum spanning tree based expansions.

We started with a reference PIN denoted as graph *G* with a set of proteins denoted as set of nodes *V*, a set of protein interactions being the set of edges *E* and a subset of proteins of interest (omics feature list) denoted as *S.*

$$G = (V, E);\ S \subseteq V \tag{7}$$

A graph expansion algorithm *Ex* applied on *S* results in a subset of nodes *R* where *S* is a subset of *R* and *R* is a subset of *V*

$$Ex(S) = R, S \subseteq R, R \subseteq V \tag{8}$$

Next neighbor expansion

A next neighbor is defined as node v_n from *V* that is not an element of *S* and is directly connected to a node v_s from *S*.

$$\exists (v_n, v_s) \in E, v_n \in V, v_n \notin S, v_s \in S \tag{9}$$

The next neighbor expansion applied on *S* results in a subset of nodes R_n containing *S* and all next neighbors of nodes from *S* (Figure 63A).

Inter-neighbor expansion of degree X

An inter-neighbor of degree *X* (*X* > 1) is defined as node v_{inX} from *V* that is not element of *S* and is directly connected to at least *X* different nodes v_s from *S*.

$$\exists (v_{inX}, v_{s1}) \in E, \dots, \exists (v_{inX}, v_{sX}) \in E, v_{inX} \in V, v_{inX} \notin S, v_{s1} \in S, \dots, v_{sX} \in S, v_{si} \neq v_{sj}, i \neq j, \\ 1 \leq i, j \leq X \tag{10}$$

The inter-neighbor expansion of degree *X* applied on *S* results in a subset of nodes R_{in} containing *S* and all *X* inter-neighbors of nodes from *S* (Figure 63B for degree 2 and Figure 63C for degree 3).

Minimum spanning tree based expansion

The minimum spanning tree based expansion applied on *S* results in a subgraph *MST* from *G* that is a tree connecting all nodes of *S.* For the calculation the algorithm of Kruskal [18] for calculating a minimum spanning tree was modified: Instead of using a singular start node, we used all nodes from *S* as a starting set for expansion. Initially this results in $|S|$ distinct subgraphs which are then steadily expanded by applying the algorithm of Kruskal on the subgraph having the smallest sum of edge weights. Picking edges adjacent to two separate subgraphs results in merging those subgraphs. Eventually all subgraphs are merged within one subgraph terminating the expansion. This subgraph is finally pruned of nodes that are connected to this subgraph via a single edge and which are not part of *S*. The resulting subset of nodes R_{mst} contains all nodes from *MST* (Figure 63D).

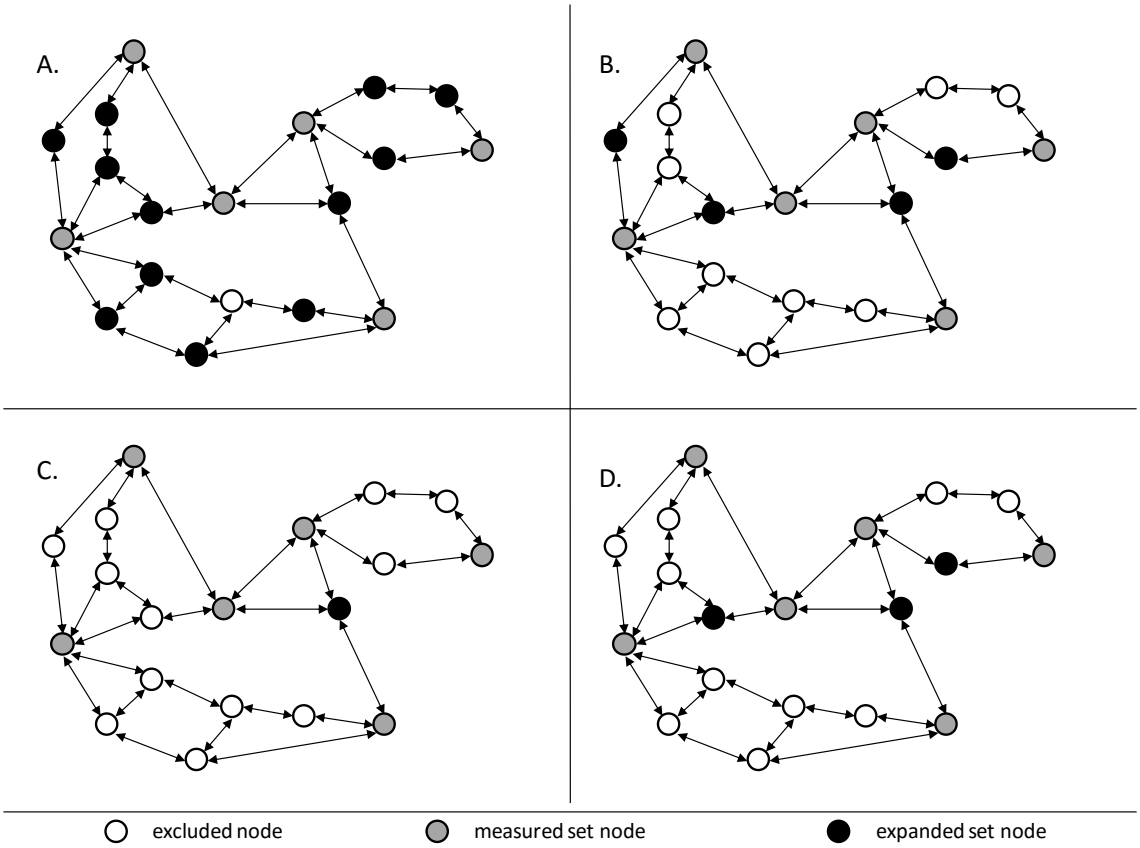


Figure 63: Schematic representation of the graph expansion algorithms on a reference PIN: A) next neighbor expansion, B) inter-neighbor expansion of degree 2, C) inter-neighbor expansion of degree 3, and D) minimum spanning tree based expansion. Grey nodes denote a selected set of nodes within G, black nodes denote expanded nodes after expansion, with white nodes being neither in the seed nor in the expanded set of nodes.

Utilizing KEGG and BIOGRID as reference PINs the two data sets from transcriptomics served as seed set for the four expansion algorithms. Results of expansion were compared to randomly generated reference data sets covering the set sizes 10 to 300, i.e. spanning typical data set sizes given for transcriptomics data sets after mapping on the reference networks.

Such randomly generated protein sets from each PIN covering the range of 10 to 300 features per data set with a step size of 10 (resulting in 30 set size intervals) were used as reference. For each set size 100 individual, random feature sets were generated (leading to in total 3,000 feature sets) for allowing computation of statistical measures. As for the two transcriptomics data sets feature set expansion was performed applying the four algorithms.

To investigate KEGG pathway enrichment of a given protein set (from the transcriptomics references or the random sets) after applying a specific expansion, the observed proteins for each pathway category were counted. Given the protein set size, the known pathway category sizes and the total protein count of a PIN, the protein count that is expected by random picking was calculated for each pathway category by

$$expected\ count = \ protein\ set\ size * \frac{pat\ hway\ \ category\ \ \ size}{total\ \ protein\ \ \ count\ \ of\ \ PIN} \tag{11}$$

The ratio from the observed protein count and the expected protein count for each pathway category of a given protein set was used for determination of pathway category enrichment.

$$enrichment\ ratio_{pat\ hway\ \ x} = \frac{observed\ \ \ protein\ \ \ count\ _{pat\ hway\ \ x}}{expected\ \ \ protein\ \ \ count\ _{pat\ hway\ \ x}} \tag{12}$$

Software tools

EXCEL 2007 (Microsoft Corporation, Redmond U.S.) was used for data handling and graphics. SPSS 13.0 (IBM, New York, U.S.) and R [424] were used for statistics and graphics. MeV was used for hierarchical clustering [422]. Network calculations were done using the Basic network analysis software suite. Cytoscape [119] was used for network graphics.

Results

We applied our set of graph expansion algorithms (next neighbor expansion, inter-neighbor expansion of degree 2 and 3, and minimum spanning tree-based expansion) on the randomized reference sets and the transcriptomic sets from Cohen et al. und Baelde et al. for KEGG and BIOGRID, respectively. For systematically analyzing the consequence of feature set expansion we first investigated the effect of expansion methods on the feature count of randomized input feature sets in dependence of applied expansion method and PIN relative to input feature set size. We describe the findings of expanded feature counts between expansion methods and PINs and compare the results with the expanded feature counts of transcriptomic data sets. Next we compare the feature overlap of expanded feature sets relative to expansion method and PINs for randomized feature sets and the transcriptomic data sets. After describing our findings regarding expanded feature count and feature overlap, we investigate the consequence of expansion on the level of functional pathway enrichment for randomized feature sets and transcriptomics data sets. We address the questions of dependency of expanded pathway enrichment on feature set size and pathway set size, and investigate the differences in enrichment of specific pathways in relation to expansion methods and underlying PIN. We identify an enrichment bias and propose a bias correction and conclude our results by investigating the difference and overlap of the count of enriched pathways in the transcriptomics data sets with and without correction.

Expansion results on feature set size

The consequence of the expansions on the number of additional features is illustrated in Figure 64. Mean increase of the number of additional features (slope of the curve) is almost constant over all set sizes for any given expansion algorithm and PIN. While standard deviation is usually in the range of 10% of mean increase in number of additional features for any given algorithm and PIN, we observed a standard deviation of up to 50% of mean increase in number of additional features for next neighbor expansion within BIOGRID (Figure 64 BIOGRID a.), suggesting the presence of highly connected hub proteins. For example we identified UBC (ubiquitin C) having 26,150 connections within BIOGRID as cause. Further comparing the applied expansions on the level of PINs we notice a similar increase of mean additional features between KEGG and BIOGRID for each expansion algorithms. Comparing the applied expansion algorithms the mean increase in the number of additional features relative to the input set size is highest within next neighbor expansion (KEGG about 8-fold, BIOGRID about 11-fold) and lowest in minimum spanning tree based expansion (KEGG about 0.6-fold, BIOGRID about 0.3-fold with respect to the input set size) with the inter-neighbor expansion algorithms, being subsets of next neighbor expansion, in between.

When applying neighborhood expansion algorithms on the signature of significant transcripts from transcriptomics data we expect higher number of features (relative to the randomized reference) due to the higher chance of proteins connecting two or more relevant proteins being in the close vicinity of each other within a PIN (local hypothesis). We notice an increase (expanded feature count of the transcriptomics data is above mean + standard deviation of randomized reference) of additional features for the data set of Cohen et al. within inter-neighbor expansion of degree 2 and 3 in KEGG and BIOGRID (Figure 64 KEGG b, c; BIOGRID b, c) and for Baelde et al. within inter-neighbor expansion of degree 2 and 3 in KEGG (Figure 64 KEGG b, c). We did not notice an increase in next neighbor expansions (Figure 64 KEGG a; BIOGRID a). For minimum spanning tree-based expansion we expect to find a reduced rate of additional features if a transcriptomics signature set is within close vicinity of each other within a PIN since less protein nodes are required to connect them. We notice a decrease (expanded feature count of the transcriptomics data is below mean - standard deviation of randomized reference) of additional features for the data set of Cohen et al. within KEGG and BIOGRID PIN (Figure 64 KEGG d; BIOGRID d), but not for the data set of Baelde et al.

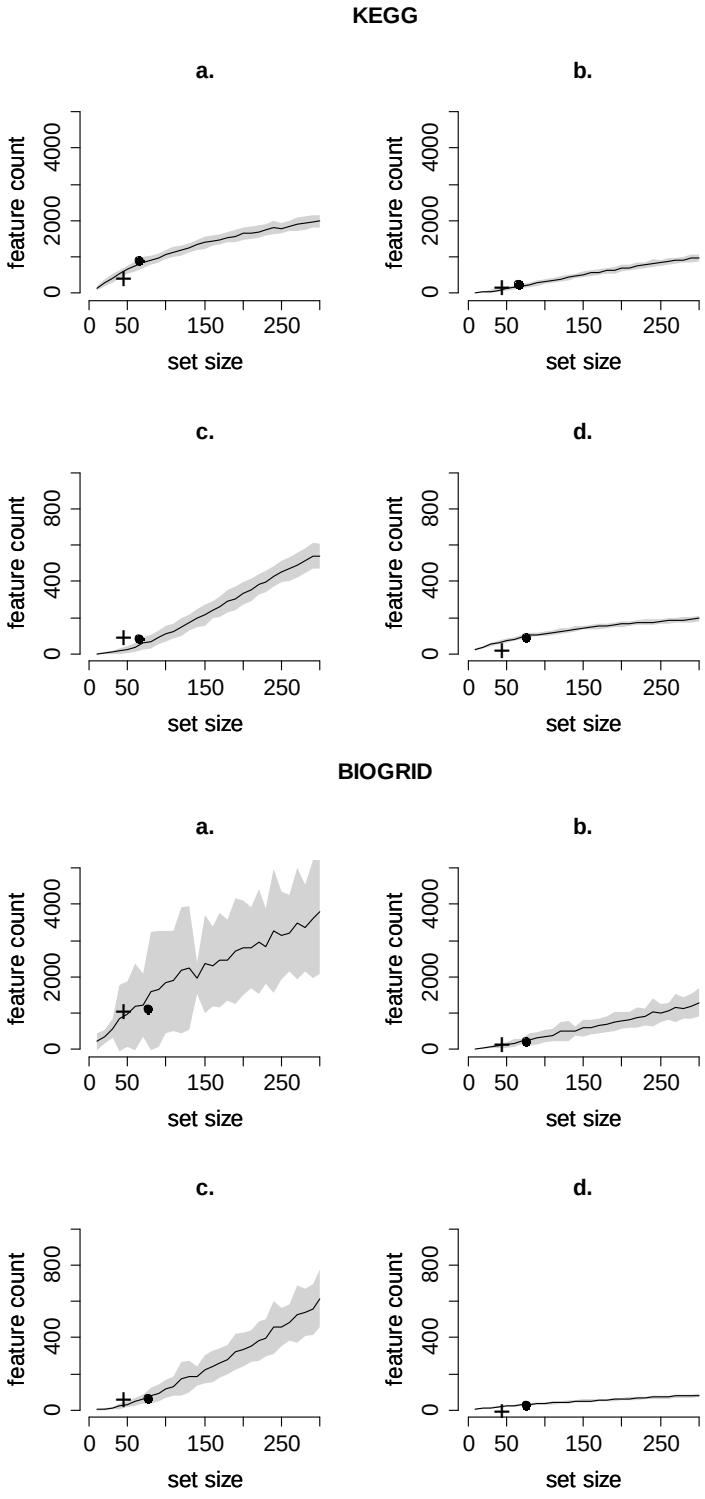


Figure 64: Random feature set size for the interval [10,300] plotted against the number of nodes included after expansion for KEGG and BIOGRID (mean, as solid line and standard deviation, as grey area) utilizing a) next neighbor expansion, b) inter-neighbor expansion of degree 2, c) inter-neighbor expansion of degree 3 and d) minimum spanning tree based expansion. Additionally, the respective values reached for the transcriptomics example data sets (solid dot for Baelde et al, cross for Cohen et al.) are displayed (feature count scale is reduced for clarity for c. and d.)

Expansion results on feature overlap

We investigated the feature overlap between expansion algorithms of randomized sets comparing them with the overlaps from transcriptomic sets from Baelde et al. and Cohen et al. Inter-neighbor expansion algorithms of degree 2 and 3 are subsets of next neighbor expansion and thus overlapping completely. Minimum spanning tree based expansion feature sets overlap with feature sets from: next neighbor expansion in the range of 67%-86% (random sets), 77% (Baelde), 86% (Cohen) within KEGG, and 100% (random sets), 100% (Baelde), 10% (Cohen) within BIOGRID; inter-neighbor expansion of degree 2 in the range of 20%-57% (random sets), 30% (Baelde), 43% (Cohen) within KEGG, and 50%-67% (random sets), 67% (Baelde), 9% (Cohen) within BIOGRID and inter-neighbor expansion of degree 3 in the range of 7%-29% (random sets), 13% (Baelde), 29% (Cohen) within KEGG, and 25%-50% (random sets), 33% (Baelde), 5% (Cohen) within BIOGRID. Expanded feature overlap for the data set of Cohen et al. within BIOGRID is significantly below the overlap from the randomized reference.

Consequences of expansion on pathway enrichment on randomized feature sets

For systematically investigating the consequence of expansion methods on the functional level of feature sets we calculated the enrichment ratio of pathways for each randomized feature set.

In a first step we investigated if the enrichment of pathways of randomized features sets is dependent on the set size by calculating linear regressions on enrichment ratios over all set sizes for each pathway at a given expansion algorithm and PIN. Mean values, minima and maxima of all regression slopes are around zero indicating that pathway enrichment by expansion algorithms is not dependent on set sizes.

In a second step we investigated if the enrichment of pathways of randomized feature sets is dependent on the pathway size by calculating the relative enrichment ratio per pathway size for all pathways of all randomized feature sets and grouping them by expansion method and PIN. We did not identify a dependency of the enrichment of pathways (data not shown).

Consequently, we investigated if the mean enrichment of pathways of randomized feature sets over all set sizes differs between expansion methods and PINs. Therefore we calculated the mean enrichment ratio for each pathway over all set sizes, for each algorithm and PIN (see Figure). Varying with expansion method we identified between 20% - 50% of all pathways as having enrichment ratios above 2.0 within KEGG, and ca. 30% of all pathways having an enrichment ratio above 1.5 within BIOGRID, clearly showing a bias for enrichment after applying expansion methods. Additionally, by ordering the pathways by the enrichment ratio (descending) using the next neighbor expansion as a reference (Figure 65 a. and Figure 66 a.) we identify a similar trend in enrichment of the same pathways for next neighbor expansion and inter-neighbor expansions of degree 2 and degree 3 (Figure 65 a., b., c.; Figure 66 a., b., c.) in each PIN. We did not identify a similar trend in minimum spanning tree based expansion (Figure 65 d.; Figure 66 d.) suggesting the enrichment of other pathways. Moreover, we notice an increase of mean enrichment ratios of inter-neighbor expansion of degree 2 and degree 3 compared to next neighbor expansion in both PINs suggesting a higher enrichment of already enriched pathways by expansion methods when picking features being connected to more already identified features. The ordering of pathways differs though between KEGG and BIOGRID PIN (data not shown). For identifying the degree of difference of pathway enrichment

between KEGG and BIOGRID we calculated the absolute difference of the mean enrichment ratio for each pathway over all randomized feature sets between KEGG and BIOGRID PIN, grouped by expansion algorithm (see Figure 67). For assuming similar pathway enrichment between KEGG and BIOGRID we would expect finding difference-distributions close to zero, which is not the case. In contrast, the majority of pathways differ in their enrichment ratio between KEGG and BIOGRID in the range 0.5 to 1.5. While minimum spanning tree-based expansion provides the least differences, inter-neighbor expansion of degree 2 and degree 3 provide the strongest differences in enriched pathways.

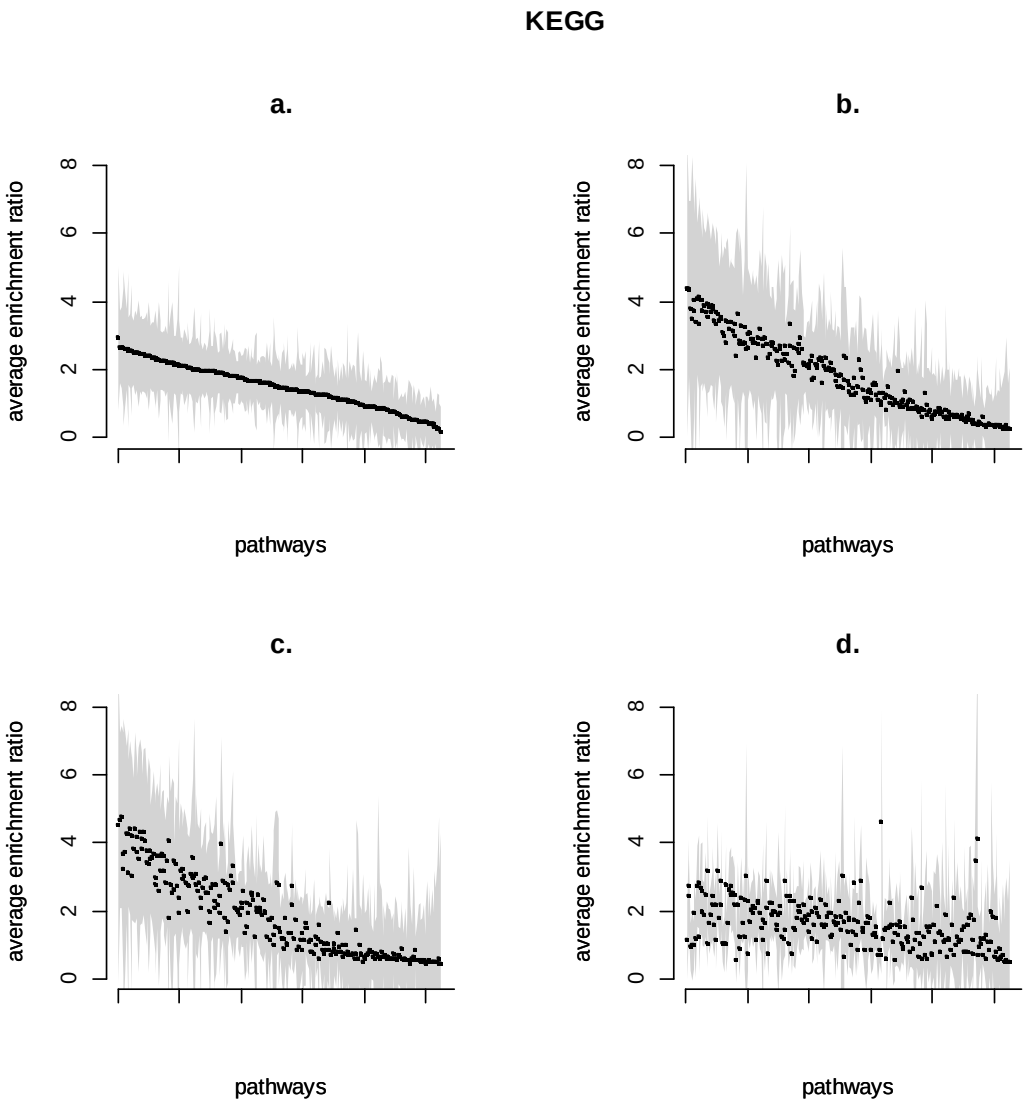


Figure 65: Pathways (x-axis) sorted by mean pathway enrichment (y-axis, solid dots, average standard deviation as grey area) of random input feature sets for a) next neighbor expansion, b) inter-neighbor expansion of degree 2, c) inter-neighbor expansion of degree 3 and d) minimum spanning tree based expansion within KEGG. Pathway ordering for all diagrams is defined by sorted enrichment order of next neighbor expansion within KEGG.

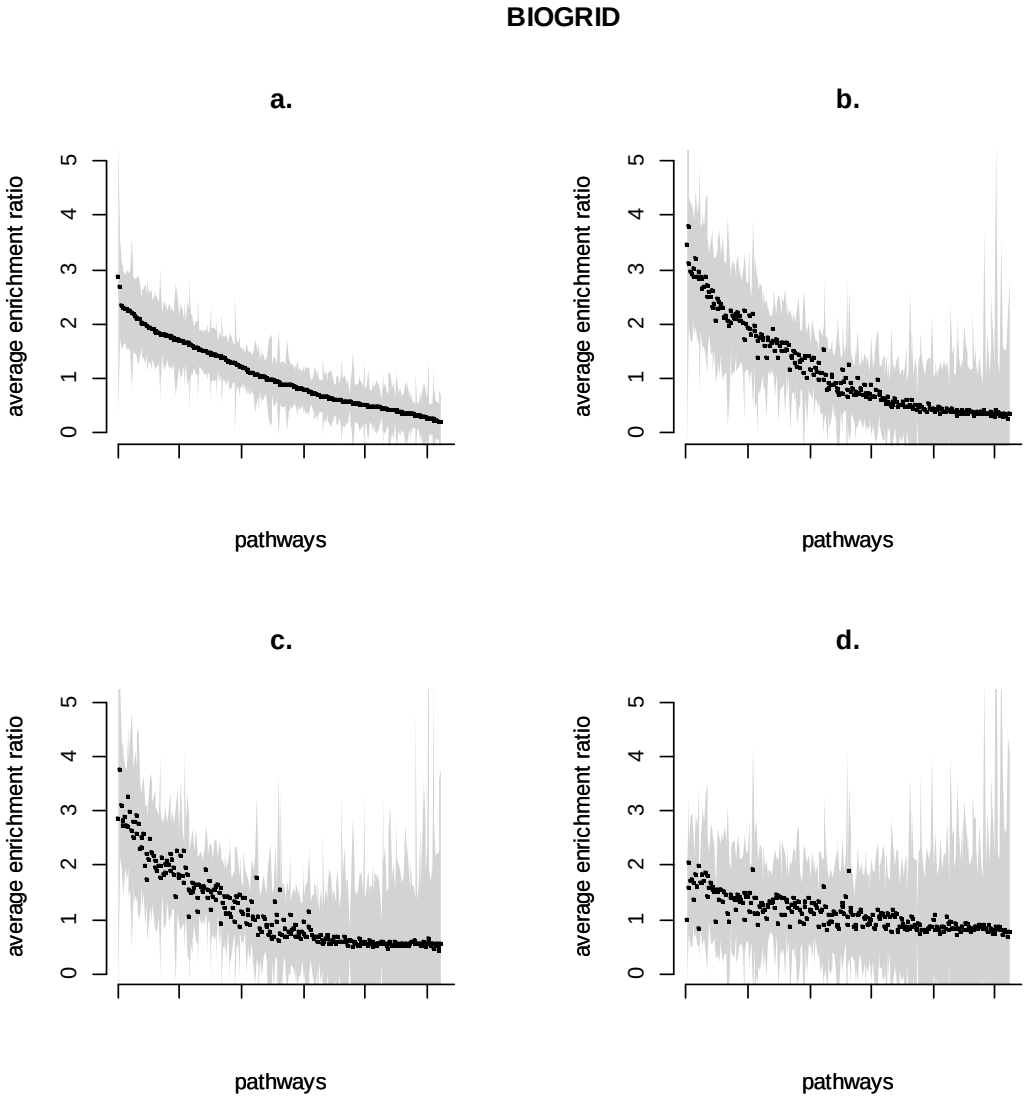


Figure 66: Pathways (x-axis) sorted by mean pathway enrichment (y-axis, solid dots, average standard deviation as grey area) of random input feature sets for a) next neighbor expansion, b) inter-neighbor expansion of degree 2, c) inter-neighbor expansion of degree 3 and d) minimum spanning tree based expansion within BIOGRID. Pathway ordering for all diagrams is defined by sorted enrichment order of next neighbor expansion within BIOGRID.

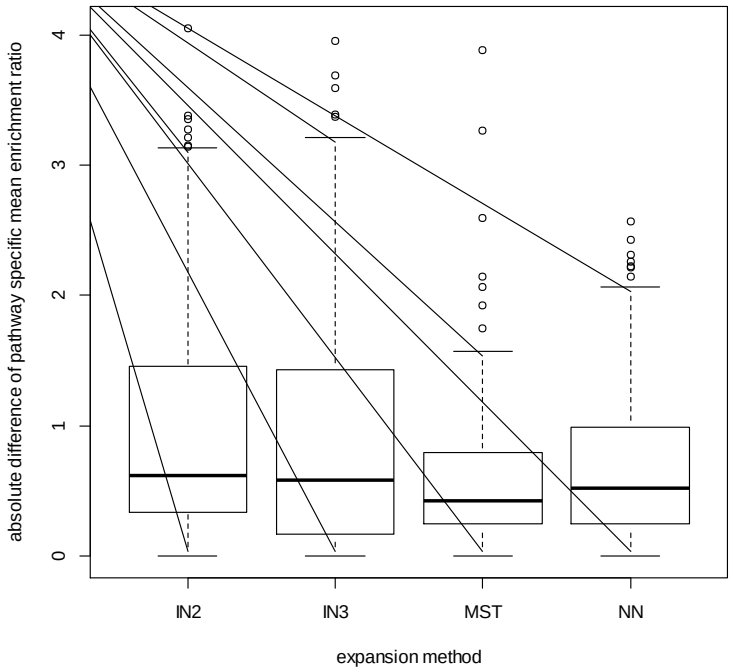


Figure 67: Box plots of the randomized feature sets distributions of absolute differences of the mean enrichment ratios for all pathways, calculated for each pathway between KEGG and BIOGRID PIN, grouped by expansion methods: inter-neighbor of degree 2 (IN2), inter-neighbor of degree 3 (IN3), minimum spanning tree based (MST) and next neighbor (NN).

Consequence of pathway enrichment on transcriptomics data sets

Finally we investigated the enrichment of pathways for the transcriptomic data sets. We calculated the pathway enrichment for the data sets from Baelde et al. and Cohen et al. identifying the number of pathways being enriched, defined as having an enrichment ratio >2 (exceeding the average standard deviations, see Figure 66). In the light of our findings regarding the bias for enrichment of pathways after applying expansion methods, we propose a corrected enrichment ratio by calculating the pathway enrichment ratio of the transcriptomics data relative to the measured enrichment ratio of randomized datasets of same input set size. We compared the numbers of enriched pathways between uncorrected and corrected pathways for each expansion method and for each PIN in Table 13. Correction results in a reduction of the number of enriched pathways in the range of 24% to 79% depending on expansion method and PIN. Additionally we identified new pathways as being enriched after correction when comparing the corrected enriched pathway count to the overlap between uncorrected and correct enriched pathway counts. Interestingly, the increase of additional pathways relative to the overlap is at most 85% for all expansion methods in all PINs except for the next neighbor expansion within BIOGRID, having an increase of 4 to 5 times.

Table 13: For each experimental data set the count of enriched pathways (enrichment ratio >2) within each expansion method was calculated for uncorrected as well as for corrected enrichment ratios. Finally the overlap of enriched pathways between uncorrected and corrected pathways was counted.

BIOGRID	uncorrected	overlap	corrected	**KEGG**	uncorrected	overlap	corrected
Cohen,N	60	5	20	Cohen,N	87	30	31
Cohen,IN2	73	19	25	Cohen,IN2	89	18	20
Cohen,MST	46	34	35	Cohen,MST	69	23	24
Cohen,IN3	71	18	20	Cohen,IN3	90	41	43
Baelde,N	36	5	25	Baelde,N	111	17	23
Baelde,IN2	69	20	37	Baelde,IN2	82	16	21
Baelde,MST	58	37	41	Baelde,MST	112	31	36
Baelde,IN3	76	24	36	Baelde,IN3	73	22	27

Discussion

We applied four graph expansion algorithms on randomized reference feature sets from KEGG and BIOGRID, as well as on two transcriptomics feature sets from Baelde et al. and Cohen et al.

For randomized feature sets we find an almost linear increase of expanded feature set size within our investigated set size range of 10-300 with standard deviations in the range of ca. 10% of mean except for next neighbor expansion within BIOGRID having a standard deviation in the range of ca. 50% of mean (Figure 64 BIOGRID a). We identified hub proteins as UBC (ubiquitin C) within BIOGRID as reason for this. Consequently when comparing the three neighborhood expansions the expanded feature set sizes of next neighbor expansions are susceptible to hub proteins, while inter-neighbor expansions requiring at least two connections from an expanded feature to the input feature set, are not afflicted by this fact. Additionally, next neighbor expansion leads to the highest increase in the number of additional features while inter-neighbor expansions lead to a lesser increase (due to the more stringent neighborhood rule). Minimum spanning tree-based expansions have the smallest increase of feature set size for KEGG and BIOGRID, i.e. a smaller number of features is required to connect random feature sets in a tree compared to a situation of having neighbors. We see this again when comparing the overlap between the expansion methods. The more stringent the neighborhood rule becomes, the less overlap we see towards minimum spanning tree based expansion. Regarding the underlying network, we find similar ranges for all expansion methods in KEGG and BIOGRID PIN, with BIOGRID having usually the higher rate in feature increases and overlaps, suggesting strong topological similarities between both PINs.

Following the local hypothesis for experimentally measured feature sets (missing features due to statistical and experimental constraints), we would expect increases of expanded feature for neighborhood expansions (especially for the stringent inter-neighbor expansions) and decreases of expanded features within minimum spanning tree based expansions. Applying the expansion methods on the transcriptomics data from Baelde et al. and Cohen et al. and

comparing them with the results from the randomized reference feature sets, we cannot identify differences in expanded feature count between random sets and transcriptomics data on the level of next-neighbor expansion, while we do see expected differences with the other expansion algorithms (stronger for Cohen et al. than for Baelde et al.) (see Figure 64). A possible explanation is that the high increase (about 10-fold) of features by next neighbor expansion will include many unspecific additional features which are found regardless of underlying feature sets.

When comparing the overlap of expansion methods between randomized feature sets and transcriptomics feature sets we notice a smaller overlap than expected between minimum spanning tree based expansions and all other neighborhood expansions for the feature set of Cohen et al. within BIOGRID. This indicates that features identified in Cohen et al. are in close vicinity of each other within BIOGRID requiring only a small number of additional features for connecting them in a tree. This cannot be seen within KEGG though suggesting that both PINs differ typologically in the subnetworks covering features as found in Cohen et al.

Functional interpretations from expanded feature sets on the level of pathway enrichment analysis have to consider a bias within the enrichment due to applied expansion methods and PIN topology. Within randomized feature sets we see a similar and expected enrichment of pathways between next neighbor and inter-neighbor expansion methods differing from minimum spanning tree based expansion. Additionally we note an increase of enrichment from next neighbor expansion to inter-neighbor expansions which emphasizes that more stringent neighborhood expansions are more likely to be functionally related to the input feature set. Interestingly this enrichment of pathways does not depend on the set size allowing a definition of a correction value for each pathway (at a given expansion method and PIN). Enrichment of pathways between KEGG and BIOGRID PIN differs though indicating topological variations in the functional subnetworks of both PINs. Applying our proposed correction on transcriptomics feature sets and comparing the count of enriched pathways with uncorrected counts, the amount of enriched pathways decreases as one would expect and fewer new pathways are identified as being enriched. Interestingly, next neighbor expansion in BIOGRID results in a very low overlap of enriched pathways between corrected and uncorrected expansions in both transcriptomics feature sets while we cannot see this within KEGG, again hinting on the influence of hubs on feature expansions.

Our analysis of applying graph expansion methods on experimental feature sets utilizing two different PINs reveals that we can address the issue of false negative features within experimental feature sets. By applying graph expansion methods we can identify expansions differing significantly from randomized feature sets thus harboring relevant additional features. Comparing the four applied expansion methods within the given PINs, next neighbor expansion is susceptible to hub proteins from BIOGRID and finds several times more features, thus increasing most likely the false positive rate. Minimum spanning tree-based expansion does find the fewest additional features but is susceptible to already highly connected feature sets and to alternative routes within a functional subnetwork which are not found per definition. Both inter-neighbor expansions seem applicable with inter-neighbor expansion of degree 3 having the more stringent rule leading to smaller feature sets. Additionally after identifying expansions of relevance performing a pathway analysis for functional interpretation can be performed when including a correction for the algorithmic and topological enrichment bias.

Angiogenesis in brain metastasis

Introduction

Brain metastases (BM) are a very common brain tumor in adults, which are usually developed by lung carcinomas (35-64%) and others [453]. The incidence of BM is estimated to about 19-25% per year based on epidemiological studies and autopsy data [454] and the prognosis of patients with BM is poor with a median survival time of less than one year. The therapy of BM is difficult and relies mainly on radiotherapy and sometimes surgery [455], [456]. Systemic chemotherapy approaches suffer from the little understanding of the molecular mechanisms that result in the growth of cancer cells within the brain [457]. Studies indicated that angiogenesis is relevant for metastasis development in types of cell lung cancer, and anti-angiogenic treatment successfully prevented metastasis development [458], [459]. The studies also indicated that BM resulting from melanoma seems to follow other angiogenic growth patterns since the anti-angiogenic treatment was not sufficient.

The goal of this project was to investigate if there are different patterns of angiogenic pathways between BM caused by lung cancer and melanoma based on real-time PCR gene expression data from a cohort of brain metastasis from patients. Expression analysis relied on gene expression based on angiogenesis RT-PCR chips. This approach is specific for angiogenesis reported genes and will not include a whole genome expression profile.

The project was supported by several contributions including the raw data preprocessing and statistics, as well as a comparative network analysis utilizing the omicsnet protein-protein interaction network [337], [441].

Material and Methods

Provided data

Data was provided for further analysis in the form of raw gene expression data given as cycle times (CT values)from a Real Time PCR System utilizing a TaqMan® pathway gene array from Applied Biosystems holding 92 genes associated to angiogenesis and lyphangiogenesis (Catalogue Number 4391016, Applied Biosystems, Forster City, CA, USA) [460] and 4 endogenous control genes. Sample setup consisted of 15 patient samples with brain metastases of which 5 patients had non-small cell lung cancer (NSCLC), 5 patients with small cell lung cancer (SCLC) and 5 patients with melanoma. The reference group consisted of one brain patient sample set who had gliosis.

Data preprocessing

Missing Values

Array raw data was preprocessed and prepared for network analysis by firstly screening for missing values. Missing values can result from either technical issues with the TaqMan array or from genes having no active expression due to down-regulation or missing activation. For

differentiating between technical and biological cause for missing values, we assumed that gene expression values had to be present in at least 5 samples. For less than 5 samples we assumed technical issues with the array for the given gene and we ruled this gene out from further analysis. For the remainder of missing values we assumed the genes were not activated and thus set the CT value to 45 at which point even a single strand should have been detected.

Outlier detection

Outliers were defined as CT values having an implausible high expression rate exceeding the overall mean CT value by three times standard deviation (mean – 3* standard deviation). Outlier correction was performed by replacing an outlier of a given samples gene with the mean CT value of the remainder gene values of the given gene within the same patient group.

Normalization and relativization

Data was normalized and relative fold change calculated according to ΔΔCt method [461] using the control sample as reference. Normalization was done calculating for a given gene *X* and Sample *Y* the dCT value against the mean CT of a samples housekeeping genes CT values (mCThg) according to:

$$\mathrm{dCT}_{\text{gene X,sample Y}} = \mathrm{CT}_{\text{gene X,sample Y}} - \mathrm{CT}_{\text{mCThg ,sample Y}} \quad (13)$$

Relative fold change was calculated against the reference gliosis sample (GLI) for each sample *Y* and gene *X* dCT value by:

$$\mathrm{ddCT}_{\text{gene X,sample Y}} = \mathrm{dCT}_{\text{gene X,sample Y}} - \mathrm{CT}_{\text{gene X,sample Y}} \quad (14)$$

Relative fold change was calculated for each sample *Y* and gene *X* by:

$$\text{relative fold change}_{\text{gene X,sample Y}} = 2\text{^}(-\mathrm{ddCT}_{\text{gene X,sample Y}}) \quad (15)$$

Data analysis

Hierarchical Clustering

Normalized data was hierarchically clustered using Pearson correlation as distance with average linkage rule using the MultiExperiment Viewer Software version 4.9 [421], [422].

Gene expression analysis

Differences in gene expression between the sample groups were identified by calculating the average relative fold change for each gene of each patient sample group. Three relative fold change intensity groups were defined for each patient sample group by grouping genes having at least 50-fold, 10-fold or 2-fold difference in their abundance compared to the reference sample. Overlaps and differences of the relative fold change between the study groups were counted for each gene.

Network analysis

The TaqMan assay gene set was mapped to omicsnet protein interaction network (version Feb. 2013) [337], [441], , and visualized using Cytoscape [119]. Proteins were highlighted according to the patient sample group and the relative fold change group thus visualizing differences

between the cancer groups. Edges from omicsnet were highlighted if the connection between the given proteins were also experimentally verified. Additionally the resulting subgraphs were connected with other proteins not part of the TaqMan assay into a single subgraph using the modified minimum spanning tree based approach from the project “Comparative analysis of expansion methods on protein interaction networks”.

Software Tools

EXCEL 2007 (Microsoft Corporation, Redmond U.S.) was used for data handling and graphics. R [424] was used for statistics and graphics. MeV was used for hierarchical clustering [422]. Network calculations were done using the Basic network analysis software suite. Cytoscape [119] was used for network graphics.

Results

Data preprocessing

Patient samples were anonymized and labeled (see Table 14).

Table 14: Sample labels after anonymization grouped by patient sample group

patient sample group	sample labels
SCLC	#11, #12, #13, #14, #15
NSCLC	#1, #2, #3, #4, #17
melanoma	#7, #9, #16, #18, #19

Initiating data preprocessing it was observed that of all the housekeeping genes only GUSB hat CT values for each patient sample. 18S had missing values in 2 samples, #4 and #15. HPT1 had missing values in samples #1, #3, #4, #11 and GAPDH had missing values in 11 samples, #1, #2, #3, #4, #7, #9, #11, #13, #13, #14, #15. For reference housekeeping genes this result is odd. Either these housekeeping genes were indeed not expressed suggesting a biological reason for this unexpected phenomenon however unlikely or a technical issue with the TaqMan assay was the cause for this. A control RT-PCR experiment was performed with the samples and the primers for the housekeeping genes. All housekeeping genes were present as would have been expected. Therefore technical issues with the TaqMan assay could not be excluded.

Missing values

A total of 8 genes had missing values in less than 5 samples and were omitted suggesting further issues with the TaqMan assay similar to the housekeeping genes. Removed genes included PLG, ANGPTL3, FGF4, SERPINB5, COL4A3, TNNI1, LECT1 and IFNG. For all other genes missing values were set to a CT value of 45. From the remaining genes gliosis reference included missing values in 4 genes (TNMD, PRL, LEP, F2).

Outliers

Outliers having CT values less than mean CT value (over all gene CT values) minus three times standard deviation were identified within sample #11 for ANGPTL1 and within sample #9 for BAI and modified accordingly to the mean CT value of those genes patient sample group CT values.

Normalization and relativization

Normalization was done with GUSB as only housekeeping gene reference since the other three housekeeping genes had to be omitted due to possible technical issues with the TaqMan assay.

Hierarchical clustering

Relative fold change data was clustered as described using Pearson correlation as distance and average linkage rule (see Figure 68). Samples #11, #12, #14, and #15 from the SCLC patient sample group forms a tight cluster as well as samples #1, #3, #4 from the NSCLC patient sample group. Two melanoma dominated groups can be identifies containing in one the samples #16, #18 and #19 as well as sample #17 from the NSCLC group and the other with the samples #7 and #9 as well as sample #13 from the SCLC group. Sample #2 from the NSCLC group had no close distance to any of the other groups.

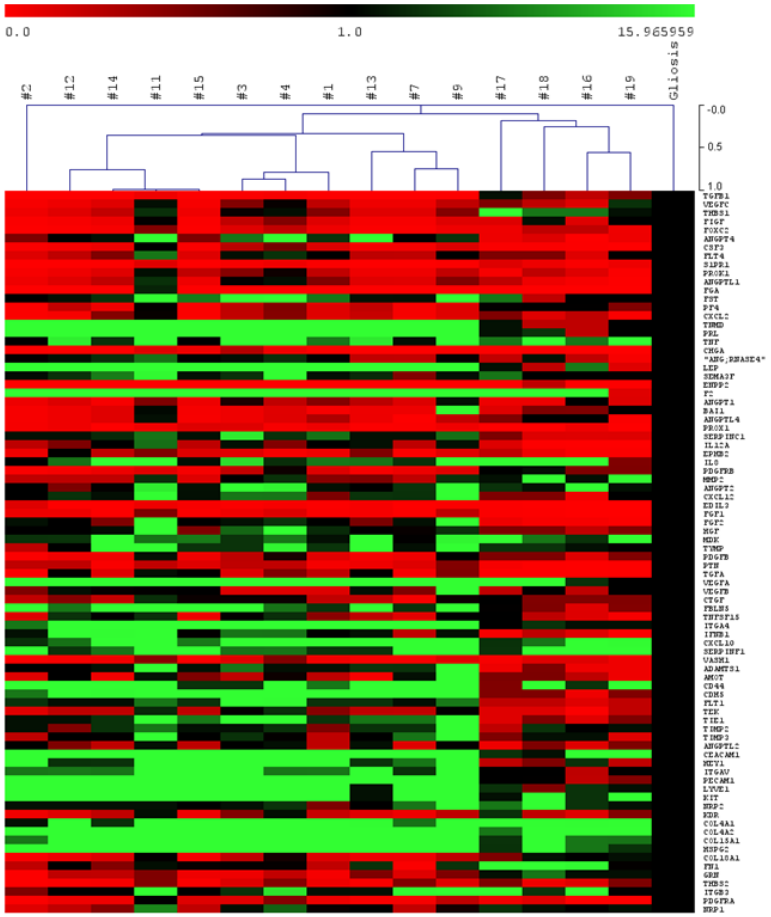


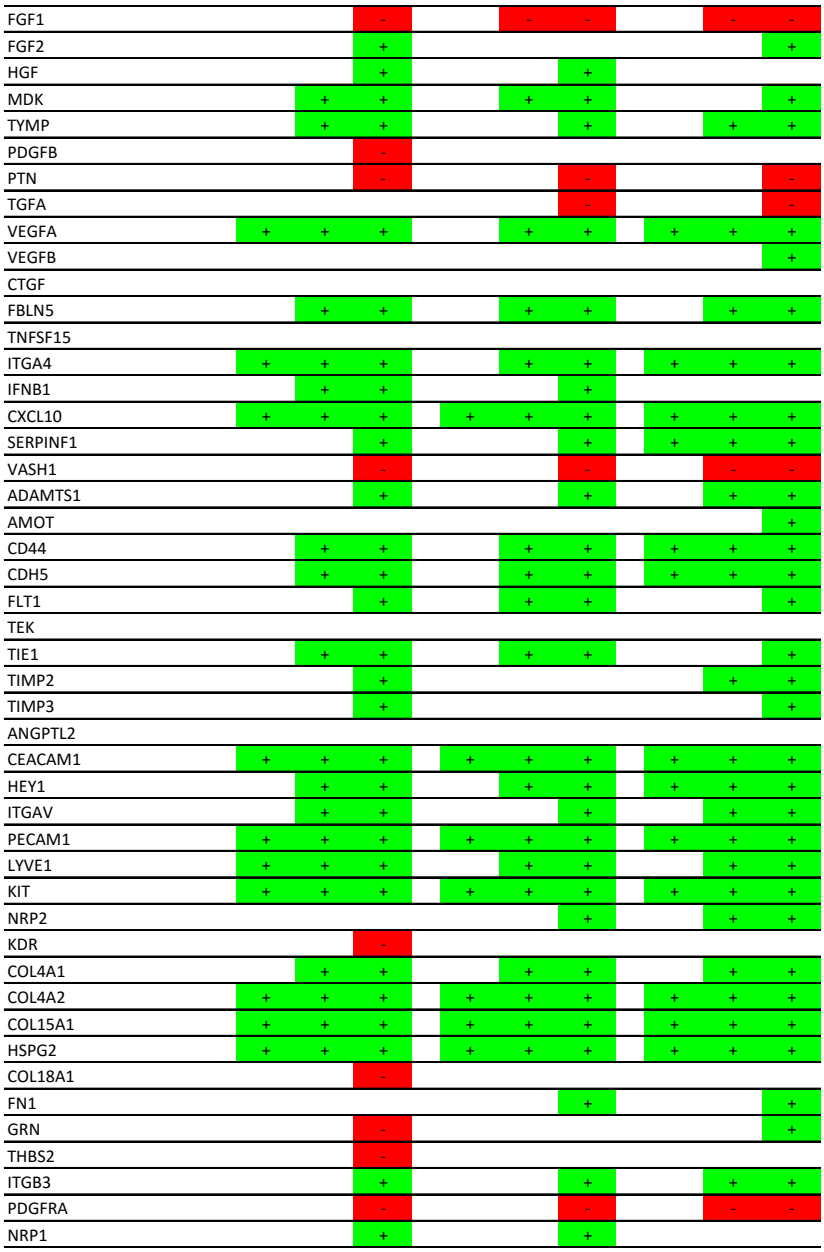
Figure 68: Hierarchical Clustering of relative fold change data between patient sample and gliosis reference using Pearson correlation and average linkage rule.

Gene expression analysis

Analysis of gene expression changes was done by firstly calculating the average relative fold change for each gene for each patient sample group (SCLC, NSCLC, Melanoma). Secondly three groups of fold change ranges are defined (50-fold, 10-fold, 2-fold) and genes are included in a group if their relative fold change to the gliosis reference is at least 50-fold, 10-fold or 2-fold respectively. Thirdly overlaps of up- or down-regulated genes between the patient sample groups are calculated (see Table 15).

Table 15: Gene expression overview of patient sample groups for small cell lung cancer (SCLC), non-small cell lung cancer (NSCLC) and melanoma (Melanoma). Genes having an average relative fold change difference for each patient group of at least 50-fold, 10-fold or 2-fold are marked '+' with green background if up-regulated, '-' with red background if down-regulated. Housekeeping genes are in italic and on top of the list. Genes having an outlier with modified value are marked with '$' and genes having a missing value within the gliosis reference sample are marked with §.

		SCLC			NSCLC			Melanoma		
relative fold change greater than:		50	10	2	50	10	2	50	10	2
count up-regulated:		15	28	43	11	25	41	17	32	44
countdown-regulated:		2	8	28	2	9	15	3	11	17
Target Gene										
GUSB			-	-						+
TGFB1		-	-	-						
VEGFC				-						
THBS1							+			+
FIGF				-						
FOXC2				-		-	-			-
ANGPT4		+	+	+			+			
CSF3				-			-		-	-
FLT4										
S1PR1			-	-		-	-		-	-
PROK1				-						-
ANGPTL1	$									-
FGA					-	-	-	-	-	-
FST			+	+		+	+		+	+
PF4				-						
CXCL2				-						-
TNMD	§	+	+	+	+	+	+	+	+	+
PRL	§	+	+	+	+	+	+	+	+	+
TNF			+	+		+	+		+	+
CHGA			-	-		-	-	-	-	-
ANG;RNASE4				+						
LEP	§	+	+	+	+	+	+	+	+	+
SEMA3F				+		+	+			
ENPP2		-	-	-	-	-	-		-	-
F2	§	+	+	+	+	+	+	+	+	+
ANGPT1				-						
BAI1	$			-		-	-		+	+
ANGPTL4				-						
PROX1			-	-		-	-		-	-
SERPINC1				+			+			
IL12A				+						
EPHB2				-			-		-	-
IL8			+	+			+		+	+
PDGFRB			-	-						
MMP2									+	+
ANGPT2				+			+		+	+
CXCL12				+			+			+
EDIL3			-	-		-	-	-	-	-



11 genes have a relative fold change difference of at least 50-fold all of which are up regulated (TNMD, PRL, LEP, F2, CXCL10, CEACAM1, PECAM1, KIT, COL4A2, COL15A1, HSPG2) of which 4 having a missing value within the gliosis reference (TNMD, PRL, LEP, F2). Since the missing values were set to 45 being interpreted as not expressed, the increase of more than 50 fold is explained. Interestingly a decrease of GUSB housekeeping gene relative to gliosis reference of at least 10-fold can be observed indicating a change in abundance which would not be expected of a housekeeping gene on that scale. One gene, BAI1 was found having up-regulation as well as down-regulation between the three patient sample groups with melanoma being up-regulated SCLC and NSCL down-regulated.

For genes with a relative fold change difference of at least 10-fold belonging to only one patient sample group Melanoma holds 11 genes (4 down-regulated, 7 up-regulated), NSCLC holds 3 genes (1 down-regulated, 2 up-regulated) and SCLC holds 4 genes (1 down-regulated, 2 up-regulated) of which 2 show also a 50-fold fold change difference. Genes showing a relative fold change difference of at least 10-fold which are part of two patient sample groups SCLC and NSCL have two joint genes (both up-regulated), SCLC and Melanoma have 3 joint genes (all up-regulated) and NSCLC and Melanoma have 2 joint genes (both down-regulated) of which one shows also a 50-fold fold change difference in both groups (see Table 16).

Table 16: Overlap matrix of genes showing at least 10-fold difference in gene expression between patient sample groups and glisosis reference (SCLC: small cell lung cancer group, NSCLC: non-small cell lung cancer group). Genes having at least 50-fold difference in gene expression are marked with '*'. Up-regulated genes are marked with '+' and down-regulated with '-'.

	SCLC					
SCLC	TGFB1 *	-				
	PDGFRB	-				
	IFNB1	+				
	ANGPT4 *	+	NSCLC			
NSCLC	MDK	+	FOXC2	-		
	TIE1	+	SEMA3F	+		
			FLT1	+	Melanoma	
Melanoma	IL8	+	FGA *	-	CSF3	-
	TYMP	+	FGF1	-	EPHB2	-
	ITGAV	+			VASH1	-
					PDGFRA *	-
					ANGPT2	+
					SERPINF1	+
					ADAMTS1	+
					TIMP2	+
					NRP2	+
					ITGB3	+
					MMP2	+

Network analysis

Mapping the TaqMan assay gene set onto omicsnet revealed 4 genes (ANGPTL1, TNMD, ANGPTL2, PECAM1) which could not be mapped to omicsnet. After connecting the resulting subgraphs with the minimum spanning tree based approach 14 genes were added to the resulting singular subgraph (EP300, MYL4, RPP14, GFI1B, CAPZB, ATXN7, DMP1, HSPA5, MYOC, MYC, SLC9A1, HDAC1, HSP90AA1, MAGI3) and visualized (see **Error! Reference source not ound.** for Melanoma, **Error! Reference source not found.** for NSCL and for SCLC).

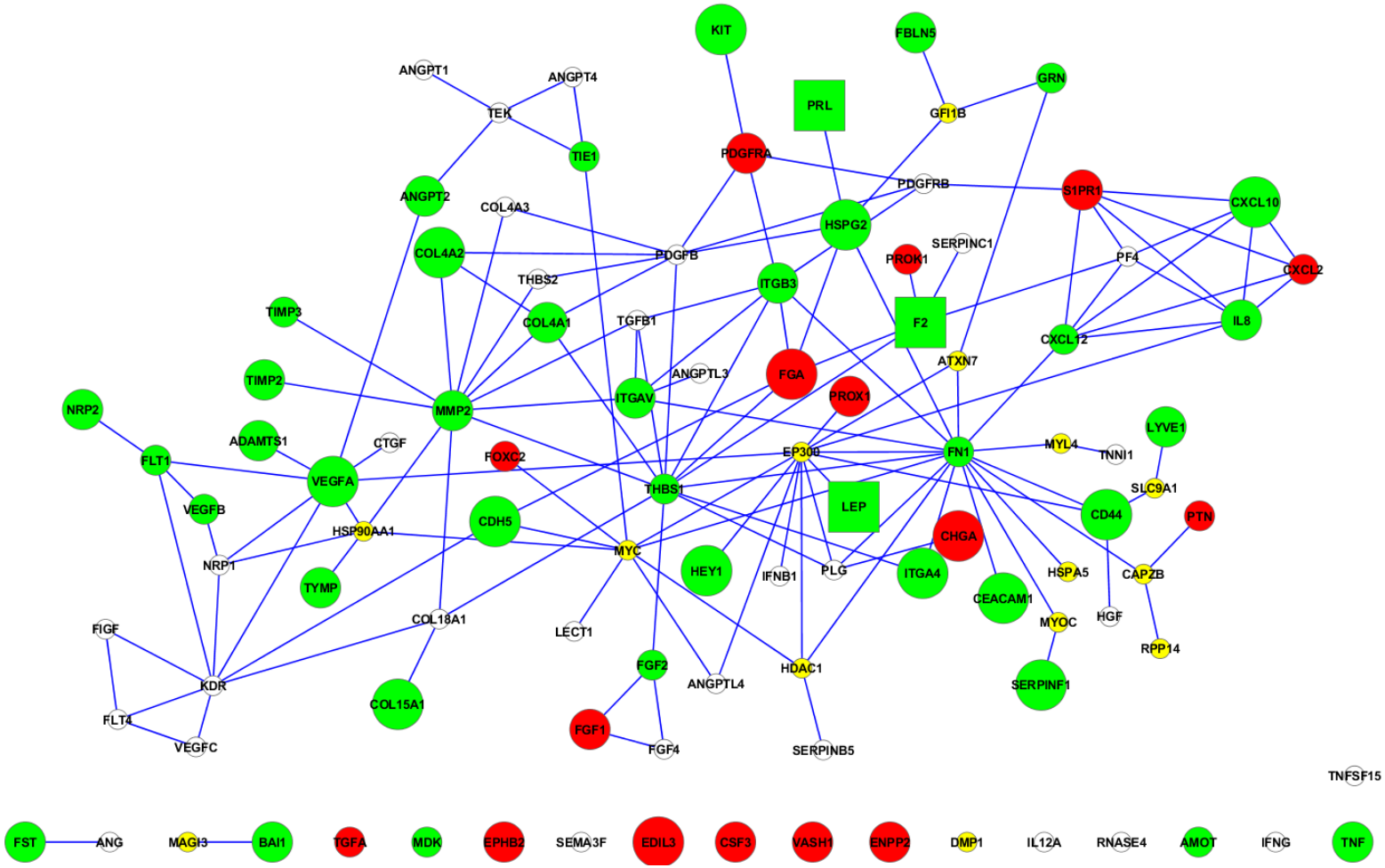


Figure 69: Network view of gene expression pattern of melanoma patient group mapped on omicsnet. Genes showing an increase of expression relative to the gliosis reference are held in green color, genes showing a decrease are held in red color and genes showing no difference are held in white color. The size of the nodes defines the amount of relative expression difference ranging from genes showing at least 50-fold difference (largest nodes), to 10-fold, 2-fold or no difference (smallest nodes). Genes having a missing value within the gliosis reference sample have a square form, all others a circle form.

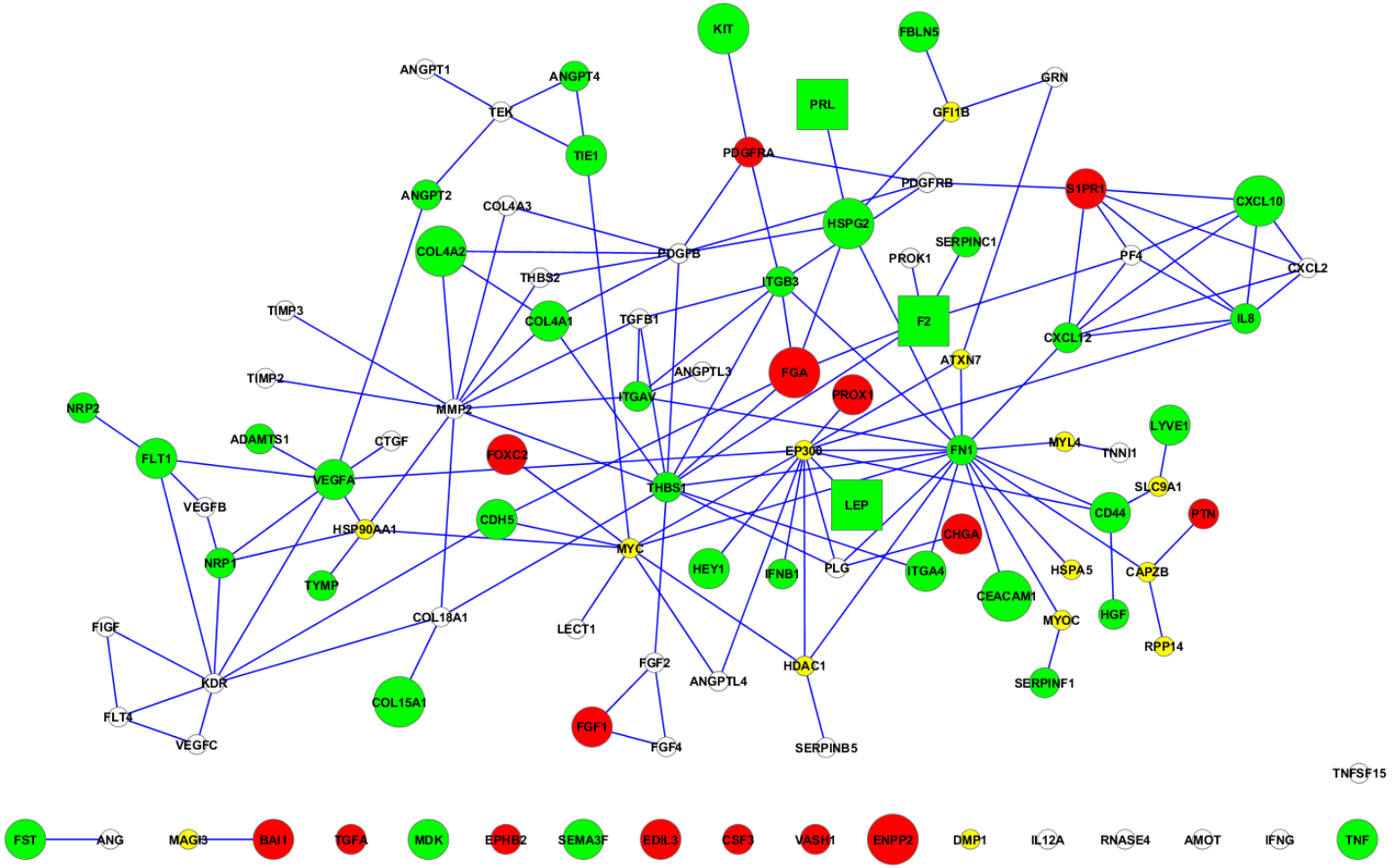


Figure 70: Network view of gene expression pattern of non-small cell lung cancer patient group mapped on omicsnet. Genes showing an increase of expression relative to the gliosis reference are held in green color, genes showing a decrease are held in red color and genes showing no difference are held in white color. The size of the nodes defines the amount of relative expression difference ranging from genes showing at least 50-fold difference (largest nodes), to 10-fold, 2-fold or no difference (smallest nodes). Genes having a missing value within the gliosis reference sample have a square form, all others a circle form.

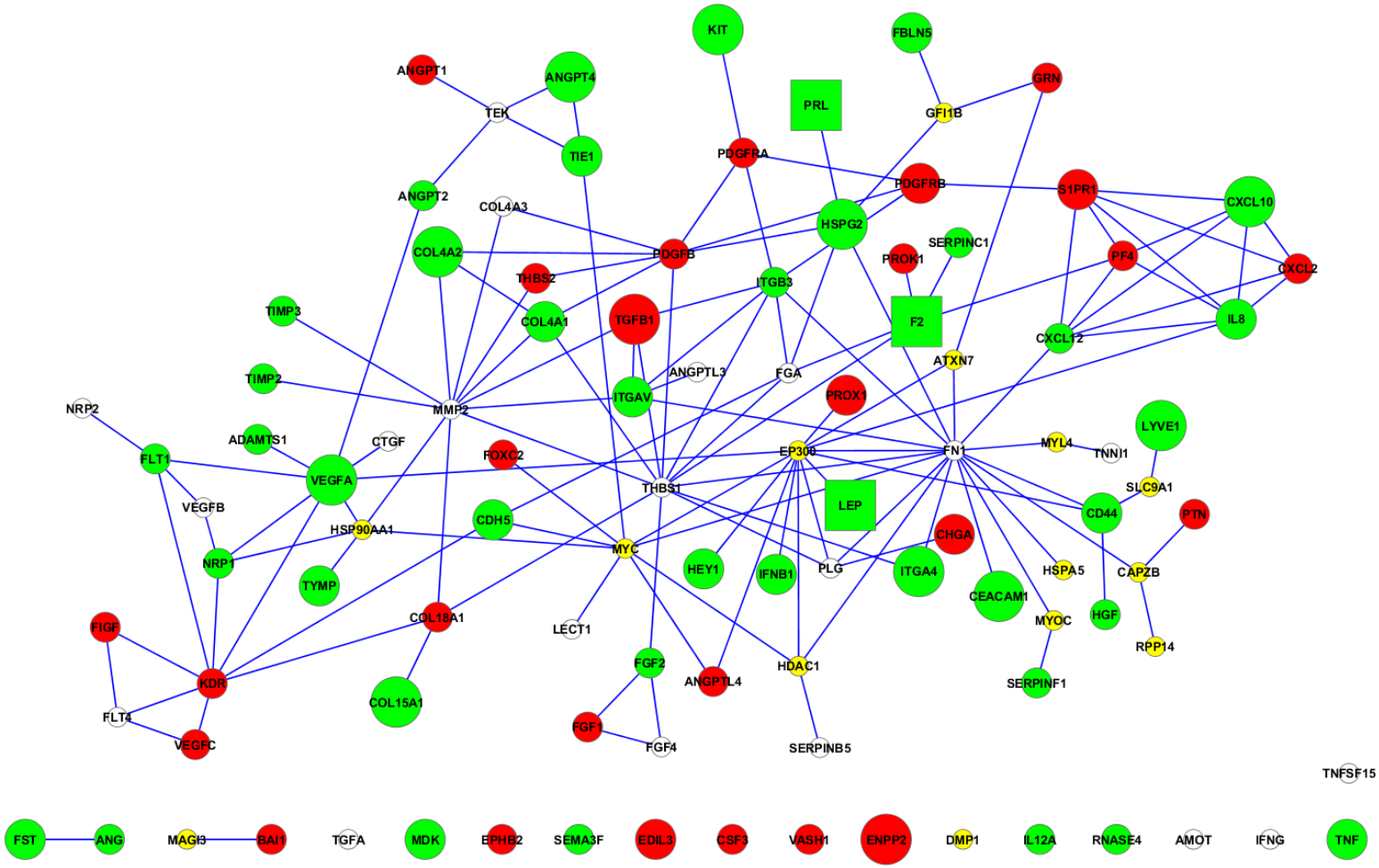


Figure 71: Network view of gene expression pattern of small cell lung cancer patient group mapped on omicsnet. Genes showing an increase of expression relative to the gliosis reference are held in green color, genes showing a decrease are held in red color and genes showing no difference are held in white color. The size of the nodes defines the amount of relative expression difference ranging from genes showing at least 50-fold difference (largest nodes), to 10-fold, 2-fold or no difference (smallest nodes). Genes having a missing value within the gliosis reference sample have a square form, all others a circle form.

Discussion

Sample collection included an epilepsy brain sample since it is not feasible to get healthy brain samples. We thus assume that a non cancerous sample would reflect the expression level of a healthy sample when compared to cancerous samples [462].

Data preprocessing revealed technical issues with the TaqMan assay resulting in an increase of missing values. 8 genes had to be removed from the data as well as 3 of the 4 housekeeping genes. We assumed that the issues were related to the primer within the chips since the overall outlier rate was very low. We observed only two values as outliers. We could not get a response on the possible cause from Applied Biosystems and got a refund of the costs for the assays.

For applying the normalization with the ΔΔCt method the amplification efficiency has to be approximately equal [461], which was not tested during the study (and seems to be frequently omitted in publications). Otherwise an absolute quantification method using standard curves is recommended.

Results from the unsupervised hierarchical clustering revealed an approximate grouping of the cancer groups suggesting a characteristic angiogenic signature of NSCLC, SCLC and melanoma. This grouping is not perfect though, which suggests either additional separation of angiogenic processes within the groups leading to further subgroups or a variance effect due to the rather low sample size per group. The rather good separation of the groups, even though there were technical issues, further indicates that these issues are most likely related to the missing values.

By identifying genes with expression patterns more than 50 and 10 times when compared to the reference, we are applying a sort of reliability filter and are thus focusing on genes which seem to have a very strong regulation pattern difference between the groups. While a weak change in regulation can be of relevance in a system as well, we were careful due to the technical issues we encountered with the assay. We also had to exclude genes where we had missing values within the reference, which usually led to such a 50 fold difference in expression patterns of those genes. Here we had to face the consequence of having only one sample as a reference.

Housekeeping genes are usually supposed to have a constant expression pattern throughout the tissues. We did notice a decrease of expression within SCLC which might indicate some biological effects on GUSB expression. This could be an artifact as well tough, due to the low sample size.

On the level of networks we expanded the genes from the chip and identified 14 further genes necessary to connect the chip gene set. These connections do not necessarily have to be related to angiogenesis yet we found that several of those like HSP90AA1, EP300 and MCY show a high degree of connections to angiogenesis related genes. This might indicate a functional relation to the angiogenesis processes. Staufer and Stoeltzing [463] discuss HSP90 as possible target for anti-angiogenic therapy. Zhang et al. [464] identified EP300 for being linked to angiogenesis. Baudino et al. [465] show that MYC is essential for vasculogenesis and angiogenesis during development and tumor progression. These studies demonstrate that the

gene set provided by the TaqMan pathway gene array is not covering all known angiogenesis related genes. This raises the question if the array is adequate at all for investigating angiogenesis specific genes since alternative approaches are necessary to complement the missing genes. This has an obvious impact on the study question. One approach to complement the current data would be to identify or propose further genes related to angiogenesis using network methods like module detection algorithms, as presented in chapter 1, or expansion algorithms as proposed previously in “Comparative analysis of expansion methods on protein interaction networks” and matching those additional findings against literature or ontologies, followed by a quantification of those specific genes in the cancer groups.

In summary the contribution to the projects question could demonstrate that there seems to be indeed a difference between brain metastases of different primary cancers on the level of angiogenic gene expression, but the picture is for sure not complete due to technical issues with the TaqMan array.

Conclusion

Molecular networks are complex and this dissertation work outlines many issues and challenges in the study of its field. Investigating complex networks and systems is already academically challenging and additionally molecular networks have to deal with highly heterogeneous and erroneous data adding another layer of difficult.

Dealing with the rather high error rates in biological data is an unavoidable challenge, as demonstrated throughout the presented scientific contributions. Consequently strategies have to be conceived to increase the reliability of analytical results using, where possible, multiple experimental setups or multiple analytical approaches and validation experiments. False negative rates from omics experiments, for example, can be successfully addressed with graph expansion algorithms as proposed in 'Comparative analysis of expansion methods on protein interaction networks' and exemplified within the 'Mesothelial cell stress response and cytoprotection in peritoneal dialysis ' project and the 'Angiogenesis of Brain Metastasis' project.

Dealing with heterogeneous biological (network) data is crucial for the analysis of biological networks and integration of such data aims at two aspects. The first aspect is providing information collected from many data sources in a redundancy free form [466]. This is especially difficult due to many different data sources and data representation standards, biological concepts and contradictory information [467]. Building representative models is necessary and has been exemplified in this work for a KEGG protein-protein interaction network model within the 'Basic network analysis software suite and KEGG interaction network modeling' project. Making sense of biological networks and systems will require scientists to include several cellular components, since regulation of the cellular system occurs at many levels [468]. The second major aspect of data integration is combining such different types of biological data for the analysis of biological networks and systems for inferring meaningful knowledge. Combining molecular data with clinical data opens additional possibilities in human health care as demonstrated recently by Soares et al. [469] for the epidemiology of tuberculosis. Integration of gene expression data, auto-antigenicity data, co-regulation data and pathway data was exemplified in the 'Ovarian cancer analysis' project. Proteomic, pathway and network data were combined in the 'Mesothelial cell stress response and cytoprotection in peritoneal dialysis' project. Gene expression data, pathway and network data were integrated in the 'Comparative analysis of expansion methods on protein interaction networks' project and gene expression data, network data and histological data were combined in the 'Angiogenesis in Brain Metastasis' project. Data integration is a current and future challenge for bioinformatic scientists on a conceptual as well as a practical scale. Many novel methodologies developed stem from data mining and knowledge management [5]. Several aspects of data integration, especially on the large amount of data, has led to novel "Big Data" strategies for working technologically and conceptually with such a huge amount of data [470]. Swarup and Geschwind illustrated the applicability of 'Big Data' in context of

Alzheimer disease [471]. The impact of advances of 'Big Data' to life sciences and health care data integration are outlined by Issa et al. [6].

The explorative analysis of biological networks has led to several findings regarding network structure, patterns and characteristics, like the modular structural organization of subnetworks and hubs [256] or the role of disease genes within the perimeter of functional modules [319]. The challenge lies in finding such biological organizational principles and patterns in context of phenotypes for furthering our understanding. The identification of those patterns has led to the development of network algorithms, like the search for modules within biological networks or graph expansion algorithms as presented in 'Comparative analysis of expansion methods on protein interaction networks'. Mapping principles of organization in biological networks on human network structures may provide possible solutions to the design of human networks. Kleinberg [472] discussed the small world phenomena of networks in the context of designing routing tables for communication and robot navigation. He states that the correlation between local structure and long-range connections provide fundamental clues to find efficiently paths through a network based on local data. If the network becomes too homogenous then the clues vanish and the goal cannot be found based on local data alone even though there are short routes available. Buldyrev et al. [473] for example investigated the cascade of failures in interdependent networks using an electrical blackout in Italy as reference. Simulations revealed that a higher degree distribution between networks actually leads to a higher probability of cascading failures, which is the opposite of how singular networks behave. A hierarchical organization of networks, as we find in biological networks, might lead to an increase of network robustness. Czaplicka et al. [474] investigated the influence of noise on information transmission based on packages shipped between nodes within networks. Interestingly they found out that noise can enhance the information transfer and that hierarchical networks perform best. Additional exploration of biological network topologies and network properties will further our understanding and assist in constructing other networks.

And finally the question of how to analyze complex biological networks and systems has led to a dispute within the scientific community. Reductionism, as a scientific methodology, while being superb for describing the building blocks of a system, has shown its limitations when describing emergent properties of systems. The complexity of system patterns makes the testing of hypotheses on the system level challenging, if not impossible. Consequently the need for alternative scientific methodologies is growing within the field of network analysis following an inductive and holistic philosophy [13]. We see for example a shift from hypothesis driven research to data driven research [15], which could be a new computerized way of inferring knowledge on a large data-based scale. The reductionist way of thinking might have clouded our view for the 'big picture', nevertheless it will remain a crucial scientific methodology yet the study of complex systems requires new scientific concepts, which can complement the classical approach as proposed by several authors [475], [476], [477].

Bibliography

[1] M. E. J. Newman, *Networks: an introduction*. Oxford ; New York: Oxford University Press, 2010.

[2] S. N. Dorogovt͡sev, *Evolution of networks: from biological nets to the Internet and WWW*, Paperback edition. Oxford ; New York: Oxford University Press, 2013.

[3] F. S. Collins, M. Morgan, and A. Patrinos, "The Human Genome Project: Lessons from Large-Scale Biology," *Science*, vol. 300, no. 5617, pp. 286–290, Apr. 2003.

[4] "Pathguide: the pathway resource list." [Online]. Available: http://pathguide.org/statistics.php. [Accessed: 17-Aug-2014].

[5] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegner, "Data integration in the era of omics: current and future challenges," *BMC Systems Biology*, vol. 8, no. Suppl 2, p. I1, 2014.

[6] N. T. Issa, S. W. Byers, and S. Dakshanamurthy, "Big data: the next frontier for innovation in therapeutics and healthcare," *Expert Review of Clinical Pharmacology*, vol. 7, no. 3, pp. 293–298, May 2014.

[7] N. F. Johnson, Neil F Johnson, *Simply complexity a clear guide to complexity theory*. 2009.

[8] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[9] E. N. Lorenz, "Deterministic Nonperiodic Flow," *J. Atmos. Sci.*, vol. 20, no. 2, pp. 130–141, Mar. 1963.

[10] S. Weinberg, *Dreams of a final theory*, 1st Vintage Books ed. New York: Vintage Books, 1994.

[11] K. L. Bertalanffy, *Kritische Theorie der Formbildung*. Berlin: Gebrüder Borntraeger, 1928.

[12] K. L. Bertalanffy, *Modern Theories of Development: An Introduction to Theoretical Biology*. London: Oxford University Press, 1933.

[13] S. A. Kauffman, *Reinventing the sacred: a new view of science, reason and religion*. New York: Basic Books, 2008.

[14] R. Gallagher, "Beyond Reductionism," *Science*, vol. 284, no. 5411, pp. 79–79, Apr. 1999.

[15] S. Leonelli, "Introduction: Making sense of data-driven research in the biological and biomedical sciences," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 43, no. 1, pp. 1–3, Mar. 2012.

[16] R. Bellman, "On a routing problem," *Quarterly of Applied Mathematics*, vol. 16, pp. 87–90, 1958.

[17] L. R. Ford Jr., "Network Flow Theory," *RAND Corporation*, 1956.

[18] J. B. Kruskal, "On the shortest spanning subtree and the traveling salesman problem," *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[19] R. C. Prim, "Shortest Connection Networks And Some Generalizations," *Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, Nov. 1957.

[20] L. Euler, “Solutio problematis ad geometriam situs pertinentis,” *Commentarii academiae scientiarum Petropolitanae*, vol. 8, pp. 128–140, 1741.
[21] A. Platzer, P. Perco, A. Lukas, and B. Mayer, “Characterization of protein-interaction networks in tumors,” *BMC Bioinformatics*, vol. 8, p. 224, 2007.
[22] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 47–97, Jan. 2002.
[23] *Network Science*. Washington, DC: The National Academies Press, 2005.
[24] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi, “The human disease network,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, May 2007.
[25] M. Newman, “Spread of epidemic disease on networks,” *Physical Review E*, vol. 66, no. 1, Jul. 2002.
[26] T. Kamada, “System biomedicine: a new paradigm in biomedical engineering,” *Front Med Biol Eng*, vol. 4, no. 1, pp. 1–2, 1992.
[27] “Systems Medicine - Large-scale data gathering and Systems Medicine - Health - Research & Innovation - European Commission.” [Online]. Available: http://ec.europa.eu/research/health/large-scale/systems-medicine/index_en.html. [Accessed: 11-Aug-2014].
[28] “CASYM Europe: Casym.” [Online]. Available: https://www.casym.eu/casym. [Accessed: 11-Aug-2014].
[29] A. S. Dhillon, S. Hagan, O. Rath, and W. Kolch, “MAP kinase signalling pathways in cancer,” *Oncogene*, vol. 26, no. 22, pp. 3279–3290, May 2007.
[30] M. F. van Delft and D. C. S. Huang, “How the Bcl-2 family of proteins interact to regulate apoptosis,” *Cell Res.*, vol. 16, no. 2, pp. 203–213, Feb. 2006.
[31] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, “Lethality and centrality in protein networks,” *Nature*, vol. 411, no. 6833, pp. 41–42, May 2001.
[32] F. Pazos, J. A. G. Ranea, D. Juan, and M. J. E. Sternberg, “Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome,” *J. Mol. Biol.*, vol. 352, no. 4, pp. 1002–1015, Sep. 2005.
[33] T. Ideker and R. Sharan, “Protein networks in disease,” *Genome Research*, vol. 18, no. 4, pp. 644–652, Apr. 2008.
[34] M. W. Gonzalez and M. G. Kann, “Chapter 4: Protein Interactions and Disease,” *PLoS Computational Biology*, vol. 8, no. 12, p. e1002819, Dec. 2012.
[35] P. F. Jonsson and P. A. Bates, “Global topological features of cancer proteins in the human interactome,” *Bioinformatics*, vol. 22, no. 18, pp. 2291–2297, Sep. 2006.
[36] H. Paik, H.-S. Heo, H. Ban, and S. Cho, “Unraveling human protein interaction networks underlying co-occurrences of diseases and pathological conditions,” *Journal of Translational Medicine*, vol. 12, no. 1, p. 99, 2014.
[37] E. Golemis and P. D. Adams, Eds., *Protein-protein interactions: a molecular cloning manual*, 2nd ed. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press, 2005.
[38] S. Fields and O. Song, “A novel genetic system to detect protein-protein interactions,” *Nature*, vol. 340, no. 6230, pp. 245–246, Jul. 1989.
[39] L. M. Brettner and J. Masel, “Protein stickiness, rather than number of functional protein-protein interactions, predicts expression noise and plasticity in yeast,” *BMC Systems Biology*, vol. 6, no. 1, p. 128, 2012.
[40] D. A. Hall, J. Ptacek, and M. Snyder, “Protein microarray technology,” *Mechanisms of Ageing and Development*, vol. 128, no. 1, pp. 161–167, Jan. 2007.

[41] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin, “A generic protein purification method for protein complex characterization and proteome exploration,” *Nat. Biotechnol.*, vol. 17, no. 10, pp. 1030–1032, Oct. 1999.
[42] T. K. Kerppola, “Design and implementation of bimolecular fluorescence complementation (BiFC) assays for the visualization of protein interactions in living cells,” *Nat Protoc*, vol. 1, no. 3, pp. 1278–1286, 2006.
[43] M. Tyagi, K. Hashimoto, B. A. Shoemaker, S. Wuchty, and A. R. Panchenko, “Large-scale mapping of human protein interactome using structural complexes,” *EMBO Rep.*, vol. 13, no. 3, pp. 266–271, Mar. 2012.
[44] H. Chen and B. M. Sharp, “Content-rich biological network constructed by mining PubMed abstracts,” *BMC Bioinformatics*, vol. 5, p. 147, Oct. 2004.
[45] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia, “Overview of the protein-protein interaction annotation extraction task of BioCreative II,” *Genome Biology*, vol. 9, no. Suppl 2, p. S4, 2008.
[46] Y. Niu, D. Otasek, and I. Jurisica, “Evaluation of linguistic features useful in extraction of interactions from PubMed; Application to annotating known, high-throughput and predicted interactions in I2D,” *Bioinformatics*, vol. 26, no. 1, pp. 111–119, Jan. 2010.
[47] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, “The large-scale organization of metabolic networks,” *Nature*, vol. 407, pp. 651–654, 2000.
[48] D. A. Fell and A. Wagner, “The small world of metabolism,” *Nat Biotech*, vol. 18, no. 11, pp. 1121–1122, Nov. 2000.
[49] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles, “Metabolic network structure determines key aspects of functionality and regulation,” *Nature*, vol. 420, no. 6912, pp. 190–193, Nov. 2002.
[50] T. I. Lee, “Transcriptional Regulatory Networks in Saccharomyces cerevisiae,” *Science*, vol. 298, no. 5594, pp. 799–804, Oct. 2002.
[51] I. Farkas, H. Jeong, T. Vicsek, A.-L. Barabási, and Z. N. Oltvai, “The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiae,” *Physica A: Statistical Mechanics and its Applications*, vol. 318, no. 3–4, pp. 601–612, Feb. 2003.
[52] N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes, “Topological and causal structure of the yeast transcriptional regulatory network,” *Nat Genet*, vol. 31, no. 1, pp. 60–63, May 2002.
[53] A. Blais, “Constructing transcriptional regulatory networks,” *Genes & Development*, vol. 19, no. 13, pp. 1499–1511, Jul. 2005.
[54] G. Michailidis, “Statistical Challenges in Biological Networks,” *Journal of Computational and Graphical Statistics*, vol. 21, no. 4, pp. 840–855, Oct. 2012.
[55] P. Perco, A. Kainz, G. Mayer, A. Lukas, R. Oberbauer, and B. Mayer, “Detection of coregulation in differential gene expression profiles,” *BioSystems*, vol. 82, no. 3, pp. 235–247, Dec. 2005.
[56] M. A. Beer and S. Tavazoie, “Predicting Gene Expression from Sequence,” *Cell*, vol. 117, no. 2, pp. 185–198.
[57] “D3c3 - Dream Initiative.” [Online]. Available: http://wiki.c2b2.columbia.edu/dream/index.php/D3c3. [Accessed: 14-Aug-2014].
[58] M. Gustafsson and M. Hörnquist, “Gene Expression Prediction by Soft Integration and the Elastic Net—Best Performance of the DREAM3 Gene Expression Challenge,” *PLoS ONE*, vol. 5, no. 2, p. e9134, Feb. 2010.
[59] V. Ambros, “The functions of animal microRNAs,” *Nature*, vol. 431, no. 7006, pp. 350–355, Sep. 2004.

[60] D. P. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function,” *Cell*, vol. 116, no. 2, pp. 281–297, Jan. 2004.
[61] L. de Pontual, E. Yao, P. Callier, L. Faivre, V. Drouin, S. Cariou, A. Van Haeringen, D. Geneviève, A. Goldenberg, M. Oufadem, S. Manouvrier, A. Munnich, J. A. Vidigal, M. Vekemans, S. Lyonnet, A. Henrion-Caude, A. Ventura, and J. Amiel, “Germline deletion of the miR-17~92 cluster causes skeletal and growth defects in humans,” *Nat. Genet.*, vol. 43, no. 10, pp. 1026–1030, Oct. 2011.
[62] P. Akçakaya, S. Ekelund, I. Kolosenko, S. Caramuta, D. M. Ozata, H. Xie, U. Lindforss, H. Olivecrona, and W.-O. Lui, “miR-185 and miR-133b deregulation is associated with overall survival and metastasis in colorectal cancer,” *Int. J. Oncol.*, vol. 39, no. 2, pp. 311–318, Aug. 2011.
[63] A. Tessitore, G. Cicciarelli, F. Del Vecchio, A. Gaggiano, D. Verzella, M. Fischietti, D. Vecchiotti, D. Capece, F. Zazzeroni, and E. Alesse, “MicroRNAs in the DNA Damage/Repair Network and Cancer,” *Int J Genomics*, vol. 2014, p. 820248, 2014.
[64] J. A. Nielsen, P. Lau, D. Maric, J. L. Barker, and L. D. Hudson, “Integrating microRNA and mRNA expression profiles of neuronal progenitors to identify regulatory networks underlying the onset of cortical neurogenesis,” *BMC Neurosci*, vol. 10, p. 98, 2009.
[65] A. Gupta, P. Nagilla, H.-S. Le, C. Bunney, C. Zych, A. Thalamuthu, Z. Bar-Joseph, S. Mathavan, and V. Ayyavoo, “Comparative expression profile of miRNA and mRNA in primary peripheral blood mononuclear cells infected with human immunodeficiency virus (HIV-1),” *PLoS ONE*, vol. 6, no. 7, p. e22730, 2011.
[66] A. Helwak and D. Tollervey, “Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH),” *Nat. Protocols*, vol. 9, no. 3, pp. 711–728, Mar. 2014.
[67] M. Maragkakis, P. Alexiou, G. Papadopoulos, M. Reczko, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, V. Simossis, P. Sethupathy, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. Hatzigeorgiou, “Accurate microRNA target prediction correlates with protein repression levels,” *BMC Bioinformatics*, vol. 10, no. 1, p. 295, 2009.
[68] A. M. McDermott, N. Miller, D. Wall, L. M. Martyn, G. Ball, K. J. Sweeney, and M. J. Kerin, “Identification and Validation of Oncologic miRNA Biomarkers for Luminal A-like Breast Cancer,” *PLoS ONE*, vol. 9, no. 1, p. e87032, Jan. 2014.
[69] S. H. Cheng, R. J. Gregory, J. Marshall, S. Paul, D. W. Souza, G. A. White, C. R. O’Riordan, and A. E. Smith, “Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis,” *Cell*, vol. 63, no. 4, pp. 827–834, Nov. 1990.
[70] G. E. Cooke, “Pharmacogenetics of multigenic disease: Heart disease as an example,” *Vascular Pharmacology*, vol. 44, no. 2, pp. 66–74, Feb. 2006.
[71] W. G. Kaelin, “The Concept of Synthetic Lethality in the Context of Anticancer Therapy,” *Nat Rev Cancer*, vol. 5, no. 9, pp. 689–698, Sep. 2005.
[72] J. Söllner, P. Mayer, A. Heinzel, R. Fechete, C. Siehs, R. Oberbauer, and B. Mayer, “Synthetic lethality for linking the mycophenolate mofetil mode of action with molecular disease and drug profiles,” *Molecular BioSystems*, vol. 8, no. 12, p. 3197, 2012.
[73] K. N.-S. Margaret Scheibner, Kara Padget, Michelle Irvine, David A. Sliwinski, Tomasz Haas, Kimberly Lee, Jaewoong Geng, Huimin Roy, Darshan Slupianek, Artur Rassool, Feyruz V. Wasik, Mariusz A. Childers, Wayne Copland, Mhairi Müschen, Markus Civin, Curt I. Skorski, Tomasz Cramer-Morales, “Personalized synthetic lethality induced by targeting RAD52 in leukemias identified by gene mutation and expression profile,” *Blood*, vol. 122, no. 7, pp. 1293–1304, Jul. 2013.

[74] A. Mora, K. Michalickova, and I. Donaldson, “A survey of protein interaction data and multigenic inherited disorders,” *BMC Bioinformatics*, vol. 14, no. 1, p. 47, 2013.
[75] J. N. Hirschhorn, “Genomewide association studies--illuminating biologic pathways,” *N. Engl. J. Med.*, vol. 360, no. 17, pp. 1699–1701, Apr. 2009.
[76] C. Ding and S. Jin, “High-throughput methods for SNP genotyping,” *Methods Mol. Biol.*, vol. 578, pp. 245–254, 2009.
[77] C. L. Koo, M. J. Liew, M. S. Mohamad, and A. H. Mohamed Salleh, “A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology,” *BioMed Research International*, vol. 2013, pp. 1–13, 2013.
[78] L. Zhang and S. Kim, “Learning Gene Networks under SNP Perturbations Using eQTL Datasets,” *PLoS Computational Biology*, vol. 10, no. 2, p. e1003420, Feb. 2014.
[79] S. L. Wong, L. V. Zhang, A. H. Y. Tong, Z. Li, D. S. Goldberg, O. D. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey, C. Boone, and F. P. Roth, “Combining biological networks to predict genetic interactions,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 44, pp. 15682–15687, Nov. 2004.
[80] J. Knowles and G. Gromo, “A guide to drug discovery: Target selection in drug discovery,” *Nat Rev Drug Discov*, vol. 2, no. 1, pp. 63–69, Jan. 2003.
[81] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, “DrugBank 4.0: shedding new light on drug metabolism,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D1091–D1097, Jan. 2014.
[82] R. Jafari, H. Almqvist, H. Axelsson, M. Ignatushchenko, T. Lundbäck, P. Nordlund, and D. M. Molina, “The cellular thermal shift assay for evaluating drug target interactions in cells,” *Nature Protocols*, vol. 9, no. 9, pp. 2100–2122, Aug. 2014.
[83] T. Takenaka, “Classical vs reverse pharmacology in drug discovery,” *BJU Int.*, vol. 88 Suppl 2, pp. 7–10; discussion 49–50, Sep. 2001.
[84] J. A. Lee, M. T. Uhlik, C. M. Moxham, D. Tomandl, and D. J. Sall, “Modern phenotypic drug discovery is a viable, neoclassic pharma strategy,” *J. Med. Chem.*, vol. 55, no. 10, pp. 4527–4538, May 2012.
[85] A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg, and E. S. Huang, “Structure-based maximal affinity model predicts small-molecule druggability,” *Nat. Biotechnol.*, vol. 25, no. 1, pp. 71–75, Jan. 2007.
[86] S. Zhu, Y. Okuno, G. Tsujimoto, and H. Mamitsuka, “A probabilistic model for mining implicit ‘chemical compound-gene’ relations from literature,” *Bioinformatics*, vol. 21 Suppl 2, pp. ii245–251, Sep. 2005.
[87] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, “Prediction of drug–target interaction networks from the integration of chemical and genomic spaces,” *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, Jul. 2008.
[88] N. Nagamine, T. Shirakawa, Y. Minato, K. Torii, H. Kobayashi, M. Imoto, and Y. Sakakibara, “Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening,” *PLoS Comput. Biol.*, vol. 5, no. 6, p. e1000397, Jun. 2009.
[89] Z. He, J. Zhang, X.-H. Shi, L.-L. Hu, X. Kong, Y.-D. Cai, and K.-C. Chou, “Predicting Drug-Target Interaction Networks Based on Functional Groups and Biological Features,” *PLoS ONE*, vol. 5, no. 3, p. e9603, Mar. 2010.

[90] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “Data, information, knowledge and principle: back to metabolism in KEGG,” *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D199–205, Jan. 2014.
[91] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders,” *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D514–D517, Jan. 2005.
[92] C. Prieto and J. De Las Rivas, “APID: Agile Protein Interaction DataAnalyzer,” *Nucleic Acids Research*, vol. 34, no. suppl 2, pp. W298–W302, Jul. 2006.
[93] G. D. Bader, D. Betel, and C. W. V. Hogue, “BIND: the Biomolecular Interaction Network Database,” *Nucleic Acids Research*, vol. 31, no. 1, pp. 248–250, Jan. 2003.
[94] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “BioGRID: a general repository for interaction datasets,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D535–539, Jan. 2006.
[95] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, “DIP: the Database of Interacting Proteins,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 289–291, Jan. 2000.
[96] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, “Human Protein Reference Database—2009 update,” *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D767–D772, Jan. 2009.
[97] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob, “The IntAct molecular interaction database in 2012,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D841–D846, Jan. 2012.
[98] A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni, “MINT: the Molecular INTeraction database,” *Nucleic Acids Research*, vol. 35, no. suppl 1, pp. D572–D574, Jan. 2007.
[99] J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø. Palsson, “BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions,” *BMC Bioinformatics*, vol. 11, no. 1, p. 213, 2010.
[100] M. M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.
[101] NCBI Resource Coordinators, “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Research*, vol. 41, no. D1, pp. D8–D20, Jan. 2013.
[102] “ASN.1 Project.” [Online]. Available: http://www.itu.int/en/ITU-T/asn1/Pages/asn1_project.aspx. [Accessed: 11-Aug-2014].
[103] “IEB homepage - NCBI.” [Online]. Available: http://www.ncbi.nlm.nih.gov/IEB/. [Accessed: 11-Aug-2014].
[104] D. Nishimura, “BioCarta,” *Biotech Software & Internet Report*, vol. 2, no. 3, pp. 117–120, Jun. 2001.
[105] “BioCarta.” [Online]. Available: http://www.biocarta.com/. [Accessed: 11-Aug-2014].

[106] R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, and P. D. Karp, "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D623–D631, Jan. 2008.

[107] "BioCyc | Pathway/Genome Databases and Pathway Tools Software." [Online]. Available: http://biocyc.org/. [Accessed: 11-Aug-2014].

[108] "BioGRID | Database of Protein and Genetic Interactions." [Online]. Available: http://thebiogrid.org/. [Accessed: 11-Aug-2014].

[109] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. N. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler, "The HUPO PSI's Molecular Interaction format - a community standard for the representation of protein interaction data," *Nat Biotech*, vol. 22, no. 2, pp. 177–183, Feb. 2004.

[110] "PHP: Hypertext Preprocessor." [Online]. Available: http://php.net/. [Accessed: 27-Aug-2014].

[111] "Welcome to The Apache Software Foundation!" [Online]. Available: http://www.apache.org/. [Accessed: 27-Aug-2014].

[112] "MySQL :: The world's most popular open source database." [Online]. Available: http://www.mysql.com/. [Accessed: 27-Aug-2014].

[113] "Java - Programming language." [Online]. Available: https://www.java.com/. [Accessed: 27-Aug-2014].

[114] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig, "ConsensusPathDB—a database for integrating human functional interaction networks," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D623–D628, Jan. 2009.

[115] A. Kamburov, U. Stelzl, H. Lehrach, and R. Herwig, "The ConsensusPathDB interaction database: 2013 update," *Nucleic Acids Research*, vol. 41, no. D1, pp. D793–D800, Jan. 2013.

[116] "MIF 2.5.0 Specification | HUPO Proteomics Standards Initiative." [Online]. Available: http://www.psidev.info/mif. [Accessed: 12-Aug-2014].

[117] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. C. Lopez-Fuentes, H. Mi, E. Pichler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, M. Syed, N. Anwar, O. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, A. Luna, P. Murray-Rust, E. Neumann, O. Ruebenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, B. Braun, M. Whirl-Carrillo, K.-H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, D. Kane, S. Krupa, M. Kutmon, J. Leonard, D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, M. Sunshine, R. Tang, R. Whaley, S. Letovksy, K. H. Buetow, A. Rzhetsky, V. Schachter, B. S. Sobral, U. Dogrusoz, S. McWeeney, M. Aladjem, E. Birney, J. Collado-Vides, S. Goto, M. Hucka, N. Le Novere, N. Maltsev, A. Pandey, P. Thomas, E. Wingender, P. D. Karp, C. Sander, and G. D. Bader, "The BioPAX community standard for pathway data sharing," *Nat Biotech*, vol. 28, no. 9, pp. 935–942, Sep. 2010.

[118] "Cytoscape.js." [Online]. Available: http://js.cytoscape.org/. [Accessed: 12-Aug-2014].

[119] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, “Cytoscape 2.8: new features for data integration and network visualization,” *Bioinformatics*, vol. 27, no. 3, pp. 431–432, Feb. 2011.

[120] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, “The Database of Interacting Proteins: 2004 update,” *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D449–D451, Jan. 2004.

[121] “DIP::Software.” [Online]. Available: http://dip.doe-mbi.ucla.edu/dip/Software1.cgi. [Accessed: 12-Aug-2014].

[122] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. A. Mayer, F. Sanz, and L. I. Furlong, “Gene-Disease Network Analysis Reveals Functional Modules in Mendelian, Complex and Environmental Diseases,” *PLoS ONE*, vol. 6, no. 6, p. e20284, Jun. 2011.

[123] A. Bauer-Mehren, M. Rautschka, F. Sanz, and L. I. Furlong, “DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene–disease networks,” *Bioinformatics*, vol. 26, no. 22, pp. 2924–2926, Nov. 2010.

[124] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, “DrugBank 3.0: a comprehensive resource for ‘Omics’ research on drugs,” *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D1035–D1041, Jan. 2011.

[125] J. Chen, S. Mamidipalli, and T. Huan, “HAPPI: an online database of comprehensive human annotated and predicted protein interactions,” *BMC Genomics*, vol. 10, no. Suppl 1, p. S16, 2009.

[126] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen, “STRING v9.1: protein-protein interaction networks, with increased coverage and integration,” *Nucleic Acids Research*, vol. 41, no. D1, pp. D808–D815, Jan. 2013.

[127] B. Snel, G. Lehmann, P. Bork, and M. A. Huynen, “STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene,” *Nucleic Acids Research*, vol. 28, no. 18, pp. 3442–3444, Sep. 2000.

[128] K. R. Brown and I. Jurisica, “Online predicted human interaction database,” *Bioinformatics*, vol. 21, no. 9, pp. 2076–2082, May 2005.

[129] “Database Software | Oracle.” [Online]. Available: http://www.oracle.com/us/products/database/overview/index.html. [Accessed: 27-Aug-2014].

[130] “The Perl Programming Language - www.perl.org.” [Online]. Available: http://www.perl.org/. [Accessed: 27-Aug-2014].

[131] A. Patil, K. Nakai, and H. Nakamura, “HitPredict: a database of quality assessed protein–protein interactions in nine species,” *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D744–D749, Jan. 2011.

[132] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, “HMDD v2.0: a database for experimentally supported human microRNA and disease associations,” *Nucleic Acids Research*, Nov. 2013.

[133] “Human Protein Reference Database.” [Online]. Available: http://www.hprd.org/. [Accessed: 11-Aug-2014].

[134] K. Kandasamy, S. Keerthikumar, R. Goel, S. Mathivanan, N. Patankar, B. Shafreen, S. Renuse, H. Pawar, Y. L. Ramachandra, P. K. Acharya, P. Ranganathan, R. Chaerkady, T. S. Keshava Prasad, and A. Pandey, “Human Proteinpedia: a unified discovery resource for

proteomics research," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D773–D781, Jan. 2009.

[135] S. Mathivanan, M. Ahmed, N. G. Ahn, H. Alexandre, R. Amanchy, P. C. Andrews, J. S. Bader, B. M. Balgley, M. Bantscheff, K. L. Bennett, E. Bjorling, B. Blagoev, R. Bose, S. K. Brahmachari, A. S. Burlingame, X. R. Bustelo, G. Cagney, G. T. Cantin, H. L. Cardasis, J. E. Celis, R. Chaerkady, F. Chu, P. A. Cole, C. E. Costello, R. J. Cotter, D. Crockett, J. P. DeLany, A. M. De Marzo, L. V. DeSouza, E. W. Deutsch, E. Dransfield, G. Drewes, A. Droit, M. J. Dunn, K. Elenitoba-Johnson, R. M. Ewing, J. V. Eyk, V. Faca, J. Falkner, X. Fang, C. Fenselau, D. Figeys, P. Gagne, C. Gelfi, K. Gevaert, J. M. Gimble, F. Gnad, R. Goel, P. Gromov, S. M. Hanash, W. S. Hancock, H. Harsha, G. Hart, F. Hays, F. He, P. Hebbar, K. Helsens, H. Hermeking, W. Hide, K. Hjerno, D. F. Hochstrasser, O. Hofmann, D. M. Horn, R. H. Hruban, N. Ibarrola, P. James, O. N. Jensen, P. H. Jensen, P. Jung, K. Kandasamy, I. Kheterpal, R. F. Kikuno, U. Korf, R. Korner, B. Kuster, M.-S. Kwon, H.-J. Lee, Y.-J. Lee, M. Lefevre, M. Lehvaslaiho, P. Lescuyer, F. Levander, M. S. Lim, C. Lobke, J. A. Loo, M. Mann, L. Martens, J. Martinez-Heredia, M. McComb, J. McRedmond, A. Mehrle, R. Menon, C. A. Miller, H. Mischak, S. S. Mohan, R. Mohmood, H. Molina, M. F. Moran, J. D. Morgan, R. Moritz, M. Morzel, D. C. Muddiman, A. Nalli, J. D. Navarro, T. A. Neubert, O. Ohara, R. Oliva, G. S. Omenn, M. Oyama, Y.-K. Paik, K. Pennington, R. Pepperkok, B. Periaswamy, E. F. Petricoin, G. G. Poirier, T. S. K. Prasad, S. O. Purvine, B. A. Rahiman, P. Ramachandran, Y. L. Ramachandra, R. H. Rice, J. Rick, R. H. Ronnholm, J. Salonen, J.-C. Sanchez, T. Sayd, B. Seshi, K. Shankari, S. J. Sheng, V. Shetty, K. Shivakumar, R. J. Simpson, R. Sirdeshmukh, K. W. Michael Siu, J. C. Smith, R. D. Smith, D. J. States, S. Sugano, M. Sullivan, G. Superti-Furga, M. Takatalo, V. Thongboonkerd, J. C. Trinidad, M. Uhlen, J. Vandekerckhove, J. Vasilescu, T. D. Veenstra, J.-M. Vidal-Taboada, M. Vihinen, R. Wait, X. Wang, S. Wiemann, B. Wu, T. Xu, J. R. Yates, J. Zhong, M. Zhou, Y. Zhu, P. Zurbig, and A. Pandey, "Human Proteinpedia enables sharing of human protein data," *Nat Biotech*, vol. 26, no. 2, pp. 164–167, Feb. 2008.

[136] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, and the rest of the SBML Forum:, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, Mar. 2003.

[137] R. Hoffmann and A. Valencia, "Implementing the iHOP concept for navigation of biomedical literature," *Bioinformatics*, vol. 21, no. suppl 2, pp. ii252–ii258, Jan. 2005.

[138] S. Balaji, C. Mcclendon, R. Chowdhary, J. S. Liu, and J. Zhang, "IMID: integrated molecular interaction database," *Bioinformatics*, vol. 28, no. 5, pp. 747–749, Mar. 2012.

[139] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D428–D432, Jan. 2005.

[140] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, "PID: the Pathway Interaction Database," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D674–D679, Jan. 2009.

[141] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000.

[142] "IMID-Web." [Online]. Available: http://integrativebiology.org/. [Accessed: 13-Aug-2014].

[143] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler, "IntAct: an open source molecular interaction database," *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D452–D455, Jan. 2004.

[144] "IntAct." [Online]. Available: http://www.ebi.ac.uk/intact/. [Accessed: 11-Mar-2014].

[145] D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 28, no. 1, pp. 10–14, Jan. 2000.

[146] "IntAct statistics." [Online]. Available: http://www.ebi.ac.uk/intact/pages/documentation/statistics.xhtml#tc01. [Accessed: 11-Aug-2014].

[147] "PostgreSQL: The world's most advanced open source database." [Online]. Available: http://www.postgresql.org/. [Accessed: 27-Aug-2014].

[148] "Welcome to the Apache Struts project." [Online]. Available: http://struts.apache.org/. [Accessed: 27-Aug-2014].

[149] "Data Visualization Software | Tulip." [Online]. Available: http://tulip.labri.fr/TulipDrupal/. [Accessed: 27-Aug-2014].

[150] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, vol. 40, no. D1, pp. D109–D114, Jan. 2012.

[151] V. G. Tarcea, T. Weymouth, A. Ade, A. Bookvich, J. Gao, V. Mahavisno, Z. Wright, A. Chapman, M. Jayapandian, A. Özgür, Y. Tian, J. Cavalcoli, B. Mirel, J. Patel, D. Radev, B. Athey, D. States, and H. V. Jagadish, "Michigan molecular interactions r2: from interacting proteins to pathways," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D642–D646, Jan. 2009.

[152] T. G. O. Consortium, "Creating the Gene Ontology Resource: Design and Implementation," *Genome Research*, vol. 11, no. 8, pp. 1425–1433, Aug. 2001.

[153] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A. N. Nikolskaya, S. Orchard, M. Pagni, C. P. Ponting, E. Quevillon, J. Selengut, C. J. A. Sigrist, V. Silventoinen, D. J. Studholme, R. Vaughan, and C. H. Wu, "InterPro, progress and status in 2005," *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D201–D205, Jan. 2005.

[154] P. J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler, "The International Protein Index: An integrated database for proteomics experiments," *Proteomics*, vol. 4, no. 7, pp. 1985–1988, Jul. 2004.

[155] Y. J. Kim, A. Boyd, B. D. Athey, and J. M. Patel, "miBLAST: scalable evaluation of a batch of nucleotide sequence queries with BLAST," *Nucleic Acids Research*, vol. 33, no. 13, pp. 4335–4344, Jan. 2005.

[156] N. Wiwatwattana and A. Kumar, “Organelle DB: a cross-species database of protein localization and function,” *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D598–D604, Jan. 2005.
[157] F. Chen, A. J. Mackey, C. J. Stoeckert, and D. S. Roos, “OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups,” *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D363–D368, Jan. 2006.
[158] R. D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, “Pfam: clans, web tools and services,” *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D247–D251, Jan. 2006.
[159] O. Sasson, A. Vaaknin, H. Fleischer, E. Portugaly, Y. Bilu, N. Linial, and M. Linial, “ProtoNet: hierarchical classification of the protein space,” *Nucleic Acids Research*, vol. 31, no. 1, pp. 348–352, Jan. 2003.
[160] “mint database statistics.” [Online]. Available: http://mint.bio.uniroma2.it/mint/statistics/statistics.do. [Accessed: 27-Aug-2014].
[161] “Apache ObJectRelationalBridge - OJB.” [Online]. Available: http://db.apache.org/ojb/. [Accessed: 27-Aug-2014].
[162] “Apache Tomcat - Welcome!” [Online]. Available: http://tomcat.apache.org/. [Accessed: 27-Aug-2014].
[163] H. W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil, “MIPS: a database for genomes and protein sequences,” *Nucleic Acids Research*, vol. 30, no. 1, pp. 31–34, Jan. 2002.
[164] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.-W. Mewes, A. Ruepp, and D. Frishman, “The MIPS mammalian protein–protein interaction database,” *Bioinformatics*, vol. 21, no. 6, pp. 832–834, Mar. 2005.
[165] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, “miRBase: microRNA sequences, targets and gene nomenclature,” *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D140–D144, Jan. 2006.
[166] A. Kozomara and S. Griffiths-Jones, “miRBase: annotating high confidence microRNAs using deep sequencing data,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D68–D73, Jan. 2014.
[167] J. Goll, S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb, and P. Uetz, “MPIDB: the microbial protein interaction database,” *Bioinformatics*, vol. 24, no. 15, pp. 1743–1744, Aug. 2008.
[168] “Cascading Style Sheets.” [Online]. Available: http://www.w3.org/Style/CSS/. [Accessed: 27-Aug-2014].
[169] “Pathway Interaction Database.” [Online]. Available: http://pid.nci.nih.gov/. [Accessed: 11-Aug-2014].
[170] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct. 1990.
[171] “IBM WebSphere software - United States.” [Online]. Available: http://www-01.ibm.com/software/websphere/. [Accessed: 27-Aug-2014].
[172] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania, “PANTHER: A Library of Protein Families and Subfamilies Indexed by Function,” *Genome Research*, vol. 13, no. 9, pp. 2129–2141, Sep. 2003.

[173] H. Mi, B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano, and P. D. Thomas, "The PANTHER database of protein families, subfamilies, functions and pathways," *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D284–D288, Jan. 2005.

[174] H. Mi, A. Muruganujan, and P. D. Thomas, "PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D377–386, Jan. 2013.

[175] "PANTHER - Gene List Analysis." [Online]. Available: http://www.pantherdb.org/. [Accessed: 11-Aug-2014].

[176] "CellDesigner." [Online]. Available: http://www.systems-biology.org/002/. [Accessed: 26-Aug-2014].

[177] N. Le Novère, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, F. T. Bergman, R. Gauges, P. Ghazal, H. Kawaji, L. Li, Y. Matsuoka, A. Villéger, S. E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T. C. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D. B. Kell, C. Sander, H. Sauro, J. L. Snoep, K. Kohn, and H. Kitano, "The Systems Biology Graphical Notation," *Nat. Biotechnol.*, vol. 27, no. 8, pp. 735–741, Aug. 2009.

[178] M. J. Cowley, M. Pinese, K. S. Kassahn, N. Waddell, J. V. Pearson, S. M. Grimmond, A. V. Biankin, S. Hautaniemi, and J. Wu, "PINA v2.0: mining interactome modules," *Nucleic Acids Research*, vol. 40, no. D1, pp. D862–D865, Jan. 2012.

[179] "AllegroMCODE | Cytoscape Clustering App – AllegroViva." [Online]. Available: http://allegroviva.com/allegromcode/. [Accessed: 27-Aug-2014].

[180] "Jersey." [Online]. Available: https://jersey.java.net/. [Accessed: 27-Aug-2014].

[181] M. D. McDowall, M. S. Scott, and G. J. Barton, "PIPs: human protein–protein interaction prediction database," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D651–D656, Jan. 2009.

[182] "PIPs: Human Protein-Protein Interaction Prediction." [Online]. Available: http://www.compbio.dundee.ac.uk/www-pips/dbStats.jsp. [Accessed: 12-Aug-2014].

[183] "JavaServer Pages Technology." [Online]. Available: http://www.oracle.com/technetwork/java/javaee/jsp/index.html. [Accessed: 27-Aug-2014].

[184] T.-W. Huang, A.-C. Tien, W.-S. Huang, Y.-C. G. Lee, C.-L. Peng, H.-H. Tseng, C.-Y. Kao, and C.-Y. F. Huang, "POINT: a database for the prediction of protein–protein interactions based on the orthologous interactome," *Bioinformatics*, vol. 20, no. 17, pp. 3273–3276, Nov. 2004.

[185] S.-A. Lee, C.-H. Chan, T.-C. Chen, C.-Y. Yang, K.-C. Huang, C.-H. Tsai, J.-M. Lai, F.-S. Wang, C.-Y. Kao, and C.-Y. Huang, "POINeT: protein interactome with sub-network analysis and hub prioritization," *BMC Bioinformatics*, vol. 10, no. 1, p. 114, 2009.

[186] U. Güldener, M. Münsterkötter, G. Kastenmüller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. García-Martínez, J. E. Pérez-Ortín, H. Michael, A. Kaps, E. Talla, B. Dujon, B. André, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H. W. Mewes, "CYGD: the Comprehensive Yeast Genome Database," *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D364–368, Jan. 2005.

[187] "NCBI Interactome download." [Online]. Available: ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz.

[188] “JSP Standard Tag Library.” [Online]. Available: https://jstl.java.net/. [Accessed: 27-Aug-2014].
[189] J. J. Garrett, “Ajax: A New Approach to Web Applications.” 18-Feb-2005.
[190] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D’Eustachio, “The Reactome pathway knowledgebase,” *Nucleic Acids Research*, Nov. 2013.
[191] R. Alcántara, K. B. Axelsen, A. Morgat, E. Belda, E. Coudert, A. Bridge, H. Cao, P. de Matos, M. Ennis, S. Turner, G. Owen, L. Bougueleret, I. Xenarios, and C. Steinbeck, “Rhea—a manually curated resource of biochemical reactions,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D754–D760, Jan. 2012.
[192] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S. M. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro, “The EcoCyc Database,” *Nucleic Acids Research*, vol. 30, no. 1, pp. 56–58, Jan. 2002.
[193] A. Morgat, E. Coissac, E. Coudert, K. B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios, and A. Viari, “UniPathway: a resource for the exploration and annotation of metabolic pathways,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D761–D769, Jan. 2012.
[194] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, “STITCH: interaction networks of chemicals and proteins,” *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D684–D688, Jan. 2008.
[195] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T. H. Blicher, C. von Mering, L. J. Jensen, and P. Bork, “STITCH 4: integration of protein–chemical interactions with user data,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D401–D407, Jan. 2014.
[196] K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and C. Schaefer, “The NCI-Nature Pathway Interaction Database: A cell signaling resource.,” *Nature Precedings*, Nov. 2007.
[197] “Protein Complexes | Gene Ontology Consortium.” [Online]. Available: http://www.geneontology.org/page/protein-complexes. [Accessed: 12-Aug-2014].
[198] S. Orchard, J.-P. Albar, E. W. Deutsch, M. Eisenacher, P.-A. Binz, and H. Hermjakob, “Implementing Data Standards: A report on the HUPOPSI Workshop September 2009, Toronto, Canada,” *Proteomics*, vol. 10, no. 10, pp. 1895–1898, May 2010.
[199] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüss, I. Reuter, and F. Schacherer, “TRANSFAC: an integrated system for gene expression regulation,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 316–319, Jan. 2000.
[200] F. Zhu, Z. Shi, C. Qin, L. Tao, X. Liu, F. Xu, L. Zhang, Y. Song, X. Liu, J. Zhang, B. Han, P. Zhang, and Y. Chen, “Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D1128–D1136, Jan. 2012.
[201] R. K. R. Kalathur, J. P. Pinto, M. A. Hernández-Prieto, R. S. R. Machado, D. Almeida, G. Chaurasia, and M. E. Futschik, “UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks,” *Nucleic Acids Research*, Nov. 2013.
[202] “Core J2EE Patterns - Data Access Object.” [Online]. Available: http://www.oracle.com/technetwork/java/dataaccessobject-138824.html. [Accessed: 27-Aug-2014].
[203] “Hibernate. Everything data. - Hibernate.” [Online]. Available: http://hibernate.org/. [Accessed: 27-Aug-2014].

[204] "UniProtKB." [Online]. Available: http://www.uniprot.org/help/uniprotkb. [Accessed: 13-Aug-2014].
[205] "The OBO Flat File Format Specification, version 1.2." [Online]. Available: http://geneontology.org/GO.format.obo-1_2.shtml. [Accessed: 13-Aug-2014].
[206] "Home / Zope Corporation." [Online]. Available: http://www.zope.com/. [Accessed: 27-Aug-2014].
[207] "Unbeatable JavaScript Tools - The Dojo Toolkit." [Online]. Available: http://dojotoolkit.org/. [Accessed: 27-Aug-2014].
[208] Z. Hu, Y.-C. Chang, Y. Wang, C.-L. Huang, Y. Liu, F. Tian, B. Granger, and C. DeLisi, "VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies," *Nucleic Acids Research*, vol. 41, no. W1, pp. W225–W231, Jul. 2013.
[209] R. B. Altman, "PharmGKB: a logical home for knowledge relating genotype to drug response phenotype," *Nat Genet*, vol. 39, no. 4, pp. 426–426, Apr. 2007.
[210] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The Genetic Association Database," *Nat Genet*, vol. 36, no. 5, pp. 431–432, May 2004.
[211] K.-C. Chen, T.-Y. Wang, H.-H. Tseng, C.-Y. F. Huang, and C.-Y. Kao, "A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae," *Bioinformatics*, vol. 21, no. 12, pp. 2883–2890, Jun. 2005.
[212] B. Teusink, J. Passarge, C. A. Reijenga, E. Esgalhado, C. C. van der Weijden, M. Schepper, M. C. Walsh, B. M. Bakker, K. van Dam, H. V. Westerhoff, and J. L. Snoep, "Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry," *European Journal of Biochemistry*, vol. 267, no. 17, pp. 5313–5329, Sep. 2000.
[213] E. Lee, A. Salic, R. Krüger, R. Heinrich, and M. W. Kirschner, "The Roles of APC and Axin Derived from Experimental and Theoretical Analysis of the Wnt Pathway," *PLoS Biology*, vol. 1, no. 1, p. e10, 2003.
[214] "Corrections: The Roles of APC and Axin Derived from Experimental and Theoretical Analysis of the Wnt Pathway," *PLoS Biology*, vol. 2, no. 3, p. e89, 2004.
[215] C. G. Moles, P. Mendes, and J. R. Banga, "Parameter Estimation in Biochemical Pathways: A Comparison of Global Optimization Methods," *Genome Research*, vol. 13, no. 11, pp. 2467–2474, Nov. 2003.
[216] J. Timmer, T. G. MüLler, I. Swameye, O. Sandra, and U. KlingmüLler, "MODELING THE NONLINEAR DYNAMICS OF CELLULAR SIGNAL TRANSDUCTION," *International Journal of Bifurcation and Chaos*, vol. 14, no. 06, pp. 2069–2079, Jun. 2004.
[217] Y. Huang, D. Liu, and H. Wu, "Hierarchical Bayesian Methods for Estimation of Parameters in a Longitudinal HIV Dynamic System," *Biometrics*, vol. 62, no. 2, pp. 413–423, Jun. 2006.
[218] B.-E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alche-Buc, "Gene networks inference using dynamic Bayesian networks," *Bioinformatics*, vol. 19, no. Suppl 2, pp. ii138–ii148, Sep. 2003.
[219] Z. Dai and L. Lai, "Differential simulated annealing: a robust and efficient global optimization algorithm for parameter estimation of biological networks," *Mol Biosyst*, vol. 10, no. 6, pp. 1385–1392, Jun. 2014.
[220] C. Siehs, B. Mayer, and C. Siehs, "Dynamical hierarchies of structure and control in chemical reaction networks *," *Nanotechnology*, vol. 10, no. 4, pp. 464–471, Dec. 1999.

[221] C. Siehs, R. Oberbauer, G. Mayer, A. Lukas, and B. Mayer, “Discrete simulation of regulatory homo- and heterodimerization in the apoptosis effector phase,” *Bioinformatics*, vol. 18, no. 1, pp. 67–76, Jan. 2002.

[222] A. A. Apte, J. W. Cain, D. G. Bonchev, and S. S. Fong, “Cellular automata simulation of topological effects on the dynamics of feed-forward motifs,” *Journal of Biological Engineering*, vol. 2, no. 1, p. 2, 2008.

[223] L. Zhang, C. A. Athale, and T. S. Deisboeck, “Development of a three-dimensional multiscale agent-based tumor model: Simulating gene-protein interaction profiles, cell phenotypes and multicellular patterns in brain cancer,” *Journal of Theoretical Biology*, vol. 244, no. 1, pp. 96–107, Jan. 2007.

[224] N. Kleinstreuer, D. Dix, M. Rountree, N. Baker, N. Sipes, D. Reif, R. Spencer, and T. Knudsen, “A Computational Model Predicting Disruption of Blood Vessel Development,” *PLoS Computational Biology*, vol. 9, no. 4, p. e1002996, Apr. 2013.

[225] S. A. Kauffman, “Metabolic stability and epigenesis in randomly constructed genetic nets,” *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, Mar. 1969.

[226] C. Darabos, M. Giacobini, J. Moore, and M. Tomassini, “Models of Gene Regulation: Integrating Modern Knowledge into the Random Boolean Network Framework,” in *Evolution, Complexity and Artificial Life*, S. Cagnoni, M. Mirolli, and M. Villani, Eds. Springer Berlin Heidelberg, 2014, pp. 43–57.

[227] M. B. Elowitz, “Stochastic Gene Expression in a Single Cell,” *Science*, vol. 297, no. 5584, pp. 1183–1186, Aug. 2002.

[228] W. J. Blake, M. KÆrn, C. R. Cantor, and J. J. Collins, “Noise in eukaryotic gene expression,” *Nature*, vol. 422, no. 6932, pp. 633–637, Apr. 2003.

[229] H. El Samad, M. Khammash, L. Petzold, and D. Gillespie, “Stochastic modelling of gene regulatory networks,” *International Journal of Robust and Nonlinear Control*, vol. 15, no. 15, pp. 691–711, Oct. 2005.

[230] D. T. Gillespie, “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions,” *Journal of Computational Physics*, vol. 22, no. 4, pp. 403–434, Dec. 1976.

[231] P. Thomas, N. Popović, and R. Grima, “Phenotypic switching in gene regulatory networks,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 19, pp. 6994–6999, May 2014.

[232] E. T. Papoutsakis, “Equations and calculations for fermentations of butyric acid bacteria,” *Biotechnology and Bioengineering*, vol. 26, no. 2, pp. 174–187, Feb. 1984.

[233] J. D. Orth, I. Thiele, and B. Ø. Palsson, “What is flux balance analysis?,” *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, Mar. 2010.

[234] N. E. Lewis and A. M. Abdel-Haleem, “The evolution of genome-scale models of cancer metabolism,” *Frontiers in Physiology*, vol. 4, 2013.

[235] K. Lotz, A. Hartmann, E. Grafahrend-Belau, F. Schreiber, and B. Junker, “Elementary Flux Modes, Flux Balance Analysis, and Their Application to Plant Metabolism,” in *Plant Metabolism*, vol. 1083, G. Sriram, Ed. Humana Press, 2014, pp. 231–252.

[236] R. Albert, “Scale-free networks in cell biology,” *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, Nov. 2005.

[237] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, “Evidence for dynamically organized modularity in the yeast protein-protein interaction network,” *Nature*, vol. 430, no. 6995, pp. 88–93, Jul. 2004.

[238] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, “Dynamic modularity in protein interaction networks predicts breast cancer outcome,” *Nat Biotech*, vol. 27, no. 2, pp. 199–204, Feb. 2009.

[239] T. Wallach, K. Schellenberg, B. Maier, R. K. R. Kalathur, P. Porras, E. E. Wanker, M. E. Futschik, and A. Kramer, “Dynamic Circadian Protein–Protein Interaction Networks Predict Temporal Organization of Cellular Functions,” *PLoS Genetics*, vol. 9, no. 3, p. e1003398, Mar. 2013.

[240] C. Sima, J. Hua, and S. Jung, “Inference of gene regulatory networks using time-series data: a survey,” *Curr. Genomics*, vol. 10, no. 6, pp. 416–429, Sep. 2009.

[241] P. C. H. Ma and K. C. C. Chan, “An effective data mining technique for reconstructing gene regulatory networks from time series expression data,” *J Bioinform Comput Biol*, vol. 5, no. 3, pp. 651–668, Jun. 2007.

[242] W. A. Schmitt, R. M. Raab, and G. Stephanopoulos, “Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data,” *Genome Res.*, vol. 14, no. 8, pp. 1654–1663, Aug. 2004.

[243] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. Theis, “Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data,” *BMC Systems Biology*, vol. 5, no. 1, p. 21, 2011.

[244] N. Friedman, M. Linial, I. Nachman, and D. Pe’er, “Using Bayesian Networks to Analyze Expression Data,” *Journal of Computational Biology*, vol. 7, no. 3–4, pp. 601–620, Aug. 2000.

[245] M. Zou and S. D. Conzen, “A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data,” *Bioinformatics*, vol. 21, no. 1, pp. 71–79, Jan. 2005.

[246] S. Martin, Z. Zhang, A. Martino, and J.-L. Faulon, “Boolean dynamics of genetic regulatory networks inferred from microarray time series data,” *Bioinformatics*, vol. 23, no. 7, pp. 866–874, Apr. 2007.

[247] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, “Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks,” *Bioinformatics*, vol. 18, no. 2, pp. 261–274, Feb. 2002.

[248] T. G. Dewey, “From microarrays to networks: mining expression time series,” *Drug Discov. Today*, vol. 7, no. 20 Suppl, pp. S170–175, Oct. 2002.

[249] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. L. Wild, and F. Falciani, “Modeling T-cell activation using gene expression profiling and state-space models,” *Bioinformatics*, vol. 20, no. 9, pp. 1361–1372, Jun. 2004.

[250] J. Cao, X. Qi, and H. Zhao, “Modeling gene regulation networks using ordinary differential equations,” *Methods Mol. Biol.*, vol. 802, pp. 185–197, 2012.

[251] F. Markowetz and R. Spang, “Inferring cellular networks - a review,” *BMC Bioinformatics*, vol. 8, no. Suppl 6, p. S5, 2007.

[252] E. N. Gilbert, “Random Graphs,” *The Annals of Mathematical Statistics*, vol. 30, no. 4, pp. 1141–1144, Dec. 1959.

[253] P. Erdős and A. Rényi, “On random graphs,” *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[254] D. J. Watts and S. H. Strogatz, “Collective dynamics of /`small-world/’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[255] G. U. Yule, “A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S.,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 213, no. 402–410, pp. 21–87, Jan. 1925.
[256] E. Ravasz and A.-L. Barabási, “Hierarchical organization in complex networks,” *Physical Review E*, vol. 67, no. 2, Feb. 2003.
[257] J. Larmouth, *ASN.1 complete*. San Diego, CA: Academic Press, 2000.
[258] “BioPAX.org.” [Online]. Available: http://www.biopax.org/. [Accessed: 26-Aug-2014].
[259] “SBML.org.” [Online]. Available: http://sbml.org/Main_Page. [Accessed: 26-Aug-2014].
[260] S. Orchard, L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, V. Stümpflen, A. Ceol, A. Chatr-aryamontri, J. Armstrong, P. Woollard, J. J. Salama, S. Moore, J. Wojcik, G. D. Bader, M. Vidal, M. E. Cusick, M. Gerstein, A.-C. Gavin, G. Superti-Furga, J. Greenblatt, J. Bader, P. Uetz, M. Tyers, P. Legrain, S. Fields, N. Mulder, M. Gilson, M. Niepmann, L. Burgoon, J. De Las Rivas, C. Prieto, V. M. Perreau, C. Hogue, H.-W. Mewes, R. Apweiler, I. Xenarios, D. Eisenberg, G. Cesareni, and H. Hermjakob, “The minimum information required for reporting a molecular interaction experiment (MIMIx),” *Nat. Biotechnol.*, vol. 25, no. 8, pp. 894–898, Aug. 2007.
[261] S. Orchard, B. Al-Lazikani, S. Bryant, D. Clark, E. Calder, I. Dix, O. Engkvist, M. Forster, A. Gaulton, M. Gilson, R. Glen, M. Grigorov, K. Hammond-Kosack, L. Harland, A. Hopkins, C. Larminie, N. Lynch, R. K. Mann, P. Murray-Rust, E. Lo Piparo, C. Southan, C. Steinbeck, D. Wishart, H. Hermjakob, J. Overington, and J. Thornton, “Minimum information about a bioactive entity (MIABE),” *Nat Rev Drug Discov*, vol. 10, no. 9, pp. 661–669, Sep. 2011.
[262] J. Bourbeillon, S. Orchard, I. Benhar, C. Borrebaeck, A. de Daruvar, S. Dubel, R. Frank, F. Gibson, D. Gloriam, N. Haslam, T. Hiltker, I. Humphrey-Smith, M. Hust, D. Juncker, M. Koegl, Z. Konthur, B. Korn, S. Krobitsch, S. Muyldermans, P.-A. Nygren, S. Palcy, B. Polic, H. Rodriguez, A. Sawyer, M. Schlapshy, M. Snyder, O. Stoevesandt, M. J. Taussig, M. Templin, M. Uhlen, S. van der Maarel, C. Wingren, H. Hermjakob, and D. Sherman, “Minimum information about a protein affinity reagent (MIAPAR),” *Nat Biotech*, vol. 28, no. 7, pp. 650–653, Jul. 2010.
[263] N. del-Toro, M. Dumousseau, S. Orchard, R. C. Jimenez, E. Galeota, G. Launay, J. Goll, K. Breuer, K. Ono, L. Salwinski, and H. Hermjakob, “A new reference implementation of the PSICQUIC web service,” *Nucleic Acids Res.*, vol. 41, no. Web Server issue, pp. W601–606, Jul. 2013.
[264] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, “The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration,” *Nat. Biotechnol.*, vol. 25, no. 11, pp. 1251–1255, Nov. 2007.
[265] M. Himsolt, “GML: A portable Graph File Format.” Universität Passau.
[266] R. Tamassia, “Graph Markup Language (GraphML),” in *Handbook of Graph Drawing and Visualization*, CRC Press, pp. 517–541.
[267] “XGMML (eXtensible Graph Markup and Modeling Language).” [Online]. Available: http://cgi7.cs.rpi.edu/research/groups/pb/punin/public_html/XGMML/. [Accessed: 27-Aug-2014].
[268] D.-Y. Cho, Y.-A. Kim, and T. M. Przytycka, “Chapter 5: Network Biology Approach to Complex Diseases,” *PLoS Computational Biology*, vol. 8, no. 12, p. e1002820, Dec. 2012.
[269] B. Lehne and T. Schlitt, “Protein-protein interaction databases: keeping up with growing interactomes,” *Human Genomics*, vol. 3, no. 3, pp. 291 – 297, 2008.

[270] A.-L. Barabasi and Z. N. Oltvai, “Network biology: understanding the cell’s functional organization,” *Nat Rev Genet*, vol. 5, no. 2, pp. 101–113, Feb. 2004.
[271] X. Zhu, M. Gerstein, and M. Snyder, “Getting connected: analysis and principles of biological networks,” *Genes Dev.*, vol. 21, no. 9, pp. 1010–1024, May 2007.
[272] A. Wagner and D. A. Fell, “The small world inside large metabolic networks,” *Proc. Biol. Sci.*, vol. 268, no. 1478, pp. 1803–1810, Sep. 2001.
[273] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási, “Functional and topological characterization of protein interaction networks,” *Proteomics*, vol. 4, no. 4, pp. 928–942, Apr. 2004.
[274] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, “The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics,” *PLoS Computational Biology*, vol. 3, no. 4, p. e59, 2007.
[275] G. Scardoni, A. Montresor, G. Tosadori, and C. Laudanna, “Node Interference and Robustness: Performing Virtual Knock-Out Experiments on Biological Networks: The Case of Leukocyte Integrin Activation Network,” *PLoS ONE*, vol. 9, no. 2, p. e88938, Feb. 2014.
[276] M. Barthélémy and L. Amaral, “Small-World Networks: Evidence for a Crossover Picture,” *Physical Review Letters*, vol. 82, no. 15, pp. 3180–3183, Apr. 1999.
[277] N. R. Zabet, “Negative feedback and physical limits of genes,” *J. Theor. Biol.*, vol. 284, no. 1, pp. 82–91, Sep. 2011.
[278] S. Mangan and U. Alon, “Structure and function of the feed-forward loop network motif,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 21, pp. 11980–11985, Oct. 2003.
[279] C. P. Bagowski and J. E. Ferrell, “Bistability in the JNK cascade,” *Curr. Biol.*, vol. 11, no. 15, pp. 1176–1182, Aug. 2001.
[280] U. S. Bhalla, P. T. Ram, and R. Iyengar, “MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network,” *Science*, vol. 297, no. 5583, pp. 1018–1023, Aug. 2002.
[281] J. E. Ferrell, “Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability,” *Curr. Opin. Cell Biol.*, vol. 14, no. 2, pp. 140–148, Apr. 2002.
[282] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network Motifs: Simple Building Blocks of Complex Networks,” *Science*, vol. 298, no. 5594, pp. 824–827, Oct. 2002.
[283] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker, “Conserved patterns of protein interaction in multiple species,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 6, pp. 1974–1979, Feb. 2005.
[284] S. Dorogovtsev, A. Goltsev, and J. Mendes, “Pseudofractal scale-free web,” *Physical Review E*, vol. 65, no. 6, Jun. 2002.
[285] B. Vogelstein, D. Lane, and A. J. Levine, “Surfing the p53 network,” *Nature*, vol. 408, no. 6810, pp. 307–310, Nov. 2000.
[286] A. G. Knudson, “Mutation and Cancer: Statistical Study of Retinoblastoma,” *Proceedings of the National Academy of Sciences*, vol. 68, no. 4, pp. 820–823, Apr. 1971.
[287] D. J. Hunter, “Gene–environment interactions in human diseases,” *Nature Reviews Genetics*, vol. 6, no. 4, pp. 287–298, Apr. 2005.
[288] L. Zhang, O. King, S. Wong, D. Goldberg, A. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, and F. Roth, “Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network,” *Journal of Biology*, vol. 4, no. 2, p. 6, 2005.

[289] A. Ma'ayan, "Formation of Regulatory Patterns During Signal Propagation in a Mammalian Cellular Network," *Science*, vol. 309, no. 5737, pp. 1078–1083, Aug. 2005.
[290] L. Babai and P. Codenotti, "Isomorphism of Hypergraphs of Low Rank in Moderately Exponential Time," 2008, pp. 667–676.
[291] E. Wong, B. Baur, S. Quader, and C.-H. Huang, "Biological network motif detection: principles and practice," *Briefings in Bioinformatics*, vol. 13, no. 2, pp. 202–215, Mar. 2012.
[292] Z. Kashani, H. Ahrabian, E. Elahi, A. Nowzari-Dalini, E. Ansari, S. Asadi, S. Mohammadi, F. Schreiber, and A. Masoudi-Nejad, "Kavosh: a new algorithm for finding network motifs," *BMC Bioinformatics*, vol. 10, no. 1, p. 318, 2009.
[293] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, no. 11, pp. 1746–1758, Jul. 2004.
[294] V. Batagelj and A. Mrvar, "Pajek— Analysis and Visualization of Large Networks," in *Graph Drawing*, vol. 2265, P. Mutzel, M. Jünger, and S. Leipert, Eds. Springer Berlin Heidelberg, 2002, pp. 477–478.
[295] F. Schreiber and H. Schwöbbermeyer, "MAVisto: a tool for the exploration of network motifs," *Bioinformatics*, vol. 21, no. 17, pp. 3572–3574, Sep. 2005.
[296] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, May 2006.
[297] A. Wagner, "How the global structure of protein interaction networks evolves," *Proc. Biol. Sci.*, vol. 270, no. 1514, pp. 457–466, Mar. 2003.
[298] E. Eisenberg and E. Y. Levanon, "Preferential attachment in the protein network evolution," *Phys. Rev. Lett.*, vol. 91, no. 13, p. 138701, Sep. 2003.
[299] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, Apr. 2010.
[300] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, Jun. 2005.
[301] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 1128–1133, Feb. 2003.
[302] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.
[303] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, Aug. 2010.
[304] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, Feb. 2010.
[305] M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao, and Q. Cui, "An Analysis of Human MicroRNA and Disease Associations," *PLoS ONE*, vol. 3, no. 10, p. e3420, Oct. 2008.
[306] M. Ray, J. Ruan, and W. Zhang, "Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases," *Genome Biology*, vol. 9, no. 10, p. R148, 2008.
[307] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, Oct. 2007.
[308] K. M. Mani, C. Lefebvre, K. Wang, W. K. Lim, K. Basso, R. Dalla-Favera, and A. Califano, "A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas," *Mol. Syst. Biol.*, vol. 4, p. 169, 2008.

[309] J. R. Managbanag, T. M. Witten, D. Bonchev, L. A. Fox, M. Tsuchiya, B. K. Kennedy, and M. Kaeberlein, "Shortest-path network analysis is a useful approach toward identifying genetic determinants of longevity," *PLoS ONE*, vol. 3, no. 11, p. e3802, 2008.

[310] Y.-K. Shih and S. Parthasarathy, "A single source k-shortest paths algorithm to infer regulatory pathways in a gene network," *Bioinformatics*, vol. 28, no. 12, pp. i49–58, Jun. 2012.

[311] Y.-A. Kim, J. H. Przytycki, S. Wuchty, and T. M. Przytycka, "Modeling information flow in biological networks," *Phys Biol*, vol. 8, no. 3, p. 035012, Jun. 2011.

[312] Y. Liu, D. A. Tennant, Z. Zhu, J. K. Heath, X. Yao, and S. He, "DiME: A Scalable Disease Module Identification Algorithm with Application to Glioma Progression," *PLoS ONE*, vol. 9, no. 2, p. e86693, Feb. 2014.

[313] A. Leung, G. D. Bader, and J. Reimand, "HyperModules: identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery," *Bioinformatics*, vol. 30, no. 15, pp. 2230–2232, Aug. 2014.

[314] P. Jiang and M. Singh, "SPICi: a fast clustering algorithm for large biological networks," *Bioinformatics*, vol. 26, no. 8, pp. 1105–1111, Apr. 2010.

[315] K. Rhrissorrakrai and K. C. Gunsalus, "MINE: Module Identification in Networks," *BMC Bioinformatics*, vol. 12, no. 1, p. 192, 2011.

[316] G. Bader and C. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 1, p. 2, 2003.

[317] B. Adamcsek, G. Palla, I. J. Farkas, I. Derenyi, and T. Vicsek, "CFinder: locating cliques and overlapping modules in biological networks," *Bioinformatics*, vol. 22, no. 8, pp. 1021–1023, Apr. 2006.

[318] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Res.*, vol. 30, no. 7, pp. 1575–1584, Apr. 2002.

[319] A.-L. Barabasi, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nat Rev Genet*, vol. 12, no. 1, pp. 56–68, Jan. 2011.

[320] J. Veenstra-VanderWeele, S. L. Christian, and E. H. Cook, Jr., "AUTISM AS A PARADIGMATIC COMPLEX GENETIC DISORDER," *Annual Review of Genomics and Human Genetics*, vol. 5, no. 1, pp. 379–405, Sep. 2004.

[321] E. E. Schadt, "Molecular networks as sensors and drivers of common human diseases," *Nature*, vol. 461, no. 7261, pp. 218–223, Sep. 2009.

[322] M. Evangelou, A. Rendon, W. H. Ouwehand, L. Wernisch, and F. Dudbridge, "Comparison of Methods for Competitive Tests of Pathway Analysis," *PLoS ONE*, vol. 7, no. 7, p. e41018, Jul. 2012.

[323] B. L. Fridley, G. D. Jenkins, and J. M. Biernacka, "Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods," *PLoS ONE*, vol. 5, no. 9, 2010.

[324] F. Dudbridge and B. P. C. Koeleman, "Rank truncated product of P-values, with application to genomewide association scans," *Genet. Epidemiol.*, vol. 25, no. 4, pp. 360–366, Dec. 2003.

[325] J. Taylor and R. Tibshirani, "A tail strength measure for assessing the overall univariate significance in a dataset," *Biostatistics*, vol. 7, no. 2, pp. 167–181, Apr. 2006.

[326] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen, "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics*, vol. 20, no. 1, pp. 93–99, Jan. 2004.

[327] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier, “Enrichment or depletion of a GO category within a class of genes: which test?,” *Bioinformatics*, vol. 23, no. 4, pp. 401–407, Feb. 2007.

[328] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “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*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005.

[329] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists,” *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, Jan. 2009.

[330] J.-H. Hung, “Gene Set/Pathway enrichment analysis,” *Methods Mol. Biol.*, vol. 939, pp. 201–213, 2013.

[331] S. J. Prohaska and P. F. Stadler, “The Use and Abuse of -Omes,” in *Bioinformatics for Omics Data: Methods and Protocols*, vol. 719, Springer Science+Business Media, 2011, pp. 173–196.

[332] J. Y. Chen, C. Shen, and A. Y. Sivachenko, “Mining Alzheimer disease relevant proteins from integrated protein interactome data,” *Pac Symp Biocomput*, pp. 367–378, 2006.

[333] A. P. Hodges, D. Dai, Z. Xiang, P. Woolf, C. Xi, and Y. He, “Bayesian Network Expansion Identifies New ROS and Biofilm Regulators,” *PLoS ONE*, vol. 5, no. 3, p. e9513, Mar. 2010.

[334] J. De Las Rivas and C. Fontanillo, “Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks,” *PLoS Computational Biology*, vol. 6, no. 6, p. e1000807, Jun. 2010.

[335] “ConsensusPathDB.” [Online]. Available: http://cpdb.molgen.mpg.de/. [Accessed: 20-Aug-2014].

[336] “APID Statistics.” [Online]. Available: http://bioinfow.dep.usal.es/apid/html/statistics14v3.htm. [Accessed: 20-Aug-2014].

[337] A. Bernthaler, I. Mühlberger, R. Fechete, P. Perco, A. Lukas, and B. Mayer, “A dependency graph approach for the analysis of differential gene expression profiles,” *Molecular BioSystems*, vol. 5, no. 12, p. 1720, 2009.

[338] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, “A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae),” *Proceedings of the National Academy of Sciences*, vol. 100, no. 14, pp. 8348–8353, Jul. 2003.

[339] Z. Tu, L. Wang, M. N. Arbeitman, T. Chen, and F. Sun, “An integrative approach for causal gene identification and gene regulatory pathway inference,” *Bioinformatics*, vol. 22, no. 14, pp. e489–496, Jul. 2006.

[340] C. Albanese, O. C. Rodriguez, J. VanMeter, S. T. Fricke, B. R. Rood, Y. Lee, S. S. Wang, S. Madhavan, Y. Gusev, E. F. Petricoin, and Y. Wang, “Preclinical Magnetic Resonance Imaging and Systems Biology in Cancer Research,” *The American Journal of Pathology*, vol. 182, no. 2, pp. 312–318, Feb. 2013.

[341] T. C. Neylan, E. E. Schadt, and R. Yehuda, “Biomarkers for combat-related PTSD: focus on molecular networks from high-dimensional data,” *European Journal of Psychotraumatology*, vol. 5, no. 0, Aug. 2014.

[342] D. Toubiana, A. R. Fernie, Z. Nikoloski, and A. Fait, “Network analysis: tackling complex data to study plant metabolism,” *Trends in Biotechnology*, vol. 31, no. 1, pp. 29–36, Jan. 2013.

[343] C. Pastrello, D. Otasek, K. Fortney, G. Agapito, M. Cannataro, E. Shirdel, and I. Jurisica, "Visual Data Mining of Biological Networks: One Size Does Not Fit All," *PLoS Computational Biology*, vol. 9, no. 1, p. e1002833, Jan. 2013.

[344] "NAViGaTOR - Network Analysis, Visualization, & Graphing TORonto." [Online]. Available: http://ophid.utoronto.ca/navigator/. [Accessed: 21-Aug-2014].

[345] E. Khurana, Y. Fu, J. Chen, and M. Gerstein, "Interpretation of Genomic Variants Using a Unified Biological Network Approach," *PLoS Computational Biology*, vol. 9, no. 3, p. e1002886, Mar. 2013.

[346] A. Heinzel, R. Fechete, J. Söllner, P. Perco, G. Heinze, R. Oberbauer, G. Mayer, A. Lukas, and B. Mayer, "Data Graphs for Linking Clinical Phenotype and Molecular Feature Space:," *International Journal of Systems Biology and Biomedical Technologies*, vol. 1, no. 1, pp. 11–25, 2012.

[347] I. Mühlberger, K. Moenks, A. Bernthaler, C. Jandrasits, B. Mayer, G. Mayer, R. Oberbauer, and P. Perco, "Integrative Bioinformatics Analysis of Proteins Associated with the Cardiorenal Syndrome," *International Journal of Nephrology*, vol. 2011, pp. 1–10, 2011.

[348] B. A. Kidd, L. A. Peters, E. E. Schadt, and J. T. Dudley, "Unifying immunology with informatics and multiscale biology," *Nat Immunol*, vol. 15, no. 2, pp. 118–127, Feb. 2014.

[349] J. P. F. Bai and D. R. Abernethy, "Systems Pharmacology to Predict Drug Toxicity: Integration Across Levels of Biological Organization*," *Annual Review of Pharmacology and Toxicology*, vol. 53, no. 1, pp. 451–473, Jan. 2013.

[350] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 719–732, Sep. 2013.

[351] E. L. Leung, Z.-W. Cao, Z.-H. Jiang, H. Zhou, and L. Liu, "Network-based drug discovery by integrating systems biology and computational technologies," *Briefings in Bioinformatics*, vol. 14, no. 4, pp. 491–505, Jul. 2013.

[352] G. Mayer, G. Heinze, H. Mischak, M. E. Hellemons, H. J. L. Heerspink, S. J. L. Bakker, D. de Zeeuw, M. Haiduk, P. Rossing, and R. Oberbauer, "Omics-bioinformatics in the context of clinical data," *Methods Mol. Biol.*, vol. 719, pp. 479–497, 2011.

[353] E. D. Levy, C. R. Landry, and S. W. Michnick, "How Perfect Can Protein Interactomes Be?," *Science Signaling*, vol. 2, no. 60, pp. pe11–pe11, Mar. 2009.

[354] L. Kiemer and G. Cesareni, "Comparative interactomics: comparing apples and pears?," *Trends Biotechnol.*, vol. 25, no. 10, pp. 448–454, Oct. 2007.

[355] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. Brinkman, G. Cesareni, A. Chatr-aryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. Hancock, L. I. Hannick, I. Jurisica, J. Khadake, D. J. Lynn, U. Mahadevan, L. Perfetto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stümpflen, M. Tyers, P. Uetz, I. Xenarios, and H. Hermjakob, "Protein interaction data curation: the International Molecular Exchange (IMEx) consortium," *Nature Methods*, vol. 9, no. 4, pp. 345–350, Mar. 2012.

[356] A. Skusa, A. Ruegg, and J. Kohler, "Extraction of biological interaction networks from scientific literature," *Briefings in Bioinformatics*, vol. 6, no. 3, pp. 263–276, Jan. 2005.

[357] K. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. Reinhold, B. Zeeberg, Ajay, and J. Weinstein, "MatchMiner: a tool for batch navigation among gene and gene product identifiers," *Genome Biology*, vol. 4, no. 4, p. R27, 2003.

[358] B. Zeeberg, J. Riss, D. Kane, K. Bussey, E. Uchio, W. M. Linehan, J. C. Barrett, and J. Weinstein, “Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics,” *BMC Bioinformatics*, vol. 5, no. 1, p. 80, 2004.

[359] M. Diehn, G. Sherlock, G. Binkley, H. Jin, J. C. Matese, T. Hernandez-Boussard, C. A. Rees, J. M. Cherry, D. Botstein, P. O. Brown, and A. A. Alizadeh, “SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data,” *Nucleic Acids Research*, vol. 31, no. 1, pp. 219–223, Jan. 2003.

[360] S. P. Wein, R. G. Cote, M. Dumousseau, F. Reisinger, H. Hermjakob, and J. A. Vizcaino, “Improvements in the protein identifier cross-reference service,” *Nucleic Acids Research*, vol. 40, no. W1, pp. W276–W280, Jul. 2012.

[361] H. Binder, T. Kirsten, M. Loeffler, and P. F. Stadler, “Sensitivity of Microarray Oligonucleotide Probes: Variability and Effect of Base Composition,” *J. Phys. Chem. B*, vol. 108, no. 46, pp. 18003–18014, Oct. 2004.

[362] H. Binder and S. Preibisch, “‘Hook’-calibration of GeneChip-microarrays: Theory and algorithm,” *Algorithms for Molecular Biology*, vol. 3, no. 1, p. 12, 2008.

[363] B. Cai, H. Wang, H. Zheng, and H. Wang, “Integrating domain similarity to improve protein complexes identification in TAP-MS data,” *Proteome Science*, vol. 11, no. Suppl 1, p. S2, 2013.

[364] K. Bunai and K. Yamane, “Effectiveness and limitation of two-dimensional gel electrophoresis in bacterial membrane protein proteomics and perspectives,” *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, vol. 815, no. 1–2, pp. 227–236, Feb. 2005.

[365] A. Brückner, C. Polge, N. Lentze, D. Auerbach, and U. Schlattner, “Yeast two-hybrid, a powerful tool for systems biology,” *Int J Mol Sci*, vol. 10, no. 6, pp. 2763–2788, Jun. 2009.

[366] A. J. R. Heck, “Native mass spectrometry: a bridge between interactomics and structural biology,” *Nat Meth*, vol. 5, no. 11, pp. 927–933, Nov. 2008.

[367] L. Salwinski and D. Eisenberg, “In silico simulation of biological network dynamics,” *Nat Biotech*, vol. 22, no. 8, pp. 1017–1019, Aug. 2004.

[368] Y. Zhou, J. Liepe, X. Sheng, M. P. H. Stumpf, and C. Barnes, “GPU accelerated biochemical network simulation,” *Bioinformatics*, vol. 27, no. 6, pp. 874–876, Mar. 2011.

[369] C. Wolfe, I. Kohane, and A. Butte, “Systematic survey reveals general applicability of ‘guilt-by-association’ within gene coexpression networks,” *BMC Bioinformatics*, vol. 6, no. 1, p. 227, 2005.

[370] J. Gillis and P. Pavlidis, “‘Guilt by Association’ Is the Exception Rather Than the Rule in Gene Networks,” *PLoS Computational Biology*, vol. 8, no. 3, p. e1002444, Mar. 2012.

[371] F. Markowetz, D. Kostka, O. G. Troyanskaya, and R. Spang, “Nested effects models for high-dimensional phenotyping screens,” *Bioinformatics*, vol. 23, no. 13, pp. i305–i312, Jul. 2007.

[372] B. Anchang, M. J. Sadeh, J. Jacob, A. Tresch, M. O. Vlad, P. J. Oefner, and R. Spang, “Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models,” *Proceedings of the National Academy of Sciences*, Mar. 2009.

[373] H. Fröhlich, P. Praveen, and A. Tresch, “Fast and Efficient Dynamic Nested Effects Models,” *Bioinformatics*, Nov. 2010.

[374] F. Emmert-Streib and G. V. Glazko, “Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases,” *PLoS Computational Biology*, vol. 7, no. 5, p. e1002053, May 2011.

[375] A. Shojaie and G. Michailidis, “Analysis of gene sets based on the underlying regulatory network,” *J. Comput. Biol.*, vol. 16, no. 3, pp. 407–426, Mar. 2009.
[376] P. Perco, R. Rapberger, C. Siehs, A. Lukas, R. Oberbauer, G. Mayer, and B. Mayer, “Transforming omics data into context: bioinformatics on genomics and proteomics raw data,” *Electrophoresis*, vol. 27, no. 13, pp. 2659–2675, Jul. 2006.
[377] R. Rapberger, P. Perco, C. Sax, T. Pangerl, C. Siehs, D. Pils, A. Bernthaler, A. Lukas, B. Mayer, and M. Krainer, “Linking the ovarian cancer transcriptome and immunome,” *BMC Systems Biology*, vol. 2, no. 1, p. 2, 2008.
[378] C. Siehs, P. Perco, R. Rapberger, A. Lukas, R. Grohmann, M. Krainer, and B. Mayer, “Autoantigenes and Differential Gene Expression in Ovarian Cancer,” presented at the RECOMB, Venice, Italy, 2006.
[379] A. Bernthaler, P. Perco, R. Rapberger, M. Haiduk, J. Söllner, C. Siehs, C. Pleban, M. Wiesinger, C. Stadler, A. Lukas, and B. Mayer, “dynaNET: A novel computational framework for deriving functional context from omics data,” presented at the ISMB & ECCB, Vienna, Austria, 2007.
[380] M. Lechner, K. Kratochwill, M. Endemann, C. Siehs, K. Herkner, B. Mayer, C. Aufricht, and A. Rizzi, “Analysis of the stress response of mesothelial cells to peritoneal dialysis fluid,” presented at the HUPO, Seoul, South Korea.
[381] A. Bernthaler, P. Perco, R. Rapberger, M. Haiduk, J. Söllner, C. Siehs, M. Wiesinger, I. Mühlberger, A. Lukas, B. Mayer, and R. Freund, “Functional distance measure and reconstruction of context sensitive protein-protein interaction networks with dynamical hierarchies,” presented at the Pacific Symposium on Biocomputing, Big Island, Hawaii, 2008.
[382] K. Kratochwill, M. Lechner, C. Siehs, H. C. Lederhuber, P. Rehulka, M. Endemann, D. C. Kasper, K. R. Herkner, B. Mayer, A. Rizzi, and C. Aufricht, “Stress Responses and Conditioning Effects in Mesothelial Cells Exposed to Peritoneal Dialysis Fluid,” *J. Proteome Res.*, vol. 8, no. 4, pp. 1731–1747, Feb. 2009.
[383] K. Kratochwill, M. Lechner, A. M. Lichtenauer, R. Herzog, H. C. Lederhuber, C. Siehs, M. Endemann, B. Mayer, A. Rizzi, and C. Aufricht, “Interleukin-1 Receptor-Mediated Inflammation Impairs the Heat Shock Response of Human Mesothelial Cells,” *The American Journal of Pathology*, vol. 178, no. 4, pp. 1544–1555, Apr. 2011.
[384] A. Ilhan-Mutlu, C. Siehs, A. Berghoff, G. Ricken, G. Widhalm, L. Wagner, and M. Preusser, “Expression profiling of angiogenesis-related genes in brain metastases of lung cancer and melanoma,” *Tumor Biol.*, pp. 1–10, Aug. 2015.
[385] J. Siek, *The boost graph library: user guide and reference manual*. Boston: Addison-Wesley, 2002.
[386] “KGML Document.” [Online]. Available: http://www.kegg.jp/kegg/xml/docs/. [Accessed: 11-Aug-2014].
[387] S. Kawashima, T. Katayama, Y. Sato, and Mm. Kanehisa, “KEGG API: A web service using SOAP/WSDL to access the KEGG system,” *Genome Informatics*, vol. 14, pp. 673–674, 2003.
[388] J. D. Zhang and S. Wiemann, “KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor,” *Bioinformatics*, vol. 25, no. 11, pp. 1470–1471, Jun. 2009.
[389] Y. Ichiki, M. Takenoyama, M. Mizukami, T. So, M. Sugaya, M. Yasuda, T. So, T. Hanagiri, K. Sugio, and K. Yasumoto, “Simultaneous cellular and humoral immune response against mutated p53 in a patient with lung cancer,” *J. Immunol.*, vol. 172, no. 8, pp. 4844–4850, Apr. 2004.

[390] M. Lu, R. M. Nakamura, E. D. Dent, J. Y. Zhang, F. C. Nielsen, J. Christiansen, E. K. Chan, and E. M. Tan, “Aberrant expression of fetal RNA-binding protein p62 in liver cancer and liver cirrhosis,” *Am. J. Pathol.*, vol. 159, no. 3, pp. 945–953, Sep. 2001.

[391] X. Wang, J. Yu, A. Sreekumar, S. Varambally, R. Shen, D. Giacherio, R. Mehra, J. E. Montie, K. J. Pienta, M. G. Sanda, P. W. Kantoff, M. A. Rubin, J. T. Wei, D. Ghosh, and A. M. Chinnaiyan, “Autoantibody signatures in prostate cancer,” *N. Engl. J. Med.*, vol. 353, no. 12, pp. 1224–1235, Sep. 2005.

[392] K. S. Anderson and J. LaBaer, “The sentinel within: exploiting the immune system for cancer biomarkers,” *J. Proteome Res.*, vol. 4, no. 4, pp. 1123–1133, Aug. 2005.

[393] N. Brass, A. Rácz, C. Bauer, D. Heckel, G. Sybrecht, and E. Meese, “Role of amplified genes in the production of autoantibodies,” *Blood*, vol. 93, no. 7, pp. 2158–2166, Apr. 1999.

[394] U. Sahin, O. Türeci, and M. Pfreundschuh, “Serological identification of human tumor antigens,” *Curr. Opin. Immunol.*, vol. 9, no. 5, pp. 709–716, Oct. 1997.

[395] “SOURCE Search.” [Online]. Available: http://smd.princeton.edu/cgi-bin/source/sourceSearch. [Accessed: 11-Aug-2014].

[396] K. Nakai and P. Horton, “PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization,” *Trends in Biochemical Sciences*, vol. 24, no. 1, pp. 34–35, Aug. 2014.

[397] “National Center for Biotechnology Information.” [Online]. Available: http://www.ncbi.nlm.nih.gov/. [Accessed: 04-Dec-2006].

[398] S. Karanam and C. S. Moreno, “CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets,” *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W475–484, Jul. 2004.

[399] “KEGG: Kyoto Encyclopedia of Genes and Genomes.” [Online]. Available: http://www.genome.jp/kegg/. [Accessed: 11-Aug-2014].

[400] H. Mi, N. Guo, A. Kejariwal, and P. D. Thomas, “PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways,” *Nucleic Acids Res.*, vol. 35, no. Database issue, pp. D247–252, Jan. 2007.

[401] K. A. Gray, L. C. Daugherty, S. M. Gordon, R. L. Seal, M. W. Wright, and E. A. Bruford, “Genenames.org: the HGNC resources in 2013,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D545–552, Jan. 2013.

[402] J. Tomfohr, J. Lu, and T. Kepler, “Pathway level analysis of gene expression using singular value decomposition,” *BMC Bioinformatics*, vol. 6, no. 1, p. 225, 2005.

[403] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, “Entrez Gene: gene-centered information at NCBI,” *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D54–58, Jan. 2005.

[404] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, “GenBank,” *Nucleic Acids Res.*, vol. 35, no. Database issue, pp. D21–25, Jan. 2007.

[405] J. U. Pontius, L. Wagner, and G. D. Schuler, “The NCBI Handbook,” in *UniGene: a unified view of the transcriptome*, Bethesda (MD): National Center for Biotechnology Information, 2003.

[406] K. Pruit, G. Brown, T. Tatusova, and D. Maglott, “The Reference Sequence (RefSeq) Project,” in *The NCBI handbook [Internet]*, Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information, 2002.

[407] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nat Protoc*, vol. 4, no. 1, pp. 44–57, 2009.
[408] The UniProt Consortium, “Activities at the Universal Protein Resource (UniProt),” *Nucleic Acids Research*, vol. 42, no. D1, pp. D191–D198, Jan. 2014.
[409] S. J. Davies, L. Phillips, A. M. Griffiths, L. H. Russell, P. F. Naish, and G. I. Russell, “What really happens to people on long-term peritoneal dialysis?,” *Kidney International*, vol. 54, no. 6, pp. 2207–2217, Dec. 1998.
[410] N. Topley, “What is the ideal technique for testing the biocompatibility of peritoneal dialysis solutions?,” *Perit Dial Int*, vol. 15, no. 6, pp. 205–209, Sep. 1995.
[411] B. Bidmon, M. Endemann, K. Arbeiter, D. Ruffingshofer, H. Regele, K. Herkner, O. Eickelberg, and C. Aufricht, “Overexpression of HSP-72 confers cytoprotection in experimental peritoneal dialysis,” *Kidney Int.*, vol. 66, no. 6, pp. 2300–2307, Dec. 2004.
[412] E. Klein, J. B. Klein, and V. Thongboonkerd, “Two-dimensional gel electrophoresis: a fundamental tool for expression proteomics studies,” *Contrib Nephrol*, vol. 141, pp. 25–39, 2004.
[413] “DECODON | Software Tools for Functional Genomics.” [Online]. Available: http://www.decodon.com/. [Accessed: 11-Aug-2014].
[414] “DECODON | Delta2D’s Statistical Methods: Overview and References.” [Online]. Available: http://www.decodon.com/delta2d-statistics-overview.html. [Accessed: 11-Aug-2014].
[415] C. E. Bonferroni, “Il calcolo delle assicurazioni su gruppi di teste,” *Studi in Onore del Professore Salvatore Ortu Carboni*, pp. 13–60, 1935.
[416] C. E. Bonferroni, “Teoria statistica delle classi e calcolo delle probabilità,” *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 1–62, 1936.
[417] K. R. Clauser, P. Baker, and A. L. Burlingame, “Role of Accurate Mass Measurement (±10 ppm) in Protein Identification Strategies Employing MS or MS/MS and Database Searching,” *Anal. Chem.*, vol. 71, no. 14, pp. 2871–2882, May 1999.
[418] C. R. Jiménez, L. Huang, Y. Qiu, and A. L. Burlingame, “Searching Sequence Databases Over the Internet: Protein Identification Using MS-Tag,” in *Current Protocols in Protein Science*, John Wiley & Sons, Inc., 2001.
[419] A. Bairoch and B. Boeckmann, “The SWISS-PROT protein sequence data bank: current status.,” *Nucleic Acids Research*, vol. 22, no. 17, pp. 3578–3580, 1994.
[420] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, Dec. 1998.
[421] A. I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and J. Quackenbush, “TM4: a free, open-source system for microarray data management and analysis,” *BioTechniques*, vol. 34, no. 2, pp. 374–378, Feb. 2003.
[422] “TM4: MeV.” [Online]. Available: http://www.tm4.org/mev.html. [Accessed: 11-Aug-2014].
[423] O. Alter, P. O. Brown, and D. Botstein, “Singular value decomposition for genome-wide expression data processing and modeling,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10101–10106, Aug. 2000.

[424] R. C. Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013.
[425] A. Korobeynikov, “svd: Interfaces to various state-of-art SVD and eigensolvers.” 2014.
[426] S. Wachi, K. Yoneda, and R. Wu, “Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues,” *Bioinformatics*, vol. 21, no. 23, pp. 4205–4208, Dec. 2005.
[427] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, “Walking the Interactome for Prioritization of Candidate Disease Genes,” *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, Aug. 2014.
[428] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, “Associating Genes and Protein Complexes with Disease via Network Propagation,” *PLoS Computational Biology*, vol. 6, no. 1, p. e1000641, Jan. 2010.
[429] D. Dunkler, F. Sánchez-Cabo, and G. Heinze, “Statistical analysis principles for Omics data,” *Methods Mol. Biol.*, vol. 719, pp. 113–131, 2011.
[430] H. T. Nguyen, M. Geens, and C. Spits, “Genetic and epigenetic instability in human pluripotent stem cells,” *Human Reproduction Update*, vol. 19, no. 2, pp. 187–205, Mar. 2013.
[431] A. Amore, G. Cappelli, P. Cirina, G. Conti, C. Gambaruto, L. Silvestro, and R. Coppo, “Glucose degradation products increase apoptosis of human mesothelial cells,” *Nephrol. Dial. Transplant.*, vol. 18, no. 4, pp. 677–688, Apr. 2003.
[432] M. P. Catalan, D. Subira, A. Reyero, R. Selgas, A. Ortiz-Gonzalez, J. Egido, and A. Ortiz, “Regulation of apoptosis by lethal cytokines in human mesothelial cells,” *Kidney Int*, vol. 64, no. 1, pp. 321–330, Jul. 2003.
[433] A. H. Yang, J. Y. Chen, Y. P. Lin, T. P. Huang, and C. W. Wu, “Peritoneal dialysis solution induces apoptosis of mesothelial cells,” *Kidney Int*, vol. 51, no. 4, pp. 1280–1288, Apr. 1997.
[434] R. I. Morimoto and M. G. Santoro, “Stress-inducible responses and heat shock proteins: New pharmacologic targets for cytoprotection,” *Nat Biotech*, vol. 16, no. 9, pp. 833–838, Sep. 1998.
[435] S. M. Keyse and E. A. Emslie, “Oxidative stress and heat shock induce a human gene encoding a protein-tyrosine phosphatase,” *Nature*, vol. 359, no. 6396, pp. 644–647, Oct. 1992.
[436] H. M. Beere and D. R. Green, “Stress management - heat shock protein-70 and the regulation of apoptosis,” *Trends Cell Biol.*, vol. 11, no. 1, pp. 6–10, Jan. 2001.
[437] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, and H. Hermjakob, “The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases,” *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D358–363, Jan. 2014.
[438] “Reactome Pathway Database.” [Online]. Available: http://www.reactome.org/. [Accessed: 11-Aug-2014].
[439] M. Milacic, R. Haw, K. Rothfels, G. Wu, D. Croft, H. Hermjakob, P. D’Eustachio, and L. Stein, “Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome,” *Cancers*, vol. 4, no. 4, pp. 1180–1211, Nov. 2012.

[440] K. Moenks, I. Muehlberger, A. Bernthaler, R. Fechete, P. Perco, R. Freund, A. Lukas, and B. Mayer, “Computational reconstruction of protein interaction networks,” in *Applied Statistics for Network Biology: Methods in Systems Biology*, Wiley-VCH, 2011, pp. 155–178.

[441] R. Fechete, A. Heinzel, J. Söllner, P. Perco, A. Lukas, and B. Mayer, “Using Information Content for Expanding Human Protein Coding Gene Interaction Networks,” *Journal of Computer Science & Systems Biology*, vol. 6, no. 2, pp. 073–082, 2013.

[442] P. Mayer, B. Mayer, and G. Mayer, “Systems biology: building a useful model from multiple markers and profiles,” *Nephrol. Dial. Transplant.*, vol. 27, no. 11, pp. 3995–4002, Nov. 2012.

[443] A.-L. Barabasi, “The network takeover,” *Nat Phys*, vol. 8, no. 1, pp. 14–16, Jan. 2012.

[444] A.-L. Barabási, R. Albert, and H. Jeong, “Mean-field theory for scale-free random networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 272, no. 1–2, pp. 173–187, Oct. 1999.

[445] M. Oti and H. G. Brunner, “The modular nature of genetic diseases,” *Clin. Genet.*, vol. 71, no. 1, pp. 1–11, Jan. 2007.

[446] M. Wiesinger, B. Mayer, P. Jennings, and A. Lukas, “Comparative analysis of perturbed molecular pathways identified in in vitro and in vivo toxicology studies,” *Toxicology in Vitro*, vol. 26, no. 6, pp. 956–962, Sep. 2012.

[447] R. Fechete, A. Heinzel, P. Perco, K. Mönks, J. Söllner, G. Stelzer, S. Eder, D. Lancet, R. Oberbauer, G. Mayer, and B. Mayer, “Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy,” *PROTEOMICS - Clinical Applications*, vol. 5, no. 5–6, pp. 354–366, Jun. 2011.

[448] P. Khatri and S. Draghici, “Ontological analysis of gene expression data: current tools, limitations, and open problems,” *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, Sep. 2005.

[449] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero, “A systems biology approach for pathway level analysis,” *Genome Res.*, vol. 17, no. 10, pp. 1537–1545, Oct. 2007.

[450] Z. Zhou, “Using Expansion Alorithm to Identify Protein from Biological Networks,” *Journal of Convergence Information Technology*, vol. 6, no. 3, pp. 63–67, Mar. 2011.

[451] H. J. Baelde, M. Eikmans, P. P. Doran, D. W. P. Lappin, E. de Heer, and J. A. Bruijn, “Gene expression profiling in glomeruli from human kidneys with diabetic nephropathy,” *Am. J. Kidney Dis.*, vol. 43, no. 4, pp. 636–650, Apr. 2004.

[452] C. D. Cohen, M. T. Lindenmeyer, F. Eichinger, A. Hahn, M. Seifert, A. G. Moll, H. Schmid, E. Kiss, E. Gröne, H.-J. Gröne, M. Kretzler, T. Werner, and P. J. Nelson, “Improved elucidation of biological processes linked to diabetic nephropathy by single probe-based microarray data analysis,” *PLoS ONE*, vol. 3, no. 8, p. e2937, 2008.

[453] Y. Kienast and F. Winkler, “Therapy and prophylaxis of brain metastases,” *Expert Rev Anticancer Ther*, vol. 10, no. 11, pp. 1763–1777, Nov. 2010.

[454] M. Preusser, D. Capper, A. Ilhan-Mutlu, A. S. Berghoff, P. Birner, R. Bartsch, C. Marosi, C. Zielinski, M. P. Mehta, F. Winkler, W. Wick, and A. von Deimling, “Brain metastases: pathobiology and emerging targeted therapies,” *Acta Neuropathologica*, vol. 123, no. 2, pp. 205–222, Feb. 2012.

[455] L. E. Gaspar, M. P. Mehta, R. A. Patchell, S. H. Burri, P. D. Robinson, R. E. Morris, M. Ammirati, D. W. Andrews, A. L. Asher, C. S. Cobbs, D. Kondziolka, M. E. Linskey, J. S. Loeffler, M. McDermott, T. Mikkelsen, J. J. Olson, N. A. Paleologos, T. C. Ryken, and S. N. Kalkanis, “The role of whole brain radiation therapy in the management of newly

diagnosed brain metastases: a systematic review and evidence-based clinical practice guideline," *J. Neurooncol.*, vol. 96, no. 1, pp. 17–32, Jan. 2010.

[456] S. N. Kalkanis, D. Kondziolka, L. E. Gaspar, S. H. Burri, A. L. Asher, C. S. Cobbs, M. Ammirati, P. D. Robinson, D. W. Andrews, J. S. Loeffler, M. McDermott, M. P. Mehta, T. Mikkelsen, J. J. Olson, N. A. Paleologos, R. A. Patchell, T. C. Ryken, and M. E. Linskey, "The role of surgical resection in the management of newly diagnosed brain metastases: a systematic review and evidence-based clinical practice guideline," *J. Neurooncol.*, vol. 96, no. 1, pp. 33–43, Jan. 2010.

[457] M. P. Mehta, N. A. Paleologos, T. Mikkelsen, P. D. Robinson, M. Ammirati, D. W. Andrews, A. L. Asher, S. H. Burri, C. S. Cobbs, L. E. Gaspar, D. Kondziolka, M. E. Linskey, J. S. Loeffler, M. McDermott, J. J. Olson, R. A. Patchell, T. C. Ryken, and S. N. Kalkanis, "The role of chemotherapy in the management of newly diagnosed brain metastases: a systematic review and evidence-based clinical practice guideline," *J. Neurooncol.*, vol. 96, no. 1, pp. 71–83, Jan. 2010.

[458] Y. Kienast, L. von Baumgarten, M. Fuhrmann, W. E. F. Klinkert, R. Goldbrunner, J. Herms, and F. Winkler, "Real-time imaging reveals the single steps of brain metastasis formation," *Nat Med*, vol. 16, no. 1, pp. 116–122, Jan. 2010.

[459] L. Holmgren, M. S. O'Reilly, and J. Folkman, "Dormancy of micrometastases: balanced proliferation and apoptosis in the presence of angiogenesis suppression," *Nat. Med.*, vol. 1, no. 2, pp. 149–153, Feb. 1995.

[460] "Real Time PCR Assays | Life Technologies." [Online]. Available: http://www.lifetechnologies.com/at/en/home/life-science/pcr/real-time-pcr/real-time-pcr-assays.html. [Accessed: 11-Aug-2014].

[461] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method," *Methods*, vol. 25, no. 4, pp. 402–408, Dec. 2001.

[462] A. Ilhan-Mutlu, A. Wöhrer, A. Berghoff, G. Widhalm, C. Marosi, L. Wagner, and M. Preusser, "Comparison of microRNA expression levels between initial and recurrent glioblastoma specimens," *J Neurooncol*, vol. 112, no. 3, pp. 347–354, May 2013.

[463] K. Staufer and O. Stoeltzing, "Implication of Heat Shock Protein 90 (HSP90) in Tumor Angiogenesis: A Molecular Target for Anti-Angiogenic Therapy?," *Current Cancer Drug Targets*, vol. 10, no. 8, pp. 890–897, Dec. 2010.

[464] B. Zhang, D. S. Day, J. W. Ho, L. Song, J. Cao, D. Christodoulou, J. G. Seidman, G. E. Crawford, P. J. Park, and W. T. Pu, "A dynamic H3K27ac signature identifies VEGFA-stimulated endothelial enhancers and requires EP300 activity," *Genome Res.*, vol. 23, no. 6, pp. 917–927, Jun. 2013.

[465] T. A. Baudino, "c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression," *Genes & Development*, vol. 16, no. 19, pp. 2530–2543, Oct. 2002.

[466] U. Leser and F. Naumann, *Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen*. Heidelberg: Dpunkt-Verl., 2007.

[467] C. Goble and R. Stevens, "State of the nation in data integration for bioinformatics," *Journal of Biomedical Informatics*, vol. 41, no. 5, pp. 687–693, Oct. 2008.

[468] J. S. Hamid, P. Hu, N. M. Roslin, V. Ling, C. M. T. Greenwood, and J. Beyene, "Data Integration in Genetics and Genomics: Methods and Challenges," *Human Genomics and Proteomics*, vol. 2009, pp. 1–13, 2009.

[469] P. Soares, R. J. Alves, A. B. Abecasis, C. Penha-Gonçalves, M. G. M. Gomes, and J. B. Pereira-Leal, “inTB - a data integration platform for molecular and clinical epidemiological analysis of tuberculosis,” *BMC Bioinformatics*, vol. 14, no. 1, p. 264, 2013.
[470] V. Mayer-Schönberger, *Big data: a revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt, 2013.
[471] V. Swarup and D. H. Geschwind, “Alzheimer’s disease: From big data to mechanism,” *Nature*, vol. 500, no. 7460, pp. 34–35, Jul. 2013.
[472] J. Kleinberg, “The small-world phenomenon: an algorithm perspective,” 2000, pp. 163–170.
[473] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, “Catastrophic cascade of failures in interdependent networks,” *Nature*, vol. 464, no. 7291, pp. 1025–1028, Apr. 2010.
[474] A. Czaplicka, J. A. Holyst, and P. M. A. Sloot, “Noise enhances information transfer in hierarchical networks,” *Scientific Reports*, vol. 3, Feb. 2013.
[475] D. B. Kell and S. G. Oliver, “Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era,” *Bioessays*, vol. 26, no. 1, pp. 99–105, Jan. 2004.
[476] N. R. Smalheiser, “Informatics and hypothesis-driven research,” *EMBO Reports*, vol. 3, no. 8, pp. 702–702, Aug. 2002.
[477] F. Mazzocchi, “Complexity and the reductionism-holism debate in systems biology,” *Wiley Interdiscip Rev Syst Biol Med*, vol. 4, no. 5, pp. 413–427, Oct. 2012.

CURRICULUM VITAE

PERSONAL INFORMATION

Name	**SIEHS, CHRISTIAN**
Academic Title	**DIPL.-ING. MAG.RER.NAT.**
Address	**KEINERGASSE 20/1/12, A-1030 WIEN**
Telephone	**+4369912554211**
E-Mail	**christian.siehs@gmail.com**
Nationality	Austria
Date of Birth	25th September, 1973, Vienna
Parents	Renate Siehs-Herusch, Karl Siehs

EDUCATION

2004 - 2015	Doctoral study program: Informatics Technical University of Vienna PhD thesis: *Simulation in Metabolic Networks*
1992 – 2004	Study program: Biology, Mayor: Microbiology & Genetics University of Vienna Diploma thesis title: *Compartment Based Biological Computation In the Apoptosis Effector Phase*
1996 – 2004	Study program: Informatics Technical University of Vienna Diploma thesis title: *Compartment Based Biological Computation*
1983 – 1991	Matura (High Scool) Bundesgymnasium und Bundesrealgymnasium Wien VIII A-1080 Wien, Albertgasse 38

LICENCES / CERTIFICATES

11/1999	ECDL European Computer Driving Licence – Trainer Certificate
10/1991	Driving Licence(s) A,B,C,E,F,G

SCIENTIFIC EXPERIENCE

RESEARCH TOPICS

Bioinformatics and Computational Biology:

- Simulations based on cellular automata
- Protein-structure prediction
- Omics data analysis
- Network algorithms and network analysis
- Systems biology
- Data Mining

PUBLICATIONS

2015 Aysegül Ilhan-Mutlu, Christian Siehs, Anna Sophie Berghoff, Gerda Ricken, Georg Widhalm, Ludwig Wagner, Matthias Preusser . Expression profiling of angiogenesis-related genes in brain metastases of lung cancer and melanoma. Tumor Biology, 2015 Aug 16, DOI 10.1007/s13277-015-3790-7

2012 H.-P. FUEHRER, C. SIEHS, R. SCHNEIDER, H. AUER. Morphometrical analysis of Taenia taeniaeformis and Taenia crassiceps in the common vole (Microtus arvalis) and the water vole (Arvicola terrestris) in Vorarlberg, Austria. HELMINTHOLOGIA, 49, 3: 169 – 173, 2012

Söllner J, Mayer P, Heinzel A, Fechete R, Siehs C, Oberbauer R, Mayer B. Synthetic lethality for linking the mycophenolate mofetil mode of action with molecular disease and drug profiles. Mol Biosyst. 2012 Oct 30;8(12):3197-207. doi: 10.1039/c2mb25256b. PubMed PMID: 23014771.

2011 Kratochwill K, Lechner M, Lichtenauer AM, Herzog R, Lederhuber HC, Siehs C, Endemann M, Mayer B, Rizzi A, Aufricht C. Interleukin-1 receptor-mediated inflammation impairs the heat shock response of human mesothelial cells. Am J Pathol. 2011 Apr;178(4):1544-55. doi: 10.1016/j.ajpath.2010.12.034. PubMed PMID: 21435443; PubMed Central PMCID: PMC3078451.

2010 Fuehrer HP, Blöschl I, Siehs C, Hassl A. Detection of Toxoplasma gondii, Neospora caninum, and Encephalitozoon cuniculi in the brains of common voles (Microtus arvalis) and water voles (Arvicola terrestris) by gene amplification techniques in western Austria (Vorarlberg). Parasitol Res. 2010 Jul;107(2):469-73. doi: 10.1007/s00436-010-1905-z. Epub 2010 May 18. PubMed PMID: 20480373.

2009 Kratochwill K, Lechner M, Siehs C, Lederhuber HC, Rehulka P, Endemann M, Kasper DC, Herkner KR, Mayer B, Rizzi A, Aufricht C. Stress responses and conditioning effects in mesothelial cells exposed to peritoneal dialysis fluid. J Proteome Res. 2009 Apr;8(4):1731-47. doi: 10.1021/pr800916s. PubMed PMID: 19231869.

2008 Wolfler, MM; Siehs, C; Schwamborn, K; Otten, D; Knuchel-Clarke, R; Rath, W Proof of Aromatase -Expression in utop endometrium through reverse-phase of protein Arraying in patients with endometriosis GEBURTSH FRAUENHEILK. 2008; 68: S124-S124.

Rapberger R, Perco P, Sax C, Pangerl T, Siehs C, Pils D, Bernthaler A, Lukas A, Mayer B, Krainer M. Linking the ovarian cancer transcriptome and immunome. BMC Syst Biol. 2008 Jan 3;2:2. doi: 10.1186/1752-0509-2-2. PubMed PMID: 18173842; PubMed Central PMCID: PMC2265674.

2006 Perco P, Rapberger R, Siehs C, Lukas A, Oberbauer R, Mayer G, Mayer B. Transforming omics data into context: bioinformatics on genomics and proteomics raw data. Electrophoresis. 2006 Jul;27(13):2659-75. Review. PubMed PMID: 16739231.

2002 Siehs C, Oberbauer R, Mayer G, Lukas A, Mayer B. Discrete simulation of regulatory homo- and heterodimerization in the apoptosis effector phase. Bioinformatics. 2002 Jan;18(1):67-76. PubMed PMID: 11836213.

1999 Siehs, Christian; Mayer, B. (1999) Dynamical Hierarchies of Structure and

Control in Chemical Reaction Networks. Nanotechnology. 10;1999;p. 464-471

POSTER

2008 D. Jurczak, A. Heinzel, M. Kutmon, S. Luger, S. Schaller, A. Schönegger, V. Hoenninger, R. Kofler, C. Siehs, RNA StructVis - A RANA Visualizer, ISMB 2008, Toronto

2007 A. Bernthaler, P. Perco, R. Rapberger, M. Haiduk, J. Söllner, C. Siehs, C. Pleban, M. Wiesinger, C. Stadler, A. Lukas and B. Mayer, dynaNET: A novel computational framework for deriving functional context from omics data, ISMB & ECCB 2007 Vienna

M. Lechner, K. Kratochwill, M. Endemann, C. Siehs, K. Herkner, B. Mayer, C. Aufricht, A. Rizzi, Analysis of the stress response of mesothelial cells to peritoneal dialysis fluid, HUPO 2007, Seoul

2006 C. Siehs, P. Perco, R. Rapberger, A. Lukas, R. Grohmann, M. Krainer, B. Mayer, Autoantigenes and differential gene expression in ovarian cancer, RECOMB 2006, Venice

PROCEEDINGS

2008 Frohner, Matthias; Urbauer, Philipp; Bauer, Martin; Schmidt, Johannes; Escorihuela Navarro, Ana; Siehs, Christian (2008) Applying standardized communication on personal health and sports devices Tagungsband der eHealth und eHealth Benchmarking / 29.-30. Mai 2008 / Wien

TALKS

2014 Introduction talk: "defining bioinformatics", bioinformatics round table discussion meets ÖGMBT, ÖGMBT annual meeting, September 15th 2014, Vienna, Austria

INTERVIEWS

2013 Bioinformatic Round Table Discussion, International Innovation, Oct. 2013 [Online] http://www.research-europe.com/magazine/HEALTHCARE2/EF25/index.html [Accessed October 18th 2014]

WORK EXPERIENCE

10/2014 – 07/2015 **Ludwig Boltzman Gesellschaft, Vienna**
Intellectual Capital Report database and analysis software development

03/2013 – 10/2013 **EUCODIS Bioscience GmbH, Vienna**
Website development

01/2010 – 12/2013 **Department of Internal Medicine IV (Nephrology and Hypertension), Medical University Innsbruck, Austria**

	Management team of the SysKid Project
10/2008 – 01/2010	**University of Applied Sciences Campus Vienna** Lecturer: • Computational Mathematics
09/2005 – 12/2007	**University of Vienna, Institute of Analytical Chemistry** phD Student • FWF Project: *Mesothelial Cell Stress Response and Cytoprotection in Peritoneal Dialysis*
09/2004 – NOW	**University of Applied Sciences Technikum Vienna** Institute of Biomedical Sciences Lecturer: • Bioinformatics • Informatics of Biological Systems • Databases and Data Mining • Structured Programming in Biomedical Sciences • Programming, Algorithms and Data Structures • Programming, Algorithms and Data Structures 2 • Programming, Algorithms and Data Structures 3 • Information Management in Medicine • Introduction into Medical Imaging and Data Management • Writing Biomedical Research Papers and Reports • Healthcare Telematics and Rehabilitation Technology Project • Warm-Up Excel • Project Coaching
03/2004 – NOW	**University of Applied Sciences Upper Austria, Hagenberg** Lecturer: • Script Programming Languages • Web Technologies • Algorithms and Data Structures • Software Project Coaching • Databases, SQL • XML
02/2004 – 12/2009	**Faltl & Krisch Immobilienverwaltung OHG, Vienna** IT Consultant
02/2002 – NOW	**emergentec biodevelopment GmbH, Vienna** Bioinformatics and IT Consultant
08/2000 – 09/2000	**Comit GmbH, Vienna** IT Consulting in Data Mining and Knowledge Discovery
02/1999 – 08/1999	**Univ. Prof. Dr. R. Oberbauer, Univ. Clinic for Internal Medicine III, Department of Nephrology, Vienna**

	Bioinformatics experts: • Structural analysis of the leader peptides within the project: *Blockade von Bcl-2 in vitro*
06/1997 – 11/2000	**Wiener Krankenanstalten Verbund, Informatik im Gesundheitsverbund** Lecturer: • Microsoft Office
01/1996 – 06/1996	**Philips Communication & Processing Services GmbH, Wien** Computer Technician • Network Administration • PC installation and maintenance • Support
05/1994 – 06/1993	**DO & CO Party Service & Catering GmbH, Vienna** Truck driver
11/1991 – 12/1991	**Verband Niederösterreichischer Winzer** Administration Assistant
07/1990 08/1989	**BILLA AG** Shop Assistant
MILITARY SERVICE	
01/1991 – 08/1991	Wilkhelmskaserne, Vienna
PERSONAL SKILLS AND COMPETENCES	
MOTHER TONGUE	**German**
OTHER LANGUAGES	**English** • Reading: excellent • Writing: good • Verbal: good **Latin** • Reading: Basic
HOBBIES AND INTERESTS	Sports: Karate, Aikido, Biking, Running Team Computer Games Nature Sciences, Technology, Psychology, Philosophy, Human Beings
REFERENCES	• Univ. Doz. Dr. rer. nat. Bernd Mayer, bernd.mayer@emergentec.com • Univ. Prof. Dr. Rainer Oberbauer, rainer.oberbauer@meduniwien.ac.at • Ao.Univ.-Prof. Mag.rer.nat. Dipl.-Ing. Dr.techn. Rudolf Freund, rudi@emcc.at