Global functional association network inference and crosstalk analysis for pathway annotation

Christoph Ogris

Academic dissertation for the Degree of Doctor of Philosophy in Biochemistry towards Bioinformatics at Stockholm University to be publicly defended on Friday 20 October 2017 at 13.00 in Magnélisalen, Kemiska övningslaboratoriet, Svante Arrhenius väg 16 B.

Abstract

Cell functions are steered by complex interactions of gene products, like forming a temporary or stable complex, altering gene expression or catalyzing a reaction. Mapping these interactions is the key in understanding biological processes and therefore is the focus of numerous experiments and studies. Small-scale experiments deliver high quality data but lack coverage whereas high-throughput techniques cover thousands of interactions but can be error-prone. Unfortunately all of these approaches can only focus on one type of interaction at the time. This makes experimental mapping of the genome-wide network a cost and time intensive procedure. However, to overcome these problems, different computational approaches have been suggested that integrate multiple data sets and/or different evidence types. This widens the stringent definition of an interaction and introduces a more general term - functional association.

FunCoup is a database for genome-wide functional association networks of *Homo sapiens* and 16 model organisms. FunCoup distinguishes between five different functional associations: co-membership in a protein complex, physical interaction, participation in the same signaling cascade, participation in the same metabolic process and for prokaryotic species, co-occurrence in the same operon. For each class, FunCoup applies naive Bayesian integration of ten different evidence types of data, to predict novel interactions. It further uses orthologs to transfer interaction evidence between species. This considerably increases coverage, and allows inference of comprehensive networks even for not well studied organisms.

BinoX is a novel method for pathway analysis and determining the relation between gene sets, using functional association networks. Traditionally, pathway annotation has been done using gene overlap only, but these methods only get a small part of the whole picture. Placing the gene sets in context of a network provides additional evidence for pathway analysis, revealing a global picture based on the whole genome.

PathwAX is a web server based on the BinoX algorithm. A user can input a gene set and get online network crosstalk based pathway annotation. PathwAX uses the FunCoup networks and 280 pre-defined pathways. Most runs take just a few seconds and the results are summarized in an interactive chart the user can manipulate to gain further insights of the gene set's pathway associations.

Keywords: *biological networks, genome wide functional association networks, global gene association networks, gene networks, protein networks, functional association, functional coupling, network biology pathway analysis, pathway annotation, pathway enrichment, network-based enrichment, enrichment.*

Stockholm 2017
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-146703

ISBN 978-91-7649-950-4
ISBN 978-91-7649-951-1

Department of Biochemistry and Biophysics

Stockholm University, 106 91 Stockholm

GLOBAL FUNCTIONAL ASSOCIATION NETWORK INFERENCE AND CROSSTALK ANALYSIS FOR PATHWAY ANNOTATION

Christoph Ogris

Stockholm
University

Global functional association network inference and crosstalk analysis for pathway annotation

Christoph Ogris

ISBN print 978-91-7649-950-4
ISBN PDF 978-91-7649-951-1

Printed in Sweden by Universitetsservice US-AB, Stockholm 2017
Distributor: Department of Biochemistry and Biophysics

Abstract

Cell functions are steered by complex interactions of gene products, like forming a temporary or stable complex, altering gene expression or catalyzing a reaction. Mapping these interactions is the key in understanding biological processes and therefore is the focus of numerous experiments and studies. Small-scale experiments deliver high quality data but lack coverage whereas high-throughput techniques cover thousands of interactions but can be error-prone. Unfortunately all of these approaches can only focus on one type of interaction at the time. This makes experimental mapping of the genome-wide network a cost and time intensive procedure. However, to overcome these problems, different computational approaches have been suggested that integrate multiple data sets and/or different evidence types. This widens the stringent definition of an interaction and introduces a more general term - functional association.

FunCoup is a database for genome-wide functional association networks of *Homo sapiens* and 16 model organisms. FunCoup distinguishes between five different functional associations: co-membership in a protein complex, physical interaction, participation in the same signaling cascade, participation in the same metabolic process and for prokaryotic species, co-occurrence in the same operon. For each class, FunCoup applies naïve Bayesian integration of ten different evidence types of data, to predict novel interactions. It further uses orthologs to transfer interaction evidence between species. This considerably increases coverage, and allows inference of comprehensive networks even for not well studied organisms.

BinoX is a novel method for pathway analysis and determining the relation between gene sets, using functional association networks. Traditionally, pathway annotation has been done using gene overlap only, but these methods only get a small part of the whole picture. Placing the gene sets in context of a network provides additional evidence for pathway analysis, revealing a global picture based on the whole genome.

PathwAX is a web server based on the BinoX algorithm. A user can input a gene set and get online network crosstalk based pathway annotation. PathwAX uses the FunCoup networks and 280 pre-defined pathways. Most runs

take just a few seconds and the results are summarized in an interactive chart the user can manipulate to gain further insights of the gene set's pathway associations.

To my family.

List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: **FunCoup 3.0: database of genome-wide functional coupling networks**
Thomas Schmitt, Christoph Ogris, Erik LL Sonnhammer. *Nucleic Acids Research*, **42**, (D1), D380-D388.
DOI: 10.1093/nar/gkt984

PAPER II: **FunCoup 4: new species, data, and visualization**
Christoph Ogris†, Dimitri Guala†, Mateusz Kaduk†and Erik LL Sonnhammer. *Submitted*

PAPER III: **A novel method for crosstalk analysis of biological networks: improving accuracy of pathway annotation**
Christoph Ogris, Dimitri Guala, Thomas Helleday, Erik LL Sonnhammer. *Nucleic Acids Research*, **45**, (2), e8.
DOI: 10.1093/nar/gkw849

PAPER IV: **PathwAX: a web server for network crosstalk based pathway annotation**
Christoph Ogris, Thomas Helleday, Erik LL Sonnhammer. *Nucleic Acids Research*, **44**, (W1), W105-W109.
DOI: 10.1093/nar/gkw356

†Contributed equaly

Contents

1. Introduction

The term bioinformatics was coined in the late 1980s with the introduction of the first whole genome sequencing techniques (Hogeweg, 2011). The field was driven to develop new algorithms capable of making sense of the biological data. Due to its success and an urge for greater understanding, a lot of more efficient experimental techniques were developed. This started a flood of biological data, confronting bioinformaticians with new challenges. In the fury to make sense of the data new and more specific fields emerged from within bioinformatics. One of those is systems biology.

In the past systems biology was referred to as one of the main prisms needed for answering the holy grail of biological questions - "What is life?"(Vidal, 2009). It attempts to analyze and understand biological processes orchestrated by interacting genes and macro molecules underlying every living organism (Vidal, 2009). Therefore, many studies in the field focus on identifying these interactions at a DNA, RNA or protein level, mapping them using biological networks. To this end, a variety of networks has been proposed, such as genetic interaction networks (Baryshnikova et al., 2013), co-expression networks (Stuart et al., 2003), protein-protein interaction networks (Krogan et al., 2006) or functional association networks (Alexeyenko and Sonnhammer, 2009) (see Figure 1.1). This thesis is focused on the generation and analysis of the last category - global functional association networks.

In global or genome-wide functional association networks, genes, or their products, are represented as nodes and edges as interactions between. Within these network, one is not differentiating between specific interaction, like regulatory interaction or physical binding. Moreover, the interaction type between two genes is also not specified as direct or indirect. Therefore the interaction is named by the more general term functional association. This generalization allows data integration of various different data types, even those being unspecific. The idea is that multiple weak signals are combined, producing strong evidence for a functional association.

The first part of this thesis introduces the theory and prediction of global functional association networks used for the FunCoup framework presented in pa-

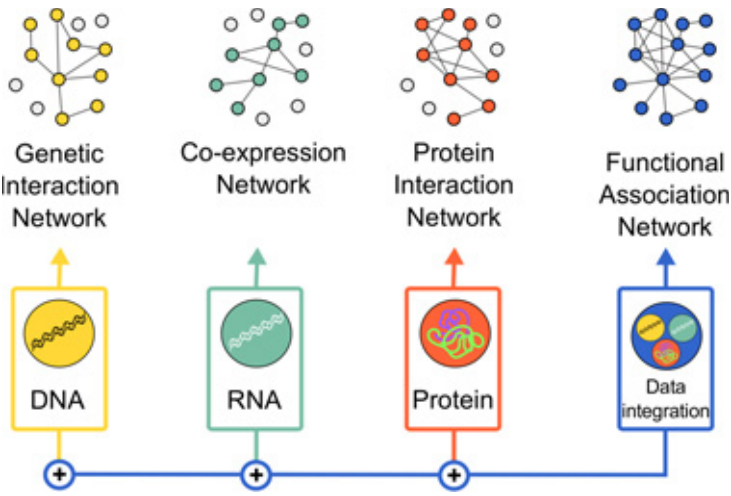

Figure 1.1: Genetic interaction network, co-expression network, protein interaction network and global functional association network - Genetic interaction networks are inferred from DNA level data (yellow). An interaction represents the combination of two or more genetic variants leading to a phenotypic change (Baryshnikova et al., 2013). Co-expression networks are based on biological data measured at a RNA level (turquoise). Here, two genes are connected if they are significantly co-expressed (Stuart et al., 2003). Most protein interaction networks link two proteins if there has been experimental evidence for physical protein interaction (red) between them (Krogan et al., 2006). Global functional association networks integrate data (blue box) measured on DNA, RNA and protein level. This leads to higher coverage, better quality and a more generalized definition of interactions between genes (Schmitt et al., 2014).

per I & II. The latest version of the FunCoup framework uses a variation of the naïve Bayesian classifier to integrate ten different data types across 17 species. FunCoup predicts up to five different networks per species, all representing different types of functional associations, participation in similar metabolic or signaling pathway, protein-protein interaction, protein complex and shared operon organization, for prokaryotic organisms.

As stressed above, complex biological processes are not controlled by single biomolecules. In fact sometimes it requires a cross play of hundreds of genes forming signaling or metabolic pathways. This is why the second part of this thesis focuses on exposing dependencies between pathways and a given set of genes. The most widely used methods rely on the number of shared genes between the two groups disregarding all dependencies between the genes. Paper III proposes the method BinoX using functional association networks as additional evidence, vastly boosting accuracy. BinoX mines information about the connectivity between a gene set and a pathway within the network and compares it to a random network model. In paper IV we present an online back-end of BinoX called pathwAX. PathwAX comes with a pre-selection of pathways and networks and all requisite user input is limited to the gene set of interest. Pre-calculated random network representations increase the speed of the algorithm resulting in an almost instant pathway annotation.

2. Background

2.1 Biological networks

Predicting a biological network on a genome-wide scale from single stand alone experiments is usually error-prone, biased and offers only incomplete snapshots (Lees et al., 2011). To improve coverage and quality, the key solution is an integration of different evidence types, like phylogenetic profiles and expression data (Lees et al., 2011; Gerstein et al., 2002; Von Mering et al., 2002). A more specific view can be provided by including regulatory information from protein phosphorylation (Ptacek et al., 2005) or global maps of promoter binding by transcription factors (Birney et al., 2007). The task of integrating and predicting networks of interacting agents has been achieved through various machine learning algorithms, such as linear (Costanzo et al., 2010) or non-linear regression models (Szklarczyk et al., 2014), random forest (Elefsinioti et al., 2011), support vector machines (Lin et al., 2009) and the most widely used, naïve Bayesian models (Lee et al., 2008; Alexeyenko and Sonnhammer, 2009; Wong et al., 2012). All of these methods require a 'gold standard' set, presenting known functionally associated gene pairs. Here, the nature of the pairs defines the context of the functional association. Commonly used gold standards include pairs extracted from KEGG pathways (Kanehisa and Goto, 2000) or selected GO terms (Ashburner et al., 2000). In some cases different gold standards are used either as an attempt to predict process specific associations, or to distinguish between different kinds of associations.

2.1.1 Naïve Bayes classifier

Bayes theorem This theorem is the backbone of the naïve Bayes classifier. Formulated during the late 18th century it builds the foundation of Bayesian statistics used by many modern algorithms. In order to understand how the classification predicts functional association we first need to get an idea how this theorem works. The Bayes' theorem estimates a conditional probability, the *posterior* probability, given a *prior*, *likelihood* and *evidence* probability and can be written in plain text as

$$posterior = \frac{prior \times likelihood}{evidence}. \tag{2.1}$$

Let us assume that we want to estimate the *posterior* probability of a functional association x between two genes, given the data d. Here, the equation 2.1 can be reformulated in mathematical terms as

$$P(x|d) = \frac{P(d|x)P(x)}{P(d)}. \tag{2.2}$$

Where $P(x|d)$ expresses the *posterior* probability and $P(d|x)$ the *likelihood*, which refers to the probability of observing d, given an association. $P(x)$, the *prior*, denotes any prior knowledge of x regardless d and $P(d)$, the *evidence* or normalization constant, is the probability of the data d.

Assumption of independence Assume multiple different data sets d such that $d \in D$. For example, D can represent different DNA microarray expression experiments and each single experiment would be defined as d. If all experiments d are naïve (strong) conditionally independent from each other, one can estimate the *likelihood* of observing D given x can be calculated through

$$P(D|x) = \prod_{d \in D} P(d|x). \tag{2.3}$$

Classification The naïve Bayes classifier uses the assumption of independence to integrate multiple data sets. Based on those, the classifier decides if a functional association is more likely to be present, x, or not, $\neg x$. To do so the algorithm applies the Bayes' theorem twice, estimating a ratio between the two mutual exclusive *posterior* probabilities, $P(x|D)$ and $P(\neg x|D)$, so that

$$\frac{P(x|D)}{P(\neg x|D)} = \frac{\frac{p(x)\prod_{d \in D} P(d|x)}{P(D)}}{\frac{p(\neg x)\prod_{d \in D} P(d|\neg x)}{P(D)}} \tag{2.4}$$

which can be simplified as,

$$\frac{P(x|D)}{P(\neg x|D)} = \frac{P(x)}{P(\neg x)} \prod_{d \in D} \frac{P(d|x)}{P(d|\neg x)}. \tag{2.5}$$

Typically, equation 2.5 is expressed in a logarithmic format, replacing the product with a sum, preventing problems which might occur due to multiplying small numbers, as follows

$$ln\frac{P(x|D)}{P(\neg x|D)} = ln\frac{P(x)}{P(\neg x)} + \sum_{d \in D} ln\frac{P(d|x)}{P(d|\neg x)}. \tag{2.6}$$

Log-prior ratio The first term on the right of the equation 2.6 is the *prior* or log-prior ratio c.

$$c = ln\frac{P(x)}{P(\neg x)} \tag{2.7}$$

As mentioned previously, the *prior* is the probability of observing a functional association in a given organism regardless of the data D. Therefore, it can be calculated through an approximation of the total number of functional associations within a species.

Log-likelihood ratio In the equation 2.6 the sum of all data points' log-likelihood ratios, LLR_x, is summarized with the log prior ratios to estimate the log posterior ratio. For many studies only the *LLR* values are used, disregarding prior and posterior. *LLR* values might be counterintuitive but are often used as a score to rank or weight predicted links,

$$LLR_x = \sum_{d \in D} ln\frac{P(d|x)}{P(d|\neg x)} \tag{2.8}$$

Probabilistic scoring To provide an easy, interpretable score, one can transform LLR_x using c to estimate a probabilistic functional score pfc within the range $[0,1]$ (Abatangelo et al., 2009; Alexeyenko and Sonnhammer, 2009).

$$pfc_x = \frac{1}{1+e^{-c-LLR_x}} \tag{2.9}$$

2.1.2 Network properties

Terminology In the field of mathematics, a network is referred to as graph composed of nodes or vertices, which are connected via edges. Depending on the biological network these nodes and edges can represent different elements of a cell/organism. For example, protein interaction networks use proteins as nodes and the edges represent physical interaction (Krogan et al., 2006). In contrast, transcriptional networks have nodes representing transcription factors and transcriptional regulation as edges in between(Lee et al., 2002).

Topology The network topology refers to the highly complex organization and structure of nodes within a graph. Almost all network properties aim to determine and quantify the topology for further analysis.

Density The density of a network is defined by the number of edges relative to the maximum possible number of edges. If every node is connected to each other the network is called full or complete, whereas low density graphs like biological networks are usually sparse networks.

Edge properties Edges can be directed and weighted. Protein-protein interaction networks have been proposed without weight and direction (Krogan et al., 2006) whereas transcriptional networks include direction and weight (Lee et al., 2002). In the case of functional association network, an undirected edge with weight represents the believe in an association (Schmitt et al., 2014).

Node degree and node distribution Another frequently used graph property is the node degree, k. The node degree refers to the number of connected adjacent nodes. If the network is directed, the node degree is further divided in in-degree for incoming edges and out-degree for edges leaving a node. A degree distribution $P(k)$ of all nodes gives information about how likely it is to observe a certain degree. Nodes with a high degree are often referred to as hub nodes.

Clustering coefficient Considering an undirected network, one can evaluate how densely a node v and its neighbors are connected by using the local clustering coefficient,

$$cc(v) = \frac{2n}{k^*(k^*-1)}$$

here k^*, denotes the degree of a node and n the amount of edges between it's neighbors. If $cc = 1$ the node and its neighbors would be fully connected whereas $cc = 0$ forms a star formation. Another formulation of the clustering coefficient evaluates the cc based on the number of triangles, closed triplets, a node builds with his neighbors in relation to the total number of possible triangles. Following this one can also formulate a global clustering coefficient in a graph G as

$$cc(G) = \frac{t}{\hat{t}}$$

Here $cc(G)$ determines the ratio between t, the number of closed and $\hat{t}$, the number of closed and open triangles.

Betweenness centrality Measuring the number of steps needed to propagate through a network can be calculated using the betweenness centrality. For a

certain node v the betweenness centrality is defined as

$$b(v) = \sum_{n \neq v \neq m} \frac{\sigma_{nm}(v)}{\sigma_{nm}}$$

the shortest path through the network, between node n to node m, is defined as σ_{nm}. $\sigma_{nm}(v)$ defines the number of shortest paths crossing node v.

Scale-free networks If a network contains a few hub nodes and lots of lower connected nodes, the network is defined as scale-free. Most biological networks are scale-free and their node distribution which follows a power law distribution,

$$P(k) \sim k^{-\gamma}$$

For biological networks the exponent γ lies within the range of $2 < \gamma < 3$ (Barabasi and Oltvai, 2004). Networks with this property are defined as scale-free. Therefore, it has a few nodes with a high degree, called hub nodes, and a lot of nodes with low degree. Another notable property of scale-free networks is their attack tolerance. Removing random nodes from a scale-free network will only have marginal effects whereas deleting specific hubs with high betweenness centrality causes the network to collapse (Maslov and Sneppen, 2002).

Small world networks In small world networks every node can be reached within a few steps by every other node in the network (Travers and Milgram, 1967). One of the most popular examples of small-world networks are social networks. Biological scale-free networks can also be referred to as ultra-small world (Cohen and Havlin, 2003).

Modularity Graph modules are groups of highly connected nodes, often named communities or subgraphs. To evaluate the modularity one can either use the clustering coefficient or compare the fraction of node degree within a module to the fraction one would expect per chance (Newman and Girvan, 2004).

Assortativity The assortativity defines the trend of nodes being connected to nodes with similar degree. If this is the case, the network is called assortative, whereas disassortative denotes the opposite scenario. Some early studies on small protein-protein interaction networks in yeast suggested that biological networks are disassortative (Maslov and Sneppen, 2002). However, it was shown more recently that genome-wide functional association networks do have assortative properties (McCormack et al., 2013).

Random networks Various methods exist to generate random networks from scratch. A common one is the Barabasi-Albert model which generates a random graph with scale-free degree distribution (Barabási and Albert, 1999). The proposed method iterates over two steps until the right network size is reached. The first step, the growth process, adds a new node with a certain degree to the graph. During the second step, known as preferential attachment, the node gets connected to pre-existing connected nodes so that $P(k) \sim k^{-3}$ holds.

Networks randomization Instead of generating a random network from scratch one can also randomize or shuffle real networks. Most often these methods aim to preserve properties of the original network as good as possible. The four example strategies to randomize large scale networks are node permutation, link permutation, link assignment and link assignment with second-order conservation (McCormack et al., 2013). Node and link permutation swap nodes with similar degree or links. Link assignment starts with an empty network randomly adding links under the constraint of preserving the node degree of the original network. Preserving the degree of the neighbors as well is done by the link assignment with second order conservation.

2.1.3 Evidence types

The performance of every classifier is determined by the underlying data. In the field of network prediction, high quality data is the key element for predicting accurate networks. Most methods apply diverse preprocessing steps to the data before applying the prediction. Usually, this consists of converting the raw experimentally derived data into a quantifiable score. This score is often referred to as the evidence score for a predicted functional association. Evidence of physical protein protein interaction or gene interaction data has been used in the past to generate single evidence networks (Costanzo et al., 2010; Costanzo et al., 2016; Maslov and Sneppen, 2002). As one can imagine, new evidences are emerging in parallel, with advances in experimental biological techniques. However, most evidences have a relatively low score due to high noise level. For predicting functional association networks, a common strategy is integrating a preferably distinct set of evidences. An overview of frequently used evidence types is listed below and visualized in Figure 2.1.

Co-expression If two genes have similar expression patterns over various conditions they are most likely functionally associated, that is the idea of mRNA co-expression (Jansen et al., 2002). Over the past few years the cost-decline of mRNA co-expression experiments caused a flood of public available

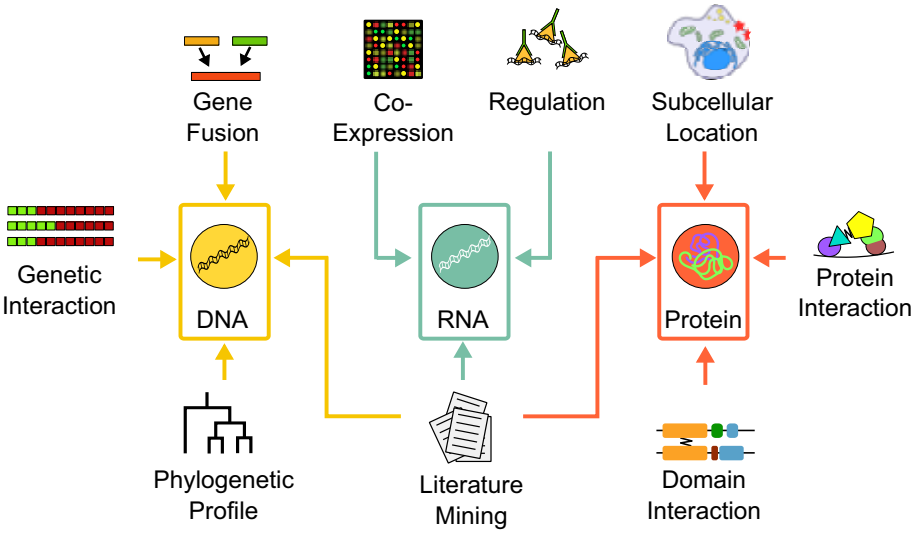

Figure 2.1: Typical evidence types - The evidence types' data relate to experiments or computational techniques evaluating either DNA (yellow), RNA (turquoise) or proteins (red).

expression data sets. The huge amount of public available data might be one of the main reasons why this is one of the most used evidence types for predicting functional associations between genes. Experiments rely on RNA-seq or DNA microarray techniques. DNA microarrays are designed to evaluate the expression of a predefined set of genes. On the other side RNA-seq experiments cover whole transcriptomes. Still DNA microarray experiments are one of the most often used data types for predicting functional association networks. This might be due to its public availability and the standardized format. But, considering current trends in biology as well as the cost efficiency and drawbacks of both techniques RNA-seq will most likely dominate the field in the future.

Protein interaction Physical protein interaction, or protein-protein interaction (PPI), can determine the cell's response to external signals or even alter the structural organizations of the cell. The most common experimental designs to identify protein interactions are yeast two-hybrid, tandem affinity purification, co-crystal structure or affinity capture-Mass Spectrometry. PPI as evidence has by far the highest signal to noise ratio, and so it should be treated with care. It was shown that the outcome of different PPI experiments has various forms of biases towards protein properties like hydrophobicity, nuclear and cytosolic localization, or even length (Jensen and Bork, 2008).

Another option is predicting PPIs on a computational basis only. Here

methods have been developed, which first predict the structures of a protein and then evaluate which other structures match the predicted binding site (Zhang et al., 2012). Another way to use protein structure for predicting functional associations is assuming that similar protein structure also causes similar functionality (Szklarczyk et al., 2017).

Genetic interaction Genes are genetically interacting if two or more genes mutate and cause an unexpected phenotype change. If this change causes reduced fitness or even synthetic lethality the interaction would be labeled as negative whereas positive interaction refer to mutations causing less severe effects than expected (Costanzo et al., 2010). One can also see this as testing and evaluating how the biological network handles cellular buffering of mutations to be more resistant against single gene defects (Boone et al., 2007). Unfortunately, current screening techniques cannot efficiently analyze large genomes like the human genome.

Phylogenetic profiles The fundamental assumption is that two functionally associated genes are more likely to be co-absent or co-present in a set of species than functionally unrelated genes. Predicting functional associations from phylogenetic profiles (PHP) was first introduced by Pellegrini et al. (Pellegrini, 2012). There are several different approaches to solve this problem. Some approaches simply compare all profiles without taking evolutionary events into account (Pellegrini, 2012; Barker and Pagel, 2005). Whereas model-based methods include this information, but are computationally more intense (Barker et al., 2007). Another possibility is to calculate the PHP score with a heuristic algorithm using a species tree to improve the quality and precision of the gained information for predicting networks (Schmitt et al., 2014).

Protein domain interactions Protein domains are structural and/or functional units from proteins (Bork, 1991). Different strategies have been suggested to use this information as a score for predicting function associations. One option is to assign functional associations between two proteins if they have the same domain (Lee et al., 2010). Another option is scoring the functional association depending on their possibility to interact with each other (Alexeyenko and Sonnhammer, 2009).

Gene fusion A gene fusion event takes place if two separate genes get combined via translocation, interstitial deletion, or chromosomal inversion. This can be detected by using orthology for comparing different genomes. The fused genes in general belong to the same functional category, a fundamental

assumption required to use this evidence type in network prediction methods (Yanai et al., 2001; Mering et al., 2003).

Subcellular location Here it is assumed that the different cellular compartments have different biological functions. Therefore, if one can find evidence for two proteins being co-localized, they might also be functionally associated (Huh et al., 2003). Experiments mostly use fluorophore tagging or antibody based labeling to detect co-localized proteins.

Genetic co-location The biological importance of gene co-localization in a genomic context are vastly different between prokaryotes and eukaryotes (Rogozin et al., 2002). In prokaryotes, operons define the genomic structure by clustering genes under the control of a single promoter. Therefore, these genes are typically co-expressed and participate in similar processes of the cell. Though it has been shown that operons exist in a few eukaryotic organisms, more complex types of gene regulations are usually needed to control and define cell functions. Nevertheless, it was shown that even in eukaryotes one can sometimes link gene functions to the genetic order of genes (Zhang and Smith, 1998; Lopez et al., 2010).

Co-regulation Two common gene regulation mechanisms used for predicting functional associations are transcriptional co-regulations of genes by similar transcription factor binding sites or post-transcriptionally interference by RNA.

Literature mining As one can imagine that the flood of experimental data leads to a flood of articles. In text mining the score for a functional association is based on the frequency of two genes co-occurring within the texts (Szklarczyk et al., 2014). What seems trivial often has major drawbacks. Beside the contradicting aspect of the English language, these methods also suffer through changing gene names, different gene identifiers and also for being bound to publicly available text alone.

2.1.4 Gold standards

The gold standard, or training set, is a set of data with known conditions used to train supervised machine learning algorithms to further predict unknown conditions in future data. Here many methods require binary conditioned data, like being functionally associated or not, which is typically referred to as positive and negative sets. As with the evidence data, the amount and quality of gold standard data directly impacts the quality of the predictions. Here, the

following rule of thumb usually applies - the bigger the coverage the better the results.

But the gold standard not only defines the quality of a prediction, it also defines how a functional association can be interpreted. In the past, networks have been created using a variety of different gold standards. For instance, tissue specific sets reveal differences of functional associations across multiple tissues types (Greene et al., 2015; Kotlyar et al., 2016). Some studies also used physical protein associations derived from PPI experiments or high quality protein complexes (Rudashevskaya et al., 2016) and other methods use the gold standards to define different network classes on a global scale to predict metabolic networks and signaling networks (Alexeyenko and Sonnhammer, 2009).

Positive sets For the positive gold standard sets, the known associations are mostly derived from pathway databases like Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa and Goto, 2000) or Gene Ontology (GO) (Ashburner et al., 2000) (see section 2.2.2). Genes are assumed to be functionally associated if they appear in the same pathway. Datasets on PPI and protein complexes can be found on databases like BioGrid (Stark et al., 2006), which gathers and curates experimentally derived protein interactions.

Negative sets The generalization of functional association helps integrating data for generating genome-wide networks. However, this makes it almost impossible to experimentally derived negative gold standards. So far it seems that only a few sets of non-interacting proteins have been clearly identified (Blohm et al., 2013) but some workarounds have been proposed to still generate negative sets. One way would be taking gene pairs which are annotated to unrelated GO terms or KEGG pathways (Mostafavi et al., 2008). Other methods generate negative sets by choosing random gene pairs, which cannot be found in the positive gold standard set (Alexeyenko and Sonnhammer, 2009).

2.1.5 Orthology

Orthologous genes, orthologs, are genes found in different species, which originated in a common ancestor (Fitch, 1970). Being in different species these genes are most likely to have similar functionality. This spawned the idea to use orthologous genes for unknown function annotation and to identify protein protein interactions in not well studied organisms (Garcia-Garcia et al., 2012; Wong et al., 2012). Besides only transferring functional annotation, it was also shown that orthologs can be used to transfer information about functional associations (Alexeyenko and Sonnhammer, 2009; Lee et al., 2008). In this

context one way of transferring information is a one to one mapping of interactions between species. A more sophisticated approach is the transferring of the data to the target species first and to treat it as a standard data set of the target species (Alexeyenko and Sonnhammer, 2009).

Resources like InParadnoid (Sonnhammer and Östlund, 2014), OMA (Altenhoff et al., 2010) and eggNog (Jensen et al., 2007) provide open access to identified orthologs between selected species.

2.1.6 Methods and databases

A diverse set of functional association networks has been proposed over the years. Beside being based on different computational solutions, every method also uses different gold standards, evidence types and data sets. This prohibits a direct comparisons between the methods. Nevertheless, the most up-to-date methods and their databases are mentioned below.

FunCoup The FunCoup framework of Functional Couplings, or functional associations, is described in detail in paper I and II.

STRINGdb The Search Tool for the Retrival of Interacting Genes/Proteins, currently contains networks for over 2000 different species based (Szklarczyk et al., 2017). The underlying data is gathered from seven different evidence types or so called 'channels' representing gene fusion, experiments, database curation, co-expression, gene neighborhood, co-occurrence and text-mined evidence. Here, the later clearly dominates all other evidence types. The channels are scored, assembled and benchmarked separately. The benchmark procedure provides a probabilistic confidence score for each functional association based on manually curated gold standards derived from KEGG (Szklarczyk et al., 2010).

IMP The IMP database, (Wong et al., 2012) , contains networks of Integrative Multi species Predictions from seven different organisms. To accomplish this the authors used a regularized Bayesian approach (Guan et al., 2008), integrating various experiments covering different tissues and developmental time points. The underlying data is specific for each species. The web interface provides the possibility to apply functional annotation with cross species annotation terms. The terms are transferred between the species using orthology.

IID The Integrated Interactions Database (IID) (Kotlyar et al., 2016) is a tissue specific protein protein interaction database. The database covers six species with up to 30 distinct tissues per species using PPI data from various

databases as well as gene and protein expression. IID uses data obtained from the Human Protein Atlas (Uhlén et al., 2015) and PaxDb (Wang et al., 2012) to assign tissue specificity to protein interactions. IID also uses orthology transfer to map interactions between orthologous proteins across different species.

GIANT The Genome-scale Integrated Analysis of gene Networks in Tissues, short GIANT, was introduce in 2015 by Greene et al. (Greene et al., 2015). The framework uses 144 human tissue and cell-linage specific gold standards to predict functional association networks for each of them. The implemented Bayesian integration combines several thousand data sets from various conditions and experiments, based largely on gene expression, but also including co-regulation, gene and protein interaction data. To improve the performance the integration pipeline automatically identifies and upweights tissue specific data sets.

GeneMANIA At the current stage GeneMANIA incorporates networks for *Homo sapiens* and five model organisms (Montojo et al., 2014) optimized for predicting gene functions. The database incorporates evidence types from subcellular localization, interactions within pathways as well as physical and genetic interactions. Similar to the other methods presented, the network nodes refer to genes or proteins, but the network edges represent co-functionality of two genes. The method predicts networks for each evidence type and applies weights depending on how much the association reflects a certain function. Afterward, a composite function specific network gets generated averaging over previously calculated weights. The gene function network is predicted by using a Gaussian field label propagation algorithm assigning a score to each node reflecting the strength of its association to a certain function (Mostafavi et al., 2008).

2.1.7 Applications

The field of application of functional association networks covers a vast amount of different topics of life science research. Beside supporting experimentally derived results, networks have also been used to study outcomes helping to interpret highly complex systems. Furthermore, the predicted associations have been used as input for sophisticated methods evaluating network properties and setting them within a biological context.

Four use cases of genome-wide functional association networks can be found below.

Network query Biological networks are mostly used by the scientific community as a look up database to find genes/gene sets of interest and identify possible interaction partner. This can help designing experiments, revealing new insight to experimental outcomes or providing additional information for drawing final conclusions (Bhatlekar et al., 2014). Various databases have been designed to provide high usability and easy access to biological networks (see section 2.1.6). Furthermore, tools like Cytoscape (Shannon et al., 2003) have been developed to provide platforms to further analyze and visualize biological networks. An example of using the *Breast Cancer Gene 1* (*BRCA1*) as FunCoup network query can be found in Figure 2.2.

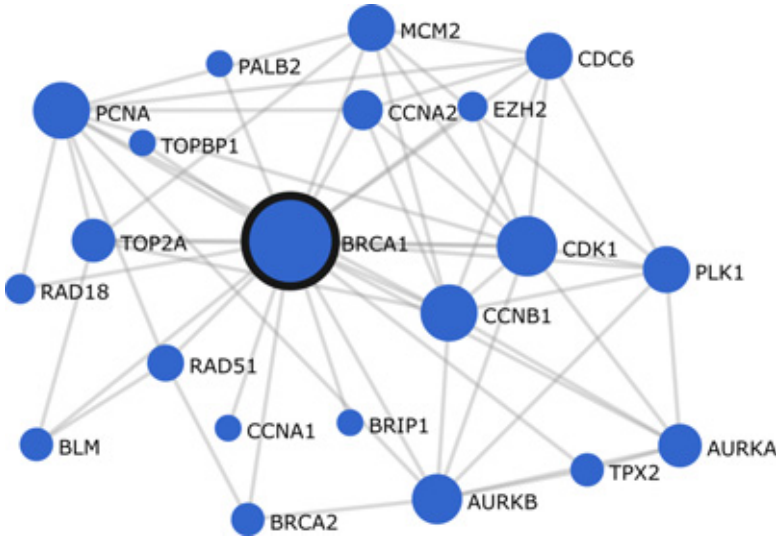

Figure 2.2: Query example using *Breast Cancer Gene 1* (*BRCA1*) - This subnetwork is the result of querying *BRCA1* in FunCoup (Paper II). The circle sizes indicate the genes' node degree (links), within the subnetwork. The interactions displayed are both known and predicted functional associations. An example of a known interaction is *BRCA1* and *BRIP1* (*BRCA1 interacting protein C-terminal helicase 1*), whereas a predicted interaction is *TOP2A* (*DNA topoisomerase II alpha*) and *BRCA1*. This prediction is likely true, considering a recent study linking *TOP2A* to breast cancer (Şahin et al., 2016).

Pathway annotation Network based pathway annotation uses the networks as additional evidence source to reveal relations between pathways and gene sets. This form of application is described in more detail in chapter 2.2.

Gene prioritization Assuming a set of input disease genes, gene prioritization algorithms yield to identify and rank related genes and candidates within

a biological network. As example, Östlund et al. (Östlund et al., 2010) used a network based gene prioritization method called MaxLink (Östlund et al., 2010; Guala et al., 2014) to detect novel cancer genes. Methods like MaxLink statistically evaluate neighboring genes based on their connectivity and rely on a paradigm called guilt by association (GBA). This paradigm assumes network adjacency is directly related to functional similarity. Other GBA approaches apply a network diffusion principle, were the algorithm propagates though the whole network seeding at the input genes (Guney and Oliva, 2012; Köhler et al., 2008; Lee et al., 2011). Benchmarks show that network diffusion methods have a higher coverage compared to neighborhood methods but are also less accurate (Guala and Sonnhammer, 2017).

Network clustering The motivation for clustering a network is the identification of individual gene modules. Gene modules can form various structures within the network and are assumed to control cellular functions (Hartwell et al., 1999; Tornow and Mewes, 2003). Network clustering is a typical unsupervised machine learning problem. The goal is to separate the network into groups. The identified groups can vary in their meaning depending on the underlying network and the applied clustering technique. A plethora of tools exists to cluster not only biological, but also social networks. One of the most common methods is Markov Clustering for graphs (MCL) (Van Dongen, 2008). It calculates the most probable clusters applying a random walk using a Markov matrix representation of the network. Other methods like MGClus (Merge Gain Clustering) (Frings et al., 2013) determine clusters by maximizing the connections within a cluster while minimizing the connections between them.

2.2 Pathway analysis

Pathway annotation tools are indispensable for the interpretation of a wide range of experiments in life sciences. The task of pathway annotation methods is straight forward - given a set of genes, or signature, identify the most relevant pathways. Even though the task seems simple, designing these methods to detect activated pathways within cells has been the target of several studies in the past and remains an ongoing challenge. Meanwhile, a plethora of algorithms exist to perform pathway analysis. In a review article Khatri et al. (Khatri et al., 2012) grouped the algorithms into three generations of methods: Over-Representation Analysis, Functional Class Scoring and Pathway Topology methods. However, due to recent developments in the field this list should be extended with a fourth generation - the Network Based Analysis (see Figure 2.3). Beside using standard inputs, signature and pathways, Network Based Analysis methods mine genome-wide functional association networks to backing up their analysis result.

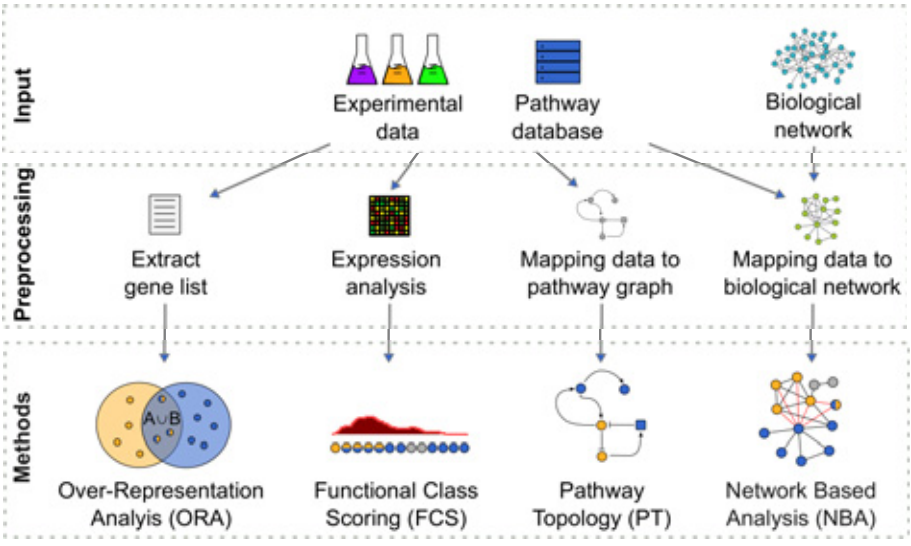

Figure 2.3: Four generations of pathway annotation methods - The first generation, Over-Representation Analysis (ORA) methods, perform set based statistics to assess significant pathway enrichment. Functional Class Scoring (FCS) methods represent the second generation, which use expression analysis results to upweight genes of more importance. Pathway Topology (PT) methods, the third generation, map the experimental data to detailed pathway maps. The latest generation, the Network Based Analysis (NBA), puts the data into context of global functional association networks. Here adjacent network genes or connections between the genes are evaluated to derive the statistical significance of a pathway being enriched.

2.2.1 Biological pathways

A cascade of interacting molecules creating a product or a change in a cell are commonly defined as biological pathway. Pathways cause structural cell changes, influence cellular signaling, vary gene expression and can even induce cell death; Other common properties are multiple entry and exit points of pathways. Therefore pathways can interact with each other on various levels. Furthermore pathway processes can adapt to cell states dynamically and are therefore not static processes. All of these properties have to be considered and increase the complexity of pathway analysis. However biological pathways can typically be divided into three groups - metabolic, signaling and genetic pathways (National Human Genome Research Institute (NIH), 2015).

Metabolic pathway These pathways are defined as a sequence of chemical reactions within a cell. Here metabolites are the intermediates of the reactions catalyzed by enzymes. Further a product of an enzymatic reaction or metabolic pathway could serve again as a substrate for another reaction.

Signaling pathway Also known as signal transduction, signaling pathways describe the biochemical cascade of physical or chemical signals transmitted through a cell. Here one distinguishes between first and second messengers. First messengers are molecules like hormones binding to the cell membrane, whereas second messengers are chemical relays which carry out the intracellular signal.

Genetic pathways These pathways can influence gene activation, gene silencing and influence mRNA and protein expression. This is achieved by a variety of molecular regulators and/or other substances. Gene regulatory networks are a common representation of genetic pathways.

2.2.2 Pathway databases

Following the resource Pathguide, there exist over 130 online databases containing over 3 billion pathway entries (Bader et al., 2006). However the most popular and commonly used databases are KEGG (Kanehisa and Goto, 2000), GO (Ashburner et al., 2000), PANTHER (Mi et al., 2017), Reactome (Joshi-Tope et al., 2005) and WikiPathways (Pico et al., 2008) which are described below.

KEGG The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a recourse incorporating databases covering not only pathways but also orthol-

ogy, genes, compounds, reactions, enzymes and many more (Kanehisa and Goto, 2000). All databases are still supported and undergoing constant updates. KEGG's pathway database covers 4973 organism and divides the pathways into 7 distinct groups: *metabolism*, *genetic information processing*, *environmental information processing*, *cellular processes*, *organismal systems*, *human diseases* and *drug development*. However, the focus of KEGG seems to lie within metabolic pathways (Mi et al., 2005). In general the database provides for most pathways a manually curated map which can also include and link to other pathways. Currently *Homo sapiens* is represented with 320 pathway entries.

GO The Gene Ontology (GO) was introduced as a joint effort of different databases to unify and provide a defined, structured and controlled vocabulary for functions of genes, pathways, and their products across all species (Ashburner et al., 2000; Consortium et al., 2015). The functions in GO are annotated via either manual or automated annotation methods. Here one should note that by 2011, 98% of the GO annotations were inferred automatically and are not reviewed by curators (du Plessis et al., 2011). GO categorizes the terms of gene functions in three distinct classes - Molecular functions, cellular component and biological process. The molecular functions class includes biochemical activities of a gene product, the cellular component refers to the place in the cell where the product is active and the biological process defines the biological objective of the gene product. The GO database is structured as directed acyclic graphs where each node represents a term and the edges relations between terms. The directed acyclic graph implies that the database can be separated in different specificity levels. For example, *signal transducer activity* would be a broad high level term, whereas *death receptor agonist activity* or *inactivation of MAPKK activity* are referred to as low level, more specific terms.

PANTHER Protein ANalysis THrough Evolutionary Relationships database, PANTHER, encompasses information about gene function and evolution of 104 organisms (Mi et al., 2017). The main purpose of the PANTHER web service is focused on analyzing new protein sequences and gene lists. The PANTHER database also includes 176 pathways, each backed up by at least three different literature sources.

Reactome The Reactome database includes reactions and pathways covering different classes like DNA replication, matabolic, signalling and many more (Joshi-Tope et al., 2005). The focus lies towards revealing these in the human organism and by 2015 the database covers 43% of all *Homo sapiens* protein

coding genes (Fabregat et al., 2015). However, inferred orthologous reactions for more than 20 other species are available as well. The provided information on Reactome is curated and cross-referenced with other databases like UCSC Genome Browser (Kent et al., 2002), KEGG (Kanehisa and Goto, 2000) or GO (Ashburner et al., 2000).

WikiPathways WikiPathways is a crowd sourced database for biological pathways with the aim of guaranteeing free and open access to their data (Pico et al., 2008). The community-based validation of the resource enabled the gathering of information about over 2300 pathways spreading across 25 species (Kutmon et al., 2015). Based on the idea of Wikipedia, any registered researcher can upload a pathway which was in some way useful for their research.

2.2.3 Over-Representation Analysis

Due to its simplicity Over-Representation Analysis (ORA) methods might be the most popular approaches to estimate pathway enrichment, and already in 2009 a study listed over 60 different ORA methods (Huang et al., 2009). ORA algorithms mostly require the following input: a gene signature i.e. significant differential expressed (DE) genes, a pathway of interest which shares at least one gene with the signature and an approximation of the genome size. Different combinatorial parameters are calculated and can be represented by a contingency table (see Table 2.1). These parameters are then used by a set based approach to evaluate the statistical significance of the amount of genes shared between the signature and pathway set. Here the main assumption is that all genes are independent and equally important for the analysis. As stressed in previous sections, genes within a pathway are often correlated. This obviously contradicts the most basic ideas of modern biology and especially systems biology (Khatri et al., 2012). As a result, ORA methods suffer from high false positive rates and low true positive rates (Gatti et al., 2010; Ogris et al., 2017).

GEA In its most naive form, ORA is often referred to as gene enrichment analysis (GEA) or Fisher exact test. It is based on hyper-geometric distribution to calculate a probability of how likely it is to obtain an overlap larger or equal to the one observed within the genome.

DAVID One of the most popular ORA method is an adaptation of GEA, calculating an EASE-score (Hosack et al., 2003) implemented within the popular tool DAVID (Database for Annotation, Visualization and Integrated Discovery) (Huang et al., 2008). The EASE score aims for lower false positive rate by

	Pathway	¬ Pathway	Row sum
Signature	k	g	n
¬ Signature	k^*	g^*	m
Column sum	K	G	N

Table 2.1: Contingency table - Over-Representation Analysis (ORA) uses this table to calculate the significance of shared/overlapping genes between a gene signature and a pathway. Following the first row k denotes the number of shared genes, g the number of genes within the pathway and n the total genes in the signature. Furthermore k^* refers to the genes in the pathway but not in the signature and g^* the genes in none of them. This results in m being all genes of a genome not in the signature, K the pathway size and G all other genes of the organism following the genome size N

subtracting 1 from the estimated overlap. This eliminates pathway enrichments based on a one gene overlap and increases the p-values of the remaining results.

2.2.4 Functional Class Scoring

Functional Class Scoring (FCS) methods are mainly designed for analyzing expression experiments derived through DNA microarray expression. FCS methods incorporate different gene expression levels instead of assuming that all signature genes are of equal importance. Therefore most FCS methods structure their analysis in three steps which may vary for different methods (Subramanian et al., 2005; Tarca et al., 2012). First, gene-level statistics evaluate the importance of each gene within the tested condition. Second, mean, median or sum of genes within a pathway are calculated representing the pathway-level. Third, assessment of the significance of the pathway-level statistics by comparison to a null-hypothesis. While ORA methods need an arbitrary cutoff to obtain significantly DE genes, FCS methods use all genes available and therefore also account for coordinated expression changes of genes. However, a recent benchmark showed that FCS based methods only have a marginal advantage over ORA tools (Dong et al., 2016).

GSEA Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was designed for analyzing DNA microarray experiments. First GSEA ranks genes within the whole experiment based on the correlation between expression and the experiment condition. Next, an enrichment score is calculated by iterating over the ranked list and comparing it to a pathway. In cases where a gene is also found within the pathway the score increases; otherwise a decrease is detected. The significance of an enrichment score is calculated as a

comparison to random data. For the random data the samples' condition labels are swapped and step 2 and 3 are repeated.

PADOG As most other FCS tools Pathway Analysis with Down-weighting of Overlapping Genes (PADOG) (Tarca et al., 2012) specializes in analyzing gene expression experiments. After computing the gene-level statistics using a t-test, PADOG uses a weighted mean of absolute t-values to calculate the pathway-level statistics. The weights of a gene are calculated depending on how often the gene occurs in other pathways. Therefore genes only present in one pathway are seen as more important/specific than genes participating in several pathways. Finally the significance of a pathway is calculated via null hypothesis testing. Here the null hypothesis is simulated the same way as for GSEA, by randomizing the input condition labels.

2.2.5 Pathway Topology

In contrast to ORA and FCS, Pathway Topology (PT) methods also take structure and dynamical changes of a pathway into account. Instead on using pathways from GO, these methods rely on detailed maps from pathways which can be obtained from resources like KEGG (Kanehisa and Goto, 2000) or Reactome (Khatri et al., 2012). The additional information is used to account for central pathway genes or various interactions between pathway elements. But their biggest strength is also the worst weakness: PT algorithms only work if a detailed, accurate picture of a pathway is known.

IFA The Impact Factor Analysis (IFA) (Draghici et al., 2007; Voichita et al., 2012) was designed specifically for analyzing signaling pathways. Here a pathway is represented in typical graph form, nodes being genes and edges the interaction between. IFA calculates a perturbation factor of a gene within a pathway by summing up its expression change and all other gene expression changes within the pathway taking their type of interaction into respect by using weights (Draghici et al., 2007).

2.2.6 Network Based Analysis

In addition to the standard input, Network Based Analysis (NBA) methods additionally use genome-wide functional association networks as input (Glaab et al., 2012; Alexeyenko et al., 2012; McCormack et al., 2013; Ogris et al., 2017), increasing coverage and power of the analysis. There are mainly two different approaches to harvest these networks for pathway annotation (see Figure 2.4). The most basic solution is to extend the input gene list using

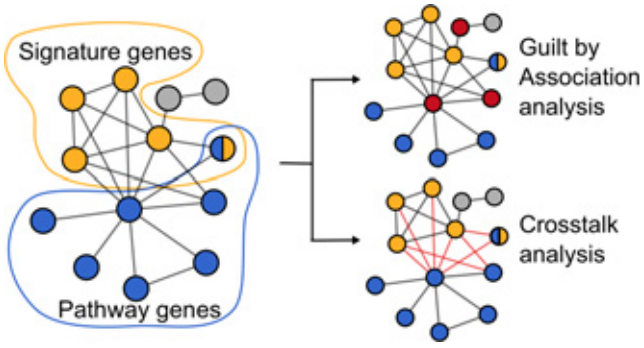

Figure 2.4: Network Based Analysis (NBA) - Network based methods assess the statistical significance of the relation between signature genes (orange nodes) and pathway genes (blue nodes) in the scope of genome-wide functional association networks. One assessment approach applies guilt by association methods to identify network neighbors (red nodes) of the signature set. The extended set is then used as input for over-representation methods. Another approach is to map signature and pathway genes within the network and evaluate the connections (red lines) between them. This is often done via alternative hypothesis testing, comparing the observed amount of connections to a random model.

neighboring network nodes or network diffusion (Dong et al., 2016). Once the list is extended one can use it as input to an ORA analysis technique. This guilt by association approach in combination with GEA is most often a standard feature of network databases like GeneMANIA (Montojo et al., 2014). A more sophisticated way to mine the networks information is to evaluate the network connections rather than shared genes. Using this concept, a pathway is assumed to be activated if a significant amount of network connections, i.e. crosstalk, between a signature and a pathway is given. Moreover using the edges makes it possible not only to estimate enrichment but also to calculate depletion. A depleted pathway would be registered if there is significantly less crosstalk than one would expect by chance. However, the fundamental assumption of NBA methods is that network properties are a product of some biological meaning and not due to random events (Barabasi and Oltvai, 2004). This follows from the fact that the performance of these methods rely on two pillars. First, the quality of the network. Networks with low coverage, no topological meaning and/or random behaviour will not give enough power to calculate statistical significance and will therefore lead to no or only a weak pathway detection. Second, the quality of the statistical model, meaning how well the method can estimate a statistical model, of the genome-wide network, to distinguish random from non-random biological behaviour (Ogris et al., 2017).

EnrichNet EnrichNet (Glaab et al., 2012) is an online service which uses genome-wide functional association network for ranking and identifying enriched pathways. EnrichNet uses a random walk with restart (Yin et al., 2010) approach to estimate a network distance between pathway and signature, taking global as well as local network information into account. Furthermore, the calculated distance is compared to a background model representing the mean distance of all other available pathways. The final score is then correlated with traditional GEA results and presented in a regression plot.

CrossTalkZ This methods evaluates the statistical significance of crosstalk enrichment between signature and pathway using z-score (McCormack et al., 2013). The algorithm creates random instances of the input network by discretizing the edge weights and preserving topological properties. The tool comes with four different randomization techniques: link permutation, node permutation, link assignment and second order link assignment. The z-score then determines how many standard deviations the observed crosstalk is from the crosstalk of the random network model. Since the crosstalk has to follow a normal distribution for using the z-score, the method also performs a chi-square test to evaluate normality, beside estimating p-value and FDR.

NEA Network Enrichment Analysis (NEA) (Alexeyenko et al., 2012) was introduced almost in parallel to CrossTalkZ. NEA is based on the same ideas as CrossTalkZ and uses the z-score to evaluate the significance of discretized edges between signature and pathway. Unfortunately the algorithm is not testing the crosstalk for normality, causing high false positive rates (Ogris et al., 2017).

BinoX The BinoX algorithm evaluates the significance of network crosstalk under the assumption that the underlying network edges are binomial distributed. BinoX is presented in detail in Paper III.

2.2.7 Performance evaluations

Performance measures from the field of machine learning are also commonly used to evaluate the performance of pathway annotation methods. Unfortunately many evaluation measures assume a binary classification problem. A commonly attempted solution is to binarize the results using a significance threshold so that a pathway is either significantly enriched, or not.

Confusion matrix Most measures are based on the confusion matrix. Assuming a test scenario with known conditions, one can create a confusion

matrix by labeling the predicted conditions as true positive (TP), false positive (FP), false negative (FN) and true negative (TN) (see Table 2.2). Once the confusion matrix is established it is easy to spot if an algorithm is biased towards a condition or mixes up prediction classes.

		Known condition	
		Known positive	Known negative
Predicted condition	Predicted positive	True Positive (TP)	False Positive (FP, Type I error)
	Predicted negative	False Negative (FN, Type II error)	True Negative (TN)

Table 2.2: Confusion matrix - Assuming a test scenario with known condition, the columns represent known positive and known negative conditions while the rows refer to predicted positive and predicted negative labels.

Sensitivity The sensitivity, recall or true positive rate (TPR) represents the ratio between true positive cases, TP, and all known positives, $TP+FN$. An algorithm is assumed to be sensitive if $TPR = 1$, meaning all conditioned positive cases are predicted as such. Considering the confusion matrix in Table 2.2, TPR can be formulated as

$$TPR = \frac{TP}{TP+FN}. \tag{2.10}$$

Specificity The specificity or true negative rate (TNR) can be seen as the opposite of the sensitivity. It determines the ration between true negative cases, TN, and all known negatives, $TN+FP$. Another measure directly related to the specificity is the false positive rate (FPR) which can be estimated by $1-TNR$. Using the terminology of Table 2.2 the specificity is determined via

$$TNR = \frac{TN}{TN+FP} \tag{2.11}$$

and the FPR is calculated by

$$FPR = 1-TNR = \frac{FP}{TN+FP}. \tag{2.12}$$

False discovery rate The false discovery rate (FDR) is the ratio between false negatives and all predicted negatives, $FN+TN$. Using the terminology of Table 2.2 the FDR can be defined as

$$FDR = \frac{FN}{FN+TN}. \tag{2.13}$$

Accuracy The accuracy (*ACC*) of a method determines the fraction of correct prediction regardless if it is positive or negative. Using the terminologies of Table 2.2, the *ACC* is defined as

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}. \tag{2.14}$$

Receiver under the operation curve The receiver under the operation curve (ROC) estimates the performance of an algorithm by evaluating the relation between true positives and false positives at various significance threshold. Regarding pathway annotation, the amount of unknown enrichment outweight those known enrichment. Therefore these data sets are called unbalanced. In these cases it is could be useful to apply a partial ROC (pROC) focusing on a fraction of the whole ROC.

Evaluation by rank Evaluation of the rank is not a typical machine learning evaluation procedure but got a common tool to evaluate pathway annotation methods (Tarca et al., 2012; Dong et al., 2016). Here the assumption is that disease specific pathways should be more enriched to the disease than other pathways. Therefore this technique ranks the result by p-value in a first step and evaluates the rank of known diseases in a second step. Sum, mean and median of the observed ranks can then be used to compare different methods.

2.2.8 Multiple comparison problem

The multiple comparisons or multiple testing problem is a statistical problem and appears while testing several null hypothesis simultaneously. It states that the more tests one applies, the higher the chance of including a false positive, also known as type I error. In pathway analysis it is common that several enrichments get tested for significance at once. Therefore the Bonferroni procedure and the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) are applied to the analysis results to counter multiple testing problem.

Bonferroni procedure The Bonferroni procedure uses the family-wise error rate (FWER) which states how likely one is to reject a true positive while accepting at least one false positive. Therefore a Bonferroni procedure corrected p-value is significant if $p - value \leq \frac{\alpha}{m}$ where m is the amount of null hypothesis tested.

Benjamini-Hochberg procedure The Benjamini-Hochberg procedure is a less strict approach and yields to control the false discovery rate (FDR) at a

given significance level α. The procedure sorts all p-values and rejects the null hypothesis if $p - value_i \leq \frac{rank_{p-value_i}}{m} * \alpha$. This approach can also be used to transform p-values to q-values which refer to the number of false positives, taking into account the amount of test performed (Benjamini and Hochberg, 1995).

3. Present investigations

3.1 Paper I & II

In paper I and II we present upgrades of the FunCoup framework/database for inferring functional associations. FunCoup was first introduced by Alexeyenko and Sonnhammer (Alexeyenko and Sonnhammer, 2009) and due to continuous framework optimization, FunCoup can be listed as one of the most comprehensive databases for functional associations. FunCoup currently contains networks for 17 species and differentiates between five classes of functional association: sharing the same metabolic pathways, sharing the same signaling pathways, participation in the same protein complex, evidence for physical protein interaction and for prokaryotic species, co-occurrence in the same operon. Via a Bayesian approach, the network integrates several evidence types, inferring networks for each class separately. FunCoup also uses ortholog based information transfer between species. Here the premise is that orthlogs of a functionally associated gene pair in species A are likely also functionally associated in species B. This approach enables us to achieve high quality as well as high coverage while restricting the input to experimentally derived data.

For release 3, which is described in paper I, we incorporate nine different evidence types: mRNA co-expression, protein co-expression, subcellular co-localization, co-miRNA regulation by shared miRNA targeting, protein interaction, domain interactions, genetic interaction profile similarity, shared transcription factor binding and phylogenetic profile similarity. These evidence types are integrated using a newly developed redundancy weighted Bayesian approach, where previous versions used a naïve Bayesian algorithm which relies on the independence of the underlying data. Since this assumption is violated by some integrated data sets, we introduced this new method to estimate the redundancy between data sets and to weight them accordingly.

Furthermore we introduce a sophisticated heuristic algorithm to determine phylogenetic profile similarity scores. Scores were previously calculated through analysis of every single combination of co-occurence between the ten Fun-

Coup species. In version 3 of FunCoup the score is estimated using a species tree derived from InParanoid v7, including 93 eukaryote species. The positive evidence for the score is derived by summing the branch length of co-conserved genes whereas the negative score is the sum of the branch length with either gene in the species.

Another feature of release 3 is the newly developed web site for FunCoup. Beside the jSquid (Klammer et al., 2008) network viewer, the user also gets a detailed list of the calculated evidence scores for each inferred link and list of enriched cell functions. This front end provides an user-friendly interface to explore the predicted networks.

All of these improvements would not have been possible without a reimplementation of the framework in Java. Previous versions were based on Perl scripts, reaching the language's limits in processing the huge amount of data used in version 2: The new java implementation is structured in modules, which not only makes future upgrades smoother but also provides the basis for increasing the amount of data processed.

In paper II we describe the latest update of FunCoup, release 4. Due to the above described framework improvements we were able to increase the amount of data points by 2.5 fold. FunCoup 4 introduces four new eukaryotic species, *Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Bos taurus*, and *Oryza sativa*. Furthermore we open FunCoup to the prokaryotic domain of life by including networks for *Escherichia coli* and *Bacillus subtilis*. This gave us the opportunity to define a new class of functional association between genes i.e. those organized in the same operon, which is only valid in prokaryotic organisms. We also supplemented the existing classes for metabolic, signaling, complex and protein interaction with up to date information. In this release we switched to InParanoid v8 as the source of orthology and calculated phylogenetic profiles. The three-fold increase of species in InParanoid v8 improved robustness and coverage. While populating all other evidence types with new data we also introduced a new evidence type based on quantitative mass spectrometry data. Here two proteins achieve high functional association scores if they have similar abundance profiles across different conditions.

Additional a new JavaScript based network viewer got incorporated in the FunCoup website which allows platform independent customization of the layout. Gene and link information are readily accessed with mouse clicks such that the tool provides the user an intuitive, responsive platform to further evaluate the query results.

3.2 Paper III

Pathway annotation tools are indispensable for the interpretation of a wide range of experiments in life sciences. Despite the tremendous growth of omics data to help understand complex biological processes and build networks, most studies still rely on pathway annotation algorithms that only consider the gene overlap to known pathways as evidence. Such methods often have very poor accuracy.

Paper III introduces the novel network based pathway annotation method BinoX. BinoX is designed to perform the analysis with vastly better accuracy. As indicated above, up to now, the most popular pathway annotation tools are still the ones based on gene overlap. Over the past few years several new methods have been suggested; however they are essentially limited to the gene overlap, which makes the improvement marginal. Recent network based methods have improved sensitivity and specificity but suffer from problems with the statistical model and excessive compute time.

The featured benchmarks show that BinoX clearly outperforms other algorithms when it comes to true and false positive rate as well as compute time. Paper III proves that BinoX improves pathway annotation by applying it to disease gene sets as well as thousands of experimental gene sets in the MSigDB collection (Liberzon et al., 2011), and compares the results to those of commonly used methods. BinoX's statistical model and implementation are designed for high throughput experiments and big data analysis to accurately and efficiently assess the statistical significance of pathway annotations for thousands of gene sets.

In summary, BinoX represents the next generation of pathway annotation tools greatly increases the likelihood of drawing correct conclusions.

3.3 Paper IV

Paper IV presents pathwAX, an online front end solution for our pathway analysis tool BinoX. PathwAX's database incorporates 1930 KEGG pathways (Kanehisa and Goto, 2000) and preprocessed FunCoup networks for 11 different species. The pipeline is designed to perform computationally heavy parts on the client side while the server is optimized for fast database queries.

PathwAX is simple. After selecting the species of your choice, all it requires is a set of input genes. A typical annotation query takes a few minutes and

visualizes the results within a chart. The chart lists all enriched pathways and aims to provide an intuition of important genes and also a rough overview of the enriched pathway classes. Furthermore the results can be manipulated by applying various filters.

4. Conclusion and outlook

The first half of this thesis introduced global gene association networks as a key strategy to infer a genome-wide picture of interactions between genes and their products, from massive amounts of biological data. The FunCoup database is one of the most comprehensive databases for global gene association networks and currently contains networks for 17 different species. In the past FunCoup networks have shown to reveal novel insights into the interplay of genes in various studies (Bhatlekar et al., 2014; Östlund et al., 2010; Hong et al., 2010; Ogris et al., 2017). In release four of FunCoup, the networks are inferred using a weighted Bayesian approach, integrating experimental data from ten different evidence types, disregarding the controversial evidence type text mining. Furthermore, the framework uses five different gold standards and predicts a network for each independently. Release three came with a major reimplementation of the FunCoup framework which improved the database, optimized the compute time of the network inference and included a newly designed website. This new framework was the foundation for release four. The smart data management enabled updates of the underlying evidence data, gold standards and the possibility for adding new species. This led to a three fold increase of the data volume used by FunCoup 4. As a result, release 4 networks report an increase in terms of coverage and average link strength for most species.

However a shortcoming of FunCoup and other global gene association networks is their incapability of reflecting the highly dynamic nature of biological systems. In these networks links and nodes are static whereas in reality they are time dependent and conditioned on processes or certain tissues. One solution might be to projecting additional process or tissue information on the inferred network filtering for tissue specific nodes and links. Another approach would be to condition the gold standards and input data on specific process or tissue. However both will result in a decreased coverage and even low confidence scores due to lack of data. Another drawback of global gene association networks is the problem of benchmarking different inference methods. One of the main issues is that the definition of the term functional association is highly dependent upon the data and gold standards used to infer the networks. Another breaking point is the fact that the test data should not be part of the data used for prediction. This in particular is very hard since all methods attempt to

gather as much data as possible leading to a circular reasoning. Nevertheless understanding the performance of different methods is a key element to interpret the networks. Therefore future efforts should aim for a joint community approach to investigate benchmark strategies evaluating gene association networks.

Pathway annotation methods are indispensable for the interpretation of a wide range of experiments in life sciences. The second half of this thesis describes a selection of pathway annotation methods with a focus on network based pathway annotation methods. The network based methods have improved sensitivity and specificity upon commonly used methods, but most existing tools suffer from problems with the statistical model and/or excessive compute time. Within the scope of this thesis we developed the state-of-the-art BinoX algorithm. BinoX uses global gene association networks and a binomial approach to assess the statistical significance of crosstalk between a pathway and gene set. At the moment BinoX neglects link weights within the network and treats all input links equally. Information of link strength might give deeper insights to the role of certain interactions and could lead to better performance. Therefore future research should aim to improve BinoX by accounting for different link weights. However, it is shown in paper III that BinoX clearly outperforms other algorithms when it comes to true and false positive rate, as well as compute time. Using the human network of FunCoup 3, BinoX improves pathway annotation by applying it to thousands of experimental gene sets in the MSigDB collection (Liberzon et al., 2011) and comparing the results to those of commonly used methods. BinoX's statistical model and implementation are designed for high throughput experiments and big data analysis to accurately and efficiently assess the statistical significance of pathway annotations for thousands of gene sets.

The most popular pathway annotation tool is probably the DAVID website (Huang et al., 2008) which only considers the gene overlap to known pathways as evidence. According to paper III, this approach carries with it a low detection rate as well as a low true positive and relatively high false positive rate. These facts raise the question: why are so many studies still relying on it? Here the main reasons might be simplicity, compute time and usability. Even though BinoX is optimized for speed it is still a command line tool which requires some Unix knowledge. Therefore we designed a light version of BinoX optimized for usability, the web server pathwAX. Due to major improvements in the computational pipeline and the statistical model it was possible to bring highly sensitive and specific network crosstalk based pathway annotation to the general public. PathwAX encompasses about 280 pathways for 11 species.

The underlying networks are obtained from FunCoup 3. The compute time takes maximal of a few seconds and summarizes the results in an interactive chart. Currently the pathwAX input is restricted to a list of significant differentially expressed genes. Future research will aim to include more species and pathway from various databases. Furthermore one might investigate the advantages of expression data as additional input. However benchmarks show that methods using expression data only have a marginal advantage over using a set of statistically significant genes (Dong et al., 2016).

Sammanfattning

Cellfunktioner styrs av komplexa interaktioner av genprodukter så som bildandet av temporära- eller stabila komplex eller förändring i genuttryck. Kartläggning av dessa interaktioner är nyckeln till att förstå biologiska processer och är således fokus för många experiment och studier. Småskaliga experiment ger högkvalitativa data men saknar en täckande totalbild medan high throughputtekniker täcker tusentals interaktioner men. Alla dessa tillvägagångssätt kan dock enbart fokusera på en typ av interaktion i taget. Detta gör experimentell kartläggning av dessa bilogiska processer till en kostsam och tidskrävande procedur. För att överkomma dessa begränsningar har olika beräkningsmetoder föreslagits vilka integrerar multipla datatyper och / eller olika bevistyper. Dessa beräkningssätt möjliggör en vidgning av den stränga definitionen av en interaktion och introducerar en mer allmän term - funktionell koppling.

FunCoup är en databas över människans och 16 olika modellorganismers nätverk av proteiner och gener som interagerar funktionellt med varandra. FunCoup skiljer på fyra olika funktionella kopplingar: medlemmar av ett proteinkomplex, fysisk interaktion, medlemmar i samma signaleringskaskad och medlemmar i samma metaboliska process. För att förutsäga nya interaktioner för vardera karaktär tillämpar FunCoup naïve Bayesian-integration av tio olika bevistyper av data. Vidare används ortologer för att överföra bevis på interaktion mellan arter. Detta ger en avsevärt bättre täckande totalbild och tillåter tolkning av omfattande nätverk även för icke välstuderade organismer.

BinoX är en ny metod för pathway-analys och bestämning av relationen mellan grupp av gener med hjälp av nätverk av funktionell koppling. Traditionellt har pathway-analys gjorts med endast genöverlappning dock återspeglar dessa metoder enbart en del av helhetsbilden. Genom att placera genuppsättningar i ett nätverksammanhang ges ytterligare bevis för pathway-analys, avslöjar en global bild baserad på hela genomet.

PathwAX är en webbserver baserad på BinoX-algoritmen. Användaren kan mata in en genuppsättning och få ut nätverkets crosstalk-baserade pathway annotation. PathwAX använder FunCoups nätverk och 280 fördefinierade pathways. De flesta körningar tar enbart några sekunder och resultaten sammanfat-

tas i ett interaktivt diagram vilket användaren kan anpassa för att få ytterligare inblick i genuppsättningens pathway-kopplingar.

Acknowledgements

Finally I am here, writing the chapter of my thesis which is often referred to as the one which matters the most. So there is basically no pressure to find the right words thanking all those people with whom I shared so many fun moments and who helped and guided me through rough times.

First of all I would like to thank my supervisor **Erik Sonnhammer** for giving me the opportunity to work as a PhD student in his group. Without your guidance I would not be here today. Your expertise and patience provided a calm environment which was perfect to become an independent scientist. I would also like to thank my co-supervisor, **Thomas Helleday**, for giving me the chance for various collaborations and helping me out in times of need. Also, many thanks to my mentor **Gunar von Heijne** for your help throughout all stages of my PhD. Further thanks go to **Stefan Nordlund** and **Pia Ädelroth** for your comments and guidance.

Furthermore, I thank all current and former members of the DBB administration, especially **Ann Nielsen**, **Malin Kamb**, **Maria Sallander**, **Charlotta Sturell** and **Alexander Tuuling**. I really don't know what PhD students would do without you. Also thanks to the DBB IT **Peter Nyberg**, **Erik Sölund** and **Stefan Fleischman**. Here, special thanks go to Stefan who helped to prevent a server apocalypse on several occasions.

The Sonnhammer group - thanks to all former and current members. **Gabriel Östlund** for being an awesome office mate and gaming buddy. **Thomas Schmitt**, I am glad that you taught me well about your little baby 'FunCoup' and all its facets before you left the group. Thanks **Matthew Studham** for the experimental validation of how to approach a polish wedding. Thanks **Andreas Tjärnberg** for all these enjoyable movie, science and political discussions. Thanks **Mateusz Kaduk** for all of your java and IT support. Thanks **Daniel Morgan** for proofreading my manuscripts and always being motivated for new adventures. Thanks **Stefanie Friedrich** for the interesting fika and lunch talks. Thanks **Dimitri & Izabela Guala** for all the fun gatherings and your friendship throughout my PhD. Also thanks to the rookies **Miguel Castresana** and **Deniz Secilmis** for all the entertainment over the past few month. **Annika Scheynius**, even if you are not a group member, as an office neighbor you got

very close to it. Thanks for all the positive thoughts while writing my thesis.

I would also like to thank all other people at SciLifeLab who I encountered over the past few years. Special thanks to **Mirco Michel**, your viral positive attitude inspires and made my day, more than once. Furthermore big thanks to **Christian Oertlin** for pushing my climbing skills to the next level. Also thanks to **Axel Rudling**, **Marco Salvatore**, **Kostas Tsirigos**, **Per Warholm** and **Walter Basile**. It was always fun with you guys.

Big thanks to **Sandra Huskanovic**, who helped with my swedish sammanfattning, and **Roman Hackl**. I am grateful for all the fun experiences we had and that I was always able to count on you no matter what. You guys rock! Thanks also to **Maja Schlittler** and **Christoph Siebenmann** for always telling me the truth and keeping my head straight during my knee rehab. Thanks **Jacopo Fontana** & **Francesca Pavan** for all the funny Italian stories. Thanks **Andre Le Lerre** for all the awesome boulder sessions and for helping me out optimizing my c++ programs. Thanks to **David Gray Lassiter** for all the interesting discussion and your help to prepare for the big exam. Also thanks to **Rocio Paublete**, **Jay Shaw** and **Arthur Cheng** for our nice Monday gatherings. Thanks to **Robert Lerner**, **Tobias Olli**, **Robin Haag**, **Frida Olofsson** and **Klas Söderlind** for taking me to all these wonderful climbing areas around Stockholm. Thanks to everyone else I have met during my PhD, I really enjoyed my time with all of you.

Thanks as well to my grandparents, sister and brother in law and my parents who always supported me without any doubts. Last but not least I would like to say, that none of this, literally none, would have been possible without the love and constant encouragement of my fiancée **Elisabeth Mucke**. I feel like I can achieve anything as long as I have you by my side.

References

Abatangelo, L., Maglietta, R., Distaso, A., D'Addabbo, A., Creanza, T. M., Mukherjee, S., and Ancona, N. (2009). **Comparative study of gene set enrichment methods**. *BMC Bioinformatics*, 10(1):1.

Alexeyenko, A., Lee, W., Pernemalm, M., Guegan, J., Dessen, P., Lazar, V., Lehtiö, J., and Pawitan, Y. (2012). **Network enrichment analysis: extension of gene-set enrichment analysis to gene networks**. *BMC Bioinformatics*, 13(1):226.

Alexeyenko, A. and Sonnhammer, E. L. (2009). **Global networks of functional coupling in eukaryotes from comprehensive data integration**. *Genome Research*, 19(6):1107–1116.

Altenhoff, A. M., Schneider, A., Gonnet, G. H., and Dessimoz, C. (2010). **OMA 2011: orthology inference among 1000 complete genomes**. *Nucleic Acids Research*, 39(suppl_1):D289–D294.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). **Gene Ontology: tool for the unification of biology**. *Nature Genetics*, 25(1):25.

Bader, G. D., Cary, M. P., and Sander, C. (2006). **Pathguide: a pathway resource list**. *Nucleic Acids Research*, 34(suppl_1):D504–D506.

Barabási, A.-L. and Albert, R. (1999). **Emergence of scaling in random networks**. *Science*, 286(5439):509–512.

Barabasi, A.-L. and Oltvai, Z. N. (2004). **Network biology: understanding the cell's functional organization**. *Nature Reviews Genetics*, 5(2):101.

Barker, D., Meade, A., and Pagel, M. (2007). **Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes**. *Bioinformatics*, 23(1):14–20.

Barker, D. and Pagel, M. (2005). **Predicting functional gene links from phylogenetic-statistical analyses of whole genomes**. *PLoS Computational Biology*, 1(1):e3.

Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B., and Boone, C. (2013). **Genetic interaction networks: toward an understanding of heritability**. *Annual review of genomics and human genetics*, 14:111–133.

Benjamini, Y. and Hochberg, Y. (1995). **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Bhatlekar, S., Fields, J. Z., and Boman, B. M. (2014). **HOX genes and their role in the development of human cancers**. *Journal of Molecular Medicine*, 92(8):811–823.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., et al. (2007). **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature*, 447(7146):799–816.

Blohm, P., Frishman, G., Smialowski, P., Goebels, F., Wachinger, B., Ruepp, A., and Frishman, D. (2013). **Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis**. *Nucleic Acids Research*, 42(D1):D396–D400.

Boone, C., Bussey, H., and Andrews, B. J. (2007). **Exploring genetic interactions and networks with yeast**. *Nature Reviews Genetics*, 8(6):437.

Bork, P. (1991). **Shuffled domains in extracellular proteins**. *FEBS Letters*, 286(1-2):47–54.

Cohen, R. and Havlin, S. (2003). **Scale-free networks are ultrasmall**. *Physical Review Letters*, 90(5):058701.

Consortium, G. O. et al. (2015). **Gene ontology consortium: going forward**. *Nucleic Acids Research*, 43(D1):D1049–D1056.

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

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., et al. (2016). **A global genetic interaction network maps a wiring diagram of cellular function**. *Science*, 353(6306):aaf1420.

Dong, X., Hao, Y., Wang, X., and Tian, W. (2016). **LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights**. *Scientific Reports*, 6.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007). **A systems biology approach for pathway level analysis**. *Genome Research*, 17(10):1537–1545.

du Plessis, L., Škunca, N., and Dessimoz, C. (2011). **The what, where, how and why of gene ontology - a primer for bioinformaticians**. *Briefings in Bioinformatics*, 12(6):723–735.

Elefsinioti, A., Saraç, Ö. S., Hegele, A., Plake, C., Hubner, N. C., Poser, I., Sarov, M., Hyman, A., Mann, M., Schroeder, M., et al. (2011). **Large-scale de novo prediction of physical protein-protein association**. *Molecular & Cellular Proteomics*, 10(11):M111–010629.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2015). **The reactome pathway knowledgebase**. *Nucleic Acids Research*, 44(D1):D481–D487.

Fitch, W. M. (1970). **Distinguishing homologous from analogous proteins**. *Systematic Zoology*, 19(2):99–113.

Frings, O., Alexeyenko, A., and Sonnhammer, E. L. (2013). **MGclus: network clustering employing shared neighbors**. *Molecular BioSystems*, 9(7):1670–1675.

Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J., and Oliva, B. (2012). **BIPS: BIANA Interolog Prediction Server. A tool for protein–protein interaction inference**. *Nucleic Acids Research*, 40(W1):W147–W151.

Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I., and Wright, F. A. (2010). **Heading down the wrong pathway: on the influence of correlation within gene sets**. *BMC Genomics*, 11(1):574.

Gerstein, M., Lan, N., and Jansen, R. (2002). **Integrating interactomes**. *Science*, 295(5553):284–287.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). **EnrichNet: network-based gene set enrichment analysis**. *Bioinformatics*, 28(18):i451–i457.

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., et al. (2015). **Understanding multicellular function and disease with human tissue-specific networks**. *Nature Genetics*, 47(6):569–576.

Guala, D., Sjölund, E., and Sonnhammer, E. L. (2014). **MaxLink: network-based prioritization of genes tightly linked to a disease seed set**. *Bioinformatics*, 30(18):2689–2690.

Guala, D. and Sonnhammer, E. L. (2017). **A large-scale benchmark of gene prioritization methods**. *Scientific Reports*, 7.

Guan, Y., Myers, C. L., Lu, R., Lemischka, I. R., Bult, C. J., and Troyanskaya, O. G. (2008). **A genomewide functional network for the laboratory mouse**. *PLoS Computational Biology*, 4(9):e1000165.

Guney, E. and Oliva, B. (2012). **Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization**. *PLoS One*, 7(9):e43557.

Hartwell, L. H., Hopfield, J. J., Leibler, S., Murray, A. W., et al. (1999). **From molecular to modular cell biology**. *Nature*, 402(6761):C47.

Hogeweg, P. (2011). **The roots of bioinformatics in theoretical biology**. *PLoS Computational Biology*, 7(3):e1002021.

Hong, M.-G., Alexeyenko, A., Lambert, J.-C., Amouyel, P., and Prince, J. A. (2010). **Genome-wide pathway analysis implicates intracellular transmembrane protein transport in Alzheimer disease**. *Journal of Human Genetics*, 55(10):707–709.

Hosack, D. A., Dennis, G., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). **Identifying biological themes within lists of genes with EASE**. *Genome Biology*, 4(10):1.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature Protocols*, 4(1):44–57.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Research*, 37(1):1–13.

Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., et al. (2003). **Global analysis of protein localization in budding yeast**. *Nature*, 425(6959):686.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). **Relating whole-genome expression data with protein-protein interactions**. *Genome Research*, 12(1):37–46.

Jensen, L. J. and Bork, P. (2008). **Not comparable, but complementary**. *Science*, 322(5898):56–57.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., and Bork, P. (2007). **eggNOG: automated construction and annotation of orthologous groups of genes**. *Nucleic Acids Research*, 36(suppl_1):D250–D254.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. (2005). **Reactome: a knowledgebase of biological pathways**. *Nucleic Acids Research*, 33(suppl_1):D428–D432.

Kanehisa, M. and Goto, S. (2000). **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Research*, 28(1):27–30.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). **The human genome browser at UCSC**. *Genome research*, 12(6):996–1006.

Khatri, P., Sirota, M., and Butte, A. J. (2012). **Ten years of pathway analysis: current approaches and outstanding challenges**. *PLoS Computational Biology*, 8(2):e1002375.

Klammer, M., Roopra, S., and Sonnhammer, E. L. (2008). **jSquid: a Java applet for graphical on-line network exploration**. *Bioinformatics*, 24(12):1467–1468.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). **Walking the interactome for prioritization of candidate disease genes**. *The American Journal of Human Genetics*, 82(4):949–958.

Kotlyar, M., Pastrello, C., Sheahan, N., and Jurisica, I. (2016). **Integrated interactions database: tissue-specific view of the human and model organism interactomes**. *Nucleic Acids Research*, 44(D1):D536–D541.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., et al. (2006). **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae**. *Nature*, 440(7084):637.

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, S. R., Miller, R., et al. (2015). **WikiPathways: capturing the full diversity of pathway knowledge**. *Nucleic Acids Research*, 44(D1):D488–D494.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., and Rhee, S. Y. (2010). **Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana**. *Nature Biotechnology*, 28(2):149–156.

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). **Prioritizing candidate disease genes by network-based boosting of genome-wide association data**. *Genome Research*, 21(7):1109–1121.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A. G., and Marcotte, E. M. (2008). **A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans**. *Nature Genetics*, 40(2):181–188.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science*, 298(5594):799–804.

Lees, J., Heriche, J., Morilla, I., Ranea, J., and Orengo, C. (2011). **Systematic computational prediction of protein interaction networks**. *Physical Biology*, 8(3):035008.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics*, 27(12):1739–1740.

Lin, M., Hu, B., Chen, L., Sun, P., Fan, Y., Wu, P., and Chen, X. (2009). **Computational identification of potential molecular interactions in Arabidopsis**. *Plant Physiology*, 151(1):34–46.

Lopez, M. D., Guerra, J. J. M., and Samuelsson, T. (2010). **Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes**. *PLoS One*, 5(5):e10654.

Maslov, S. and Sneppen, K. (2002). **Specificity and stability in topology of protein networks**. *Science*, 296(5569):910–913.

McCormack, T., Frings, O., Alexeyenko, A., and Sonnhammer, E. L. (2013). **Statistical assessment of crosstalk enrichment between gene groups in biological networks**. *PloS One*, 8(1):e54945.

Mering, C. v., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). **STRING: a database of predicted functional associations between proteins**. *Nucleic Acids Research*, 31(1):258–261.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2017). **PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements**. *Nucleic Acids Research*, 45(D1):D183–D189.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., et al. (2005). **The PANTHER database of protein families, subfamilies, functions and pathways**. *Nucleic Acids Research*, 33(suppl_1):D284–D288.

Montojo, J., Zuberi, K., Rodriguez, H., Bader, G. D., and Morris, Q. (2014). **GeneMANIA: Fast gene network construction and function prediction for Cytoscape**. *F1000Research*, 3.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). **GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function**. *Genome Biology*, 9(1):S4.

National Human Genome Research Institute (NIH) (2015). **Biological Pathways**. *Retrieved from https://www.genome.gov/27530687/*.

Newman, M. E. and Girvan, M. (2004). **Finding and evaluating community structure in networks**. *Physical review E*, 69(2):026113.

Ogris, C., Guala, D., Helleday, T., and Sonnhammer, E. L. (2017). **A novel method for crosstalk analysis of biological networks: improving accuracy of pathway annotation**. *Nucleic Acids Research*, 45(2):e8–e8.

Östlund, G., Lindskog, M., and Sonnhammer, E. L. (2010). **Network-based Identification of novel cancer genes**. *Molecular & Cellular Proteomics*, 9(4):648–655.

Pellegrini, M. (2012). **Using phylogenetic profiles to predict functional relationships**. In *Bacterial Molecular Networks*, pages 167–177. Springer New York.

Pico, A. R., Kelder, T., Van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008). **WikiPathways: pathway editing for the people**. *PLoS Biology*, 6(7):e184.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., et al. (2005). **Global analysis of protein phosphorylation in yeast**. *Nature*, 438(7068):679–684.

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A., and Koonin, E. V. (2002). **Connected gene neighborhoods in prokaryotic genomes**. *Nucleic Acids Research*, 30(10):2212–2223.

Rudashevskaya, E. L., Sickmann, A., and Markoutsa, S. (2016). **Global profiling of protein complexes: current approaches and their perspective in biomedical research**. *Expert Review of Proteomics*, 13(10):951–964.

Şahin, S., Işık Gönül, İ., Çakır, A., Seçkin, S., and Uluoğlu, Ö. (2016). **Clinicopathological Significance of the Proliferation Markers Ki67, RacGAP1, and Topoisomerase 2 Alpha in Breast Cancer**. *International Journal of Surgical Pathology*, 24(7):607–613.

Schmitt, T., Ogris, C., and Sonnhammer, E. L. (2014). **FunCoup 3.0: database of genome-wide functional coupling networks**. *Nucleic Acids Research*, 42(D1):D380–D388.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Research*, 13(11):2498–2504.

Sonnhammer, E. L. and Östlund, G. (2014). **InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic**. *Nucleic Acids Research*, 43(D1):D234–D239.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Research*, 34(suppl_1):D535–D539.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). **A gene-coexpression network for global discovery of conserved genetic modules**. *Science*, 302(5643):249–255.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., et al. (2010). **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored**. *Nucleic Acids Research*, 39(suppl_1):D561–D568.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. (2014). **STRING v10: protein–protein interaction networks, integrated over the tree of life**. *Nucleic Acids Research*, 43(D1):D447–D452.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., et al. (2017). **The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible**. *Nucleic Acids Research*, 45(D1):D362–D368.

Tarca, A. L., Draghici, S., Bhatti, G., and Romero, R. (2012). **Down-weighting overlapping genes improves gene set analysis**. *BMC Bioinformatics*, 13(1):1.

Tornow, S. and Mewes, H. (2003). **Functional modules by relating protein interaction networks and gene expression**. *Nucleic Acids Research*, 31(21):6283–6289.

Travers, J. and Milgram, S. (1967). **The small world problem**. *Phychology Today*, 1:61–67.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). **Tissue-based map of the human proteome**. *Science*, 347(6220):1260419.

Van Dongen, S. (2008). **Graph clustering via a discrete uncoupling process**. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141.

Vidal, M. (2009). **A unifying view of 21st century systems biology**. *FEBS Letters*, 583(24):3891–3894.

Voichita, C., Donato, M., and Draghici, S. (2012). **Incorporating gene significance in the impact analysis of signaling pathways**. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on Machine Learning and Applications*, volume 1, pages 126–131. IEEE.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). **Comparative assessment of large-scale data sets of protein–protein interactions**. *Nature*, 417(6887):399–403.

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., and von Mering, C. (2012). **PaxDb, a database of protein abundance averages across all three domains of life**. *Molecular & Cellular Proteomics*, 11(8):492–500.

Wong, A. K., Park, C. Y., Greene, C. S., Bongo, L. A., Guan, Y., and Troyanskaya, O. G. (2012). **IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks**. *Nucleic Acids Research*, 40(W1):W484–W490.

Yanai, I., Derti, A., and DeLisi, C. (2001). **Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes**. *Proceedings of the National Academy of Sciences*, 98(14):7940–7945.

Yin, Z., Gupta, M., Weninger, T., and Han, J. (2010). **A unified framework for link recommendation using random walks**. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 152–159. IEEE.

Zhang, Q. C., Petrey, D., Lei Deng, L. Q., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., et al. (2012). **Structure-based prediction of protein-protein interactions on a genome-wide scale**. *Nature*, 490(7421):556.

Zhang, X. and Smith, T. F. (1998). **Yeast "operons"**. *Microbial & Comparative Genomics*, 3(2):133–140.