Functional association networks for disease gene prediction

Dimitri Guala

Academic dissertation for the Degree of Doctor of Philosophy in Biochemistry towards Bioinformatics at Stockholm University to be publicly defended on Friday 10 November 2017 at 14.00 in Magnélisalen, Kemiska övningslaboratoriet, Svante Arrhenius väg 16 B.

Abstract

Mapping of the human genome has been instrumental in understanding diseasescaused by changes in single genes. However, disease mechanisms involvingmultiple genes have proven to be much more elusive. Their complexityemerges from interactions of intracellular molecules and makes them immuneto the traditional reductionist approach. Only by modelling this complexinteraction pattern using networks is it possible to understand the emergentproperties that give rise to diseases.The overarching term used to describe both physical and indirect interactionsinvolved in the same functions is functional association. FunCoup is oneof the most comprehensive networks of functional association. It uses a naïveBayesian approach to integrate high-throughput experimental evidence of intracellularinteractions in humans and multiple model organisms. In the firstupdate, both the coverage and the quality of the interactions, were increasedand a feature for comparing interactions across species was added. The latestupdate involved a complete overhaul of all data sources, including a refinementof the training data and addition of new class and sources of interactionsas well as six new species.Disease-specific changes in genes can be identified using high-throughputgenome-wide studies of patients and healthy individuals. To understand theunderlying mechanisms that produce these changes, they can be mapped tocollections of genes with known functions, such as pathways. BinoX wasdeveloped to map altered genes to pathways using the topology of FunCoup.This approach combined with a new random model for comparison enables BinoXto outperform traditional gene-overlap-based methods and other networkbasedtechniques.Results from high-throughput experiments are challenged by noise and biases,resulting in many false positives. Statistical attempts to correct for thesechallenges have led to a reduction in coverage. Both limitations can be remediedusing prioritisation tools such as MaxLink, which ranks genes using guiltby association in the context of a functional association network. MaxLink'salgorithm was generalised to work with any disease phenotype and its statisticalfoundation was strengthened. MaxLink's predictions were validatedexperimentally using FRET.The availability of prioritisation tools without an appropriate way to comparethem makes it difficult to select the correct tool for a problem domain.A benchmark to assess performance of prioritisation tools in terms of theirability to generalise to new data was developed. FunCoup was used for prioritisationwhile testing was done using cross-validation of terms derived fromGene Ontology. This resulted in a robust and unbiased benchmark for evaluationof current and future prioritisation tools. Surprisingly, previously superiortools based on global network structure were shown to be inferior to a localnetwork-based tool when performance was analysed on the most relevant partof the output, i.e. the top ranked genes.This thesis demonstrates how a network that models the intricate biologyof the cell can contribute with valuable insights for researchers that study diseaseswith complex genetic origins. The developed tools will help the researchcommunity to understand the underlying causes of such diseases and discovernew treatment targets. The robust way to benchmark such tools will help researchersto select the proper tool for their problem domain.

Keywords: *network biology, biological networks, network prediction, functional association, functional coupling, network integration, functional association networks, genome-wide association networks, gene networks, protein networks, fret, functional enrichment analysis, network cross-talk, pathway annotation, gene prioritisation, network-based gene prioritization, benchmarking.*

Stockholm 2017
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-147217

ISBN 978-91-7649-976-4
ISBN 978-91-7649-977-1

Department of Biochemistry and Biophysics

Stockholm University, 106 91 Stockholm

FUNCTIONAL ASSOCIATION NETWORKS FOR DISEASE GENE PREDICTION

Dimitri Guala

Stockholm
University

Functional association networks for disease gene prediction

Dimitri Guala

ISBN print 978-91-7649-976-4
ISBN PDF 978-91-7649-977-1

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

Abstract

Mapping of the human genome has been instrumental in understanding diseases caused by changes in single genes. However, disease mechanisms involving multiple genes have proven to be much more elusive. Their complexity emerges from interactions of intracellular molecules and makes them immune to the traditional reductionist approach. Only by modelling this complex interaction pattern using networks is it possible to understand the emergent properties that give rise to diseases.

The overarching term used to describe both physical and indirect interactions involved in the same functions is functional association. FunCoup is one of the most comprehensive networks of functional association. It uses a naïve Bayesian approach to integrate high-throughput experimental evidence of intracellular interactions in humans and multiple model organisms. In the first update, both the coverage and the quality of the interactions, were increased and a feature for comparing interactions across species was added. The latest update involved a complete overhaul of all data sources, including a refinement of the training data and addition of new class and sources of interactions as well as six new species.

Disease-specific changes in genes can be identified using high-throughput genome-wide studies of patients and healthy individuals. To understand the underlying mechanisms that produce these changes, they can be mapped to collections of genes with known functions, such as pathways. BinoX was developed to map altered genes to pathways using the topology of FunCoup. This approach combined with a new random model for comparison enables BinoX to outperform traditional gene-overlap-based methods and other network-based techniques.

Results from high-throughput experiments are challenged by noise and biases, resulting in many false positives. Statistical attempts to correct for these challenges have led to a reduction in coverage. Both limitations can be remedied using prioritisation tools such as MaxLink, which ranks genes using guilt by association in the context of a functional association network. MaxLink's algorithm was generalised to work with any disease phenotype and its statistical foundation was strengthened. MaxLink's predictions were validated experimentally using FRET.

The availability of prioritisation tools without an appropriate way to com-

pare them makes it difficult to select the correct tool for a problem domain. A benchmark to assess performance of prioritisation tools in terms of their ability to generalise to new data was developed. FunCoup was used for prioritisation while testing was done using cross-validation of terms derived from Gene Ontology. This resulted in a robust and unbiased benchmark for evaluation of current and future prioritisation tools. Surprisingly, previously superior tools based on global network structure were shown to be inferior to a local network-based tool when performance was analysed on the most relevant part of the output, i.e. the top ranked genes.

This thesis demonstrates how a network that models the intricate biology of the cell can contribute with valuable insights for researchers that study diseases with complex genetic origins. The developed tools will help the research community to understand the underlying causes of such diseases and discover new treatment targets. The robust way to benchmark such tools will help researchers to select the proper tool for their problem domain.

Посвящаю моей семье
и памяти бабушки Али

List of Papers

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

PAPER I: **Comparative interactomics with Funcoup 2.0**
Alexeyenko, A., Schmitt, T., Tjärnberg, A., Guala, D., Frings, O., and Sonnhammer, ELL. (2011). *Nucleic Acids Research*, **40(November)**:821-828.

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

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

PAPER IV: **A large-scale benchmark of gene prioritization methods**
Guala, D., and Sonnhammer, ELL. (2017). *Scientific Reports*, **7**:46598.

PAPER V: **FunCoup 4: new species, data, and visualization**
Ogris, C.†, Guala, D.†, Kaduk, M.†, and Sonnhammer, ELL. *submitted.*, (2017).

PAPER VI: **Experimental validation of predicted cancer genes using FRET**
Guala, D., Bernhem, K., Ait Blal, H., Lundberg, E., Brismar, H., and Sonnhammer, ELL. *manuscript.*, (2017).

†Contributed equally

Contents

1. Introduction

"Bioinformatics is the queen AND servant of biology."
–Pavel Pevznev

This chapter provides a short description of the research area, purpose, scope, limitations, and outline of the thesis.

1.1 Background

Throughout history, diseases have played a decisive role in shaping cultures and societies (Borsch, 2005; Boutayeb, 2010). Humanity was already combating disease in the neolithic era, pre-dating even the most ancient, known civilisations, e.g. those of the Egyptians and Babylonians. Early medicine involved shamans and healers using primitive tools and mostly superstitious practices due to lack of knowledge or even concept of disease pathologies. Many scientific breakthroughs and revolutions later, the advent of modern sequencing techniques has provided the research community with the key piece of knowledge for understanding human disease. This piece was the sequence of the human genome. Further advances in high-throughput experimental techniques such as Next-Generation Sequencing (NGS) and genome-wide association studies (GWAS) have produced enormous amounts of data with the potential to generate insights into disease mechanisms and cures.

Despite improved methodology resulting in massive improvements in performance, new high-throughput techniques are still flawed when it comes to accuracy (Bromberg, 2013), producing numerous false positive results (Oti et al., 2011). At the moment, biological data is produced at an impressive pace of 2-40 exabytes (10^{18}) per year (Stephens et al., 2015) and can only be utilised with the help of bioinformatics and other computational approaches.

If properly analysed, the generated data can inform us of the complex interplay between the intracellular molecules. The cell is a complex system. Interactions between its parts and the environment give rise to properties such as non-linearity, emergence, and feedback structures, which cannot be understood in a deterministic way, without grasping the topology and dynamics of

the underlying interactions. Before the high-throughput era, we could focus on mapping and understanding individual interactions within this complex system. Now we have the opportunity to establish how these interactions, acting on a genomic scale, produce complex, non-deterministic behaviours and characteristics, and which perturbations lead to the diseases and disorders we are trying to treat.

1.2 Problem statement

Despite successes in combating monogenic disorders, the research community struggles with the causes of disorders with complex genetic origins, e.g. cancer and Alzheimer's, that have a significant negative impact on human society (Seuring et al., 2015; Siegel et al., 2017; Winblad et al., 2016). The pathophysiology of multigenic diseases is difficult to approach with traditional high-quality, small-scale experiments due to the emergent features and other non-deterministic properties at the cellular level. High-throughput techniques suffer from high rates of false positives. Additionally, different types of experimental techniques come with unique biases. A way to circumvent these obstacles is to integrate different types and sources of data into networks of functional associations, where molecules sharing functions are connected. If implemented properly, this approach can cancel out the individual biases and overcome false positives (Oti et al., 2011). Available networks, have various degrees of false positives due to e.g. use of noisy data types like text mining or failure to account for redundancies arising from the better-studied associations. Besides tackling that, new networks need to be user-friendly to have an impact on the research community.

The next step on the path of drug discovery is to use the predicted networks to identify genes associated with the diseases of interest. The past decade has seen the advent of many such "prioritisation" tools able to provide ranked lists of candidate genes (Guala et al., 2014; Köhler et al., 2008; Tranchevent et al., 2016). However, these tools are usually only supported by proof of concept experiments or simple validations. A solid experimental validation is a pre-requisite but in no way a guarantee that the method will be able to generalize to new data. Neither does it provide information on how a tool would perform in comparison with others. Some attempts at benchmarking gene prioritisation tools have been made (Börnigen et al., 2012), but they have so far been unable to show robust statistical differences between tools or have suffered from knowledge cross-contamination. There is an urgent need for a systematic and unbiased way to test the performance of prioritisation tools.

When an experimental technique in a perturbation setting identifies a set of genes that deviate in terms of expression or other traits, it is natural to wonder

what functions these genes have in common. Traditionally, simple statistical methods have been used to analyse the deviating genes by looking at their overlap with pathways of known function. The currently used methods have several limitations: some generate many false positives, while others lack in coverage (Khatri et al., 2012). New tools for gene analysis and functional enrichment are therefore desperately needed.

In this thesis a network of functionally associated genes and proteins is further developed (paper I and V) together with tools for gene prioritisation (paper II) and identification of new functions by functional enrichment analysis (paper III). Additionally, both experimental validation (paper VI) of the discovered potential disease genes and a generalised benchmarking approach (paper IV) are established to enable an unbiased assessment of prioritisation tools.

1.3 Purpose

The purpose of this thesis is to utilise the wealth of genomical and proteomical data produced by high-throughput experiments to aid disease mechanism and disease gene discovery.

1.4 Scope and limitations

Ideally, networks developed and used to uncover disease mechanisms would reflect the dynamics of cellular interactions with respect to time and different experimental conditions, e.g. stress, starvation, tissue specificity, etc. This may potentially provide even greater insights into the complex cellular interactions. Unfortunately, such a level of detail is still unavailable for genome scale networks, and thus falls outside of the scope of this thesis.

The type of networks described in this thesis are composed of genes and proteins. There are other intra-cellular molecules, e.g. metabolites and non-coding RNA, with important roles in disease pathogenesis (Esteller, 2011). Methods for data integration, prioritisation, and functional enrichment could potentially also be applied to these other molecules, but this too remains outside of the scope of this thesis.

Several crucial steps need to be taken before going from high-throughput data, via networks, functional enrichment and disease gene discovery, to drug targets and treatments. After the discovery of potential disease genes, further experimental efforts in model organisms and humans are required to validate the usefulness of the proposed candidates. Subsequently, an entire sequence of drug discovery needs to happen after that, even before potential treatments can

be designed, tested and approved. This thesis offers only the ability to take the first steps, up to the prediction of potential drug targets and their function, on this long journey.

1.5 Outline

The theoretical background needed for this thesis is covered in chapters 2, 3 and 4, which describe biological networks and functional associations networks as well as their applications. The papers and manuscripts forming the basis of this thesis are covered in chapter 5.

2. Biological networks

"It is not complicated. It is just a lot of it."
–Richard Feynman, about the universe

2.1 Why networks?

One of the key methodologies of modern science has been to reduce a studied phenomenon to its smallest constituents and then to study these building blocks in order to understand the phenomenon in question. This is known as the "reductionist" approach, and it has had tremendous success in the past centuries. One of its accomplishments was the discovery of the laws of thermodynamics and how they govern the motion of gas molecules, which enabled an understanding of the behaviour of ideal gases. However, many phenomena cannot be explained simply by studying their smallest components. Complex systems, like the human brain, the Earth's climate, and the living cell, are examples of such phenomena. The non-linear interactions between their fundamental parts give rise to features, e.g. emergent properties, feedback loops, and self-organisation, which cannot be predicted from an understanding of the activity of a system's constituents.

A way to address the shortcomings of the reductionist approach in understanding complex systems has been to model these systems using networks. In the example of the internet, the nodes in the network represent internet servers, while the links depict the connections between the servers (Siganos et al., 2006). Another complex system that requires a network-based approach in order to understand it and model its behaviour is the human cell. This is an example of a biological network, where the nodes represent intra-cellular molecules and links depict interactions between them. Networks are found in biological systems at various scales, ranging from the evolutionary tree of life, spanning eons of evolutionary history of living organisms, to predator-prey networks of local ecological environments, and down to the nano-scale of the intra-cellular interactions of the living cells.

The remainder of this chapter will focus on the key properties of biological networks and the approach used to model them.

2.2 Network representation

Graph theory lends itself well to the study of the properties of biological and other networks. A network can be represented by a graph $G(V,E)$, composed of N number of vertices, $V = \{v_1, v_2, \ldots, v_N\}$, connected by M number of edges, $e = (v_j, v_k)$, where $e_i \in E$. Graphs can be directed, with edges forming ordered pairs, i.e. $(v_j, v_k) \neq (v_k, v_j)$; or un-directed, where $(v_j, v_k) = (v_k, v_j)$. Both vertices and edges can have labels associated with them. When these labels are numbers, they are called "weights" (Fig. 2.1). Vertices have other properties (see section 2.4) e.g. degree, $d(v)$, which is defined as the number of edges that connect to that particular vertex.

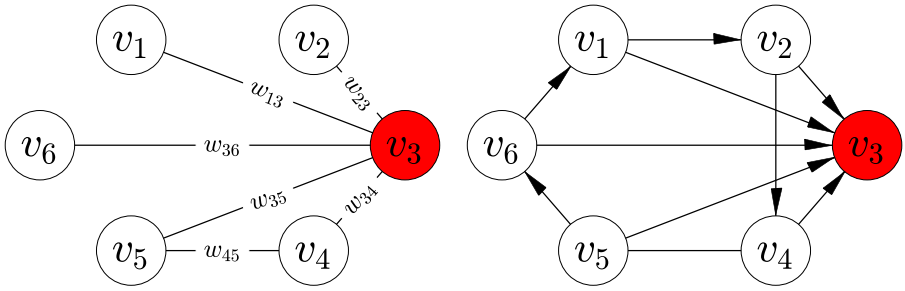


Figure 2.1: Graph structure - The basic structure of an undirected, weighted graph (on the left) and a directed, unweighted graph (on the right). Node v_3 (in red) is an example of a hub node.

2.3 Network characteristics

Degree distribution

One of the most important topological properties of a graph is the degree distribution $P(d)$, of its vertices. Classical graph theory, predicts a Poisson degree distribution, i.e. $P(d) \sim \lambda^d / d!$, in real-world graphs (Fig. 2.2). The assumptions behind this can be traced to the Erdős-Rényi model (Erdös and Rényi, 1959), which assumed a fixed number of vertices with randomly connected pairs of vertices, producing a random graph topology (Fig. 2.3). Experimental evidence has shown that real-world graphs, including biological networks, do not exhibit random topology. Instead, they have a power law degree distribution, i.e. $P(d) \sim d^{-\gamma}$, with $2 < \gamma < 3$ (Barabasi and Albert, 1999), later shown

to be closer to $2 < \gamma < 5$ (Barabasi, 2013). This is an example of an emergent property that arises from the complex non-linear interactions of the nodes in a real-world graph or a network.

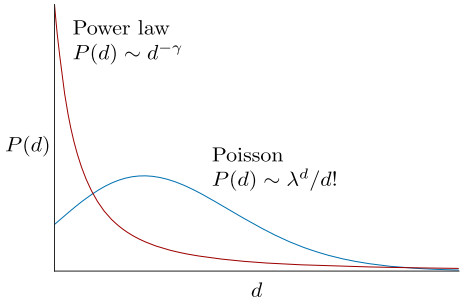


Figure 2.2: Degree distribution - Degree distribution of real-world networks follows a power law distribution instead of the Poisson distribution predicted by random network models (image adapted from Barabasi (2013))

Scale-free networks

For random networks, the addition of new nodes is random and uniform. Barabasi and Albert (1999) showed that biological networks and other real-world graphs, e.g. the World Wide Web, the electrical power grid of the USA, form through preferential attachment of new nodes to an initial small set of nodes; i.e., the probability of adding a new node is $p(d_i) = d_i / \sum_j d_j$, where, d_i and d_j are the degrees of nodes i and j, respectively (Barabasi and Albert, 1999). This probability is thus directly proportional to node degree, with higher probabilities for nodes having many links, so called "hubs", compared to nodes with a few links (Fig. 2.3). Since power law functions have the same shape at all scales, biological networks are said to be *scale-free*. A direct consequence of the scale-free topology is that these networks have few hubs and many nodes with very few links (Fig. 2.2). This makes such networks robust to random node failures (Albert et al., 2000), which for intracellular networks can be represented by a cell functioning despite many random mutations. However, it also makes networks vulnerable to non-random mutations targeting the hubs.

Small-world networks

The scale-free property of biological networks gives rise to the "small-world" behaviour of a network. The small-world phenomenon implies that it takes surprisingly few links to transition from any node in the network to any other node (Fig. 2.3). This property arises from the fact that the number of neighbouring

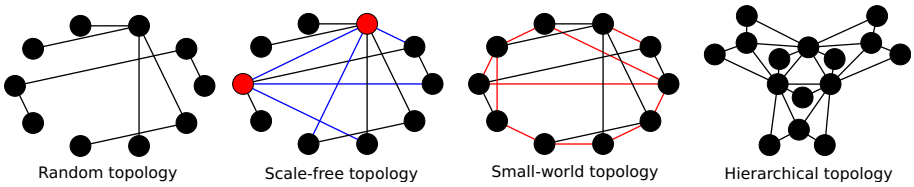


Figure 2.3: Different network topologies. Examples of hubs are marked in red. The blue and red links are the links that differ between the scale-free and random topology, and between the small-world and random topology, respectively.

nodes increases exponentially with the increasing distance from the starting node; i.e., at distance $D = 1$, the number of neighbours of node v is $d(v)^1$, at distance $D = 2$ the number of neighbours is $d(v)^2$, and at distance $D = x$ the number is $d(v)^x$, resulting in the average distance, $\langle D \rangle \approx ln\, N / ln\, d(v)$ (de Sola Pool and Kochen, 1978). This means that $\langle D \rangle$ is orders of magnitude smaller than the network size, N, and that it gets smaller with increased connectivity, i.e. larger $d(v)$. This is true for both random graphs and the scale-free real-world networks. However, not accounting for the scale-free properties of large real-world networks can result in a misjudgement of $\langle D \rangle$, because the hubs in a scale-free network act as bridges between many nodes with small $d(v)$. This shortens the $\langle D \rangle$, making the network ultra small-world.

Clustering coefficient

Another important network property is the clustering coefficient (Fig. 2.1). It helps to describe how the network is connected beyond its degree distribution. The local clustering coefficient $c(v)$ of a node v describes the number of links between the immediate neighbours of v. This is shown in (2.1) for undirected networks, with $n_e(v)$, being the number of edges between the neighbours of node v (Chalancon et al., 2013).

$$c(v) = \frac{2n_e(v)}{d(v)(d(v)-1)} \tag{2.1}$$

The network-wide measure of clustering is the average clustering coefficient, $\langle C \rangle$, described in (2.2).

$$\langle C \rangle = \frac{1}{N} \sum_{v \in G} c(v) \tag{2.2}$$

Both the random Erdős-Rényi model and the scale-free Barabasi-Albert model fail to describe the high degree of clustering observed in real-world networks. Instead, it is a hierarchical organisation of real-world networks, where

nodes with a small $d(v)$ also have a small $c(v)$ while the hubs have a large $c(v)$, which accurately predicts the clustering property of real-world networks. This clustering can be described by the scaling law, where a node with degree d has the clustering coefficient $c(v)(d) \sim d^{-1}$ (Ravasz and Barabasi, 2003). An interpretation of this hierarchical property of real-world networks can be viewed as many small communities that cluster into larger communities, which in turn cluster into even larger ones, and so on (Fig. 2.3).

2.4 Node-centric characteristics

Many applications of networks involve determining the most important nodes in the network. This can be achieved by taking into account one or several of the properties described in the following sections (Fig. 2.4).

Degree centrality determines the ability of a node to spread information to its local neighbourhood. It is measured by the degree of a node, sometimes normalised by the maximum possible degree in the network.

Shortest path (sp), between two nodes, is defined as the minimum number of edges required to transition from one node to the other.

Closeness centrality (CC) determines how quickly information is able to spread to nodes in the network and depends on the average length of sp_{ij} between a node v_i and all the other nodes v_j (Bavelas, 1950). $CC(v_i) = \frac{1}{\sum_{j \neq i} sp_{ij}}$

Betweenness centrality (CB) determines the amount of information that runs through a node. It is defined as the probability of a shortest path between two nodes v_i and v_j in the network containing the node v_k. It is often described as the fraction of shortest paths that pass through node v_k, $sp_{ij}(v_k)$ and all shortest paths sp_{ij} (Freeman, 1977), where $(i \neq k \neq j) \in V$.

$$CB(v_k) = \sum_i \sum_j \frac{sp_{ij}(v_k)}{sp_{ij}} \tag{2.3}$$

PageRank centrality (CR) is based on the PageRank algorithm (Page et al., 1999), which assigns weights to a node proportional to the number and weight of the nodes connected to it. The weights are a measure of how often one would arrive at a node v_i by randomly moving between the nodes (2.4).

$$CR(v_i) = \beta \sum_j a_{ji} \frac{CR(v_j)}{\sum_i a_{ji}} + \frac{1-\beta}{N} \tag{2.4}$$

N is the total number of nodes. The parameter controlling the restart at the starting node during the random transition between the nodes is $0 \leq \beta \leq 1$. The number of outgoing links from node v_j and v_i are $\sum_j a_{ji}$ and $\sum_i a_{ji}$, respectively.

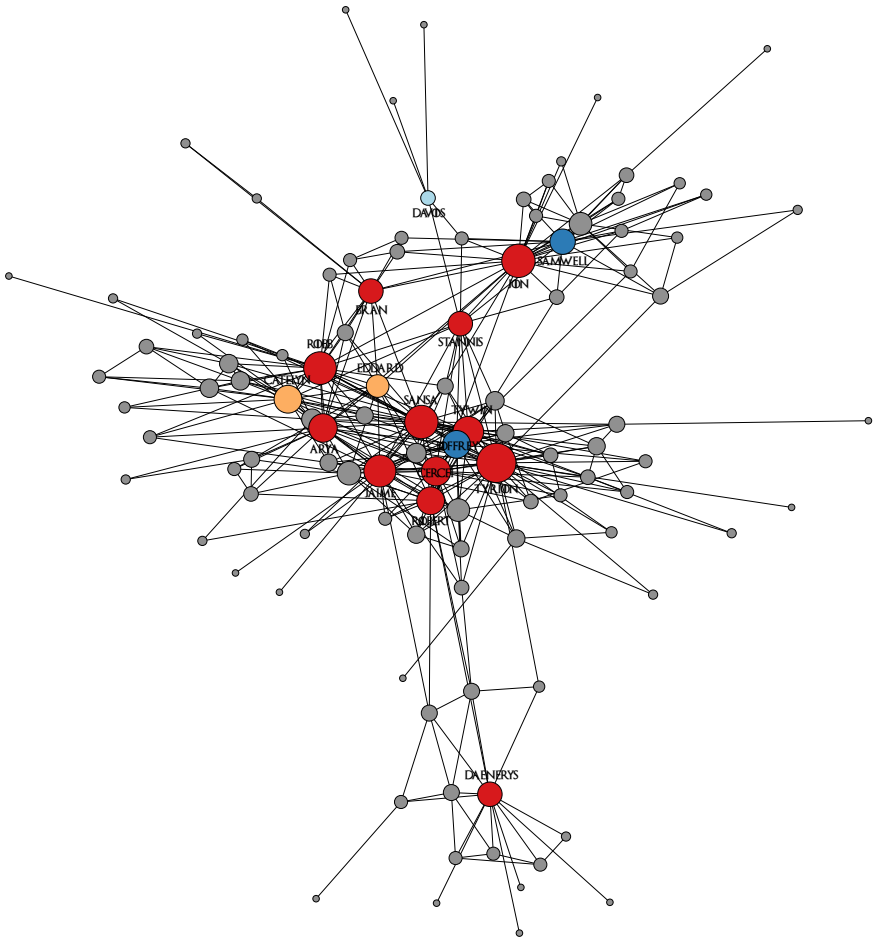


Figure 2.4: Node-centric network properties - The size of the nodes is proportional to their degree centrality. Colouring is based on the other centrality measures, with the top ten nodes for each of the centrality measures coloured as follows: *orange* for closeness, *light blue* for betweenness, *blue* for PageRank, and *red* if the node is in the top 10 nodes for two or more centrality measures.The network is based on interactional data generated using text-mining of "A Storm of Swords", the third book in the "Game of Thrones" book series by George R. R. Martin (Beveridge and Shan, 2016).

3. Functional association networks

"All models are wrong. Some are useful."
–George Box

3.1 Overview

High-throughput experimental techniques, such as the yeast-2-hybrid (Y2H) system (Fields and Song, 1989), affinity purification (AP) (Rigaut et al., 1999), and microarrays of mRNA expression, provide a wealth of gene and protein interaction data. Historically, single type of experimental data was used for network prediction. However, all experimental techniques, come with limitations; e.g., Y2H falsely predicts interactions between proteins that never co-localise in the same cellular compartment, whilst being unable to identify interactions between proteins after post-translational modification. Another aspect affecting coverage is the lack of overlap in the resulting protein-protein interactions (PPIs) discovered by different experimental techniques (Yu et al., 2008). Purely PPI-based networks are therefore incomplete and sometimes of low quality contributing to a bias towards the more well-studied genes (Piro and Di Cunto, 2012). Combining all experimental evidence of physical interactions provides about 20% coverage of all possible PPIs (Sharma et al., 2005). Therefore, studying pure PPI-based networks may not be sufficient to understand the complex biological processes in a cell.

Other experimental techniques such as microarrays are prone to noise and bias stemming from the different inherent levels of mRNA expression. There are also studies showing that even consistently co-expressed, genetically linked genes can stem from different pathways (Yu et al., 2009), suggesting a risk of false positives in the use of co-expression data. Combining data from different experimental techniques, into networks, compensates for missing interactions, random noise, and other artefacts, since the true signal is more likely to be enhanced while the uniformly distributed noise is dampened.

Additionally, molecular interactions forming the underlying biological networks arise from different modes of communication, such as signalling cascades, protein complexes, etc., between genes, proteins, and other molecules. This communication takes place on genetic, genomic, and proteomic levels of regulation. Therefore, by combining heterogeneous types of data stemming from diverse experimental sources, it is possible to provide a more comprehensive view of the underlying interplay of intracellular activities. This in turn enables the mapping and prediction of mechanisms behind complex phenotypes like diseases and other genomic traits. Networks of functional association, where nodes represent genes or proteins and links depict interactions, present a flexible model able to capture the complexity of the underlying data (Hassani-Pak and Rawlings, 2017).

This chapter will describe how functional association networks are constructed as well as the most common types of data used for this purpose. At the end some examples of such networks will be presented.

3.2 Evidence of functional association

Intracellular molecules rely on interaction partners to perform complex cellular processes such as signal transduction, transcription, replication, and membrane transport. The nature of these processes can be direct e.g. stable physical interactions or transient binding; as well as indirect, e.g. metabolic cascades or propagation of information. Functional association refers both to direct and indirect ways in which intracellular molecules, involved in the same function, interact.

Different types of evidence of functional association require customized handling to ensure strength and quality of the resulting signal. This can be achieved by evidence-specific metrics, producing scores for the information content of a particular piece of data, as well as a confidence score or a weight of the dataset.

The rest of this section is dedicated to describing different types of evidence and examples of metrics.

3.2.1 Protein-protein interaction

PPI data is readily available from high-throughput experimental techniques such as Y2H, co-immunoprecipitation (coIP), mass spectrometry (MS) preceded by affinity purification (AP-MS), and tandem affinity purification (TAP-MS) (Gavin et al., 2002).

Various databases such as database of interacting proteins (DIP) (Salwinski et al., 2004), biomolecular interaction network database (BIND) (Alfarano

et al., 2005), and human protein reference database (HPRD) (Keshava Prasad et al., 2009) store the results from such experiments. More comprehensive databases, e.g. biological general repository for interaction datasets (BioGrid) (Chatr-Aryamontri et al., 2017) and iRefIndex (Razick et al., 2008), further aggregate collections of primary databases and experimental data sets.

Metrics for PPI data usually assign higher confidence to PPIs confirmed in several experiments. Scoring is based on the number of interactions found between a pair of proteins, contrasted with the total number of interactions found for the individual proteins. The direct *bait-prey* interactions are generally regarded as more reliable than the indirect *prey-prey* interactions due to the higher rate of false positives (FPR) in the latter (Krogan et al., 2006). When this aspect is taken into account it may result in higher scores for or even exclusive use of *bait-prey* interactions.

3.2.2 Domain-domain interaction

A protein domain is a well-defined structural and functional component of a protein, with a distinct folding, independent from the folding of other parts of the protein. Protein domains are highly conserved across evolutionary time, implying conservation of function. Domain-domain interaction (DDI) have therefore been shown to be more reliable than PPIs (Mrowka et al., 2001). A way to handle the problem of high FPRs and low coverage in high-throughput experimental PPI techniques is to consider protein domains. Unfortunately, data on DDIs is scarce calling for computational prediction techniques. Several databases encompassing both experimental and predicted DDIs, such as UniDomInt (Björkholm and Sonnhammer, 2009) and InterPro (Finn et al., 2017), provide comprehensive collections of DDIs that can be mapped to proteins using e.g. Pfam (Punta et al., 2012).

A score for a DDI can be obtained by dividing the observed domain pair co-occurence by the expected co-occurence, based on the individual frequencies of domains in the pair (Rhodes et al., 2005). Alternatively, a profile for each domain can be calculated based on the presence of the domain in the genome. Then the mutual information (MI) (Cover and Thomas, 2005) between the profiles of two domains would provide the score for the DDI (Lee et al., 2010).

3.2.3 Gene fusion and neighbourhood

The ability of single-domain proteins in one organism to fuse into multi-domain proteins in another organism can be utilised as evidence of functional association. This type of evidence is termed "gene fusion" (Marcotte, 1999). Due to the conserved structural and functional aspects of individual domains, the

multi-domain proteins containing the constituent single domains tend to interact (Enright et al., 1999). Gene fusion events are identified in whole-genome comparisons and stored in protein domain databases mentioned above. Scoring can be done by assigning a constant score for the predicted interaction partners. This evidence type has high accuracy but suffers from poor coverage.

Whole-genome comparisons also contain information about genes' immediate neighbourhood. In prokaryotic genomes, genes with the same or similar functions are grouped together into simultaneously transcribed units called "operons" in order to facilitate regulation. This feature of prokaryotic genomes lends itself well to the prediction of functional associations between proteins (Dandekar et al., 1998). In eukaryotes, operon structures are not common because of recombination events. However, several studies have found the order of genes in eukaryotes not to be completely random; instead, genes with common expression patterns are seen to cluster together as a result of regulation at the chromatin level (Ghanbarian and Hurst, 2015). This allows the use of gene neighbourhood as evidence of functional association also in eukaryotes. Both the order and the intergenic distance can be used in the scoring of this evidence type.

3.2.4 Co-evolution

Functional associations tend to be conserved during species evolution (Pellegrini et al., 1999). This feature can be exploited by comparing the evolution of different proteins where interacting proteins tend to exhibit co-evolution of molecular changes. The modified "mirror tree" method (Sato et al., 2005), involving calculation of Pearson correlation between the distance matrices of the studied protein pairs can be used for scoring.

Another way in which evolutionary conservation can be used to determine the level of functional association is by comparing the phylogenetic profiles of different proteins, i.e. the vectors summarising the presence or absence of a protein across multiple species (Pellegrini et al., 1999). Similarity between phylogenetic profiles can serve as a simple metric for evolutionary evidence of functional association. A somewhat more elaborate but computationally feasible method (Schmitt et al., 2014) involves reconstruction of a species tree, by "neighbour joining" from orthologs (see section 3.2.10) retrieved from the InParanoid database (Sonnhammer and Östlund, 2015). This is followed by the identification of sub-trees present or absent in both species trees to assign positive or negative evidence, respectively. The score is obtained separately for the positive and negative evidence by dividing the branch lengths of sub-trees by the branch length of the full tree. Other more complex, albeit computationally expensive, metrics have also been proposed.

The increasing number of sequenced genomes provides larger phylogenetic profiles, potentially increasing the accuracy of the co-evolution-based evidence. However, there is the caveat that gene duplications, loss of function and other genomic events could introduce noise in this evidence type.

3.2.5 Genetic interaction

Two or more genes are said to interact genetically when mutations in them produce an unexpected phenotype. In an extreme case, two benign mutations can result in cell death (Costanzo et al., 2016). This can also be considered evidence of functional association. The most common high-throughput experimental technique to determine genetic interactions is "synthetic genetic array" analysis (Tong, 2001), which allows evaluation of systematically induced pairwise mutations in a model organism of interest, e.g. *S. cerevisiae* (Boone et al., 2007). The most common metric for this evidence type is the profile similarity of gene pairs, calculated using Pearson correlation.

3.2.6 Co-regulation

Many signalling cascades include a step where one or several transcription factors (TFs) are affected, resulting in changes in the expression of genes regulated by the affected TF(s). This is an example of functional association through co-regulation.

The two most widely used high-throughput experimental techniques to obtain data on TF binding sites are chromatin immunoprecipitation–on–chip (ChIP-chip) and chromatin immunoprecipitation sequencing (ChIP-seq).

Data from ChIP-chip and ChIP-seq experiments is considered direct evidence of regulatory activity of TFs on their gene targets, since it provides information about the TF binding to the target gene's promotor region. The analysis of changes in the expression pattern of the target gene as a result of changes in the expression of the TF is considered indirect evidence. Both types of evidence are tracked in databases (Abdulrehman et al., 2011; Brown and Celniker, 2015; Myers et al., 2011) and can be scored by studying the shared fraction of target sites for a gene pair multiplied by the cardinality of the subset of shared target sites.

Another important regulatory mechanism is the post-translational control of protein coding genes by microRNA (miRNA). miRNA is a 23 nucleotide (nt) short RNA which stems from transcribed DNA. miRNA forms hairpin-like structures and can basepair to the mRNA of a target gene, affecting the gene's post-translational levels (Bartel, 2009). Regulatory control by other types of non-coding RNA, e.g. small interfering RNAs (siRNAs), Piwi-interacting

RNAs (piRNA), etc. also known as "RNA interference", has also been demonstrated. miRNAs are however of particular interest, since each miRNA usually has a high number of specific targets, providing it with the ability to profoundly impact functional association networks in a cell.

Experimental data on regulation by miRNA is still scarce and most of it comes from predictions based on sequence-base-pair complementarity of miRNA and its target mRNAs (Betel et al., 2008). Profiles of regulating mRNAs can be assembled for each gene; then a similarity score of a gene-pair can be calculated in the same way as for TF profiles.

3.2.7 Co-expression

Genes with similar co-expression profiles tend to interact more frequently than would be expected by random chance (Ge et al., 2001; Grigoriev, 2001; Lee et al., 2010), because co-expression is a proxy for co-regulation. This feature has been exploited in prediction of functional associations of proteins coding for the co-expressed genes.

Co-expression of genes across many experimental conditions (multiple time points or states) can be scored using Pearson correlation or Spearman's rank correlation, but also MI can be applied to discretised expression profile values.

Microarrays

Microarray experiments are still the main source of mRNA expression data. One of the strengths of microarray experiments is the large amount of data they yield providing high genome coverage for analysis. Tens of thousands of microarrays, in thousands of species and experimental conditions are collected in the gene expression omnibus (GEO) (Barrett et al., 2013) and ArrayExpress (Kolesnikov et al., 2015).

RNA sequencing

In order to assess the quantity of RNA in a sample, a whole transcriptome shotgun sequencing technique can be used. This technique is called "RNA sequencing" (RNA-seq). The rapid evolution of sequencing techniques in recent years has produced enough RNA-seq data for use in co-expression analysis (Hong et al., 2013).

There are several advantages of RNA-seq over microarray technology, including a more comprehensive and unbiased coverage of the transcriptome, with the ability to detect alternative gene splicing, post-translational modifications, gene fusions, SNPs, and various small RNAs e.g. miRNA, ncRNA,

tRNA, etc. (Morin et al., 2008). Compared to microarrays, RNA-seq has therefore higher specificity and sensitivity in detection, and lacks cross-hybridisation issues.

Limitations of the RNA-seq technology, beyond its relative immaturity, include higher costs compared to microarrays, making it expensive for larger studies. The short sequence reads produce high uncertainties when it comes to paralogs (see section 3.2.10) and sequences of low complexity or high similarity, making the correct assignment of reads to genes problematic. Additionally, there are biases and limitations common to all next-generation sequencing approaches, such as non-uniformly-distributed read depths, primer and tag composition effects, etc. (Zhao et al., 2014). This is why RNA-seq has not yet been fully embraced as the primary source of gene expression data and microarrays are still dominate.

Protein expression

Levels of mRNA expression do not always correspond with the levels of protein expression due to post-translational modifications and regulation (Katagiri and Glazebrook, 2009). Protein co-expression can therefore be studied separately providing a more direct measure of functional association, albeit with lower coverage.

The human protein atlas (HPA) (Uhlen et al., 2015) is a comprehensive database of protein expression in a collection of 32 human tissues and 36 human cell lines. Protein expression profiles in HPA use an antibody-based quantification of target proteins in tissue and cell microarrays.

Another way to quantify the level of proteins in a cell is using different proteomic techniques, e.g. quantitative MS (Deutsch et al., 2008). PaxDB(Wang et al., 2012) hosts a large collection of protein abundance data across different species, tissues, cell types, and experimental conditions, e.g. samples, environmental conditions and developmental stages.

Protein expression profiles based on tissue, cell types, and experimental condition, can be generated for each protein and used for scoring. Such a metric would consist of pairwise similarity between protein abundance profiles, measured by MI or the adapted Jaccard index (Schmitt et al., 2014).

3.2.8 Co-localisation

Two proteins in the same signalling cascade do not have to share the same physical sub-cellular location, i.e. co-localise, in order to be functionally associated. Physically interacting proteins, on the other hand, have to co-localise at some point. Sub-cellular co-localisation can therefore be used as evidence of functional association (Huh et al., 2003).

There are several resources containing experimental and predicted subcellular localisation data for single (Orfanoudaki and Economou, 2014; Uhlen et al., 2015) and multiple organisms (Negi et al., 2015; Rastogi and Rost, 2011), where the cellular component of the Gene Ontology (GO) is the largest.

Scoring of this evidence type increases in reliability when co-localisation is measured across many cell compartments. It also strongly depends on the total number of proteins occupying the compartment, with lower confidence associated with the largest compartments in terms of number of localised proteins, due to the higher probability of random co-localisation. Therefore, a suitable metric for this evidence type can be co-occurrence in or MI of the pattern of compartments, with terms weighted by the size of the compartment (Alexeyenko and Sonnhammer, 2009).

3.2.9 Co-annotation

Data in the scientific literature can be mined for PPIs using natural language processing and other techniques (Marcotte et al., 2001). Irrespective of the approach, two fundamental problems need to be solved: finding the gene names and identifying how they are related. A popular approach involves searching abstracts from databases of scientific literature such as Medline for co-occurrence of gene and protein names. The co-occurrence is compared to the frequencies and distributions of gene and protein names in the entire database. A challenge of such approaches is that co-occurrences that are only present in the main body of the articles are not available for discovery, which may lead to high rates of false negatives. Another challenge is the ability to identify genes and proteins correctly despite the plethora of gene and protein identifiers, e.g. NCBI (Acland et al., 2014), RefSeq (Pruitt et al., 2005), etc. (Arighi et al., 2011; Morgan et al., 2008), leading to an increased risk for high false positive rates. Distinguishing between positive and negative interactions from term co-occurence may be even more problematic.

3.2.10 Orthology transfer

Genes in different species that stem from the same gene in an ancestral species are called orthologs and have been shown to be functionally conserved (Matthews et al., 2001). Evidence of functional association in one species mapped via orthologs to another species is known as "orthology transfer". This technique can be used to increase the coverage of genes in a network to the extent of reconstructing the entire network for a species without any available large-scale datasets (Alexeyenko and Sonnhammer, 2009).

3.3 Network prediction

When all the evidence of functional association has been gathered, it can be integrated into a cohesive network representation. Unfortunately, the collected data stems from experimental techniques or prediction efforts seldom intended for integration into networks. This poses challenges such as scale, both in terms of differences between the datasets easily adding up to orders of magnitude, and in general, as seen with the millions of data points of microarray data. The data also tends to be heterogeneous in terms of metrics, directness of interaction, and quality. Additionally, all data sources have inherent biases and are sometimes overlapping, especially for the well-studied genes and proteins. Such overlaps can produce enough supportive data to generate false positives.

Different network prediction techniques, accounting for these factors have been proposed. One of the first techniques developed to reduce noise was a majority-voting approach, where only links that appear in the majority of the underlying data sources were kept in the network (Marcotte et al., 1999).

Another approach is to generate individual networks for each data source and let the resulting composite network be the unweighted sum of the individual networks (Pavlidis et al., 2002). The individual networks can also be assigned weights based on the type of function being predicted and to optimise the machine learning technique used for integration, e.g. SVM (Lanckriet et al., 2004) or Gaussian random fields (GRFs) (Tsuda et al., 2005). Additionally, individual datasets can be weighted using ridge regression (Mostafavi et al., 2008) based on the abundance of examples in the training set, i.e. the gold standard. Another popular technique that uses gold standards to evaluate each type of evidence and combine them into a composite network is naïve Bayesian classification (Myers and Troyanskaya, 2007).

The rest of this section will focus on aspects related to naïve Bayesian classification and its application to network prediction. It also discusses some of the challenges of network prediction.

3.3.1 Gold standards

Many of the network prediction techniques require positive and negative training examples, i.e. positive and negative gold standards, in order to assess performance and quality of the data. A suitable positive gold standard should consist of verifiable, high-quality associations. Physical interactions between proteins, membership in the same protein complex, part of the same signalling or metabolic pathway, or the same functionally transcribed union, i.e. shared operon, are all examples of positive gold standard classes. Training on each of these classes produces a network that is enriched in the corresponding class of

functional association. Combining the resulting networks ensures that the final network accounts for all the trained classes of functional associations.

Gene pairs co-annotated in a set of GO terms curated by experts (Wong et al., 2015) or reliable examples, from e.g. HPRD (Rhodes et al., 2005), have been directly used for the positive gold standard. However, these approaches do not account for the high rates of false positives in PPI experimental techniques. To minimise the risk of false positives, the PPIs in the gold standard need to be supported by multiple independent assays or be present in multiple gold standard classes (Schmitt et al., 2014). Experimental origins of PPIs and protein complexes are readily annotated in iRefIndex, which makes this database a suitable source for *PPI* and *protein complex* gold standard classes.

Kyoto Encyclopaedia of Genes and Genomes (KEGG) is commonly used as the source of gold standards (Alexeyenko and Sonnhammer, 2009; Szklarczyk et al., 2017; Wong et al., 2015). A positive gold standard pair in the *signalling* class can be a gene pair, where both genes are members of a small[1] signalling pathway or a signalling pathway of any size if the gene pair is also a member of several[2] other pathways in KEGG (Schmitt et al., 2014). This applies similarly to the *metabolic* gold standard. Another database used as the source of gold standard examples is OperonDB (Pertea et al., 2009), where gene co-membership in the same operon is a requirement for assignment to the *shared operon* gold standard class (Ogris et al., 2017b).

Ideally, verifiable, high-quality data should also be the source of negative gold standard examples. However, 'absence of evidence' does not guarantee 'evidence of absence' making it challenging to find high-quality negative examples. Approaches of assigning protein pairs to the negative gold standard are based on selecting pairs annotated to distinctly different sub-cellular localisations, e.g. plasma membrane and nucleus, or on the lack of co-annotation in pathways or terms of either KEGG or GO, respectively (Rhodes et al., 2005; Szklarczyk et al., 2017). However, none of these guarantee a complete lack of association. Even proteins with distinctly different localisations may occasionally interact, especially when functional association is considered instead of pure physical interaction. To avoid polluting the training set with falsely assigned negative examples, a set of randomly selected genes can be used in the negative gold standard to form a baseline of random interactions (Alexeyenko and Sonnhammer, 2009). The set can be made sufficiently large to avoid sampling errors and connections between the genes can be randomly generated.

Not all network prediction methods rely on external gold standard data for training and evaluation. Some use topological properties of the network, such as modularity, to assign a 'confidence' of an interaction or a dataset (Kamburov

[1] Species dependant e.g. for *H. sapiens* < 20

[2] Species dependant e.g. for *H. sapiens* ≥ 2

et al., 2012). Such approaches can be a good alternative when external gold standard data is scarce or biased. However, they are not able to address other fundamental limitations of gold standards, such as inability to represent the full range and complexity of functional associations in a cell and lack of the less frequently studied proteins and unusual classes of interactions (Yu et al., 2012).

3.3.2 Bayes' theorem

Probability is a central concept in statistics. There are two main schools of thought when it comes to the interpretation of probability. The *frequentist* interpretation of probability $P(A)$ of an event A is the proportion of times A is observed in a large number of experiments. The *Bayesian* approach sees probability $P(A)$ as a *degree of belief* in A, i.e. how likely A is to occur. Bayesian interpretation gets its name from reverend Thomas Bayes (1701-1761), who derived an equation that provides a way of combining new evidence with prior beliefs (3.1). This is opposed to the *frequentists* approach, which only relies on the observed evidence.

$$P(A \mid E) = P(A)\frac{P(E \mid A)}{P(E)} \tag{3.1}$$

We can use Bayes' theorem to calculate the probability of a functional association A given the evidence E, i.e. $P(A \mid E)$, also known as the "posterior". The initial probability of observing a functional association is called the "prior", $P(A)$. The probability of observing the evidence when a functional association exists is referred to as the "likelihood", $P(E \mid A)$, while the probability of observing the evidence is known as the "marginal probability", $P(E)$. Therefore, the posterior can be interpreted as the product of the prior belief $P(A)$ and the new evidence, comprised of the ratio between the likelihood and the marginal probability $P(E \mid A)/P(E)$.

Bayesian classification

Another way to study the degree of belief in an association is by viewing it as a classification problem, between the presence of an association A or its absence A'. This can be done by computing the odds of the posterior and its complement, i.e. $P(A \mid E)/P(A' \mid E)$ and is known as the "Bayes' factor" (3.2).

$$\frac{P(A \mid E)}{P(A' \mid E)} = \frac{P(E \mid A)}{P(E \mid A')}\frac{P(A)}{P(A')} \tag{3.2}$$

When considering multiple pieces of evidence $\{e_1, e_2, \ldots, e_n\} \in E$ for an association, Bayes' factor can be expressed as the product of likelihood ratios

between the marginal probabilities for each piece of evidence (3.3). This is known as the "naïve Bayes' classifier" and is only possible when evidence is conditionally independent, i.e. the presence of one piece of evidence is unrelated to the presence of any other piece.

$$\frac{P(A \mid E)}{P(A' \mid E)} = \frac{P(A)}{P(A')} \prod_i \frac{P(e_i \mid A)}{P(e_i \mid A')} \tag{3.3}$$

Taking the logarithm on both sides of (3.3) provides a more convenient form of the classifier (3.4), where log-posterior odds, on the left, is described as the sum of the log-prior odds and the natural logarithm of likelihood ratios (LLRs) on the right.

$$ln\frac{P(A \mid E)}{P(A' \mid E)} = ln\frac{P(A)}{P(A')} + \sum_i ln\frac{P(e_i \mid A)}{P(e_i \mid A')} \tag{3.4}$$

Discretisation

Before integration of evidence can proceed, scores based on the metrics of each evidence type need to be processed. Continuous scores are not particularly suitable for Bayesian network integration due to the difficulty of estimating joint probability densities over all the nodes in a network. In the naïve Bayesian approach, joint probabilities are not required, and dependencies between continuous scores can be fit using regression. However, such a fit requires huge computational efforts and still involves a level of discretisation, albeit on very small intervals. If scoring data is unequally spread, missing, or too sparse, regression may not be applicable and proper discretisation, i.e. binning, can be performed.

A manually curated binning procedure or a fixed number of bins are unsuitable for an automated network construction approach, particularly when it involves multiple large species networks. Various adaptive discretisation techniques have thus been proposed. One such technique uses conditional entropy (Butterworth et al., 2004). However, it requires a parameter being set empirically after integration, making the technique computationally and practically difficult. Another approach (Alexeyenko et al., 2011) is unsupervised and involves sorting of the metric scores for individual interactions, which are labelled as positive, negative, or unknown in accordance with their membership in the gold standard. The dataset is then split based on the χ^2 (chi-square) test, comparing labels on the right with those on the left of a split. A split is only made for significant χ^2 tests and only if the number of labels in each class in a potential partition exceeds some minimum value. The number of splits can be limited by having a maximum number of allowed bins.

Confidence score

By applying discretisation and using the gold standards an evidence-specific LLR score for each bin can be calculated (3.5). The likelihood of association $P(e_i \mid A)$ is assessed by the occurrence of e_i in the positive gold standard, while the likelihood of lack of association $P(e_i \mid A')$ is estimated by the occurrence of e_i in the negative gold standard. The log-prior odds can be left out since it is constant.

$$LLR_i = \sum_i ln \frac{P(e_i \mid A)}{P(e_i \mid A')} \tag{3.5}$$

Each of the classified associations receives an LLR_i score corresponding to the bin it falls in and the piece of evidence e_i used. A more convenient measure of confidence, denoted p, with $0 \leq p \leq 1$, summarises all the evidence E for an association (3.6). The log-prior odds is denoted as the constant c and the sum of $LLR_i(A)$ across all evidence $e_i \in E$ is shown as LLR_E.

$$p = \frac{1}{1 + e^{-c - LLR_E}} \tag{3.6}$$

Redundancy reduction

The integration of different types of evidence in a naïve Bayesian fashion, either directly or in a network form, carries the assumption of conditional independence between datasets. This assumption is not always valid because of occasional overlaps in the interactions assessed in different studies. Violation of this assumption may generate false positives, especially when multiple small LLRs from separate but non-independent evidence types are combined into a large LLR, falsely producing a high confidence link.

One way to address this issue is to compute the added shared information U_k that a dataset D_k contributes and assign the result as the weight for that dataset (Park et al., 2013; Wong et al., 2015). The shared information, depends on the MI of the unrelated gene pairs $I_{pairs \in negative}$ of D_k and other datasets D_i and the entropy $H(D_k)$ of the D_k.

$$U_k = \frac{\sum_{i=k} I_{pairs \in negative}(D_k, D_i)}{H(D_k)} \tag{3.7}$$

The weight w_k for D_k is then set to decrease exponentially with decreased shared information, i.e. $w_k = 2^{U_k} - 1$.

Another approach is to weight every link in a dataset of the same evidence type proportionally to the amount of novel information it adds with respect to other datasets of the same type. In this approach, the different datasets $e_{kt} \in E_t$

of the same evidence type t are ranked by their LLR_{kt} in decreasing order. Then the information distance between two datasets e_{kt} and e_{it} is calculated, based on the Spearman's rank correlation r_{ki} as $\alpha(1-max(0,r_{ki}))$. The correction parameter α is estimated from the noise in the data. The LLR for the evidence type t is the sum of the LLRs from different datasets, weighted by the product of the distances between the LLRs (3.8).

$$LLR_t = \sum_{k}^{|E_t|} LLR(e_{kt}) \prod_{i<k} \alpha(1-max(0,r_{ki})) \tag{3.8}$$

The use of naïve Bayesian classification in integration of multiple heterogeneous types of evidence is a popular approach to network prediction. It has been shown to outperform other statistical classifier-based approaches (Rhodes et al., 2005) for genome-wide data as long as its independence criteria are fulfilled or properly addressed. It is particularly tolerant to missing values and is practically unbiased. The convenience of its automatic provision of confidence scores is complemented by its robustness and lack of sensitivity to noise.

The most apparent limitation of naïve Bayesian prediction is the previously mentioned assumption of conditional independence. This assumption is practically impossible to fulfil, making the redundancy managing techniques a prerequisite.

3.3.3 Limitations of network prediction

One of the limitations of current network prediction approaches is the loss of contextual information, e.g. disease state and organ or tissue of interaction, when different types of data are projected onto a single network representation (Cho et al., 2016). For instance, tissues are hierarchically organised, with proteins in more biologically similar tissues having similar functions; a fact that may be lost in the network integration process, when tissue-specific information is omitted.

Another limitation inherent to all available, comprehensive networks is their static view of functional associations in a cell without considering the dynamical aspects of interactions, such as e.g. stable interactions in a complex and transient interaction events in a signalling cascade; or the fact that genes, proteins, and even pathways can have different time-dependent functions (Hu et al., 2016). Despite advances in differential networks (Ideker and Krogan, 2012) such as the application of context specific interactions to a static network; and quantitative techniques such as time-resolved MS (Chen and Urban, 2013) aiming to provide a more dynamic representation of the complexities of interactions, the coverage needed for a comprehensive approach is still limited.

Most of the network prediction methods, including naïve Bayesian integration, utilise supervised learning techniques that rely on a gold standard. The quality and coverage of the gold standard data as well as the general lack of negative gold standards therefore have a huge impact on the accuracy of the current methods. Proposed unsupervised techniques, such as "multiple dataset integration" (Kirk et al., 2012), which assigns functional associations based on similarities in the clustering of genes in different datasets and other network characteristics, have yet to be implemented on a larger scale.

Finally, none of the current methods is immune to annotation bias stemming from unequal investigations of different genes and phenotypes.

3.3.4 Example networks

This section contains a non-exhaustive selection of functional association networks of note, which represent different approaches to network prediction and are being actively maintained.

ConsensusPathDB (CPDB) integrates interaction data into a global network for *H. sapiens* from over 30 databases (Kamburov et al., 2013). The types of evidence included are PPIs, metabolic and signaling pathways, as well as regulatory, genetic, and drug-target interactions. The data has been extracted from databases such as BIND, KEGG, IntAct, etc., and add up to over 500,000 interactions. Instead of relying on external gold standard data, each interaction is assigned a confidence score based on the modularity properties of the network (Kamburov et al., 2012). This method applies Markov clustering (Dongen, 2000) on an interaction network to identify modules of functional interactions. Proteins in the same module are more likely to interact with each other, resulting in a higher interconnectedness. The confidence of an interaction is determined based on how specific both proteins are to the respective module.

Multiple association network integration algorithm (GeneMANIA) contains networks for 9 model organisms covering almost 300 million interactions between 150,000 genes (Zuberi et al., 2013). GeneMANIA uses PPIs and genetic interactions from iRefIndex and BioGRID, as well as interactions predicted using protein domains extracted from InterPro, co-expression from selected GEO datasets, and other manually curated datasets to construct weighted, dataset-specific networks. Various different weighting algorithms can be applied, with regularised ridge regression (Hastie et al., 2009) being the default together with co-annotation in GO as the gold standard. Link weights can be uploaded directly by the user, and are then used as a prior before the ridge-regression procedure. The dataset-specific networks are combined into a

composite network. The link weights in the composite network are the product of the degree-normalised link weights from the dataset network, and the corresponding network's *network weight*.

GeneMANIA's real-time predictions require fast computations meaning sacrifices in terms of data coverage. Therefore, only the 20 'most informative' GEO datasets are used for prediction.

FunCoup is described in sections 5.1 and 5.5.

Integrated interactions database (IID) specialises in tissue-specific PPIs for *H. sapiens* and five model organisms. Experimentally obtained PPIs from e.g. BioGRID, IntAct, etc. are integrated with predicted, high-confidence PPIs from four PPI prediction studies (Elefsinioti et al., 2011; Kotlyar et al., 2015; Rhodes et al., 2005; Zhang et al., 2012) for each species and transferred using orthologs from HomoloGene across all species.

The resulting 1.8 million interactions are mapped to up to 30 different tissues using mRNA expression datasets from GEO and the protein expression databases HPA and PaxDB. An interaction is said to be present in a tissue only if both of its constituents are expressed in that tissue.

Integrative multi-species prediction (IMP) consists of networks of functional association for *H. sapiens* and six model organisms (Wong et al., 2015). It uses PPIs and genetic interactions from BioGRID, IntAct, MINT, and MIPS, scoring each interaction based on the number of assays in which it appears. Associations based on co-regulation by similar patterns of TFs are extracted from Yeastract (Abdulrehman et al., 2011) and Jaspar (Mathelier et al., 2016) and scored using Pearson correlation between the TF pattern vectors. Microarray data for co-expression-based associations is obtained from GEO, and scored using Pearson correlation. Protein sequence similarity from BioMart (Kinsella et al., 2011) and the presence of protein domains in PfamA and PROSITE are the last types of association evidence used. TreeFam is used for orthology transfer.

A set of approximately 1,000 expert-curated GO terms from the biological process (BP) ontology are used as the gold standard. The positive training examples are defined as co-annotation of a gene pair among the GO terms, while the negative training examples are gene pairs lacking co-annotation in GO, KEGG, PID or Biocyc.

For each GO BP term, a network is predicted using a regularised naïve Bayesian approach. Regularisation is performed in order to reduce redundancy. It consists of each dataset receiving a weight, based on the conditional MI between datasets, where an increase in shared information exponentially

decreases the weight.

Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) is one of the most widely used resources of functional associations (Szklarczyk et al., 2017). It integrates experimentally produced 'known' interactions from primary databases, e.g. BIND, IntAct, PID, and curated databases such as: GO, KEGG, and Reactome, with interactions predicted using gene fusions and conserved gene neighbourhood, and phylogenetic co-occurrence, text mining of Medline and mRNA co-expression data.

Predicted interactions are benchmarked using KEGG pathways, providing each dataset with a score that represents the probability of finding the tested interacting partners within the same KEGG pathway. Interactions predicted in one species are transferred to other species using orthologs from the eggNOG database. Orthology transfer is weighted by the evolutionary distance between species in a phylogenetic tree. The scores from each data source s_i, are assumed to be independent and are integrated in a naïve Bayesian fashion into a 'combined' score S for the interaction i.e. $S = 1 - \prod_i (1 - s_i)$.

There are clear advantages to STRING compared to other networks of functional associations. It has an application programming interface (API) access and uses a comprehensive set of more than 2,000 species for orthology transfer. It also stands out by incorporating all GEO microarray platforms and experiments.

There are also limitations in the form of the majority of species being prokaryotic where species definitions are not always clear, which can bias the interactions obtained. Some of the data sources are imported directly without being benchmarked against the gold standard, meaning that high FPRs of experimental methods (see section 3.2.1) allow for large amounts of false positive interactions to be incorporated into the network. Additional noise and unintentional bias may also be added by interactions predicted using text-mining (Harmston et al., 2010).

3.3.5 Other important databases

The Gene Ontology project is an effort to provide consistent, species-agnostic descriptions of gene products as a controlled vocabulary (Ashburner et al., 2000). GO consists of three ontologies: biological process (BP), molecular function (MF), and cellular compartment (CC). Each ontology is a hierarchical structure of gene product attributes in the form of a directed acyclic graph (DAG), i.e. a graph with no loops with directed edges. The CC ontology contains terms describing gene product locations within a cell, e.g. nucleus, lysosome, etc. The MF ontology terms describe molecular-level activities of

individual gene products or complexes, without specifying where, when, nor in what context an activity takes place; e.g. insulin binding, kinase activity. The BP ontology terms describe organised series of molecular functions, e.g. the protein modification process, histone-tyrosine phosphorylation.

The process of assigning a particular term to a gene product is called "annotation" and is performed by a number of research groups of the GO consortium (Blake et al., 2015). Each annotation has an evidence code assigned to it, depending on the type of evidence behind the annotation. Evidence codes are divided into experimental, stemming from physical characterisations of gene products, e.g. inferred from direct assay (IDA); and computational, based on sequence similarity analysis, e.g. inferred from sequence alignment (ISA). These evidence codes have varying degrees of curation behind them. An additional evidence code, inferred from electronic annotation (IEA), is reserved for annotations assigned by automatic methods without any curation step.

The Kyoto Encyclopaedia of Genes and Genomes is a suite of 15 manually curated databases and computational resources containing information on pathways, chemical reactions, genomes, expression, diseases, and drugs (Kanehisa et al., 2017). The information in KEGG is divided into four categories, namely *system* (pathways, modules, and brite), *genomic* (genome, genes, orthology), *chemical* (compound, reaction, enzyme) and *health* (disease, drug environ). KEGG provides functional annotation of genes in completely annotated genomes, reconstructs pathways based on literature of experimental annotations, and computes biological networks from molecular interaction data.

3.4 Experimental validation

In a perfect world, all newly predicted associations would be experimentally validated. However, due to the size of functional association networks, this is practically impossible. In addition, the definition of a functional association reaching beyond pure physical interaction poses further challenges in experimental validation, demanding different types of experimental techniques to be used for different classes of associations.

A reasonable approach is therefore to determine if a representative sample of the predicted associations provides biologically meaningful insights. Sample selection is however usually neither trivial nor independent of context and could involve a set of genes with a particular importance for a certain phenotype, e.g. cancer, or a cellular process (Snider et al., 2015).

Experimental techniques for validation are usually based on determining a physical association between proteins. Generally, *in situ* methods are preferred

for validation, due to their ability to provide tissue- and cell-type-specific direct measurements of protein associations instead of relying on averaging across cell types, tissues and sub-cellular localisations. In order to enable the detection of interactions, experimental techniques either use modification of the studied proteins, e.g. Y2H, (T)AP-MS (section 3.2.1), etc.; utilise other molecules to detect them, e.g. coIP (section 3.2.1) or fluorescent resonance energy transfer (FRET) (section 3.4.1); or rely on the physical properties of the studied proteins for detection, e.g. proximity ligation assay (Söderberg et al., 2006). All methods have their strengths and limitations and the choice of experimental technique is usually determined by the problem domain.

3.4.1 FRET

FRET is one of the most widely used experimental techniques for the *in situ* study of interactions between molecules. It relies on transfer of energy from a "donor" fluorophore to a molecule located in its proximity ($\leq 10nm$), usually also a fluorophore termed the "acceptor" (Lakowicz, 1999). This technique involves excitation of the donor fluorophore and detection of the light emitted by the acceptor using fluorescence microscopy (Fig. 3.1). The efficiency of the energy transfer E, is the fraction of energy absorbed by the donor and transferred to the acceptor. E is inversely proportional to the sixth power of the distance between the fluorophores, R_{DA} (3.9). The distance at which E is 50% is called the Förster radius, R_0, and has been determined for every pair of fluorophores.

$$E = \frac{R_0^6}{R_0^6 + R_{DA}^6} \tag{3.9}$$

Proteins that make up the interaction studied by FRET can be directly labelled by donor and acceptor fluorophores or using fluorophore-conjugated antibodies against the studied proteins. Another common setup that allows for larger studies is where proteins are first labelled by so-called "primary antibodies", then fluorophore-conjugated secondary antibodies against the primary antibodies are used to introduce the fluorophores (Fig. 3.1) (Guala et al., 2017). Despite its convenience, the latter technique may introduce false negatives, due to the size of antibodies potentially increasing the distance between fluorophores.

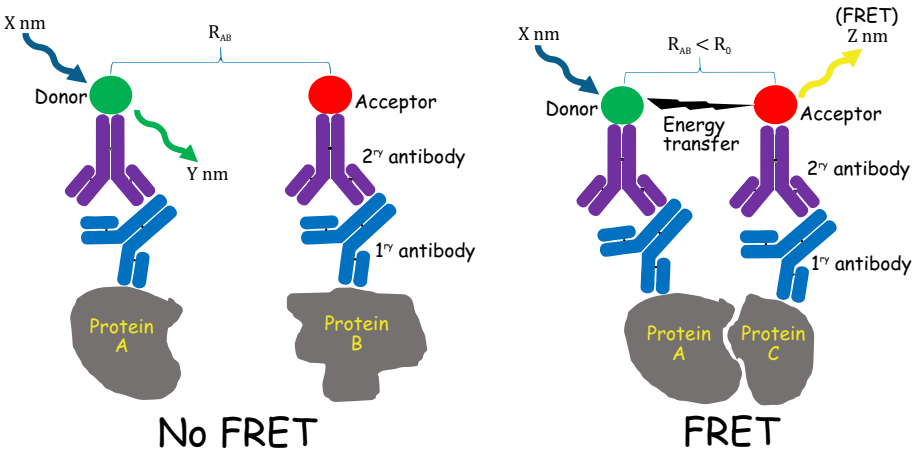


Figure 3.1: Fluorescent resonance energy transfer - Proteins *A* and *B* are labelled with primary antibodies. Fluorophore-conjugated secondary antibodies against the primary ones are applied. The sample is illuminated by light of excitation wavelength Xnm and the emission light is studied using fluorescent microscopy. Transfer of energy occurs between the donor and the acceptor fluorophores if the distance between them R_{AB} is sufficiently small. When energy transfer occurs it can be observed by the emission of light of Znm, otherwise the emitted light will have the wavelength of Ynm. (Guala et al., 2017)

4. Network applications

"Neurons that fire together, wire together."
–Donald Olding Hebb

Functional association networks encode topology of biomolecular interactions. This information can be used to assign function to previously unannotated genes in the investigation of disease mechanisms (Goh et al., 2007).

Most functions cannot be established experimentally due to the rapid growth of new sequences and limitations in the experimental methods. Thus, there is a need for other ways in which to determine the function of genes. Topological information embedded in functional association networks lends itself well to the *guilt by association* (GBA) principle (Aravind, 2000), where annotations from genes with known functions are propagated to associated genes lacking established annotations (Lee et al., 2006; Letovsky and Kasif, 2003; Oti and Brunner, 2007). GBA is based on the assumption that associated or interacting proteins are more likely to share function.

An understanding of the molecular mechanisms behind diseases is imperative for the selection of the right drug targets in the early stages of drug development. Prioritisation methods (see section 4.2) apply information about the already known disease genes on the topology of functional association networks using GBA to identify new disease genes and potential disease mechanisms (Ideker and Sharan, 2008).

Survival prediction in patients with kidney cancer has demonstrated much better results when pathway- instead of gene-centric information is used. This can be explained by the fact that survival mechanisms of certain tumours are tied to a certain pathway rather than to exact molecular alterations (Hofree et al., 2013). In order to combat such tumours, techniques that search for destabilised pathways can be useful instead of those identifying individual genes. However, different results have been seen for different cancer types, demonstrating the need for both gene- and pathway-centric approaches.

The remainder of this chapter will describe functional enrichment analysis and prioritisation methods together with a way to assess the performance of the latter using benchmarking.

4.1 Functional enrichment analysis

Measuring whole genome expression under different experimental conditions, different perturbations, or in time series has become a relatively inexpensive way to map the regulatory elements of the studied specimen. Results from gene expression studies usually contain a set of differentially expressed genes. Without the functional context of the identified genes, it is difficult to understand their connection to the applied perturbations and the molecular mechanisms that drive them. The first step in understanding the results is to map them to the critical level of biological organisation, i.e. the underlying functional modules (Hartwell et al., 1999). This process is called "functional enrichment analysis" or "pathway annotation" and involves the identification of pathways with known annotations or modules in a functional association network to which the differentially expressed genes belong.

Whenever multiple tests are performed, e.g. testing whether a gene set belongs to a number of different pathways or terms, there is an increased risk that some tests produce statistically significant results purely by chance, i.e. an increased "Type I error" rate (Tab. 4.1). This warrants corrections for multiple hypothesis testing. One common way of compensating for the increased Type I error rate is by decreasing the significance threshold α by the number of tested hypotheses m. This means that smaller p-values, i.e. $p_k \leq \frac{\alpha}{m}$ are needed to reject the corresponding null hypotheses. This is known as the Bonferroni procedure (Benjamini and Hochberg, 1995).

Unfortunately, the conservative Bonferroni procedure can lead to a loss of statistical power, increasing the number of false negatives. Another popular approach, which retains statistical power, is to limit the "false discovery rate" (FDR) to a desired level α. This is known as the Benjamini-Hochberg (B-H) procedure and involves sorting of all produced p-values in a descending order, i.e. $p_1, p_2, \ldots, p_m$, and then keeping only the p-values for all k, where (4.1) is fulfilled (Benjamini and Hochberg, 1995).

$$p_k \leq \frac{k}{m}\alpha \tag{4.1}$$

4.1.1 Gene enrichment analysis

Traditionally, pathway annotation has been carried out using gene enrichment analysis (GEA), which measures the direct gene overlap of the gene set of interest with pathways or terms from KEGG or GO, respectively. The most commonly used method for functional enrichment analysis is the GEA-based "Database for Annotation, Visualisation, and Integrated Discovery" (DAVID) (Huang et al., 2009). In GEA, the proportion of genes from the gene set of

interest, annotated to a pathway or a GO term, is contrasted with the proportion of genes expected to be annotated with that pathway or GO term by chance. To assess the significance of the overlap, statistical tests such as Fisher's exact test, the χ^2, or the hypergeometric test are performed, producing a p-value for each pathway or term. Following correction for multiple hypothesis testing, the pathways or terms are ranked by their p-values.

4.1.2 Functional class scoring

Functional class scoring (FCS) tools, e.g. gene set enrichment analysis (GSEA) (Subramanian et al., 2005), builds on assumptions and gene overlap testing strategies similar of GEA, with the addition of gene expression levels in terms of e.g. fold-change or t-statistic. This requires a slightly different set of statistical tests, e.g. the Kolmogorov-Smirnov statistic or Wilcoxon rank sum test, to assess the pathway overlap. The only clear advantages of FCS are that information about the importance of the differentially expressed genes is not discarded, as in GEA, and the dependence of genes in a pathway is assessed.

4.1.3 Pathway topology-based methods

Some of the limitations of GEA and FCS are remedied when interactions of genes in a pathway are added to the overlap methods, resulting in pathway topology-based (PTB) methods, e.g. Pathway-Express (Draghici et al., 2007), SPIA (Tarca et al., 2009). Besides the information used in FCS, PTB methods utilise information about the position of the differentially expressed genes in the pathway with respect to down- and upstream genes as well as the type of interaction they are involved in, e.g. repression, induction.

4.1.4 Network cross-talk analysis

Partly utilising the underlying functional association network by extending the target gene set with neighbouring genes in order to later use overlap assessment is one way to use the underlying network.

Another way to utilize the underlying functional association network is by network cross-talk analysis (NCA), where "cross-talk" refers to the connectivity between two gene sets in the network. This procedure considers the topology of the whole underlying network and is utilised by e.g. network enrichment analysis (NEA) (Alexeyenko et al., 2012), network enrichment analysis test (NEAT) (Signorelli et al., 2016), CrossTalkZ (McCormack et al., 2013) and BinoX (Ogris et al., 2017a).

Hypergeometric link distribution

One way to utilize network cross-talk is to assume hypergeometric distribution of links between two gene sets A and B in the context of a network i.e. $N_{AB} \sim \mathcal{H}(d_A, d_B, d_V)$, where d_A, d_B and d_V are the total degrees of sets A, B and the set of all network nodes, respectively. Then the link- instead of gene-overlap between the target gene set and different pathways or GO terms, can be tested for statistical significance, as it is done in NEAT.

Random model of the network

Other methods to ensure that the observed cross-talk is not due purely to chance involve comparison of the observed cross-talk to a distribution of expected cross-talk generated by a random model of the network. A random model can be produced by randomly rewiring connections or relabelling the nodes in the network. However, such a procedure does not preserve the topological network information, e.g. a hub may be rewired to a leaf node and vice versa. Instead, a procedure that shuffles the links in a way that preserves both the network degree distribution and the degrees of neighbouring nodes, such as *link assignment with second order conservation* (LP+S) (McCormack et al., 2013), can be used. In LP+S, all network nodes are ranked by their degree d and divided into bins $bin(d)$ corresponding to the logarithm of their degree $bin(d) = \lfloor log(d) + 1 \rfloor$. Then the node labels are permuted in such a way that the new labels are selected at random from the same bin.

Normally distributed model

NEA and CrossTalkZ assume that links in a functional association network are normally distributed $\mathcal{N}(\mu'_{AB}, \sigma'^2_{AB})$, with the random model having the mean number of links μ'_{AB} and standard deviation σ'_{AB}. A Z-score can then be calculated for the observed number of links n_{AB} between gene sets A and B (4.2).

$$Z_{AB} = \frac{n_{AB} - \mu'_{AB}}{\sigma'_{AB}} \tag{4.2}$$

The Z-score is later transformed into a p-value, which is corrected for multiple hypothesis testing using the B-H procedure.

Binomially distributed model

Unfortunately, the normality assumption does not always hold, in particular for pathways that are too densely or too sparsely connected. Since all network connections are binary, the resulting cross-talk between two gene sets A and B,

P_{AB}, is binomially distributed, i.e. $P_{AB} \sim \mathcal{B}(n_{AB}, p_{AB})$, where n_{AB} is the maximum number of connections and p_{AB} is the probability of observing k number of connections, between gene sets A, and B. The exact values of parameters n_{AB} and p_{AB} are unknown, but they can be estimated from a random model of the network, obtained using the LP+S shuffling method. The expected number of links $E(k')_{AB}$ between A and B can be estimated from the random model generated using N randomisation runs as in (4.3), where $k'(i)_{AB}$ is the number of links between A and B observed in the i-th randomisation run.

$$E(k')_{AB} = \frac{1}{N}\sum_{i}^{N} k'(i)_{AB} \tag{4.3}$$

4.2 Prioritisation of gene products

The ability of prioritisation methods to focus further experimental efforts on a set of the most important genes for a particular phenotype is imperative in order to capitalise on the findings of large-scale experimental studies. Large volumes of data require rigorous correction for multiple hypothesis testing, which leads to a contradictory loss of statistical power. For example in GWAS, the weaker associations are discarded during the Bonferroni correction, favouring only the strongest associations (Lee et al., 2011). In such situations, prioritisation methods are used to expand on the genes already associated to a particular phenotype, recovering the missed associations, which are weaker, but potentially equally important. In clinical applications, prioritisation methods can serve as a support in assessments of clinical phenotypes and other patient outcomes.

The rest of this section will cover the most common types of prioritisation algorithms as well as ways in which to evaluate them, including the performance metrics that are used.

4.2.1 Methods and algorithms

All current prioritisation methods rely on prior knowledge in the form of a query set of genes known to be associated with a phenotype of interest. Different strategies and combinations of data types are then used to apply this prior knowledge to the data in order to select and rank a set of candidate genes by the strength of their association to the phenotype of interest. Approaches that do not rely on prior knowledge are still lacking (Ideker and Sharan, 2008).

Most of the current prioritisation algorithms use either text mining, functional annotation, network analysis, or combinations of these techniques (Moreau and Tranchevent, 2012).

Intrinsic gene-property-based methods

Disease genes and corresponding proteins seem to have intrinsic properties that separate them from their non-disease counterparts (Piro and Di Cunto, 2012). In addition of having a preference for certain protein domains, fewer closer paralogs and a broader phylogenetic range, disease genes seem to be longer, to produce longer proteins, to have longer intergenic distances and regulatory regions (such as 3'-UTR), ant to have a higher proportion of promoter CpG islands and lower mutation rates (Tiffin, 2011). These characteristics can be exploited in statistical analysis or by classifier algorithms such as random forests, e.g. as in PROSPECTR (Adie et al., 2005).

Functional annotation-based methods

The majority of tools based on functional annotation use GO, KEGG, etc. Annotation similarity-based; expression pattern-based; and gene list enrichment methods, such as ToppGene and ToppFun (Chen et al., 2009), are among the most prominent in this tool class. Gene list enrichment methods utilise techniques described in section 4.1.1, while expression-based methods compare patterns of co-expression between the query set and new candidates, ranking the most similar candidates higher in the output. Annotation similarity refers to, e.g. when a set of GO terms annotating a candidate gene is compared with the GO terms annotating the query set. Similarity is determined based on e.g. information content or distance (Lin, 1998; Resnik, 1995) in the ontology between the term sets, or by term overlap using the Jaccard index. In the end, the candidates are ranked based on their GO term similarity measure.

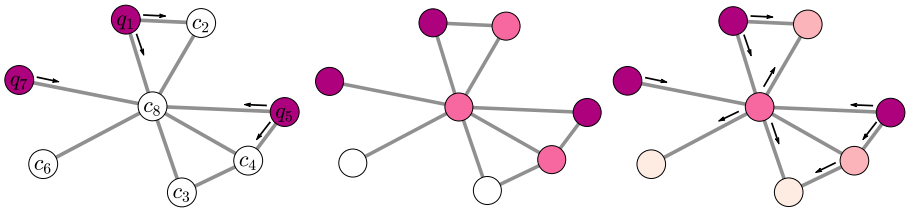
Text-mining-based methods

Tools based on the data mining of scientific literature, e.g. Beegle (ElShal et al., 2016), use natural language processing, often through a controlled vocabulary e.g. MeSH and UMLS (Piro and Di Cunto, 2012), to extract relational information from biomedical literature databases, such as Medline. Later, term co-occurrence statistics (Bromberg, 2013) or heuristic ranking (Masoudi-Nejad et al., 2012) are used for prioritisation.

Local network-information-based methods

Data integration into network representations using machine learning techniques and statistical methods (Doncheva et al., 2012), e.g. FunCoup, STRING, etc., has been described in section 3.1. Since such combinations of orthogonal data yield better predictions, many network-based prioritisation algorithms

have emerged. Such methods can be divided into those based on *local* and *global* network information (Fig. 4.1).



(a) Query **(b)** Local methods **(c)** Global methods

Figure 4.1: Local and global network-based prioritisation methods - (a) Query genes q_1, q_5 and q_7 in purple are applied to the network. (b) In local methods, e.g. direct neighbours of the query genes are selected using some criteria. (c) Global methods diffuse information along the links between all genes directly or indirectly connected to the query. Image inspired by Lee et al. (2011).

Local network methods use the direct neighbourhood of query genes (Fig. 4.1b) through node properties such as node degree and the shortest path to make predictions about the node's involvement in a phenotype. A relatively straight forward approach is to count the number of links a candidate gene v has to the query set Q or to take into account the sum of the weights of all direct links (Linghu et al., 2009; Shim et al., 2015), or the shortest path from a candidate gene to the genes in the query set (Doncheva et al., 2012; Lage et al., 2007; Oti and Brunner, 2007). A more sophisticated approach employs a statistical test to verify that a candidate's connections to the query set d_Q are not simply due to chance or to the large degree d of a particular candidate gene. This approach is implemented in MaxLink (Guala et al., 2014) (see section 5.2), which uses a hypergeometric test to assess the enrichment of connections to the query set $Q = \{q_1, q_2, \ldots, q_n\}$ compared to all N network genes, according to (4.4) and sets a threshold on candidates selected for output, ranked by their d_Q.

$$p(v) = \frac{\binom{n}{d_Q}\binom{N-n}{d-d_Q}}{\binom{N}{d}} \tag{4.4}$$

Global network-information-based methods

Global network methods take into account properties of the whole network instead of just the immediate neighbourhood of a node. Centrality-based methods (Browne et al., 2015; Cickovski et al., 2017) use social-network-like analysis, where node-centric network characteristics such as closeness centrality,

betweenness centrality, etc. are calculated (see section 2.4) in order to identify nodes with greater "importance" to the set of disease genes. One way in which this can be done is by ranking candidate nodes using one or a combination of node-centric characteristics, potentially adding functional-based information from e.g. GO, as in Browne et al. (2015). A more elaborate approach, used in ATria (Cickovski et al., 2017), involves an iterative procedure, where the most 'important' node is identified first. It is then removed together with all its links, and centrality measures are recalculated for all remaining nodes. The removal procedure is repeated until all nodes are ranked.

Another approach is to use the 'flow of information' in the network (Fig. 4.1c). An example of this is the Random Walk with Restart (RWR) method (Köhler et al., 2008), which simulates a random walk along network edges. It starts with information distributed in the query nodes, represented by an initial vector of probabilities $\mathbf{p}^0 = p_1, p_2, \ldots, p_n$, where $p_i = 1/N$ for all nodes $v_i \in Q$ and $p_i = 0$ otherwise. At each time step, information is transferred from one node to another along the edges connecting the nodes. The amount of information flow over an edge is proportional to the edge weight w_{ij} and there is a probability of restarting the walk $0 \leq \beta \leq 1$. All the weights are collected in a weight matrix $\mathbf{W}$ where each entry w'_{ij} is the degree-normalised weight of the original edge, i.e. $w'_{ij} = w_{ij}/d_j$. At each time point, the current node probabilities are contained in $\mathbf{p}^t$ and an iterative procedure (4.5) is applied to compute the final probabilities $\mathbf{p}^{t+1}$. The iteration is usually stopped when $\mathbf{p}^{t+1} - \mathbf{p}^t \leq 10^{-6}$.

$$\mathbf{p}^{t+1} = (1-\beta)\mathbf{W}\mathbf{p}^t + \beta\mathbf{p}^0 \tag{4.5}$$

The final probabilities are used to rank the nodes in the output from the tool. Variations of the RWR algorithm include the popular tool "Prioritisation and complex elucidation" (PRINCE) (Vanunu et al., 2010), where entries in $\mathbf{W}$ are normalised by the sum of all weights in the row, $W'_i = \sum_i w'_{ij}$, see (4.6).

$$w''_{ij} = \frac{w'_{ij}}{\sqrt{W'_i W'_j}} \tag{4.6}$$

Network propagation techniques or diffusion methods such as hyper-induced topic search (HITS) (Kleinberg, 1999) and PageRank-based (Page et al., 1999) methods like ToppNet (Chen et al., 2009) are similar to RWR with the information flow being normalised in slightly different ways; e.g. both by outgoing and incoming edges or by assigning similar scores to adjacent nodes. These methods have also been successfully adapted for prioritisation (Masoudi-Nejad et al., 2012; Piro and Di Cunto, 2012).

More recently developed methods utilise RWR indirectly, e.g. as a measure of topological similarity (Cho et al., 2016). The VAVIEN (Erten et al., 2011) tool computes RWR for each node with respect to all other nodes in the network. The node-specific vector of final probabilities is then used as the node's topological profile, describing the RWR-based proximity of the node to all other nodes in the network. The similarity of each candidate node's profile with the averaged profile of all query nodes is then calculated using Pearson correlation in order to produce a score for each candidate node, which is used for the ranking of the output.

Combined approaches

Nothing stops the network-based prioritisation approaches from being used in conjunction with other previously mentioned types of prioritisation methods. A straightforward approach is to combine centrality measures with similarities derived from functional annotations in e.g. GO or KEGG into a combined score used for ranking candidates based on their combined 'importance' for the query set (Browne et al., 2015).

Another approach is to use phenotypic information from e.g. OMIM to construct a multi-layered network, where diseases are connected in one layer and genes in another. Connections between the two layers would go through the known disease-gene associations. Prioritisation can then be performed using one of the network-based methods, e.g. RWR (Li and Patra, 2010). This approach has been extended to work on multi-layered networks with an arbitrary number of layers, where each layer is a network constructed from a specific data type, e.g. diseases, PPIs, and complexes (Yang et al., 2011) or DDIs as in ProphNet (Martínez et al., 2014).

A different take on prioritisation is to use patient data, e.g. gene expression, DNA and RNA abberations, etc., to construct patient- or phenotype-centric networks instead of functional association networks made up of genes. An example of this approach is similarity network fusion (SNF) (Wang et al., 2014).

There are also methods that use data sources directly without first constructing networks. In Endeavour (Tranchevent et al., 2016), the query set is used to train a model for each selected dataset. The training is conducted using a simple statistical technique, e.g. GO term enrichment, in order to extract predictive features for that data source. Candidate genes are then scored using the data-source-specific model and their rankings are combined into a global ranking using order statistics (Aerts et al., 2006).

4.2.2 Performance metrics

Evaluating performance of prioritisation methods usually involves identifying the instances in a ranked list as either positive or negative. Hence, all the metrics discussed in this section are described from the point of view of a binary classification problem. However, many of the described metrics can be adapted to work on multi-class classification problems.

Confusion matrix

Genes in the resulting ranked lists are labelled as TP, FP, TN, or FN and used to calculate different performance metrics.

Table 4.1: Confusion matrix - Performance of a classification model is summarized in a table with rows depicting predicted class membership and columns demonstrating the true class.

		True class	
		Positive	Negative
Predicted class	Positive	TP	FP *(type I error)*
	Negative	FN *(type II error)*	TN

Row- and column-wise metrics can be calculated based on the confusion matrix (Tab. 4.1) e.g. the row-wise $precision = TP/(TP+FP)$. Column-wise metrics are exemplified by the fraction of accurately recovered true labels or the *true positive rate* (TPR) also known as the *sensitivity* or *recall*, with $recall = TP/(TP+FN)$, and the fraction of false predictions among all negatively labelled cases, the $specificity = TN/(TN+FP)$. Since these measures only use part of the information provided by the labels - e.g. *recall* and *precision* only focus on the positive examples and ignore the negative cases - their utility is limited. Combinations of such incomplete metrics, e.g. the *F-measure*, which is the harmonic average of *precision* and *recall* (4.7), have the same limitation.

$$F-measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{4.7}$$

Additionally, the row-wise metrics, e.g. *precision*, are prone to severe bias when one class dominates the data. Class imbalance does not affect the column-wise metrics, which for ranked outputs, e.g. *recall*, can still be valu-

able when looking at a certain percentage, e.g. 1%, 10%, of the top-ranked genes, instead of looking at the whole list (Börnigen et al., 2012).

Accuracy

Accuracy (ACC) is a performance metric that takes all four labels into account, (4.8), making it more comprehensive and informative. It represents the fraction of all correct predictions, both positive and negative, among all predictions. Accuracy is easy to understand and is usually one of the first performance metrics to be applied. However, it is sensitive to class imbalance, always favouring the overrepresented class, which is why it can be misleading and should be used with caution.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.8}$$

Averaging the class-specific accuracies into a "balanced accuracy" (BCC) metric (Brodersen et al., 2010) solves the class imbalance problem (4.9).

$$BCC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{4.9}$$

Matthews correlation coefficient

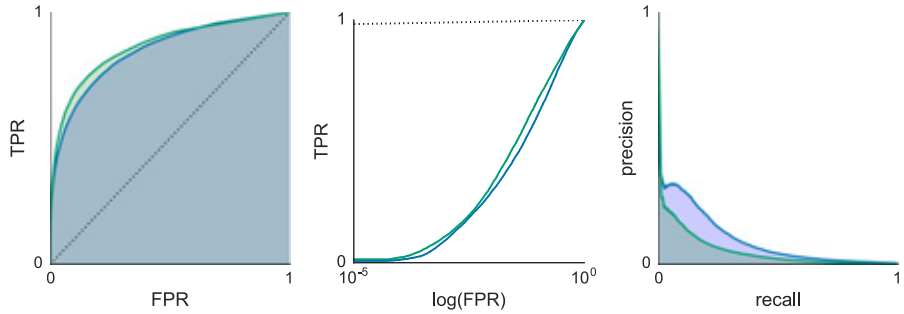
Another performance metric that takes into account all four labels but is well balanced, except in the extreme cases when only one class is present, is the Matthews correlation coefficient (MCC) (4.10).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.10}$$

The MCC measures correlation between the observed and predicted labels, $-1 \leq MCC \leq 1$, where values close to 0 represent random predictions, while values closer to the extremes, 1 and -1, represent either strong correlation or anti-correlation, respectively. MCC is readily used for classifiers where the output contains numeric values representing the strength of an association to a particular label. However, due to outputs from prioritisation methods being ranked lists based on normalised values, MCC may not be the most suitable performance measure for this task.

Receiver operating characteristics

Other combinations of metrics are also possible and encouraged. Receiver operating characteristics (ROC) (Fawcett, 2006) is a combined look at performance, describing the change in *recall* and the FPR, where $FPR = 1 - specificity$, at various thresholds. It is visualised in the form of a ROC curve (Fig. 4.2a). A visual inspection of performance on a ROC curve can yield a direct appreciation of the difference in performance between two methods. A steeper rise of the curve and a higher reach signify better performance. ROC curves illustrate the ranking potential of a classifier, making this metric very useful for assessing ranked outputs of prioritisation methods.



(a) ROC **(b)** Semi-log ROC **(c)** Precision-recall

Figure 4.2: ROC and PR-curves - The solid lines in all the figures represent performance of two classifiers visualised in the (a) ROC, (b) semi-log ROC, and (c) precision-recall curves. The dotted lines in (a) and (b) represent the performance of a random classifier. The shaded regions in (a) and (c) represent the area under the curve, AUC_{ROC} and AUC_{PR}, respectively. Better performance is generally depicted by the largest AUC. The semi-log ROC curve can be used when emphasising performance in the top of the output.

Alternative ROC curves e.g. the partial ROC (pROC) representing a part of the ranked output, usually focus on the first part of the curve. Another option for the top of the output is to replace the linear ROC curve with a semi-logarithmic one (semi-log ROC) (Fig. 4.2b) (Clark and Webster-Clark, 2008).

Precision recall curve

Another visualisation combining two metrics is the precision recall (PR) curve, which illustrates the change in *precision* (PPV) when *recall* (TPR) is changed (Fig. 4.2c). The PR curve is less sensitive to class imbalance than the ROC (Saito and Rehmsmeier, 2015), hence its popularity.

Area under the curve

The area under the ROC curve (AUC), AUC_{ROC}, is a convenient way to summarise the information presented in a ROC curve in one number. It is calculated by integrating over a thresholding parameter, T (4.11), and can be interpreted as the probability of ranking a randomly chosen positive example, X_P, higher than a randomly chosen negative, X_N (Bradley et al., 1997), where a random classification result has $AUC_{ROC} = 0.5$ while perfect performance results in $AUC_{ROC} = 1$.

$$AUC_{ROC} = \int_{-\infty}^{\infty} TPR(T)FPR'(T)dT = p(X_P > X_N) \tag{4.11}$$

Area under the PR curve is also a convenient summary metric for prioritisation and can be calculated similarly (4.12) (Keilwagen et al., 2014).

$$AUC_{PR} = \int_{0}^{1} PPV(TPR)dTPR \tag{4.12}$$

Other ranking-based metrics

If the resulting gene list is ranked, then the average rank of TPs can be studied. Since the results of non-random prioritisation methods enrich the top of the list for TPs, the output may be skewed, and median rank TPs, $\tilde{R}(TP)$, normalised by the length of the output, n, is a more appropriate measure, called "median rank ratio", $MRR = \tilde{R}(TP)/n$, where $0 \leq MMR \leq 1$, means better performance for values closer to zero and poorer for those closer to one.

A measure developed for ranking results from recommender systems can be applied to the assessment of performance of prioritisation methods. This measure is based on the discounted cumulative gain (DCG) (4.13), which represents the relevance of the retrieved instances at position i, rel_i, automatically penalising TPs retrieved late in the list.

$$DCG_P = rel_1 + \sum_{i=2}^{P} \frac{rel_i}{log_2(i+1)} \tag{4.13}$$

When DCG is normalised by an ideal version of DCG, $iDCG_P$, where all positive instances are returned first, it is referred to as the "normalised DCG" (nDCG), $nDCG = DCG_P/iDCG_P$.

Prioritisation methods bear an additional constraint when it comes to performance measures. In order for their predictions to be useful, they need to be able to retrieve as many of the true positives at the top of the candidate list as possible, because an experimental validation of the whole output or its further

investigations are constrained to the top-ranked candidates. Therefore, performance measures, explicitly looking at the ranking of the held out genes at the top of the output list, e.g. MRR, pROC, pAUC, and nDCG, have more practical importance than the measures considering performance across the whole output.

Other metrics, such as MI, Hamming distance, Euclidean distance, Bayesian information criteria, relative entropy, and other correlation-based metrics, are not readily used in the assessment of prioritisation tools. Since they are already well covered in Baldi et al. (2000) they fall outside of the scope of this thesis.

4.2.3 Benchmarking

Many prioritisation methods are available, making the selection of one suitable for the prioritisation problem at hand a very difficult problem. Robust and unbiased benchmarking is therefore needed to assess method performance. Unfortunately, it is difficult to find a universal benchmarking strategy. When published, most of the tools only come equipped with some kind of validation, which is usually just a proof of concept or a case study rather than a real estimation of performance in the context of its tools (Tranchevent et al., 2011; Zhu et al., 2014). Many comparisons are highly heterogeneous in terms of setup, performance measures, data sources, and test/training sets. In an ideal situation, predictions from a prioritisation tool would be verified by experimental techniques, assessing the true prediction classes as TP, FP, TN, or FN. In reality, experimental validations are difficult, and retrospective data from OMIM or GWAS studies are often used.

Statistical methods

Different flavours of cross-validation (Kohavi, 1995) are the most common ways to evaluate classification and prediction algorithms. More precisely, cross-validation estimates the future error rate, e.g. how well a prediction algorithm generalises to new data. This technique involves holding out some portion of data for testing while the rest is used for training. An example for prioritisation methods involves *3-fold* cross-validation, where a gene set associated with a phenotype is randomly divided into three parts of roughly equal size. Two parts are combined and used as the query input to a prioritisation tool, which tries to recover the third part among all the nodes in the network (Guala and Sonnhammer, 2017). The process is usually repeated, varying the part that is left out, until all parts of the data have been tested. The data can be divided into as many parts as necessary, denoted as *X-fold* or *leave-1/X-out*.

A less stringent approach is to let a prioritisation method recover the testing set not from all the remaining genes in the network, but from the testing set

mixed with some number of network genes selected at random, i.e. *leave-one-out* cross-validation, where the testing set of size 1 is mixed in with 99 randomly selected genes from the network (Börnigen et al., 2012).

GO term enrichment

Another class of performance evaluation methods, where good performance is necessary for good prioritisation but insufficient to guarantee good real-world results, is GO term enrichment. It is most often used as a "sanity check" for the method, examining whether it is able to produce results correlated in function to the seed set. Good performance, here, is when candidate genes returned by the studied method are highly enriched for the same GO term annotations as the query list. The same argument made with respect to cross-validation regarding the use of already available knowledge can be applied here. Hence, there is a lack of guarantees for real-world performance, even when enrichment results are good.

Time-stamped benchmarking

There have been attempts to simulate real world situations and use prospective instead of retrospective knowledge in benchmarking prioritisation tools known as "time-stamp" benchmarks. In Börnigen et al. (2012), literature mining was used to identify recently discovered disease genes. Then disease genes from OMIM, not yet updated with the new genes for the corresponding diseases, were used as seed lists to several prioritisation tools. Later, some statistical performance measures could be calculated and compared between the tools.

Challenges in benchmarking

In order to assess the true performance of the available techniques and potentially select the best technique for a particular problem, it is imperative to conduct good and objective comparisons. One of the obvious practical challenges when comparing computational tools in a field so clearly diverse is the large variation of the ways in which different tools are meant to be used. Many tools are not available for download or at all, making them difficult or impossible to test in large or repetitive benchmarking assessments.

Because of the heterogeneity of used data sources by different prioritisation tools it is difficult to avoid knowledge cross-contamination, e.g. selecting benchmark data that has not been used in the design of the tool. The overuse of disease databases, e.g. OMIM, risks overestimating the true predictive power of a prioritisation tool. A solution using time-stamped benchmarks suffers from lack of coverage because of new data's rapid incorporation into

the databases used by the prioritisation tools. This leads to a reduced ability of time-stamped benchmarks to make statistically sound distinctions between tools.

A new benchmarking approach addressing challenges of insufficient power and knowledge cross-contamination is part of this thesis (see section 5.4).

There are other limitations inherited from the use of functional association networks where topological properties favour some methods over others. An example of this is that tightly connected networks favour RWR methods, while in sparser networks, other methods perform better (Hsu et al., 2011). This could be the reason for lack of dominance of one particular approach across different data sources and performance measures (Oti et al., 2011).

4.2.4 Local vs global

The issue of which class of methods is the best still remains to be settled. There is not even consensus when it comes to which of the network-based methods outperforms the other. There are claims that genes cannot be fully represented by their shortest path, suggesting that global methods could be more discriminative than local methods, because the former consider the whole network structure (Köhler et al., 2008; Wang et al., 2012). The closeness centrality-based method InterConnectedNess (Hsu et al., 2011) and network-propagation-based techniques (Lee et al., 2011) have been shown to outperform local network-based methods. RWR has been shown to outperform local network-based methods (Köhler et al., 2008; Navlakha and Kingsford, 2011) as well as centrality-based ones (Wu et al., 2008) and network-propagation-based techniques (Navlakha and Kingsford, 2011). Somewhat contradictorily, in another study, a network-propagation-based method has been shown to outperform RWR (Vanunu et al., 2010).

However, superior performance on the whole output does not guarantee better performance on the more biologically relevant, top-ranked predictions (Shim et al., 2015). This has been substantiated by the superior performance of a local network-based method "naïve Bayes" compared to a diffusion-based technique (Shim et al., 2015). Recently, it was also confirmed when another local network-based method, MaxLink, outperformed both RWR and network-propagation-based techniques in an unbiased, large-scale benchmark (Guala and Sonnhammer, 2017).

RWR on a heterogeneous network with combined genotype and phenotype seems to outperform the simple RWR (Li and Patra, 2010), suggesting that phenotypic information can further complement the interactional landscape around the genes being studied. The performance of a method is strongly linked to the type of data and network setup, thus influencing the outcomes of

comparisons. However, when local and global methods are compared, they all identify genes not found by the other methods, suggesting that a combination of techniques may yield the most comprehensive results (Navlakha and Kingsford, 2011).

4.2.5 Challenges

Although prioritisation techniques have come a long way during the last decade, the field is still faced with significant challenges. It is a daunting task to provide good-quality data for prioritisation methods. Gathering, normalising, and integrating heterogeneous data from multiple data sources while keeping it up to date is crucial for performance of all the prioritisation methods mentioned. There are techniques for overcoming sparsity and redundancy in the data, but they usually come at the cost of performance or quality of the output.

Another obvious challenge is the unequal availability of annotated information for the whole genome. This bias towards the more studied and well-characterised genes is present in almost all data sources (Oti et al., 2011; Piro and Di Cunto, 2012). Its effect on the output is similar to that of the self-inflicted limiting of the search space by prioritisation methods that only work on a certain locus or genomic region instead of using the whole genome for prioritisation. There are benefits to narrowing the search in terms of computational cost, but this approach unfairly disregards both the potential target genes not yet allocated to that gene locus and the potential target genes outside the constriction.

Finally, there have been studies suggesting that most gene function prediction and prioritisation that relies on GBA is at the very least heavily influenced, if not almost entirely caused, by the multi-functionality of genes (Gillis and Pavlidis, 2012). This came about after seeing how prioritisation based solely on node degrees, i.e. ranking candidates based on their degree, could yield comparable performances to sophisticated prioritisation methods such as SVMs and GeneMANIA. Additionally, the distribution of performance exhibited by the node-degree-based method and prioritisation algorithms had a similar shape and variance. It may well be that nodes with high degree are biologically relevant for many phenotypes, but caution is advised when the output from a prioritisation method is too highly correlated to node degrees, since the contribution of the multifunctional nodes may be unspecific at best and spurious at worst. Attempts to correct for multi-functionality are thus necessary. This can be done by testing each candidate's enrichment for the query set, as in MaxLink (see section 4.2.1), thereby avoiding high prioritisation purely based on degree. For the diffusion-based methods, an approach where the observed diffusion, i.e. the final probabilities $\mathbf{p}^{t+1}$, is compared to diffusion scores when

random starting gene sets are utilised as queries (Mazza et al., 2016) can be used to generate *p-values* which correct for the hub genes' multi-functionality.

5. Present investigations

5.1 FunCoup 2.0 (Paper I)

FunCoup uses naïve Bayesian integration of different evidence types (PPI, DDI, mRNA and protein co-expression, phylogenetic profile similarity, subcellular co-localisation, TF and miRNA co-regulation, and genetic interaction profile similarity) and transfer of orthology from several model organisms to predict functional association between genes and proteins.

The algorithm employs a unique scoring function for each data source e.g Pearson correlation for mRNA co-expression, PPI scoring for PPI, etc. The raw metric scores are then mapped to LLRs for functional association for each dataset, species, and type. Orthology information is transferred using InParanoid. The summed LLRs from individual data sets are converted into a confidence value for each link in the network.

Purpose - The purpose of this paper was to update the underlying framework improving quality and coverage of FunCoup. Additionally, we wanted to demonstrate how a user may leverage the built-in network analysis capabilities when it comes to comparative interactomics and other applications.

Methodology - Three new model organisms, *C. familiaris* (dog), *G. gallus* (chicken), and *D. rerio* (zebra fish), were included. Most of the data sources were updated in order to enhance the quality of predictions. New, more comprehensive data sources were added; notably, IntAct (Björkholm and Sonnhammer, 2009) as the source of PPI information and UniDomInt (Aranda et al., 2010) as the source of DDIs, since both incorporate all reliable information previously collected in other sources.

In order to increase network coverage, the core framework was extended with a new evidence type "genetic interaction", based on the correlation of the genetically interacting profiles of two genes. An improved scoring function for PPI evidence was also introduced to increase the quality of the PPI scores.

The comparative interactomics feature introduced in the previous version received a complete overhaul, involving a much stricter and more reliable way to use orthology in visualising conserved clusters and functional modules.

Findings - The resulting species networks continued to exhibit scale-free and small-world properties observed in the previous version, despite a dramatic (2-10 fold) increase in the number of inferred links. High-confidence links i.e. $p \geq 0.99$ from the previous version were conserved up to 90% in this release

Novelty - The markedly increased number of links due to the volume of newly incorporated high-throughput data made FunCoup 2.0 one of the most comprehensive networks of functionally associated genes. The comparative interactomics feature enabled its users to receive more utility from this resource.

Further research - New updates will be required in order to incorporate newly available data. Addition of relevant species should increase transfer of orthology data as well as allow more researchers to work with model organisms of their preference. A mechanism for balancing similar evidence from different data sources in order to reduce the bias from highly studied genes would further improve quality.

5.2 Network-based prediction of disease genes (Paper II)

MaxLink - a prioritisation tool based on the GBA paradigm - utilises a functional association network in order to identify new disease genes. Due to the increasing amount of genomic data, there is a constant need for improvement of *in silico* tools such as MaxLink in order to extract useful information from the generated data and direct future experimental efforts.

Purpose - The primary goal of this paper was to generalise the underlying algorithm and strengthen its statistical basis. Additionally we wanted to improve MaxLink's usability by introducing a web resource.

Methodology and Findings - MaxLink uses FunCoup to extract all the direct neighbours of the query genes. After this step, a series of filters are applied in order to retain the most relevant candidates. If not controlled properly, hubs have a much higher probability of being selected as candidates, purely by chance, because of their high number of connections. In order to address this, a "connectivity filter" was applied to retain candidates with proportionally more connections to the query genes compared to all the other genes. In the new version, a hypergeometric test is used instead of the proportion-based approach. Previously, there was an "annotation filter" removing genes with annotations to cancer, in order to retain potentially novel cancer genes. The new version saw this filter removed to allow the algorithm to operate on any gene set. The

new version was validated using GO term enrichment and cross-validation on orphan disease genes from OrphaNet (Maiella et al., 2013).

To improve usability of MaxLink, the original Perl program was re-implemented in C++, increasing the speed of execution. Additionally, the tool was made accessible through an easy-to-use web resource `http://MaxLink.sbc.su.se`, in addition to its standalone application.

Novelty - The main contributions of this research were the novel, validated, statistically robust connectivity filter and the increased usability delivered by a user-friendly web resource.

Further research - Future developments for this tool would lie in the further refinement of the underlying algorithm. Another avenue might be to add a global network algorithm, e.g. a type of network diffusion, in order to leverage the strengths while minimising the weaknesses of both algorithms in an ensemble approach.

5.3 Analysing cross-talk in biological networks (Paper III)

The most popular pathway annotation methods use gene overlap between a gene set of interest and pathways with known function. This approach assumes equal importance and independence of all genes in a pathway and requires the pathways to be completely mapped. Both of these assumptions are flawed. Another way to approach the problem is to use the topology of a functional association network and consider the cross-talk between the gene set of interest and the known pathways.

Purpose - The purpose was to introduce and evaluate a novel way to analyse cross-talk between gene sets, called BinoX, for pathway annotation.

Methodology - BinoX assesses the cross-talk between two gene sets, e.g. a set of differentially expressed genes and a KEGG pathway, by comparing the cross-talk between the studied gene sets with the cross-talk measured in a random model of the underlying network.

Two benchmarks were constructed to assess the performance of BinoX in comparison with other pathway annotation methods: GEA, NEA, and CrossTalkZ. Sensitivity was assessed using KEGG pathways, each divided in half, with a typical amount of gene overlap. The FPR was measured on the same KEGG pathways but randomised, preserving degree distribution. Additional benchmarks were constructed from 25% of the smallest KEGG pathways in order to simulate a more challenging problem.

The four tested tools were also compared in real-world pathway annotation scenarios involving finding enriched and depleted KEGG pathways for gene sets from MSigDB (Subramanian et al., 2005), catalogue of somatic mutations in cancer (COSMIC) (Futreal et al., 2004), and OMIM.

Findings - Network-based pathway annotation methods (NEA, CrossTalkZ, and BinoX) clearly outperformed GEA in all of the benchmarks. However, there were significant differences noted even among the network-based methods, where BinoX was shown to be clearly superior to others in terms of overall performance and computational speed. The differences in performance were magnified when the more challenging, smaller KEGG pathways were used.

As for the real-world tasks of identifying KEGG pathway enrichment and depletion, BinoX demonstrated its ability to uniquely identify annotations supported in the literature.

Novelty - BinoX applies a sounder statistical approach to the analysis of crosstalk in a functional association network, resulting in superior performance.

Further research - Despite that BinoX's network randomization procedure is able to preserve second-order network topology it could be further improved using a less heuristic approach. Another future task is to combine BinoX with MaxLink, to leverage the predictive power of both.

In its current form, BinoX uses a threshold cut-off for links in the underlying network, instead of incorporating link weights into its framework. This results in a loss of information and should be addressed in future updates.

5.4 Evaluating performance of prioritisation methods (Paper IV)

Most new prioritisation tools are usually introduced together with some form of proof of concept or a validation, but there is no good way to assess their performance or even compare them in a systematic manner. Prospective ways of comparing such tools have been underpowered to make statistically sound comparisons, and there is no good strategy for retrospective approaches.

Purpose - The purpose of this study was to construct a generalised benchmark for prioritisation tools that is both robust and sufficiently sensitive to provide meaningful distinctions of performance supported by statistical analysis.

Methodology - Four prioritisation tools, MaxLink, NetRank, NetWalk, and

NetProp, were tested using 3-fold cross validation on GO terms. Since prioritisation tools work best when provided with a cohesive set of genes with some underlying association pattern, only terms with 10 to 300 genes associated to them were used. This avoided too small gene sets, where association patterns may be incomplete, and large sets where any association pattern would be obscured by the noise. All tested tools used FunCoup as the source of functional associations. Performance of each tool was assessed using a set of suitable metrics i.e. pROC, pAUC, MedRR, and NDCG, and the results compared.

Findings - We found that MaxLink demonstrated significantly better performance than NetRank for most of the tests, as measured by NDCG and MedRR. NetRank outperformed NetWalk and NetProp, while NetWalk had a statistically significant advantage over the very similar NetProp for these measures. With respect to pROC and pAUC, MaxLink outperformed the other tools, which ranked in the same way as for the other performance measures, i.e. NetRank, NetWalk, and NetProp.

Novelty - Combination of a subset of GO terms with the used performance measures provides a robust and objective benchmark for prioritisation tools. This is the largest benchmark to date, with the ability to distinguish performance differences between even very similar tools, e.g. NetWalk and NetProp. Previously, network diffusion methods such as NetRank, NetWalk, and NetProp were shown to outperform methods operating on the direct network neighbourhood, such as MaxLink. However, looking at the most relevant part of the returned output, i.e. the top 1-10% of the genes, we were able to show that MaxLink outperforms the network diffusion algorithms.

Further research - The benchmark architecture is under constant development, where there is an emphasis on trying out new bases for functional association to see if the results can be reproduced. GO terms may be substituted by KEGG pathways as the testing set. To obtain a more complete picture of performance of the available prioritisation tools, even non-network-based tools should be assessed, as long as they do not already use data from GO or FunCoup for their predictions. The benchmark could also be extended to work on heterogeneous networks to allow for evaluation of the newest algorithms.

5.5 FunCoup - the fourth generation (Paper V)

In version 3.0 (Schmitt et al., 2014) FunCoup was reimplemented in Java and included data updates, a new interactive web portal, and a new procedure for redundancy reduction.

Despite the number of competing networks, it is imperative to continue the development of FunCoup due to its unique combination of characteristics. FunCoup uses a broader definition of functional association, allowing it to infer a more comprehensive set of ways in which proteins are able to influence complex processes in a cell. This is achieved without resorting to text mining, which is prone to noise and biases. Species included in FunCoup are carefully selected based on quality, coverage, and potential for orthology transfer. Finally, FunCoup has an intuitive web interface with a contemporary and interactive way to exploit and visualise produced networks.

Purpose - The purpose of this version was to make a complete overhaul of the underlying data and to extend the database in terms of additional species, data sources, evidence types, and gold standard.

Methodology - Six, carefully selected new species, including four eucaryotic - *Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Bos taurus*, *Oryza sativa* - and two prokaryotic - *Escherichia coli* and *Bacillus subtilis* - were added in this release. All existing data sources were updated if possible. Additionally, the explosion of microarray data since the previous release was put to use for all existing and newly added species. A dramatic increase in quantitative mass spectrometry (QMS) data allowed for addition of a new type of evidence. When it comes to the gold standard, essential for the quality of the predicted links, we were able to purify the "protein complex" class and add a new class "shared operon". It relies on similar function in genes that belong to the same transcriptional unit or operon. Finally, we replaced the obsolete Java-based viewer with a new, interactive viewer based on d3.

Findings - The refinement and update of the gold standard, together with the complete overhaul of the supporting data and orthology transfer from the newly added species, has resulted in an increase in both the sizes and the quality of the species' networks. The newly added evidence source, QMS, has seen a successful integration into the framework, judging by its contribution of evidence to the produced networks. The updated, interactive network viewer is an essential resource for the community of FunCoup users.

Novelty - The new gold standard class, shared operon, in conjunction with the orthology transferred from the newly included, relevant prokaryotic species and the novel source of evidence, QMS, have already contributed to the increase in both the quality and sizes of the FunCoup networks.

Further research - The mapping of different types of identifiers, e.g. En-

sembl, UniProt, HGNC, etc. has been and remains a challenge, resulting in loss of genome coverage for all the species. Improved identifier mapping will need to be addressed in future releases. Automatic data updates could potentially be resolved using a web crawler. Programmatic access via an API is another potential feature that could be implemented in the framework to facilitate batch uses of the networks. If there is sufficient data for it, the links in the FunCoup networks could be made directional, producing more dynamic networks and capturing important complex processes in the cell.

Relatively small changes to the current framework would be required in order to introduce a more dynamic view of the networks. For tissue specificity, this can be achieved by filtering out the inferred interactions in tissues where none or only one of the proteins are expressed, as evident by e.g. HPA. Another approach can be to integrate data by its tissue specificity into tissue-specific networks. These networks can be used as they are or combined into an aggregated network.

5.6 Experimental validation of predicted cancer genes (Paper VI)

MaxLink performance has been assessed but it has not had a proper experimental validation. Previously it has been validated *in silico* using data from GO and OrphaNet (Guala et al., 2014) and a set of differentially expressed cancer genes (Östlund et al., 2010). However, these validations need to be complemented by a direct experimental validation.

Purpose - The purpose was to use FRET to experimentally validate Maxlink's predictions based on a set of known cancer genes.

Methodology - The cancer gene set was compiled using COSMIC and UniProt and submitted to MaxLink. One of the cancer genes with the highest number of predicted candidates was chosen together with its predicted interacting partners for experimental validation by FRET. Fluorescent microscopy was used to detect FRET and its efficiency was calculated for all predicted PPIs.

Findings - Janus kinase 2 (JAK2), a non-receptor tyrosine kinase important in different leukaemias, had the highest number of predicted candidates and was selected for validation, together with 18 of its candidates. Experimental validation was successful for 15 interactions, which showed a mean efficacy of 12% of the maximum determined for this experiment. This was orders of magnitude above the simulated negative control values of 0.3% and indicated

a successful validation of the candidates. A follow-up scan of the literature for the validated, but previously unknown interaction partners showed potential co-implication with JAK2 in pathogenesis of breast cancer and Chronic Lymphocytic Leukemia and as potential drug targets.

Novelty - Since MaxLink uses FunCoup to make its predictions, this validation is not only the first direct experimental evidence of MaxLink's performance, but also of FunCoup's. Parts of the experimental setup such as the positive control construct were developed for this approach.

Further research - Future developments of the validation techniques could benefit from a direct negative control, although it is difficult to show evidence of lack of interaction. FRET could be applied to the validation of predictions from other prioritisation techniques that currently lack proper validation. A large-scale validation of FunCoup could be attempted in the future.

Sammanfattning

Den moderna gentekniken har gjort det möjligt att fastställa orsakerna bakom flera ärftliga sjukdomar, som beror på isolerade genförändringar. Dock har det varit svårare att hitta grunden till många av de vanliga folksjukdomarna, som exempelvis cancer, diabetes och Alzheimers då dessa orsakas av ett komplext samspel mellan förändringar i olika gener och proteiner. Kan modeller av det komplexa samspelet användas för att underlätta sökandet efter orsakerna bakom dessa sjukdomar och identifiera mål för potentiella behandlingar? Hur pålitliga är de verktyg vi använder för att utföra sökandet och kan deras prestanda jämföras på ett objektivt sätt?

I denna avhandling vidareutvecklades det bioinformatiska verktyget FunCoup, som avbildar samspelet mellan olika gener och proteiner, som ett ”social nätverk”. Till skillnad från ett nätverk av personer och vänskapsband, består FunCoup av olika gener och proteiner samt interaktioner mellan dem. FunCoup använder Bayesiansk statistik för att lägga samman experimentellt bestämda interaktioner från människa och 16 andra arter inklusive tarmbakterie, jästsvamp, ris, mus, etc. i ett nätverk. Avhandlingens första delarbete beskriver FunCoups första vidareutveckling, som resulterade i en ökning av kvalitén och täckningsgraden av interaktionerna, samt tillägget av en ny funktion för att jämföra olika arters delnätverk. Det femte delarbetet avhandlar en total översyn av underliggande data och en ökning av nätverket med sex nya arter. Vidare inkluderades en ny typ av interaktioner baserad på mängden samverkande proteiner inne i cellen.

FunCoup kan användas för att visualisera samspelet mellan gener av intresse tillsammans med andra gener i dess omgivning i en eller flera modellorganismer. Utöver detta fungerar FunCoup som bas för olika analysverktyg såsom BinoX och MaxLink, som nyttjar nätverket för sina beräkningar.

Utvecklandet av BinoX beskrivs i den tredje artikeln. BinoX använder strukturella egenskaper hos FunCoup-nätverket för att urskilja oväntat många kopplingar till samlingar av gener med kända funktioner. Detta gör att BinoX kan användas för att bestämma en tidigare okänd funktion hos en grupp gener, som t.ex. skiljer sjuka patienter från friska kontroller. En sådan funktion kan vidare avslöja den studerade sjukdomens genetiska orsaker och på sikt leda till nya behandlingar mot sjukdomen.

MaxLink beskrivs i den andra artikeln. Verktyget använder FunCoup-nätverket

för att upptäcka och prioritera individuella gener med statistiskt starkare koppling till en vald grupp sjukdomsgener från t.ex. cancer, Alzheimer och andra folksjukdomar. De nyupptäckta generna kan i sin tur bli mål för nya behandlingar. Algoritmen bakom MaxLink vidareutvecklades till att kunna användas på vilken sjukdom som helst och förstärktes statistiskt.

Det fjärde delarbetet resulterade i ett sätt att utvärdera prestandan hos MaxLink-liknande verktyg med avseende på deras förmåga att generalisera till nya data. Prioriteringarna gjorda med FunCoup som bas testades på funktionsbeskrivningar från Gene Ontology, som grupperar gener kring gemensamma funktioner i cellen. Resultatet var ett robust och objektivt sätt att utvärdera existerande och framtida prioriteringsverktyg. För att öka tilltron till MaxLinks resultat ytterligare, validerades det experimentellt med FRET. Detta beskrivs i det sjätte delarbetet.

Utvecklingarna som beskrivs i denna avhandling visar på hur nätverk som modellerar cellens komplexa biologi kan bidra med värdefull kunskap för forskare som studerar orsakerna bakom sjukdomar med komplicerad ärftlig bakgrund respektive letar efter nya sjukdomsgener, som kan användas som mål för nya behandlingar. Det nya robusta sättet att jämföra prestanda hos olika prioriteringsmetoder kommer att hjälpa forskare att välja rätt verktyg för sin problemdomän.

Acknowledgements

At last, you have reached (or skipped to) one of the most read part of any doctoral thesis, the chapter where everyone who has provided their support in this scientific endeavour is remembered and properly thanked.

First of all I would like to thank my supervisor *Erik Sonnhammer*, who saw my curiosity and potential for research, and gave me the opportunity to follow it through as a PhD student in his group. You provided me with the needed flexibility and support without which this endeavour might not have been possible. I would also like to thank my co-supervisor *Hjalmar Brismar*, who's creative approach got me out of various apparent dead-ends on the not-so-straight road to scientific discovery. Additionally, I would like to thank the staff at DBB for their administrative help throughout my PhD studies and all my research collaborators.

Next, my gratitude goes out to important people from my "other life", i.e. my day-to-day job in the field of clinical research and the pharma industry. Here I would first like to thank my manager at RPS, who failed to realize my full potential and left me unchallenged and bored, which triggered my initial PhD application. When I started at Merck AB, my manager *Viveka Åberg*, did the opposite and allowed me the creative freedom to pursue my PhD studies. Thank you, Viveka! I am also deeply grateful to my latest manager, *Ingela Hallberg*, who made sure that my flexibility was maintained, and supported me in my research endeavours in every way possible. I would also like to thank other *former and current colleagues at Merck* for their support throughout the years.

A few colleagues at Science for Life Laboratory also deserve my gratitude. Thank you *Walter* and *Mirco* for your companionship and *Axel* for your help in preparation for our Big exam. This journey would not have been possible without the excellent company of former and present colleagues in the Sonnhammer group. Thank you *Kristoffer* and *Dave* for the always-insightful discussions; *Gabriel* for kick-starting one of the central projects of my thesis, MaxLink, and for being a fellow "Northener" raising the team spirit with your presence. The North remembers! I am grateful to *Olli* and *Matt* for relaxing discussions about everything else other than work. *Thomas*, thank you for sharing your amazing adventure tips and stories, and for one of the most memorable photos from my wedding. Thank you, *Stefanie* and *Denize* for providing

perspective in our coffee break discussions. Thank you, *Daniel* for the sense of humour and positive attitude you bring to the group. *Andreas*, your vast ungoogleable knowledge of the universe of comics has always inspired me to be a more knowledgeable nerd. I hope to see you at the next Sci-fi convention. *Mateusz* you have always been our Linux and Java guru. Thank you and *Kasia* for your friendship and cheering up of our gatherings. *Christoph*, your friendship and collaboration mean a lot to me. Thank you for providing valuable scientific insights when it comes to work and for your diligent scrutiny of my thesis. It has been an absolute pleasure to share many of our common interests with you and *Lisi*.

The old Russian saying "Не имей сто рублей а имей сто друзей." (eng. Don't have a hundred rubel, have a hundred friends instead) was true before the 1990s inflation and is even more true now. I would therefore like to thank my friends who have supported me with their friendship and companionship throughout the years.

I probably would have made it through my undergraduate studies at UTH, but it would have been so much duller if I did not have *Kalle* and *Olle* constantly at my side. Together with Kalle and Olle; *Lina*, *Lotta*, *Anton* and *Karin* were also instrumental in making my student years in Uppsala one of the most exciting times of my life. Kalle and I were since then united in our shared nerdist interests only partly approved by his dear wife *Anna*.

A rich social life has been the key to overcoming all obstacles and maintaining my motivation throughout the years. For this I am grateful to many of my friends: *Ola*, *Pedram*, *Rutan*, *Maryam* and *Anders*, *Jop* and *Virginia*, *Linda* and *Erik*, *Imane*, *Keban*, *Igor* and *Aliaksandra*, *Tobbe* and *Anna*, *Danne* and *Angelica*, *Maria* and *Jocke*, and *Frida* and *Magnus*. You have brought me laughter and joy throughout the years and a great way to clear my head. A perfect way to release stress is to enjoy the company of good friends during a "Grabbweekend". For this I would in particular thank the three of my oldest friends *Danne*, *Jimmy* and *Floose*. A special thanks goes out to my oldest friend Danne, who triggered my interest in languages by teaching me the apparently random, proper use of "en" and "ett" in the Swedish language.

At some point the line between close friends and family becomes blurred and some friends like *Frida* and *Magnus* become part of the extended family that you know you can always count on. The others in my extended family like my *father in-law*, *Bernt*, *Gerd* and sisters *Johanna*, *Lina* and *Julia* also deserve a word of gratitude; In particular *Julia*, who helped me make the Swedish summary at the end of the thesis hopefully more clear and understandable. Another special thanks goes out to my *Teściowa* who would fly in from Poland to lovingly take care of my daughter, whenever I needed extra time to focus on my research. Dziękuję! I am also grateful for all the support and the knowledge

that I can always count on Alicia's Godfather, *Krystian* and Godmother, *Maria*. Despite lack of blood relation, Юля is family and I am grateful to her for a lot, Спасибо Юля! *Abuelita Amancia*, my *aunts*, my *uncle*, all my *cousins*, and Катя и Соня also deserve a big thanks for all their love and support. Gracias & Спасибо!

I am forever grateful to my *parents*, who have created the perfect, safe and loving environment for me to be the best I can possibly be. I know that they had to sacrifice their careers, their comforts and their status, when we moved to Sweden, in order for me and my brother to have the best possible opportunities to succeed. I can also never express the level of gratitude I feel towards my *brother* who is also my best man and best friend, and the joy his sun, my nephew, *Leonel* bring to my life. Спасибо Мама, Папа и Вова!

A little more than a year ago, my life changed. I got a new source of inspiration, my daughter Алисия. You inspire me to be the best I can be and in your name you also carry the reminder of my initial source of inspiration, my late grandmother Алиса Яковлевна. It was my grandmother's stories about the dinosaurs, volcanoes and the second World War, which triggered my initial curiosity for science. Спасибо бабушка!

Finally, none of this would be possible without the constant support and endless love from my precious wife, *Izabela*. I am deeply grateful for all you had to sacrifice to allow me the time I needed to conduct my research and to support me in every way that you could. You also brought me my new source of inspiration and motivation, our daughter Alicia. Kocham Cię, always!

References

Abdulrehman, D., Monteiro, P. T., Teixeira, M. C., Mira, N. P., Lourenço, A. B., Dos Santos, S. C., Cabrito, T. R., Francisco, A. P., Madeira, S. C., Aires, R. S., Oliveira, A. L., Sá-Correia, I., and Freitas, A. T. (2011). YEASTRACT: Providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Research*, 39(Database issue):D136–D140.

Acland, A., Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., Bryant, S. H., Canese, K., Church, D. M., Clark, K., Dicuccio, M., Dondoshansky, I., Federhen, S., Feolo, M., Geer, L. Y., Gorelenkov, V., Hoeppner, M., Johnson, M., Kelly, C., Khotomlianski, V., Kimchi, A., Kimelman, M., Kitts, P., Krasnov, S., Kuznetsov, A., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Karsch-Mizrachi, I., Murphy, T., Ostell, J., O'Sullivan, C., Panchenko, A., Phan, L., Pruitt, D. P. K. D., Rubinstein, W., Sayers, E. W., Schneider, V., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Siyan, K., Slotta, D., Soboleva, A., Soussov, V., Starchenko, G., Tatusova, T. A., Trawick, B. W., Vakatov, D., Wang, Y., Ward, M., John Wilbur, W., Yaschenko, E., and Zbicz, K. (2014). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 42(D1):D7–D17.

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6(1):55.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. Nature, 406: 378–482, 2000. *Nature*, 406(6794):378–382.

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., Schmitt, T., Tjärnberg, A., Guala, D., Frings, O., and Sonnhammer, E. L. L. (2011). Comparative interactomics with Funcoup 2.0. *Nucleic Acids Research*, 40(November 2011):821–828.

Alexeyenko, A. and Sonnhammer, E. L. L. (2009). Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Research*, 19(6):1107–16.

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P., Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F. F., and Hogue, C. W. V. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, 33(Database issue):D418–D424.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, 38(Database issue):D525–31.

Aravind, L. (2000). Guilt by association: Contextual information in genome analysis.

Arighi, C. N., Roberts, P. M., Agarwal, S., Bhattacharya, S., Cesareni, G., Chatr-Aryamontri, A., Clematide, S., Gaudet, P., Giglio, M. G., Harrow, I., Huala, E., Krallinger, M., Leser, U., Li, D., Liu, F., Lu, Z., Maltais, L. J., Okazaki, N., Perfetto, L., Rinaldi, F., Sætre, R., Salgado, D., Srinivasan, P., Thomas, P. E., Toldo, L., Hirschman, L., and Wu, C. H. (2011). BioCreative III interactive task: an overview. *BMC Bioinformatics*, 12 Suppl 8:S4.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–9.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424.

Barabasi, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375–20120375.

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

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research*, 41(D1):D991–D995.

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289 – 300.

Betel, D., Wilson, M., Gabow, A., Marks, D. S., and Sander, C. (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Research*, 36(Database issue):D149–53.

Beveridge, A. and Shan, J. (2016). Nerwork of Thrones. *Math Horizons*, (April):18–22.

Björkholm, P. and Sonnhammer, E. L. L. (2009). Comparative analysis and unification of domain-domain interaction networks. *Bioinformatics*, 25(22):3020–3025.

Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., Sitnikov, D., Burgess, S., Buza, T., Gresham, C., McCarthy, F., Pillai, L., Wang, H., Carbon, S., Dietze, H., Lewis, S. E., Mungall, C. J., Munoz-Torres, M. C., Feuermann, M., Gaudet, P., Basu, S., Chisholm, R. L., Dodson, R. J., Fey, P., Mi, H., Thomas, P. D., Muruganujan, A., Poudel, S., Hu, J. C., Aleksander, S. A., McIntosh, B. K., Renfro, D. P., Siegele, D. A., Attrill, H., Brown, N. H., Tweedie, S., Lomax, J., Osumi-Sutherland, D., Parkinson, H., Roncaglia, P., Lovering, R. C., Talmud, P. J., Humphries, S. E., Denny, P., Campbell, N. H., Foulger, R. E., Chibucos, M. C., Giglio, M. G., Chang, H. Y., Finn, R., Fraser, M., Mitchell, A., Nuka, G., Pesseat, S., Sangrador, A., Scheremetjew, M., Young, S. Y., Stephan, R., Harris, M. A., Oliver, S. G., Rutherford, K., Wood, V., Bahler, J., Lock, A., Kersey, P. J., McDowall, M. D., Staines, D. M., Dwinell, M., Shimoyama, M., Laulederkind, S., Hayman, G. T., Wang, S. J., Petri, V., D'Eustachio, P., Matthews, L., Balakrishnan, R., Binkley, G., Cherry, J. M., Costanzo, M. C., Demeter, J., Dwight, S. S.,

Engel, S. R., Hitz, B. C., Inglis, D. O., Lloyd, P., Miyasato, S. R., Paskov, K., Roe, G., Simison, M., Nash, R. S., Skrzypek, M. S., Weng, S., Wong, E. D., Berardini, T. Z., Li, D., Huala, E., Argasinska, J., Arighi, C., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Bateman, A., Bely, B., Blatter, M. C., Bonilla, C., Bougueleret, L., Boutet, E., Breuza, L., Bridge, A., Britto, R., Casals, C., Cibrian-Uhalte, E., Coudert, E., Cusin, I., Duek-Roggli, P., Estreicher, A., Famiglietti, L., Gane, P., Garmiri, P., Gos, A., Gruaz-Gumowski, N., Hatton-Ellis, E., Hinz, U., Hulo, C., Huntley, R., Jungo, F., Keller, G., Laiho, K., Lemercier, P., Lieberherr, D., Macdougall, A., Magrane, M., Martin, M., Masson, P., Mutowo, P., O'Donovan, C., Pedruzzi, I., Pichler, K., Poggioli, D., Poux, S., Rivoire, C., Roechert, B., Sawford, T., Schneider, M., Shypitsyna, A., Stutz, A., Sundaram, S., Tognolli, M., Wu, C., Xenarios, I., Chan, J., Kishore, R., Sternberg, P. W., Van Auken, K., Muller, H. M., Done, J., Li, Y., Howe, D., and Westerfeld, M. (2015). Gene ontology consortium: Going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056.

Boone, C., Bussey, H., and Andrews, B. J. (2007). Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, 8(6):437–449.

Börnigen, D., Tranchevent, L.-C. C. L., Bonachela-Capdevila, F., Devriendt, K., De Moor, B., De Causmaecker, P., Moreau, Y., Bornigen, D., and Tranchevent, L.-C. C. L. (2012). An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28(23):3081–8.

Borsch, S. J. (2005). *The black death in Egypt and England: A comparative study*.

Boutayeb, A. (2010). *The Impact of Infectious Diseases on the Development of Africa*, pages 1171–1188. Springer New York, New York, NY.

Bradley, A. P., Use, T. H. E., The, O. F., Under, A., Roc, T. H. E., In, C., Of, E., and Learning, M. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceedings - International Conference on Pattern Recognition*, pages 3121–3124.

Bromberg, Y. (2013). Chapter 15: disease gene prioritization. *PLoS Computational Biology*, 9(4):e1002902.

Brown, J. B. and Celniker, S. E. (2015). Lessons from modENCODE. *Annual Review of Genomics and Human Genetics*, 16(1):31–53.

Browne, F., Wang, H., and Zheng, H. (2015). A computational framework for the prioritization of disease-gene candidates. *BMC Genomics*, 16 Suppl 9:S2.

Butterworth, R., Simovici, D. A., Santos, G. S., and Ohno-Machado, L. (2004). A greedy algorithm for supervised discretization. *Journal of Biomedical Informatics*, 37(4):285–292.

Chalancon, G., Kruse, K., and Babu, M. M. (2013). *Clustering Coefficient*, pages 422–424. Springer New York, New York, NY.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B. J., Dolinski, K., and Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379.

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37(Web Server issue):W305–11.

Chen, Y. C. and Urban, P. L. (2013). Time-resolved mass spectrometry.

Cho, H., Berger, B., and Peng, J. (2016). Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Systems*, 3(6):540–548.e5.

Cickovski, T., Peake, E., Aguiar-Pulido, V., and Narasimhan, G. (2017). ATria: a novel centrality algorithm applied to biological networks. *BMC Bioinformatics*, 18(Suppl 8):239.

Clark, R. D. and Webster-Clark, D. J. (2008). Managing bias in ROC curves. *Journal of Computer-Aided Molecular Design*, 22(3-4):141–146.

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., van Leeuwen, J., van Dyk, N., Lin, Z.-Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., Srikumar, T., Bahr, S., Chen, Y., Deshpande, R., Kurat, C. F., Li, S. C., Li, Z., Usaj, M. M., Okada, H., Pascoe, N., San Luis, B.-J., Sharifpoor, S., Shuteriqi, E., Simpkins, S. W., Snider, J., Suresh, H. G., Tan, Y., Zhu, H., Malod-Dognin, N., Janjic, V., Przulj, N., Troyanskaya, O. G., Stagljar, I., Xia, T., Ohya, Y., Gingras, A.-C., Raught, B., Boutros, M., Steinmetz, L. M., Moore, C. L., Rosebrock, A. P., Caudy, A. A., Myers, C. L., Andrews, B., and Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306):aaf1420–aaf1420.

Cover, T. M. and Thomas, J. A. (2005). *Elements of Information Theory*.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–8.

de Sola Pool, I. and Kochen, M. (1978). Contacts and influence. *Social Networks*, 1(1):5–51.

Deutsch, E. W., Lam, H., and Aebersold, R. (2008). PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO reports*, 9(5):429–34.

Doncheva, N. T., Kacprowski, T., and Albrecht, M. (2012). Recent approaches to the prioritization of candidate disease genes. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 4(5):429–42.

Dongen, S. V. (2000). A cluster algorithm for graphs. Technical Report R 0010.

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.

Elefsinioti, A., Sarac, O. S., Hegele, A., Plake, C., Hubner, N. C., Poser, I., Sarov, M., Hyman, A., Mann, M., Schroeder, M., Stelzl, U., and Beyer, A. (2011). Large-scale De Novo Prediction of Physical Protein-Protein Association. *Molecular & Cellular Proteomics*, 10(11):M111.010629–M111.010629.

ElShal, S., Tranchevent, L.-C. C., Sifrim, A., Ardeshirdavani, A., Davis, J., and Moreau, Y. (2016). Beegle: From literature mining to disease-gene discovery. *Nucleic Acids Research*, 44(2):e18.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90.

Erdös, P. and Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297.

Erten, S., Bebek, G., and Koyutürk, M. (2011). Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *Journal of Computational Biology*, 18(11):1561–74.

Esteller, M. (2011). Non-coding RNAs in human disease. *Nature Reviews Genetics*, 12(12):861–874.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H. Y., Dosztanyi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, S. C., Wu, C. H., Xenarios, I., Yeh, L. S., Young, S. Y., and Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research*, 45(D1):D190–D199.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews. Cancer*, 4(3):177–83.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147.

Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature Genetics*, 29(4):482–6.

Ghanbarian, A. T. and Hurst, L. D. (2015). Neighboring genes show correlated evolution in gene expression. *Molecular Biology and Evolution*, 32(7):1748–1766.

Gillis, J. and Pavlidis, P. (2012). "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3):e1002444.

Goh, K.-i., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–90.

Grigoriev, A. (2001). A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Research*, 29(17):3513–3519.

Guala, D., Bernhem, K., Ait Blal, H., Lundberg, E., Brismar, H., and Sonnhammer, E. L. L. (2017). Experimental validation of predicted cancer genes using FRET (manuscript).

Guala, D., Sjölund, E., and Sonnhammer, E. L. 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. L. (2017). A large-scale benchmark of gene prioritization methods. *Scientific Reports*, 7:46598.

Harmston, N., Filsell, W., and Stumpf, M. P. H. (2010). What the papers say: text mining for genomics and systems biology. *Human Genomics*, 5(1):17–29.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52.

Hassani-Pak, K. and Rawlings, C. (2017). Knowledge Discovery in Biological Databases for Revealing Candidate Genes Linked to Complex Phenotypes. *Journal of Integrative Bioinformatics*, 14(1).

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. *Elements*, 1:337–387.

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115.

Hong, S., Chen, X., Jin, L., and Xiong, M. (2013). Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Research*, 41(8):e95.

Hsu, C.-L. C.-T., Huang, Y.-H., Hsu, C.-L. C.-T., and Yang, U.-C. (2011). Prioritizing disease candidate genes by a gene interconnectedness-based approach. *BMC Genomics*, 12 Suppl 3:S25.

Hu, J. X., Thomas, C. E., and Brunak, S. (2016). Network biology concepts in complex disease comorbidities. *Nature Reviews. Genetics*, 17(10):615–629.

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., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, 8(565):1–9.

Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Research*, 18(4):644–52.

Kamburov, A., Grossmann, A., Herwig, R., and Stelzl, U. (2012). Cluster-based assessment of protein-protein interaction confidence. *BMC Bioinformatics*, 13(1):262.

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 Update. *Nucleic Acids Research*, 41(D1):D793–D800.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361.

Katagiri, F. and Glazebrook, J. (2009). Overview of mRNA expression profiling using DNA microarrays. *Current Protocols in Molecular Biology*, Chapter 22(SUPPL. 85):Unit 22.4.

Keilwagen, J., Grosse, I., and Grau, J. (2014). Area under precision-recall curves for weighted and unweighted data. *PLoS ONE*, 9(3).

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Research*, 37(Database issue):D767–D772.

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.

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., Flicek, P., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., and Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database: The Journal of Biological Databases and Curation*, 2011(0):bar030.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297.

Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(May 1997):668–677.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, pages 1137–1143. Morgan Kaufmann Publishers Inc.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, 82(4):949–58.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., Megy, K., Pilicheva, E., Rustici, G., Tikhonov, A., Parkinson, H., Petryszak, R., Sarkans, U., and Brazma, A. (2015). ArrayExpress update-simplifying data submissions. *Nucleic Acids Research*, 43(D1):D1113–D1116.

Kotlyar, M., Pastrello, C., Pivetta, F., Lo Sardo, A., Cumbaa, C., Li, H., Naranian, T., Niu, Y., Ding, Z., Vafaee, F., Broackes-Carter, F., Petschnigg, J., Mills, G. B., Jurisicova, A., Stagljar, I., Maestro, R., and Jurisica, I. (2015). In silico prediction of physical protein interactions and characterization of interactome orphans. *Nature Methods*, 12(1):79–84.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643.

Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–16.

Lakowicz, J. R. (1999). Energy Transfer. In *Principles of Fluorescence Spectroscopy*, pages 367–394. Springer US, Boston, MA.

Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635.

Lee, H., Tu, Z., Deng, M., Sun, F., and Chen, T. (2006). Diffusion kernel-based logistic regression models for protein function prediction. *Omics : a Journal of Integrative Biology*, 10(1):40–55.

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):1109–21.

Letovsky, S. and Kasif, S. (2003). Predicting protein function from protein/protein interaction data: A probabilistic approach. In *Bioinformatics*, volume 19, pages i197–i204. Oxford University Press.

Li, Y. and Patra, J. C. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224.

Lin, D. (1998). An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304.

Linghu, B., Snitkin, E. S., Hu, Z., Xia, Y., and Delisi, C. (2009). Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biology*, 10(9):R91.

Mackay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*.

Maiella, S., Rath, A., Angin, C., Mousson, F., and Kremp, O. (2013). Orphanet and its consortium: where to find expert-validated information on rare diseases. *Revue Neurologique*, 169(SUPPL.1):S3–S8.

Marcotte, E. M. (1999). Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science*, 285(5428):751–753.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–6.

Marcotte, E. M., Xenarios, I., and Eisenberg, D. (2001). Mining literature for protein-protein interactions. *Bioinformatics*, 17(4):359–63.

Martínez, V., Cano, C., and Blanco, A. (2014). ProphNet: a generic prioritization method through propagation of information. *BMC Bioinformatics*, 15 Suppl 1:S5.

Masoudi-Nejad, A., Meshkin, A., Haji-Eghrari, B., and Bidkhori, G. (2012). Candidate gene prioritization. *Molecular Genetics and Genomics*, 287(9):679–98.

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2016). JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126.

Mazza, A., Klockmeier, K., Wanker, E., and Sharan, R. (2016). An integer programming framework for inferring disease complexes from network data. *Bioinformatics*, 32(12):i271–i277.

McCormack, T., Frings, O., Alexeyenko, A., and Sonnhammer, E. L. L. (2013). Statistical Assessment of Crosstalk Enrichment between Gene Groups in Biological Networks. *PLoS ONE*, 8(1):e54945.

Moreau, Y. and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews. Genetics*, 13(8):523–36.

Morgan, A. A., Lu, Z., Wang, X., Cohen, A. M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.-h., Torres, R., Krauthammer, M., Lau, W. W., Liu, H., Hsu, C.-N., Schuemie, M., Cohen, K. B., and Hirschman, L. (2008). Overview of BioCreative II gene normalization. *Genome Biology*, 9 Suppl 2:S3.

Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J. M., and Marra, M. A. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45(1):81–94.

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 Suppl 1:S4.

Mrowka, R., Patzak, A., and Herzel, H. (2001). Is there a bias in proteome research? *Genome Research*, 11(12):1971–1973.

Myers, C. L. and Troyanskaya, O. G. (2007). Context-sensitive data integration and prediction of biological networks. *Bioinformatics*, 23(17):2322–30.

Myers, R. M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R. C., Bernstein, B. E., Gingeras, T. R., Kent, W. J., Birney, E., Wold, B., Crawford, G. E., Epstein, C. B., Shoresh, N., Ernst, J., Mikkelsen, T. S., Kheradpour, P., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Lin, M. F., Kellis, M., Davis, C. A., Kapranov, P., Dobin, A., Zaleski, C., Schlesinger, F., Batut, P., Chakrabortty, S., Jha, S., Lin, W., Drenkow, J., Wang, H., Bell, K., Bell, I., Gao, H., Dumais, E., Dumais, J., Antonarakis, S. E., Ucla, C., Borel, C., Guigo, R., Djebali, S., Lagarde, J., Kingswood, C., Ribeca, P., Sammeth, M., Alioto, T., Merkel, A., Tilgner, H., Carninci, P., Hayashizaki, Y., Lassmann, T., Takahashi, H., Abdelhamid, R. F., Hannon, G., Fejes, K. T., Preall, J., Gordon, A., Sotirova, V., Reymond, A., Howald, C., Graison, E. A. Y., Chrast, J., Ruan, Y., Ruan, X., Shahab, A., Poh, W. T., Wei, C. L., Furey, T. S., Boyle, A. P., Sheffield, N. C., Song, L., Shibata, Y., Vales, T., Winter, D., Zhang, Z., London, D., Wang, T., Keefe, D., Iyer, V. R., Lee, B. K., McDaniell, R. M., Liu, Z., Battenhouse, A., Bhinge, A. A., Lieb, J. D., Grasfeder, L. L., Showers, K. A., Giresi, P. G., Kim, S. K., Shestak, C., Pauli, F., Reddy, T. E., Gertz, J., Partridge, E. C., Jain, P., Sprouse, R. O., Bansal, A., Pusey, B., Muratet, M. A., Varley, K. E., Bowling, K. M., Newberry, K. M., Nesmith, A. S., Dilocker, J. A., Parker, S. L., Waite, L. L., Thibeault, K., Roberts, K., Absher, D. M., Mortazavi, A., Williams, B., Marinov, G., Trout, D., King, B., McCue, K., Kirilusha, A., DeSalvo, G., Fisher, K. A.,

Amrhein, H., Pepke, S., Vielmetter, J., Sherlock, G., Sidow, A., Batzoglou, S., Rauch, R., Kundaje, A., Libbrecht, M., Margulies, E. H., Parker, S. C., Elnitski, L., Green, E. D., Hubbard, T., Harrow, J., Searle, S., Parker, S. C., Aken, B., Frankish, A., Hunt, T., Despacio-Reyes, G., Kay, M., Mukherjee, G., Bignell, A., Saunders, G., Boychenko, V., Brent, M., van Baren, M. J., Brown, R. H., Gerstein, M., Khurana, E., Balasubramanian, S., Lam, H., Cayting, P., Robilotto, R., Lu, Z., Derrien, T., Tanzer, A., Knowles, D. G., Mariotti, M., Haussler, D., Harte, R., Diekhans, M., Lin, M., Valencia, A., Tress, M., Rodriguez, J. M., Raha, D., Shi, M., Euskirchen, G., Grubert, F., Kasowski, M., Lian, J., Lacroute, P., Xu, Y., Monahan, H., Patacsil, D., Slifer, T., Yang, X., Charos, A., Reed, B., Wu, L., Auerbach, R. K., Habegger, L., Hariharan, M., Rozowsky, J., Abyzov, A., Weissman, S. M., Struhl, K., Lamarre-Vincent, N., Lindahl-Allen, M., Miotto, B., Moqtaderi, Z., Fleming, J. D., Newburger, P., Farnham, P. J., Frietze, S., O'Geen, H., Xu, X., Blahnik, K. R., Cao, A. R., Iyengar, S., Kaul, R., Thurman, R. E., Wang, H., Navas, P. A., Sandstrom, R., Sabo, P. J., Weaver, M., Canfield, T., Lee, K., Neph, S., Roach, V., Reynolds, A., Johnson, A., Rynes, E., Giste, E., Vong, S., Neri, J., Frum, T., Nguyen, E. D., Ebersol, A. K., Sanchez, M. E., Sheffer, H. H., Lotakis, D., Haugen, E., Humbert, R., Kutyavin, T., Shafer, T., Noble, W. S., Dekker, J., Lajoie, B. R., Sanyal, A., Rosenbloom, K. R., Dreszer, T. R., Raney, B. J., Barber, G. P., Meyer, L. R., Sloan, C. A., Malladi, V. S., Cline, M. S., Learned, K., Swing, V. K., Zweig, A. S., Rhead, B., Fujita, P. A., Roskin, K., Karolchik, D., Kuhn, R. M., Wilder, S. P., Sobral, D., Herrero, J., Beal, K., Lukk, M., Brazma, A., Vaquerizas, J. M., Luscombe, N. M., Bickel, P. J., Boley, N., Brown, J. B., Li, Q., Huang, H., Sboner, A., Yip, K. Y., Cheng, C., Yan, K. K., Bhardwaj, N., Wang, J., Lochovsky, L., Jee, J., Gibson, T., Leng, J., Du, J., Harris, R. S., Song, G., Miller, W., Suh, B., Paten, B., Hoffman, M. M., Buske, O. J., Weng, Z., Dong, X., Wang, J., Xi, H., Tenenbaum, S. A., Doyle, F., Chittur, S., Penalva, L. O., Tullius, T. D., White, K. P., Karmakar, S., Victorsen, A., Jameel, N., Bild, N., Grossman, R. L., Collins, P. J., Trinklein, N. D., Giddings, M. C., Khatun, J., Maier, C., Wang, T., Whitfield, T. W., Chen, X., Yu, Y., Gunawardena, H., Feingold, E. A., Lowdon, R. F., Dillon, L. A., Good, P. J., and Risk, B. (2011). A user's guide to the Encyclopedia of DNA elements (ENCODE). *PLoS Biology*, 9(4):e1001046.

Navlakha, S. and Kingsford, C. (2011). Network archaeology: uncovering ancient networks from present-day interactions. *PLoS Computational Biology*, 7(4):e1001119.

Negi, S., Pandey, S., Srinivasan, S. M., Mohammed, A., and Guda, C. (2015). LocSigDB: a database of protein localization signals. *Database: The Journal of Biological Databases and Curation*, 2015:bav003.

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

Ogris, C., Guala, D., Kaduk, M., and Sonnhammer, E. L. L. (2017b). FunCoup 4: new species, data, and visualization (manusript).

Orfanoudaki, G. and Economou, A. (2014). Proteome-wide subcellular topologies of E. coli polypeptides database (STEPdb). *Molecular & Cellular Proteomics*, 13(12):3674–87.

Östlund, G., Lindskog, M., and Sonnhammer, E. L. L. (2010). Network-based Identification of novel cancer genes. *Molecular & Cellular Proteomics*, 9(4):648–55.

Oti, M., Ballouz, S., and Wouters, M. A. (2011). Web tools for the prioritization of candidate disease genes. *Methods in Molecular Biology*, 760:189–206.

Oti, M. and Brunner, H. G. (2007). The modular nature of genetic diseases. *Clinical Genetics*, 71(1):1–11.

Page, L., Brin, S., Motwani, R., Winograd, T., and Page, Lawrence; Brin, Sergey; Motwani, Rajeev; Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. *World Wide Web Internet And Web Information Systems*, pages 1–17.

Park, C. Y., Wong, A. K., Greene, C. S., Rowland, J., Guan, Y., Bongo, L. A., Burdine, R. D., and Troyanskaya, O. G. (2013). Functional Knowledge Transfer for High-accuracy Prediction of Under-studied Biological Processes. *PLoS Computational Biology*, 9(3):e1002957.

Pavlidis, P., Weston, J., Cai, J., and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. *Journal of Computational Biology*, 9(2):401–411.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Biochemistry*, 96(8):4285–4288.

Pertea, M., Ayanbule, K., Smedinghoff, M., and Salzberg, S. L. (2009). OperonDB: A comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Research*, 37(Database issue):D479–D482.

Piro, R. M. and Di Cunto, F. (2012). Computational approaches to disease-gene prediction: rationale, classification and successes. *The FEBS journal*, 279(5):678–96.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database issue):D501–4.

Punta, M., Coggill, P., Eberhardt, R., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E., Eddy, S., Bateman, A., and Finn, R. (2012). The Pfam protein families databases. *Nucleic Acids Research*, 30(1):1–12.

Rastogi, S. and Rost, B. (2011). LocDB: Experimental annotations of localization for homo sapiens and arabidopsis thaliana. *Nucleic Acids Research*, 39(Database issue):D230–4.

Ravasz, E. and Barabasi, A. L. (2003). Hierarchical organization in complex networks. *Physical Review E*, 67(2 Pt 2):26112.

Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1):405.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1 - IJCAI'95*, 1:6.

Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology*, 23(8):951–959.

Rigaut, G., Shevchenko, A., Rutz, B., Matthias, W., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17(10):1030–1032.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3).

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):D449–51.

Sato, T., Yamanishi, Y., Kanehisa, M., and Toh, H. (2005). The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, 21(17):3482–3489.

Schmitt, T., Ogris, C., and Sonnhammer, E. L. L. (2014). FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Research*, 42(1):D380–8.

Seuring, T., Archangelidi, O., and Suhrcke, M. (2015). The Economic Costs of Type 2 Diabetes: A Global Systematic Review.

Sharma, A., Kitsak, M., Ghiassian, S., and Vidal, M. (2005). Uncovering disease-disease relationships through the human interactome. *Science*, 310(5751):1122–1123.

Shim, J. E., Hwang, S., and Lee, I. (2015). Pathway-Dependent Effectiveness of Network Algorithms for Gene Prioritization. *PLoS ONE*, 10(6):e0130589.

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(1):7–30.

Siganos, G., Tauro, S. L., and Faloutsos, M. (2006). Jellyfish: A conceptual model for the as Internet topology. *Communications and Networks, Journal of*, 8(3):339–350.

Signorelli, M., Vinciotti, V., Wit, E. C., Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Kanehisa, M., Goto, S., Huang, D., Sherman, B., Lempicki, R., Robinson, M., Grigull, J., Mohammad, N., Hughes, T., Beißbarth, T., Speed, T., Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., Kim, S., Volsky, D., Smet, R. D., Marchal, K., Marbach, D., Prill, R., Schaffter, T., Mattiussi, C., Floreano, D., Stolovitzky, G., Lauritzen, S., Friedman, J., Hastie, T., Tibshirani, R., Abegaz, F., Wit, E. C., Marbach, D., Costello, J., Küffner, R., Vega, N., Prill, R., Camacho, D., Allison, K., Kellis, M., Collins, J., Stolovitzky, G., Kim, H., Shin, J., Kim, E., Kim, H., Hwang, S., Shim, J., Lee, I., Schmitt, T., Ogris, C., Sonnhammer, E., Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., Valencia, A., Alexeyenko, A., Lee, W., Pernemalm, M., Guegan, J., Dessen, P., Lazar, V., Lehtiö, J., Pawitan, Y., McCormack, T., Frings, O., Alexeyenko, A., Sonnhammer, E., Gibbons, J., Pratt, J., Blaker, H., Agresti, A., Csardi, G., Nepusz, T., Goffeau, A., Barrell, B., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., and Brown, P. (2016). NEAT: an efficient network enrichment analysis test. *BMC Bioinformatics*, 17(1):352.

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., and Stagljar, I. (2015). Fundamentals of protein interaction network mapping. *Molecular Systems Biology*, 11(12):848.

Söderberg, O., Gullberg, M., Jarvius, M., Ridderstråle, K., Leuchowius, K.-J., Jarvius, J., Wester, K., Hydbring, P., Bahram, F., Larsson, L.-G., and Landegren, U. (2006). Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nature Methods*, 3(12):995–1000.

Sonnhammer, E. L. L. and Östlund, G. (2015). InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, 43(D1):D234–D239.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, 13(7):e1002195.

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

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., and von Mering, C. (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., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., Kim, C. J., Kusanovic, J. P., and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82.

Tiffin, N. (2011). Conceptual thinking for in silico prioritization of candidate disease genes. *Methods in Molecular Biology*, 760:175–87.

Tong, A. H. Y. (2001). Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science*, 294(5550):2364–2368.

Tranchevent, L.-C., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 12(1):22–32.

Tranchevent, L.-C. C., Ardeshirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Auboeuf, D., and Moreau, Y. (2016). Candidate gene prioritization with Endeavour. *Nucleic Acids Research*, 44(W1):W117–W121.

Tsuda, K., Shin, H., and Schölkopf, B. (2005). Fast protein classification with multiple networks. *Bioinformatics*, 21 Suppl 2:ii59–65.

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, 6(1):e1000641.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–7.

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.

Winblad, B., Amouyel, P., Andrieu, S., Ballard, C., Brayne, C., Brodaty, H., Cedazo-Minguez, A., Dubois, B., Edvardsson, D., Feldman, H., Fratiglioni, L., Frisoni, G. B., Gauthier, S., Georges, J., Graff, C., Iqbal, K., Jessen, F., Johansson, G., Jönsson, L., Kivipelto, M., Knapp, M., Mangialasche, F., Melis, R., Nordberg, A., Rikkert, M. O., Qiu, C., Sakmar, T. P., Scheltens, P., Schneider, L. S., Sperling, R., Tjernberg, L. O., Waldemar, G., Wimo, A., and Zetterberg, H. (2016). Defeating Alzheimer's disease and other dementias: A priority for European science and society.

Wong, A. K., Krishnan, A., Yao, V., Tadych, A., and Troyanskaya, O. G. (2015). IMP 2.0: A multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Research*, 43(W1):W128–W133.

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Molecular Systems Biology*, 4(189):189.

Yang, P., Li, X., Wu, M., Kwoh, C. K., and Ng, S. K. (2011). Inferring Gene-Phenotype associations via global protein complex network propagation. *PLoS ONE*, 6(7):e21502.

Yu, C. L., Louie, T. M., Summers, R., Kale, Y., Gopishetty, S., and Subramanian, M. (2009). Two distinct pathways for metabolism of theophylline and caffeine are coexpressed in Pseudomonas putida CBB5. *Journal of Bacteriology*, 191(14):4624–32.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–10.

Yu, J., Murali, T., and Finley, R. L. (2012). Assigning confidence scores to protein-protein interactions. *Methods in Molecular Biology*, 812:161–174.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., and Honig, B. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):1–6.

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, 9(1):e78644.

Zhu, C., Wu, C., Aronow, B. J., and Jegga, A. G. (2014). Computational approaches for human disease gene prediction and ranking. *Advances in Experimental Medicine and Biology*, 799:69–84.

Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C. T., Bader, G. D., and Morris, Q. (2013). GeneMANIA prediction server 2013 update. *Nucleic Acids Research*, 41(Web Server issue):W115–W122.