Data integration for robust network-based disease gene prediction

Gabriel Östlund

Data integration for robust network-based disease gene prediction

Gabriel Östlund

ISBN 978-91-7447-629-3

Printed in Sweden by US-AB, Stockholm 2013
Distributor: Department of Biochemistry and Biophysics

Till mina älsklingar

List of publications

Publications included in this thesis

Paper I: Berglund, A. C., Sjolund, E., Ostlund, G., and Sonnhammer, E. L. 2008. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res. 36: D263-D266.

Paper II: Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., Frings, O., and Sonnhammer, E. L. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38: D196-D203.

Paper III: Ostlund, G., Lindskog, M., and Sonnhammer, E. L. 2010. Network-based Identification of novel cancer genes. Mol. Cell. Proteomics 9: 648-655.

Paper IV: Ostlund, G. and Sonnhammer, E. L. 2012. Quality criteria for finding genes with high mRNA-protein expression correlation and coexpression correlation. Gene 497: 228-236.

Paper V: Ostlund, G. and Sonnhammer, E. L. 2013. Pitfalls in gene (co)expression meta-analysis. *Manuscript*.

Abstract

For many complex diseases the cause/mechanism can be tied not to a single gene and in order to cope with the complexity a systems wide approach is needed. By combining evidence indicative of functional association it is possible to infer networks of protein functional coupling. The reliability of these networks is dependent on having sufficient data and on the data being informative.

By combining evidence from multiple species, functional coupling networks can reach higher coverage and accuracy. Genes in different species derived from the same gene by a speciation event are orthologous and likely to have a conserved function. In order to enable the transfer of information across species we inferred orthology with the InParanoid algorithm and made the inferences available to the public in the associated database.

Identification of genes involved in diseases is an important biomedical goal. Based on the "guilt by association" principle, we implemented an approach, Maxlink, for identifying and prioritizing novel disease genes. By searching the FunCoup network for genes functionally coupled to cancer genes we identified some 1800 novel cancer gene candidates showing characteristics of cancer genes.

Inferred networks of functional coupling give a large coverage but do so at the cost of losing information about the specifics of the couplings. Combining the network with gene expression data from patient/control studies could give context by highlighting the couplings and genes relevant for a specific disease.

While proteins are the active components, mRNA is often used as a proxy due to the difficulty of measuring protein abundance. We examined the relationship between mRNA and protein, using properties of expression profiles to identify subsets of genes with higher mRNA-protein concordance.

If technical and biological differences between patient/control studies of gene expression have a large impact, the results of studies of the same disease might be inconsistent. To determine this impact we examined the consistency in differential (co)expression between different studies of cancer, as well as non-cancer studies. Such consistency could generally be found, even between studies of different diseases, but only when common pitfalls of gene expression analysis are avoided.

Contents

List of publications vii
Publications included in this thesis vii
Abstract ix
Introduction 15
Background 18
Orthology 18
A primer on gene evolution 18
Orthology concepts 19
Orthology conjecture 20
Orthology interference 20
Comparison of orthology inferences 25
The need for standardization 25
Gene/protein networks 27
Reliable predictions based on many weak evidences 27
Evidence sources 28
Inferred networks of protein functional coupling 31
Functional coupling networks and disease 36
Differential coexpression and network analysis 37
Expression analysis 38
Measuring mRNA abundance 38
Measuring protein abundance 40
Differential expression and coexpression 41
Present investigations 44

InParanoid (Papers I and II) 44
Reflections 44
Maxlink (Paper III) 47
Reflections 48
Concordance between mRNA and protein expression (Paper IV) 49
Reflections 50
Consistency between gene expression studies (Paper V) 50
Reflections 51
Future Perspectives 52
Sammanfattning på svenska 55
Acknowledgements 58
References 60

Abbreviations

3D	Three-Dimensional
BLAST	Basic Local Alignment Search Tool
cDNA	complementary DNA
DNA	Deoxyribonucleic Acid
MAS5	Microarray Analysis Suite 5
RMA	Robust Multi-Array Average
GCRMA	Guanine Cytosine Robust Multi-Array Analysis
GO	Gene Ontology
HPA	Human Protein Atlas
HPLC	High-Performance Liquid Chromatography
KEGG	Kyoto Encyclopedia of Genes and Genomes
MCL	Markov Cluster
MIAME	Minimum Information About a Microarray Experiment
mRNA	messenger RNA
MS	Mass Spectrometry
PAM	Point Accepted Mutation
PrEST	Protein Epitope Signature Tag
qPCR	quantitative Polymerase Chain Reaction
RBH	Reciprocal Best Hit
RNA	Ribonucleic Acid
RNAi	RNA interference
RSD	Reciprocal Smallest Distance
siRNA	small interfering RNA

Introduction

For many complex diseases the cause/mechanism cannot be tied to a single gene. Rather there are a number of distinct changes that are necessary. This has the effect that the mechanisms can differ between patients depending on the specific changes and contributing factors for each individual. In order to cope with such complex diseases, a systems-wide approach is needed.

With the advances in sequencing enabling the identification of all genes, the main active components of the cellular machinery, the proteins, are now largely known. In order to fully exploit this knowledge we need to determine how proteins are functionally coupled. That is, we need to map the functions as well as members of the processes, pathways and interactions necessary to perform those functions.

Mapping the functional couplings of all proteins could be done by examining how they interact and under what conditions. Unfortunately, experimentally determining all possible functions and interactions is not feasible in the foreseeable future. There are many modes of interaction and specific conditions could be required for the determination of each one.

However, an alternative path is to try to infer functional coupling by integrating evidence indicative of functional association. The recent years has seen a number of such approaches being suggested, combining evidence in the form of e.g. physical interaction or coexpression and weighing the evidence together to construct inferred networks of functional coupling. These networks reach a coverage of functional coupling above and beyond that of the single types of evidence used to build them. It is however often at the cost of details, the mode and conditions of functional coupling is generally lost.

The focus of this thesis is on the starting- and end points of inferring such networks. Working at one end towards ensuring that data used for inference is of high quality (Paper IV) and can be transferred from multiple species (Papers I and II). At the other end the work has been focused on trying to exploit the information available in the networks to facilitate drawing conclusions about disease genes (Paper III) as well as working towards adding context to the network by examining the consistency of results from expression studies of disease (Paper V).

Evidence from any one organism is limited. It is thus crucial for network inference to be able to combine evidence from multiple species. In order to reliably do this it is necessary to identify corresponding pairs of genes between/across species, orthologs. Papers I and II presents the ongoing work on InParanoid, an algorithm for orthology inference and associated database. InParanoid provides high quality orthology inferences of importance not only for network inference but also to any other application where information needs to be transferred between species.

The impairment of any one protein part of an essential protein complex can have adverse effects. Disease genes have in fact been found to often be functionally coupled to other disease genes and this guilt by association can be exploited to find novel disease genes. In paper III we identify genes that are functionally coupled to cancer genes and have no prior association to cancer, resulting in a list of some 1800 candidates. These candidates show characteristics of cancer genes, more pronounced for candidates linked to many cancer genes.

As mentioned, inferring functional coupling is often done at the cost of losing details. While there might be strong support that two proteins are functionally coupled the context might be limited and not be of importance for disease.

A way towards making an inferred network more context specific would be to combine it with gene expression data. There is a high availability of patient/control studies contrasting gene expression of healthy and diseased individuals in an effort to facilitate increased understanding and diagnosis of a disease. As protein abundance is difficult to measure this is generally done using mRNA abundance as a proxy for proteins.

Using mRNA as a proxy for protein faces the issue that while mRNA abundance is directly linked to protein abundance there is uncertainty as to the exact relationship. In paper IV we examine the concordance between mRNA and protein (co)expression and show that it is possible to utilize properties of mRNA expression profiles to identify genes for which concordance is higher.

Another issue facing patient/control studies of gene expression is that technical and biological differences between studies can impact the consistency of their results. Moreover, the same pathways and processes could potentially be impacted in multiple diseases and a study of a single disease thus be insufficient to draw disease specific conclusions. Paper V examines the consistency of gene expression studies and identifies pitfalls of gene (co)expression analysis.

In conclusion the work presented in this thesis provides improvements to and evaluation of input data for network inference as well as adding crucial piec-

es towards being able to combining functional networks and differential gene (co)expression to make context guided predictions and prioritization of disease genes.

Background

Orthology

A primer on gene evolution

Due to infeasibility of performing many experiments directly on humans, many research findings are discovered using animal models. In order to transfer that information to human, or to other species, when the findings concern genes, one needs to be able to say which gene(s) of one species corresponds to which gene(s) of the other species. In order to do this one needs to make inferences about how the genes have evolved.

Evolution works on a genetic level by genes acquiring random mutations. When a mutation results in changes that are advantageous to the organism's ability to survive and procreate, it will be selected for, when the mutation is disadvantageous, it will be selected against. Mutations can, besides altering function, change signals for expression, leading to truncated products, or a complete stop of expression, a loss of the gene or gene deletion.

In order to obtain a change of function in one gene multiple mutations might be required, where initial mutations result in impaired function. Consequently, some genes crucial for survival will be very stable, having a high selective pressure. Conversely, genes with little benefit can be changed/lose function without overly large disadvantages and thus be more prone to change, experiencing a low selective pressure (Andersson and Andersson 1999).

For small organisms, replicating the genome constitutes a notable part of the energy needed for procreation. It is thus advantageous to keep the genome as small as possible, deleting unneeded genes in order to minimize this cost. At the same time this makes the organism more specialized and less able to adapt to changes in the environment. Gene deletions can happen for higher organisms too, with a mutation of the sequences needed for function and/or expression of unneeded genes increasing the efficiency of the cells (Andersson and Andersson 1999).

It is possible for an organism to acquire genes from another organism, horizontal gene transfer, thus rapidly acquiring new functionality that might have taken generations to evolve. Mitochondria likely originally having been a prokaryote incorporated into early eukaryote cells (Gray 2012), and bacteria commonly exchanging genetic material through the transfer of plasmids are examples of this. Other mechanisms enabling evolutionary leaps include gene fusion, multiple genes or parts of genes being combined, and gene fission, genes being split (Koonin 2005).

Unequal crossing over during homologous recombination, retrotransposition as well as chromosomal or genome duplication can result in multiple identical copies of a gene (Zhang 2003). Such redundancy results in reduced selective pressure, where one copy of the gene can acquire repeated mutations without adversely affecting the organism (Taylor and Raes 2004).

Orthology concepts

Two sequences are said to be homologous when they are derived from the same evolutionary origin. Fitch further refined this with the concepts of orthology and paralogy (Fitch 1970), based on the nature of the event separating the sequences from each other. If the event was a speciation, the sequences are said to be orthologs whereas they are paralogs if the event was a gene duplication. In other words, paralogs have coexisted in the same species whereas orthologs have not.

The concept of orthology/paralogy was further extended by Sonnhammer and Koonin to separate two different classes of paralogs; in-paralogs from out-paralogs (Sonnhammer and Koonin 2002). The difference between in-paralogs and out-paralogs is based on if the gene duplication took place before or after a speciation event of reference, with in-paralogs having diverged after the speciation and out-paralogs before, see Fig. 1. Since out-paralogs have coexisted in the same species, they have likely been under reduced selective pressure and thus may be more likely to have diverged functions compared to orthologs, this is the ortholog conjecture

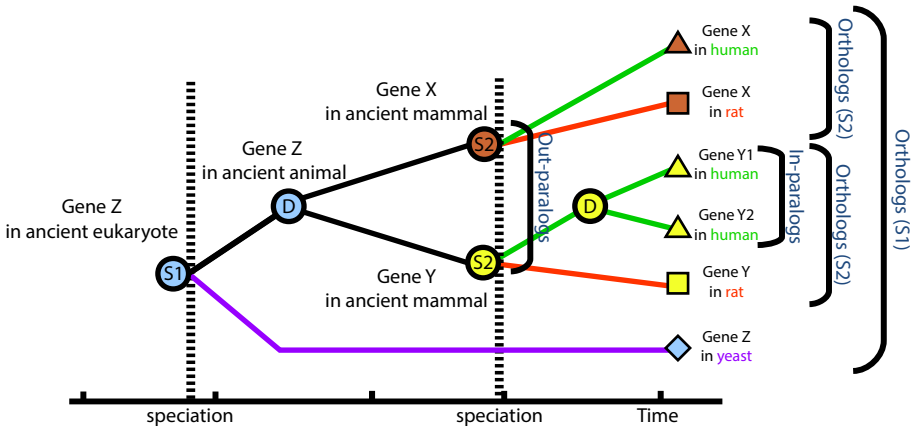

Fig 1: Orthology Concepts. *The evolution of a hypothetical gene family with D denoting a duplication event and S1/S2 denoting the speciation events splitting human from rat and yeast from animals, respectively. Orthologs are derived from one gene by a speciation event while paralogs are derived from one gene by a duplication event. Paralogs are in-paralogs, and co-orthologs, if the duplication event took place after a speciation event of reference and out-paralogs if the duplication event took place prior to the speciation event of reference. While gene X and gene Y are out-paralogs in reference to S2, they are in-paralogs and co-orthologous to gene Z with reference to S1.*

Orthology conjecture

The orthology conjecture states that since paralogs have coexisted in the same species, thus having been under reduced selective pressure, one would expect paralogs to have more divergent function than orthologs. This is used as a basis for functional annotation transfer through orthology (Koonin 2005).

Nehrt et al. (Nehrt et al. 2011) set out to test the conjecture on a broader scale by examining the agreement of Gene Ontology (GO) (Ashburner et al. 2000) functional annotations as well as gene coexpression for orthologs and paralogs, respectively. Their results indicated that there was no basis for the conjecture. However, the study by Nehrt et al. has been criticized by Thomas et al. (Thomas et al. 2012) as a difference in annotation does not necessarily exclude a similarity of function, and also that experimental and curation bias leads to a greater similarity within species than between. Further, Altenhoff et al. (Altenhoff et al. 2012) performed an evaluation of the conjecture using GO functional annotations, compensating for potential bias due to differing expected similarity due to species and homology type as well as authorship- and propagated annotation bias. When these biases are accounted for, orthologs show a higher functional agreement than paralogs.

Examining the comparison using coexpression of Nehrt et al., there seems to be a similar bias as for the functional annotation comparison in that similarities are larger for within species than between comparisons. In fact, orthologs show a higher degree of coexpression than cross-species paralogs and therefore, these results of Nehrt et al. lend support for the orthology conjecture.

There are also other studies where orthologs have been shown to have a higher degree of domain conservation (Forslund et al. 2011), higher intron position conservation (Henricson et al. 2010), lower tissue expression divergence (Huerta-Cepas et al. 2011) and higher structural similarity (Peterson et al. 2009), all lending support to the orthology conjecture.

Orthology interference

Since we can only observe the current point of evolution, and not the events leading up to it, orthology cannot be exactly determined, only inferred. This is mainly carried out in two fashions, constructing evolutionary trees or through clustering based methods. It is also possible to combine the two main approaches and/or to add additional information in hybrid approaches.

Common to all approaches is the attemp to identify the genes most likely to have a common ancestor, something that can be difficult in the presence of

differential gene deletions, horizontal gene transfers as well as gene fusion/fission (Koonin 2005).

Tree based methods

Briefly, tree based methods (Kuzniar et al. 2008; Kristensen et al. 2011) generally start out from a multiple sequence alignment, using a tree reconstruction method to construct a tree, optionally guided by a species tree. Speciation and duplication events can be identified using the constructed tree and orthology/parology thereby be determined. Tree based methods carry the advantage of accurately depicting broader relations and the possibility for higher accuracy. The accuracy is dependent on the quality of the initial multiple sequence alignment as well as the tree reconstruction method. When a guide tree is used, the quality of the guide tree can be an issue, especially for lineages where it is uncertain. The main disadvantage of tree based methods is scalability, it is somewhat unfeasible to obtain high accuracy trees for large numbers of sequences without sacrificing accuracy.

Clustering based methods

Clustering based methods (Kuzniar et al. 2008; Kristensen et al. 2011) are much more scalable than tree based methods and are thus more suited for full genome comparisons of large numbers of species. Most current methods can be summarized in four broad steps; sequence comparison, initial clustering, filtering and merging (see Fig. 2).

Hybrid methods

Briefly, hybrid methods (Kuzniar et al. 2008; Kristensen et al. 2011) can involve the use of clustering and tree building as well as using additional information. This can be the case of e.g. clustering results from sequence comparisons guided by a species tree or using information about synteny (gene order).

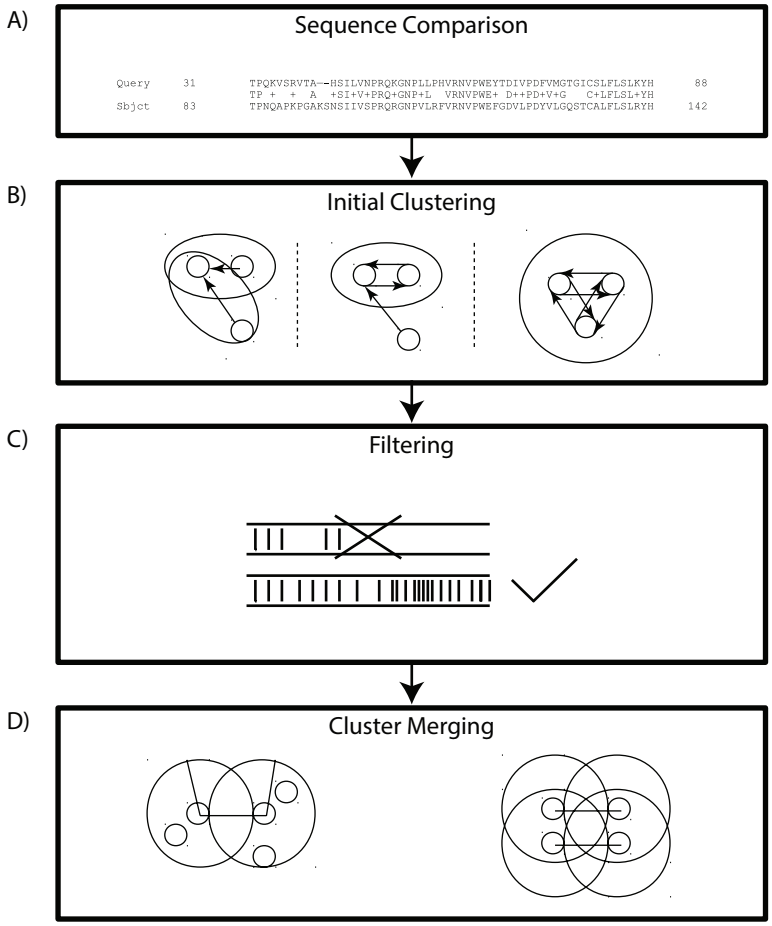

Fig 2: Clustering based orthology inference. *A) The ortholog inference start out using a sequence comparison, generally using BLAST or Smith-Waterman, comparing all sequences of two organisms with the comparison generally performed using protein sequences. For efficiency, most approaches do not compare all protein sequences but only the translated sequence for one selected transcript per gene, generally the longest. B) The scores or the p-values from the sequence comparison are then used for initial clustering, based on either the best one-directional hit, reciprocal (bi-directional) best hit (RBH) or triangles of reciprocal best hits. The reasoning behind this is that the more similar two sequences are, the more likely are they to be derived from one ancestor before the event separating them. C) The selection of initial clusters can then be filtered with removal of pairs failing to fulfil certain criteria based on e.g. score, alignment length or evidence of non-orthology found in other comparisons. D) A final step of merging clusters can be performed using e.g. single linkage- or Markov clustering. This can involve adding in-paralogs as well as resolving conflicting cluster assignments.*

Clustering based databases

The following sections describe the algorithms of a selection of clustering based databases, focusing on alternative options for different steps of orthology inference. A more complete listing of databases of all types can be found at: http://questfororthologs.org/orthology_databases

InParanoid

Inparanoid (Remm et al. 2001; Ostlund et al. 2010b) uses BLAST for sequence comparison, running two passes, one with masking of low complexity regions for homology detection and a second one to obtain non-truncated alignments. Initial clustering is done with reciprocal best hits, and in-paralogs are added if they have a stronger hit within their species than to the other species. Rules for merging, splitting or deleting groups are used to resolve conflicting assignments during the addition of in-paralogs and filters for score and proportion of sequence that is aligned are used for all sequence comparisons.

orthoMCL

OrthoMCL (Chen et al. 2006) uses BLAST for sequence comparisons and perform identification of putative orthologs and in-paralogs by reciprocal best hits, similarly to InParanoid, using a p-value cutoff to avoid spurious assignments.

For the main clustering step, a graph is constructed with proteins as nodes and log-transformed p-values of the BLAST results as edges. Edges normalized to correct for systematic differences between species and flow simulation through the MCL algorithm are then used to construct ortholog clusters.

OMA

OMA (Roth et al. 2008; Altenhoff et al. 2010) performs sequence comparison in two steps using optimized Smith-Waterman (Szalkowski et al. 2008), both steps include filtering on alignment score. The initial step is performed with a fixed PAM matrix and in the second step scores are refined by finding the PAM matrix that gives the highest score for the pair, thus obtaining an evolutionary distance. Further filtering is performed based on alignment length, and the cutoff is set as to reach a balance between minimizing triangle inequality violations and maximizing domain content agreement while retaining a sufficient number of orthologous relationships.

Clustering is performed by identifying stable pairs, pairs closer to each other than to any other sequence in the other species, using the evolutionary dis-

tances. A tolerance interval based on distance and variance is used to allow for the inclusion of in-paralogs, and it is optimized using the out-group genome closest to the divergence of the two species being compared.

A final filter is applied where clusters are examined for witness of non-orthology in order to cope with differential gene losses.

COG/KOG/arCOG

COG/KOG/arCOG (Tatusov et al. 2000, 2003; Makarova et al. 2007) uses BLAST for sequence comparison with masking of both low complexity regions as well as widespread, typically repetitive domains. Initial clustering is done constructing triangles of best reciprocal hits between 3 species and triangles with a common side are merged. Manual curation based on domain composition is performed to eliminate false positives and multidomain proteins causing artificial bridging of groups are split and the parts treated as separate genes.

eggNOG

EggNOG (Jensen et al. 2008; Muller et al. 2010; Powell et al. 2011) uses sequence comparisons from the SIMAP database (Rattei et al. 2010). It is based on COG/KOG/arCOG (Tatusov et al. 2003; Makarova et al. 2007) and if a gene has a best hit to a COG/KOG/arCOG group, it is assigned to that group.

Genes without a COG/KOG/arCOG hit are automatically assembled into clusters. In-paralogous groups are created by first assembling highly related genomes, such as different strains of the same species or closely related species such as human and chimp, into clades. Proteins more similar within the clade than outside of the clade are joined into in-paralogous groups, with the requirement of having at least 20 similar residues aligned.

Orthology is assigned by triangles of reciprocal best hits, using the best matching member for each in-paralogous group. Clusters with an abundance of reciprocal hits between groups are joined and proteins that bridge otherwise unrelated clusters are split and the parts assigned to the respective clusters. Finally, unassigned proteins are assigned to existing clusters by reciprocal best hits to a cluster member.

Roundup

Roundup (Deluca et al. 2006, 2012) compares sequences using the reciprocal smallest distance (RSD) algorithm (Wall et al. 2003) with some modifications. BLAST is used, for two species, with a significance cutoff followed by alignments with kalign (Lassmann and Sonnhammer 2005; Lassmann et al. 2009), for each query protein and its hits, and finally maximum likelihood

estimation of evolutionary distance with PAML (Yang 2007). Filtering is performed with 4 cutoffs for the BLAST p-values as well as 3 divergence cutoffs, resulting in 12 sets of ortholog inferences. The main clustering is performed by a deterministic single linkage clustering, initiating with genes as single clusters and merging clusters if one gene in one cluster is orthologous to one gene in another cluster.

Comparison of orthology inferences

As mentioned, it is impossible to determine true orthologs and as such, comparing orthology inferences is inherently difficult.

Comparisons have been made based on agreement with consensus inferences (Chen et al. 2007), correlation of expression profiles (Hulsen et al. 2006; Altenhoff and Dessimoz 2009), functional annotations (Hulsen et al. 2006; Altenhoff and Dessimoz 2009) as well as agreement with high quality phylogenetic trees (Altenhoff and Dessimoz 2009; Trachana et al. 2011; Boeckmann et al. 2011).

Results in general are not overly surprising, inferences made with stringent criteria, e.g. OMA, give a high specificity and low coverage while less stringent criteria, e.g. eggNOG, give the opposite. Somewhat surprisingly, clustering based methods often give as good results as tree based methods (Altenhoff and Dessimoz 2009).

The need for standardization

An issue that confounds these comparisons is that very often different databases use different proteome versions. Unfair benchmark results may be obtained when some databases more closely match the reference standard. Examining the comparisons to phylogenetic trees of Altenhoff et al. (Altenhoff and Dessimoz 2009), where other orthology databases have been mapped to OMA, it is quite easy to see that the variation of results for OMA, depending on what other database it is compared to, seems to be greater than the difference with the closest competitors. I.e. there is a non-negligible variation dependent on the set of orthologs used for comparison. The issue is so important that Boeckmann et al. (Boeckmann et al. 2011) notes that mapping genes is not possible to perform based on sequence identity, due to the large differences between the sequence data used in the different databases. Another issue is that input and output for orthology databases lacked reliable standards. The Fasta format was generally used for input which due to the openness of the format generally required specific parsing for each genome source and faced the problem that corruption of files was difficult to detect. Output orthology inferences were presented in different formats for different

databases. The orthology inference formats might also be difficult to read computationally and collecting inferences from multiple databases entailed substantial manual labor. These issues have been recognized by the orthology community (Gabaldon et al. 2009; Dessimoz et al. 2012) and suggested standard species (Reference Genome Group of the Gene Ontology Consortium 2009) as well as formats for both input sequences (seqXML) and orthology inferences (orthoXML) (Schmitt et al. 2011, http://seqxml.org) have been put forward. seqXML and orthoXML provides both standardized formats and also tools for simplifying computational use.

Gene/protein networks

Each human cell operates through a multitude of processes and pathways. Proteins, the active components, do not act in isolation. Rather, they affect each other through e.g. direct interaction through binding, regulating each other or being part of the same complex. Proteins could also be linked by more indirect functional associations, e.g. two proteins being part of the same pathway. These functional couplings can be described as a network, with proteins as nodes and the functional couplings as links.

For multi-factorial diseases such as cancer, there is not a single causative gene. Multiple different sets of changes can give rise to the same symptoms and as such, gaining an understanding about which proteins are functionally coupled, and under what conditions, would be of immense worth.

Studies to assay direct protein interactions face problems in that the conditions required to identify a true interaction differs between protein pairs and interaction types. Low-throughput experiments are costly and time consuming while high-throughput experiments suffer from low accuracy and are known to have many false positives.

Reliable predictions based on many weak evidences

While experimental evidence could be used to directly infer functional coupling, the available experimental data is limited and coverage would thus be low. A way to get around the problem of low accuracy of high-throughput experiments as well as a generally low coverage of functionally coupled protein pairs, especially low for high-quality low-throughput experiments, is to combine evidence of functional coupling from many sources. These sources include high-throughput experiments as well as other sources of evidence indicative of functional coupling.

Networks are constructed to describe protein functional coupling. However, due to limitations of data sources and complexity it is generally done on the level of genes.

Combining evidence can be done e.g. using a naïve Bayesian framework, where evidence sources are assumed to be independent. The naïve Bayesian framework revolves around training by comparing the probability of functionally coupled protein pairs and random protein pairs having a certain feature such as being highly coexpressed.

If a feature is more prevalent among functionally coupled protein pairs, the feature can be used as evidence of functional coupling. Conversely, if a feature is less prevalent among functionally coupled pairs, e.g. different subcellular co-localization, it can be used as evidence against functional coupling.

The methodology is thus heavily reliant on the known functionally coupled pairs used for training. Data used cannot by itself give information about

functional coupling. It is only through training with known functional couplings that the strength of association between properties of the data and functional coupling can be estimated. Any given feature from one data source could either give evidence for or against functional coupling, depending on the type of functional coupling, defined by the training set, being trained for.

Continuous variables are generally binned, and bin memberships used as properties, e.g. two genes having a low/moderate/high degree of coexpression. This is done for each evidence source separately. By combining the probabilities of all evidence sources one can then predict whether two given proteins are functionally coupled.

Other approaches extend the naïve Bayesian framework by trying to account for data interdependence or by applying other machine learning techniques such as support vector machines (Lin et al. 2009).

Evidence sources

Coexpression (Co-regulation)

Two genes having the same pattern of expression across tissues/conditions, having a high degree of coexpression, could be an indication of co-regulation. They are thus more likely to be involved in the same processes, and their proteins more likely to be functionally coupled. Conversely, if their expression patterns are entirely disparate, the opposite is true. Indeed, it has been shown that protein complex members tend to be coexpressed (Jansen et al. 2002), that genes with similar expression profiles are more likely to interact (Ge et al. 2001) and also that coexpression is conserved for functionally related genes (Bergmann et al. 2004). Using the wide availability of mRNA expression studies it is possible to examine coexpression over multiple tissues and conditions, generally using Pearson- or Spearman correlation as a measure of coexpression.

Protein interactions

Protein interactions can be assayed through various techniques including complementation and affinity purification (Panchenko and Przytycka 2008). Complementation involves linking proteins to a reporter sequence that has been split into complementary pieces. If two proteins interact the complementary pieces will come together and the reporter reconstructed. Affinity purification involves immobilizing a bait protein using antibodies or by fu-

sion with a tag, alternatively immobilization can be achieved by printing onto a slide. By exposing the bait to a protein solution and subsequent washing, only protein that directly, or indirectly, bind to the bait are retained and can be identified through mass spectrometry.

These techniques, depending on setup, will give information about one-to-one interactions or complex interactions, where it is sometimes uncertain what the exact interactions are. When the interactions are uncertain it is common to use either the spoke-, where all preys are assumed to interact with the bait, or matrix model, where all preys are also assumed to interact with each other. In reality the case could even be that only one of the identified proteins directly interacts with the bait and the others indirectly through that. Using the spoke model is correct in that all the preys are functionally coupled, even if not through direct interaction, to the bait but likely results in false negatives due to not accounting for the possibility of prey-prey coupling. The matrix model is also correct in that all the preys are functionally coupled, it is however less specific since many of the interactions might be indirect.

Many known interactions are collected in databases focusing on interactions (Kerrien et al. 2012; Keshava et al. 2009; Pagel et al. 2005; Salwinski et al. 2004; Licata et al. 2012; Chatr-Aryamontri et al. 2013), pathways (Kanehisa et al. 2012; Kanehisa and Goto 2000), combining both interactions and pathways (Alfarano et al. 2005), consolidating results of multiple databases (Razick et al. 2008; Kamburov et al. 2009; Cerami et al. 2011) or providing an overview of existing resources (Bader et al. 2006). The databases are not fully complete in regard to experimentally verified interactions and literature mining can be used to reach fuller coverage.

Phylogeny

Certain sets of functions and processes are specific to certain subsets of species. Consequently, two genes being present/absent in the same species, i.e. having a shared evolution, could be an indication of functional coupling. By comparing the evolution and diversification, phylogeny, of pairs of genes through phylogenetic profiling (Pellegrini et al. 1999; Marcotte 2000; Zheng et al. 2002; Glazko and Mushegian 2004; Date and Marcotte 2005), gene pairs with a similar phylogeny can be separated from those with a differing phylogeny.

Domain Content and Interactions

There are structural forms (domains) that are common to multiple proteins. These domains are often highly conserved and can be found in many different combinations, and the composition of domains is of importance for the function of the protein.

This can be exploited by either examining interacting domains (Björkholm and Sonnhammer 2009), domain-function associations (Hegyi and Gerstein 1999) or the degree of shared domains between proteins (Hegyi and Gerstein 2001; Forslund and Sonnhammer 2008).

Sub-cellular co-localization

Proteins can be localized in different sub-cellular compartments. If two proteins have no common localization the possibilities of direct interaction is low. On the other hand, if two proteins have a very similar pattern of localization, this could be indicative of them being more likely to interact.

Transcription factor binding (Co-regulation)

Transcription factor binding sites can be used to link a transcription factor to genes it regulates. They can also be used to determine associations between non-transcription factors by examining the extent of overlapping binding sites for pairs of genes.

Orthology transfer

By using orthology to transfer information across species, the available data can be massively increased. Orthology transfer can even be used to predict functional couplings for a species where there is no available data (Alexeyenko and Sonnhammer 2009). While simple conceptually, there are issues e.g. when orthology assignments are uncertain or when there are widely differing degrees of divergence.

There are also different options for when and how to use orthology. It could be used after training e.g. by transferring evidence from pairs to their ortholog groups and then from groups back to to the individual genes (Franceschini et al. 2012), i.e. transferring evidence of coupling. Alternatively it could be used to transfer data used in the training, e.g. examining the relative likelihood of functionally coupled genes in the species of interest having orthologs with a certain feature in some other species (Alexeyenko and Sonnhammer 2009).

Inferred networks of protein functional coupling

There have been multiple initiatives to predict protein functional coupling, the results of which can be found in online databases. They differ in how evidence is processed and combined, what evidence types are included and importantly, the definition of coupling as defined by the training data. Below and in Table 1 follows a brief description of a selection of such databases.

GeneMania

GeneMania (Mostafavi et al. 2008; Warde-Farley et al. 2010) focuses on speed and usability. While it contains evidence sources (see Table 1), pre-computed mainly using correlation, training is done on the fly using a list of genes given by the user. The genes in the list are assumed to be known interacting proteins or being related through a GO biological process. A composite network is constructed from the evidence sources by ridge regression. During the network construction redundant information is weighted down, and GO co-annotation patterns or alternatively connectivity between given genes is maximized, for short and long lists, respectively. The user can decide what data sources to include or even upload data sources of his own. For further details see Table 1.

Hefalmp

By constructing networks for 229 biological processes of interest for human cellular biology, Hefalmp (Huttenhower et al. 2009) enables the analysis of protein functional coupling with context awareness. Functionally related/unrelated pairs are selected from manually curated databases of protein interaction. This selection is based on marking pairs as functionally related if they are annotated as such in at least one of the databases, and marking them as unrelated if they are annotated to different processes in at least two of the databases, where that is possible. Process-specific networks as well as one context independent network are constructed with human data using regularized naïve Bayesian classification, differing from naïve Bayesian classification in that datasets that contain similar information are weighted down, in order to remove redundancy. For further details see Table 1.

String

String (von Mering et al. 2003; Franceschini et al. 2012) exploits the propensity, mainly of prokaryotes, to group genes with a similar function on the genome. This propensity can be evaluated for any species that has been sequenced and through this String manages to cover a substantial number of species. Additionally, full text literature mining with natural language processing is employed to retrieve interactions not present in data bases. String employs orthology to transfer evidence about coupling between species, this is done in a hierarchical two-step procedure where evidence of functional coupling is transferred to ortholog groups for increasing taxonomic levels, and subsequently transferred back to individual protein pairs. For further details see Table 1.

FunCoup

In order to capture different modes of functional coupling, FunCoup (Alexeyenko and Sonnhammer 2009; Alexeyenko et al. 2012) is trained on four different gold standards; metabolic pathway co-membership, signalling pathway co-membership, protein complex co-membership and physical interaction. In order to avoid potential biases from ad hoc binning of continuous variables, a dynamic binning is employed where bin borders are adjusted to separate known functionally coupled pairs from the background. FunCoup does not require the presence of large scale data for a species in order to reconstruct the network. As long as there are known functionally coupled pairs, orthology can be used to transfer data for use in training. For further details see Table 1.

FunctionalNet

YeastNet (Lee et al. 2007), WormNet (Lee et al. 2010b), MouseNet (Kim et al. 2008), AraNet (Lee et al. 2010a) and HumanNet (Lee et al. 2011) are five networks with separate databases with a common framework for network inference. Continous variables are initially binned to estimate probabilities of functional coupling. This is however followed by a regression allowing continuous mapping from variables to probabilities. Bootstrapping is applied to evaluate dataset accuracy and overall probability of functional coupling for each dataset taken as the average between the training- and the withheld test set. Integration of probabilities is done using optimized weighting with parameters for minimum probability thresholding and overall dependence among datasets. The optimized weighting is compared to naïve Bayesian integration and the most accurate chosen. Integration is first done within evidence sources for which there are multiple datasets and then to combine different evidence sources.

While the framework for constructing the networks is the same, there are some differences between the individual networks. The gold standard for HumanNet is restricted to annotations supported by experimental evidence. WormNet also uses functional annotations from KEGG (Kanehisa et al. 2012; Kanehisa and Goto 2000) as a gold standard. Evidence sources are not identical for all networks, e.g. gene fusions are only included in YeastNet which is also the only species where domain co-occurence is not included.

Database	*FunCoup*	*GeneMania*
Included Species	10 + 1 inferred	7
Context specificity	4 types of protein interaction	Network constructed for given list of genes
Gold Standard	4 separate sets, pathways or interactions	Based on input list of genes or GO
Network construction	Naïve Bayesian	Ridge regression + label propagation
Binning of Continuous data	Dynamic binning to maximize separation between known interactions and background	NA
Coexpression	X	X
Subcellular coloc	X	X
Phylogeny	Phylogenetic profiling	-
TF	X	-
PPI	X	X
Litterature mining	-	-
Domains	Domain interactions	Domain content
Orthology	Transfer of data before training	Indirect by incorporating other networks
Genetic Interactions	X	X
Genomic Neighborhood	-	-
Additional	miRNA targeting; Protein coexpression	Predicted Interactions

Table 1: Overview of selected functional coupling networks. The first column contains row headers with evidence sources in italic text. A "X" indicates that a database uses evidence of that type whereas a "-" indicates that it does not.

Hefalmp	*String*	*FunctionalNet*
1	1133	5
229 biological processes	-	-
Mix of physical interaction and pathways	Mix of pathways	GO biological process
Regularized Naïve Bayesian	Naïve Bayesian	Optimized weighting or Naïve Bayesian
Z-transformation followed by static binning	Normalization followed by static binning	Static binning followed by regression and continuous scoring
X	Conserved Coexpression	X
-	-	-
-	Mutual information	Co-inheritance
X	-	-
X	X	X
X	X	X
Domain content	-	Domain co-occurence
-	Transfer of evidence after training	Transfer of data before training
-	-	X
-	X	X
Sequence similarity	Gene Fusions	Gene Fusions (YeastNet)

Functional coupling networks and disease

One promising prospect for inferred functional coupling networks is to use them to characterize disease genes as well as to find novel disease genes. This is sometimes done using available inferred functional coupling networks, but is often done by a small scale integration of protein interaction data possibly complemented by a few other evidence sources.

Taylor et al. (Taylor et al. 2009) examined coexpression of hubs with their neighbours, in a network constructed by integrating high- and low-throughput interaction data, and found that changes in coexpression are associated with poor outcome in breast cancer. The findings of Taylor et al. is in line with prior work by Jonsson and Bates (Jonsson and Bates 2006), who examined the connectivity of cancer genes in a network constructed by homology transfer of protein interactions, and found that cancer genes tend to be central hubs. The tendency of disease genes to be hubs does however not seem to be replicated when the network is constructed using only gene-disease associations for a wide range of diseases (Goh et al. 2007).

Protein functional coupling networks have been applied to rank genes in a genomic region linked to the disease, as well as to rank all human genes. This has been carried out by ranking genes using simple network descriptors as well as more complex methods. Examples of simple network descriptors include ranking genes according to the number of linked known disease genes (Ostlund et al. 2010a), shortest path length to a disease gene (George et al. 2006), sum of phenotypic similarities to the disease of neighbours (Lage et al. 2007), correlation of phenotypic similarity with distance in the network (Wu et al. 2008).

Complex approaches includes random walks and diffusion models as well as other methods modelling information propagation through the network (Lee et al. 2010b; Vanunu et al. 2010; Chen et al. 2011).

Validation of the soundness of the approach has generally been carried out through a combination of cross-validation, functional analysis or expression analysis. Although somewhat rare, there are also examples where experimental validation has shown the soundness of using protein networks to guide prioritizing novel disease genes. These examples include siRNA knockdown of candidates resulting in reduced cell viability of a colorectal cell line (Li et al. 2009), using network-based inference to identify new targets for approved drugs (Cheng et al. 2012), predicting yeast sporulation (Shen et al. 2010) and recovering false negatives from RNAi screening (Wang et al. 2009). Also see Wang and Marcotte (Wang and Marcotte 2010) for further examples and a review of the field.

Differential coexpression and network analysis

Multiple studies have examined differences in coexpression between networks for healthy and diseased cohorts, somewhat similar to the work by Taylor et al. (Taylor et al. 2009), but with functional couplings defined purely by coexpression rather than experimentally determined protein interactions. Network construction has been done selecting a set fraction of links with a high coexpression (Choi et al. 2005), applying a threshold for coexpression (Reverter et al. 2006) or keeping links with a coexpression above the threshold in at least one (Yu et al. 2011) network. Alternatively, all links are kept with weighting proportional to coexpression (Fuller et al. 2007). Analysis has been done collapsing the network down to GO categories (Choi et al. 2005) or genes (Yu et al. 2011; Reverter et al. 2006; Fuller et al. 2007) and analyzing differences in connectivity, defined by counting the number of links or as a function of the weights of the links, between the two networks. Comparisons to random permutations of class labels or statistical models are used to identify genes/modules with significant changes and literature study or functional enrichment analysis used to show relevance to the diseases.

Expression analysis

The abundance of mRNAs and proteins shows a fingerprint of the current state of the cell, what genes and processes are active. Taking it a step further, comparing multiple such fingerprints can shed light on differences between healthy and diseased individuals. Two genes having similar patterns of expression could be due to co-regulation and the genes thus more likely to be involved in the same processes.

While proteins generally are the entities of interest, measuring protein abundance is made difficult by proteins being folded into very stable structures and it being difficult to construct a complementary structure. mRNA by comparison is easier to measure, due to higher order structure being less difficult to determine or dissolve and the ease of constructing complementary sequence.

Consequently, mRNA abundance are often used as a proxy for protein abundance, as well as giving information about cell state and regulation by itself. There is not a 1-to-1 relationship between mRNA and protein, other factors such as translational efficiency, mRNA and protein decay, translocation and sequestering come into play. Therefore, using mRNA as a proxy might not be feasible for all genes for all applications.

Measuring mRNA abundance

Measurements of mRNA abundance have seen immense technical improvements in the last couple of decades. Starting with northern blot, introduced by Alwine et al. in 1977 (Alwine et al. 1977), and stepping on to methods based on qPCR (Abbott et al. 1988; Syvänen et al. 1988) in the late 80s (VanGuilder et al. 2008). Contemporary qPCR works through reverse transcription of mRNA coupled with simultaneous amplification and quantification. While it still remains the "gold standard" for accurate mRNA measurements the scope is limited to a few hundred genes. The scope was greatly increased when microarrays, and more recently direct sequencing, enabled efficient measurement of the entire transcriptome.

Oligo based / Microarray

The initial microarrays were made by printing oligo probes, short stretches of complementary DNA, on glass slides. Fluorescent labelled cDNA, reverse transcribed from mRNA, could then be hybridized to the microarry and multiple genes measured in parallel (Schena et al. 1995). This has been improved upon in modern commercial platforms where probes density can be massively increased e.g. by synthesizing directly on the chip (Affymetrix, Agilent, Roche Nimblegen) or on small beads (Illumina).

A technical limitation of microarrays is that probes must be designed in advance and that there can be issues of cross-reactivity. Moreover, the measurements are generally in the form of light intensities and need to be processed into abundance. Lastly, there might be systematic differences be-

tween samples and additional normalization might be prudent. As the expression analysis in this thesis has been made with data from Affymetrix arrays the following sub sections will mainly focus on processing data of that type.

Data preprocessing

In order to obtain accurate measurements of mRNA abundance using microarrays, corrections need to be made for noise added in the measurement itself, as well as systematic differences within each array, due to e.g. different base compositions of probes. Moreover, it is crucial to discern between specific and non-specific hybridization. Finally, multiple probes forming a probe set must be summarized into an expression value for the sequence they correspond to. There are multiple options for each of these steps (see Zhu et al. 2010) and implementations of common approaches such as RMA (Irizarry et al. 2003), GCRMA (Wu et al. 2004) and MAS5 (Hubbell et al. 2002) generally also include a normalization step.

Normalization

Since there can be systematic differences between arrays, e.g. due to different amounts of mRNA from different samples, arrays are generally normalized to alleviate this problem. The normalization can take place either before, or after, probe summarization and carried out using different approaches such as scaling arrays by median centring, quantile normalization (Bolstad et al. 2003), loess (Bolstad et al. 2003) or variance stabilizing normalization (Huber et al. 2002)).

The resulting data after normalization is then composed of relative abundance rather than absolute. This can be a problem, e.g. if examining regulation. In such cases normalizing using carefully controlled spike-ins, adding known quantities of unique mRNAs to the sample, would be preferable (Lovén et al. 2012).

Sequencing based / RNA-seq

Advances in sequencing technology (Ronaghi 1998; Margulies et al. 2005) has enabled measuring mRNA abundance directly through sequencing (Bainbridge et al. 2006; Nagalakshmi et al. 2008; Mortazavi et al. 2008; Bentley et al. 2008; Cloonan et al. 2008).

While RNA-seq generally has comparable accuracy to microarrays at the same cost, it is not dependent on selecting what to measure in advance and doesn't face problems with cross hybridization (Malone and Oliver 2011). Also, a higher resolution for low abundance transcripts can be achieved

through increasing sequencing depth and high resolution measurements have been used e.g. to examine transcription at a level lower than one transcript per cell (Hebenstreit et al. 2011). The technology is new though, and issues such as transcript length bias (Oshlack and Wakefield 2009) exists.

Measuring protein abundance

Proteins are folded into complex 3D structures that are very stable and higher order structure for protein is considerably more difficult to predict than structure for mRNA. Moreover, the accessible protein surface generally has quite specific binding properties due to a higher complexity in the form of higher number of amino acids than nucleotides and the possibility of post-translational modifications. While mRNA measurements can be carried out utilizing complementary sequence, a similar approach is not feasible for proteins. Consequently, it is comparably more difficult to measure protein abundance than mRNA abundance, and is generally carried out with either Mass Spectrometry (MS) or antibody-based methods.

Mass Spectrometry based

Mass spectrometry is based on ionizing peptides, or full length proteins, and subsequent detection resulting in mass spectra of relative abundance versus mass-to-charge ratios (m/z) (Yates 1998; May et al. 2011). In Mass Spectrometry based methods proteins are first separated, followed by ionization and detection. Separation can be carried out using electrophoresis in gels (Klose 1975; O'Farrell 1975; Bjellqvist et al. 1982; Görg et al. 1988) or HPLC (Mitulovic and Mechtler 2006) where proteins, or peptides, are separated based on e.g. size, charge or affinity to an immobilized ligand. While it is possible to identify full length proteins, digesting proteins into peptides and analyzing those is more efficient (Pappin et al. 1993). Peptide masses can then be measured, through e.g. electrospray ionization (Fenn et al. 1989) coupled with ion trap (Busch and Paul 1961) detection or laser adsorption ionization (Zaluzec et al. 1995) and time of flight detection (Wollnik 1993). Fingerprints of resulting peptide masses can then be used to search a database of protein sequences and their fingerprints for identification

In order to measure abundance, isotope labelling (Gygi et al. 1999; Schmidt et al. 2005; Ross 2004; Ong et al. 2002) or spectral counting (Carvalho et al. 2008) can be used to obtain relative abundance and spike-ins can be used to obtain absolute abundance (Gerber et al. 2003).

Antibody based

Foreign proteins entering the body can subsequent to fragmentation be presented as antigens. This will cause the activation of B-cells which will then

produce antibodies targeted at a specific sequence (epitope) of the antigen. Antibodies are roughly Y-shaped proteins with a constant part as well as a highly variable part, capable of specifically binding one epitope. Consequently, multiple antibodies targeting different antigens, corresponding to different fragments of the foreign protein, will be produced (Goldsby et al. 2002).

This mechanism can be exploited to produce antibodies targeting a protein of interest by introducing the protein into an animal (inoculation) and harvesting the animal for antibodies. These antibodies are then polyclonal, directed at different parts of the protein. It is also possible to generate monoclonal antibodies, targeted at one specific epitope, by isolating a B-cell from a inoculated animal and fusing it with a cancer cell into a hybridoma, in order to immortalize the B-cell (Köhler and Milstein 1975). Since monoclonal antibodies are uniform, assays using these will be more reproducible while assays using polyclonal antibodies can vary more, due to targeting multiple epitopes. Monoclonal antibodies, while more reliable, require more time and effort to generate.

Protein abundance measurements can be accomplished e.g. by having a primary antibody targeted at a protein of interest and a secondary antibody targeted at the primary antibody (Bock 2000). The secondary antibody is linked to a fluorophore or an enzyme resulting in staining and abundance can be measured by measuring fluorescence or evaluating the amount of staining.

HPA

The Human Protein Atlas (HPA) is an ambitious project aimed at generating antibodies to all human proteins (Uhlen et al. 2010). By selecting a stretch of sequence as unique to a protein as possible (PrEST) (Lindskog et al. 2005) and using the PrEST both for antibody generation as well as affinity purification of generated antibodies, polyclonal antibodies specific to the PrEST can be generated (Agaton et al. 2004).

As a part of the project, antibodies are used to stain human normal as well as cancer tissues and stained tissues are manually curated for abundance, thus composing a library of protein expression in humans (Berglund et al. 2008; Uhlén et al. 2012).

Differential expression and coexpression

Examining the differences in expression and coexpression between healthy and diseased states can give information about the disease. Something that is substantially up- or down regulated could be central to the mechanisms and/or a potentially good target for drugs. Similarly, a loss or gain of coexpression could indicate regulatory changes or changed potential for interac-

tion of two proteins. Also, such differences could potentially be used, where conventional methods fail, to diagnose patients at an early stage or to discriminate between sub-types of the disease.

Patient/control studies

In order to examine the differences between healthy and diseased, biologically relevant samples are needed. While cell lines can be representative in some cases (Kasperkovitz et al. 2005), analysis is generally carried out using patient/control studies. Samples will commonly be derived from from biopsies and as such the composition of cell types might differ. This will introduce uncertainty as expression patterns might be different for different cell types (Holloway et al. 2006), this can however be controlled by isolating individual cells through e.g. laser capture microdissection (Bonner et al. 1997).

As obtaining biopsy samples from healthy humans can be quite difficult, control samples are often derived from patients undergoing biopsy for some other disease. Samples can in many cases be said to be based on a convenience sample of patients (Simon 2005) and demographic differences are generally not controlled, sometimes not even annotated. Initiatives such as MIAME (Brazma et al. 2001) aimed at improving the standard of expression study annotations work towards reducing this problem.

Differential expression

The aim of differential expression analysis is to determine what genes have the largest changes to expression. However, this is not trivial to define and there are multiple different metrics of measuring differential expression. Fold change is calculated by dividing the average expression of one state with that of the other In order to consider up- and down regulation equally fold change is often transformed by taking the absolute values of logarithms. Fold change carries the draw-back that when expression is highly varying in both states, a statistically insignificant difference can still receive a high fold change (Allison et al. 2006). A simple way to avoid this is to perform a t-test and rank genes according to the resulting p-values, this however carries the opposite problem to fold change in that genes with a biologically insignificant change in expression can still receive low p-values. In order to avoid these issues, ranking of fold change is often coupled with applying a t-test based p-value cutoff (e.g. in Guo et al. 2006), or vice versa. These issues have also prompted new metrics based on evaluating expression change relative to expression variation with comparison to empirical distributions generated from the data (Tusher et al. 2001) and empirical bayes approaches (Smyth 2004; Baldi and Long 2001).

Differential coexpression

Differential coexpression examines the agreement of expression profiles for pairs of genes and has been shown to enable identifying causative genes missed by only examining differential expression (Hudson et al. 2009). There is comparably much less work done using differential coexpression than differential expression, likely due to relatively higher requirements for amount of samples. Differential coexpression is commonly defined as the difference in Pearson- or Spearman correlation although there are some cases where other definitions are used, such as an extension of the F-statistic (Lai et al. 2004) or using Bayesian approaches (Freudenberg et al. 2010; Dawson and Kendziorski 2011).

Present investigations

InParanoid (Papers I and II)

InParanoid is a widely used and well performing (Hulsen et al. 2006; Chen et al. 2007; Altenhoff and Dessimoz 2009) source for orthology predictions. Papers I and II present our continued work to provide this service with an increased number of species included and improvements to the quality of orthology inferences.

The recent advances in sequencing technology have resulted in a massive increase in the number of species sequenced. Throughout papers I and II we have introduced steps taken in order to cope both with the increase in the computations required and to improve the quality of the inferences. We examine the consistency of genome annotations. If genome annotation is uncertain and changing between versions this will affect inferences. Providing this information enables the users to determine how stable inferences are. Also, new species are only included if they are sufficiently different from already included species.

The algorithm was adjusted by introducing a two pass approach, which enables the handling low complexity sequences while not losing inferences due to truncated alignments. We have also performed background work to streamline the pipeline and improve validation, to better cope with increasing demands due to larger amounts of data.

Inparanoid 7 saw the move to standardized formats for input and output; seqXML and orthoXML (Schmitt et al. 2011). Also, we now provide inferences for standardized input data (Reference Genome Group of the Gene Ontology Consortium 2009) in order to facilitate comparison between orthology inference databases.

Reflections

While the increased availability of sequenced species is a good thing, it also presents a problem in that it increases computational demands, the computations needed scales squared with the number of species.

There are a number of potential avenues towards dealing with this increased demand. Examining genome consistencies could potentially be used to re-

strict sequence comparisons, not unnecessarily comparing sequences that are unchanged between genome versions. One promising approach is to conduct inferences in a hierarchical manner rather than pairwise. This has been implemented in hieranoid (Schreiber and Sonnhammer 2013) which has been shown to provide accurate inferences as well as a vastly reduced compute time.

The orthology conjecture

The validity of the orthology conjecture, if orthologs are functionally more similar than (out-)paralogs, is of great importance for the use of orthology inferences for functional transfer. It has been heavily debated and some important issues for how it should be evaluated, such as annotation bias (Altenhoff et al. 2012), has been brought into light. Current attempts to evaluate the conjecture do not fully account for dependence on sequence similariy, the impact of duplications and correct definition of out-paralogs. Consequently, the validity of the conjecture is still an open question.

The logical basis for the conjecture is that a duplication would lead to a reduced selective pressure and thereby higher likelihood of functional differentiation. Orthologs separated by a speciation would instead be under higher selective preassure to retain function. While this certainly is reasonable when considering only 1-to-1 orthologs with a single un-duplicated out-paralog in each species, that is not always the case. In the presence of one-to-many or many-to-many orthology relations, functional equivalence is however non-trivial to determine (Remm et al. 2001). Using the logic the conjecture is based upon, the degree of duplication would be of importance. Functional similarity could very well be higher for an out-paralog pair without additional duplications compared to an ortholog pair with a high number of duplications. Therefore, evaluation of the conjecture should preferably take differing degrees of duplications into account.

The issue of taking duplications into account ties closely into another issue, how out-paralogs are defined. Including distant out-paralogs would introduce bias towards lower functional similarity whereas only considering one very close out-paralogs, even if there are other out-paralogs decended from the duplication separating them from the ortholog pair, will bias towards higher functional similarity.

There is a known dependence between functional similarity and sequence similarity. In the evaluations of the orthology conjecture by Nehrt et al. (Nehrt et al. 2011) as well as by Altenhoff et al. (Altenhoff et al. 2012) this is compensated for by binning sequence pairs according to similarity. Using orthology inferences based upon sequence comparisons, for every pair of between-species outparalogs there will exist a more similar ortholog pair. As such, the distributions will be disparate, with a shift to higher similarities for

ortholog pairs. Binning will thus at most reduce, but not fully remove, sequence similarity bias. Comparing within species out-paralogs and in-paralogs carries the same problem, for each in-paralog pair there exists less similar within species out-paralog pairs.

Comparing different ortholog inferences

Current comparisons have been flawed in that they have been based on different genome versions. This problem has been so profound as to necessitate mapping genes through identifiers rather than sequence (Boeckmann et al. 2011), adding uncertainty due to it not being known to what extent differing results are due to differing genome versions. While using standardized genome versions (Reference Genome Group of the Gene Ontology Consortium 2009) should resolve this issue, there is a tradeoff between sensitivity, capturing true orthology relations, and specificity, not predicting non-orthologs as being orthologous (Hulsen et al. 2006; Chen et al. 2007). If a comparison is done at different sensitivities for different databases, it is uncertain if differences in specificities reflect true differences or if they are an artifact of differing coverage.

It is not surprising that a methodology with very stringent criteria, such as OMA, would reach a high specificity while having a comparably low sensitivity. Similarly one would expect less stringent criteria as in eggNOG resulting in a high coverage with a low specificity. Many of the orthology databases report confidence criteria for the inferences and sensitivity/specificity can be varied even when using default parameters. A database with high sensitivity but relatively low specificity can have a higher specificity when restricted to higher confidence inferences. This is evident in the orthology benchmark results (Dessimoz et al. 2012; Ortholog Benchmarking Webservice) for InParanoid and the higher confidence InParanoidCore. InParanoidCore has lower sensitivity than InParanoid but higher specificity.

Additionally, what methodology performs best might be dependent on the species used for comparison, and different methodologies might be better at different evolutionary distances.

Parameters of different steps in the orthology inference can be tuned for each methodology, which is quite evident in the twelve sets of inferences for Roundup. It would thus be possible to compare the current methodologies not only using default parameters, but tuning parameters to vary sensitivity.

Existing databases certainly do not cover all possible methodologies, e.g. it would certainly be possible to run InParanoid using Smith-Waterman sequence comparisons rather than BLAST. Comparing inferences for a larger set of algorithm components than exists in current methodologies, it might be possible to reach an understanding of what impact the possible choices at

each step of the inference has. This could help improving inferences as a whole.

Maxlink (Paper III)

Complex diseases without a single causative gene, such as cancer, might require a systems wide approach to fully understand the possible causes and mechanisms. Paper III presents Maxlink, an approach for using inferred gene/protein networks to facilitate prioritizing candidate disease genes.

Prior work (George et al. 2006; Lage et al. 2007; Wu et al. 2008) has mainly focused on ranking candidates in a linked genomic region, starting from candidates and ranking them based on their links being prevalent for disease genes. We wanted to examine if it was viable to do this from the opposite direction, starting from the disease genes, considering all human genes and focusing on truly novel candidates.

The Maxlink approach revolves around taking an inferred network and checking all genes for links to a given set of disease genes. These initial candidates are then filtered for genes with spuriously high numbers of links as well as for genes with annotations suggesting that they might already be associated with the disease.

The approach was applied to cancer, using cancer genes from the Cancer Gene Census (Futreal et al. 2004) and UniProt (The UniProt Consortium 2012) with the FunCoup inferred network (Alexeyenko and Sonnhammer 2009; Alexeyenko et al. 2012). This resulted in a list of close to 2000 novel candidates ranked by the number of functionally coupled cancer genes. The approach was validated through cross-validation, differential expression analysis as well as functional enrichment analysis and showed increased cancer association for highly ranked genes.

Work on Maxlink subsequent to the publication has resulted in a webpage, maxlink.sbc.su.se, providing a user interface to an improved version of the algorithm. Improvements include optimization of the algorithm as well as replacing the connectivity filter with a hypergeometric test.

The Cancer Gene Census has grown beyond the genes used for the initial set of candidates. This makes it possible to add an additional validation by examining what initially proposed candidates has been shown to in fact be cancer genes. Out of 102 new cancer genes, 20 are to be found among the maxlink candidates. This is a significant 2-fold enrichment and is increased to a close to 4-fold enrichment for more highly ranked genes.

Reflections

There has been much activity in the field with a multitude of related publications (see Tiffin et al. 2009; Baudot et al. 2009; Linghu et al. 2009; Wang and Marcotte 2010; Chan et al. 2012). These publications vary both in the base network used, what disease(s) is examined, the definition of disease genes as well as the algorithms used for prioritizing candidates. Surprisingly, despite using inferred networks having been shown to be superior to using only protein interaction data (Gonçalves et al. 2012), most studies are based on simpler networks, being constructed using one or a few data sources.

Independent comparisons (Wang and Marcotte 2010; Navlakha and Kingsford 2010) have shown more complex methodologies, such as diffusion methods, outperforming using simple network properties, such as neighbor counting.

Many publications include a comparison with prior methodologies, generally showing the proposed methodology performing better than the prior ones. Vanunu et al. (Vanunu et al. 2010) compared PRINCE with CIPHER and showed PRINCE to perform better, recovering the relevant disease gene in leave one out cross-validation for 34% of the cases compared to 24.7% for CIPHER. However, CIPHER recovered the relevant disease gene for 49% of the cases when the equivalent validation was performed in the original CIPHER publication (Wu et al. 2008). This shows that the choice of network has a profound impact on the results.

Additionally what algorithm is most suitable would reasonably vary between diseases due to differing pathogenesis. For example, if a disease is related to a certain complex, where direct interactions are of utmost importance, compared to a disease related to signaling or regulation, where interactions are more indirect.

It might not be optimal to consider links to all types of cancer genes as this might result in selecting genes too vital and central to be of use. It is motivated for a proof of concept, where having sufficient coverage for sound statistical evaluation is paramount. However, a more practical approach would be to focus more narrowly, e.g. at a single cancer (sub)type.

The annotation filter used for Maxlink removes all genes with annotations that might suggest relation to cancer. While this facilitates ensuring that the final candidates are indeed novel, it would have been prudent to evaluate the filtered genes for cancer association in the validation steps.

Are all cancer genes equally cancerous? The amount of support for cancer association can vary widely among known cancer genes. While some might have strong support, where animal models show that a mutation directly leads to cancer, others do not have as clear association. Taking care that the

set of disease genes only contains genes with strong associations to the disease, or handling genes with weaker association differently, would likely help improve priorization.

While inferring networks through integrating diverse data sources yields a good coverage of interacting pairs, they are generally of a more general context. Using the abundantly available mRNA expression data from patient/control studies could capture the specific changes related to a disease, but constructing the network directly from that data wouldn't produce a network as reliable as the inferred ones. Combining the two could potentially better identify impacted genes and functional couplings to better prioritize candidate disease genes. However, before that is possible, the two hurdles presented in papers IV and V need to be overcome.

Concordance between mRNA and protein expression (Paper IV)

While proteins are the active components, mRNA is comparably easier to measure, and is often used as a proxy. However, protein abundance is not directly proportional to mRNA abundance and the impact of translational efficacy, degradation and other factors is not fully understood (De Sousa Abreu et al. 2009).

mRNA-protein concordance could vary between genes, with some genes having protein abundance more directly regulated by mRNA abundance. In paper IV we examined if it was possible to identify subsets of genes with a higher concordance. While this wouldn't solve the problem of mRNA-protein discordance, if successful, it would help minimizing it.

Where previous work done examining mRNA-protein concordance had generally been carried out on a single celltype, we utilized protein abundance data from the Human Protein Atlas. This data had abundance for multiple tissues, and thus enabled examining the concordance for separate genes, rather than tissues, as well as enabling examining the concordance of coexpression.

Quality descriptors for mRNA data was constructed based on the reasoning that profiles with distinct expression in different tissues are more likely to have protein abundance more directly controlled by mRNA abundance, and be less sensitive to noise.

We first successfully determined that such a descriptor did indeed separate genes/gene pairs with a higher concordance from those with lower. This was especially true when comparing concordance of mRNA and protein for the same gene with randomly pairing mRNA and protein for different genes.

Subsequently we utilized FunCoup to evaluate coexpression only for pairs with evidence of functional coupling, reasoning that this would reduce noise from spuriously high coexpression. Perhaps not surprisingly, this resulted in a substantially higher mRNA-protein concordance for coexpression.

Reflections

Other studies of mRNA-protein concordance have shown a large variation in the results (De Sousa Abreu et al. 2009). This might in part be due to differing conditions and differing sets of genes. It would be very interesting, given additional large-scale protein data sets, to examine if genewise concordance is more stable.

It would presently be possible to examine the uncertainty of the degree of concordance. Using the multiple studies performed in yeast or human the concordance for the set of genes common to multiple studies could be examined. Calculating the concordance for this set of genes using data from each study separately would give multiple measures of concordance and the variation of concordance across studies would give an indication of the uncertainty, at least for that set of genes.

Several recent studies complement mRNA measurements with additional factors to better predict protein abundance. These factors include sequence features associated with translation and degradation (Vogel et al. 2010) or directly measuring mRNA and protein turnover (Schwanhausser et al. 2011). When adding additional variables one must however be careful not to overtrain the model, e.g. the model of Schwanhäusser et al. suffers a large drop in predictive power when applied to a technical replicate, despite low variation between the replicates.

Consistency between gene expression studies (Paper V)

Gene expression is dynamically adapted as the cell responds to its surroundings and to incoming signals. On a higher level, tissues are composed of many different cell types and the specific composition obtained in a biopsy can affect expression measurements. Different individuals suffering from the same disease can differ in both the cause and the mechanism of the disease as well as demographics. In the face of such potential diversity, one might ask how consistent results from one study might be with those from another. Perhaps not surprising, studies examining the agreement of differentially expressed genes have reported generally low overlaps.

Measuring gene expression with microarray requires processing measured light intensities into mRNA abundance and there are numerous choices to be

made for different steps during data processing. The choices made for these steps as well as how differential expression is defined have impact on the results. While several studies have examined optimal choice of processing parameters and differential expression definition, the optimal combination can vary between different data sets.

Consistency across studies reflecting the same clinical condition has previously not been evaluated taking method selection into account. Paper V fills this gap and extends previous work to a systematic study of in vivo data and examines consistency of results taking parameter choices into account. This was performed not only for expression but also for coexpression, for which neither consistency nor impact of parameter choices had previously been evaluated.

In addition to this we also try to answer a somewhat controversial question: is differential expression or differential coexpression from a single study informative about the disease or specific to the data used? In the same manner that very distinct symptoms can be common to multiple diseases, large (co)expression changes might also be shared. In order to test this we compared studies from different cancer types with each other as well as with studies of non-cancer diseases. We examined the agreement of differential (co)expression and evaluated if it is possible to draw conclusions about a specific disease from a single study.

The results of the analysis highlights pitfalls of gene (co)expression analysis. Some of the results are not very surprising, such as that coexpression analysis is unreliable with small sample sizes, while others are quite noteworty. There is no agreement for differential expression between studies using different generations of the same microarray platform, begging the question how reliable any of them are. While there is agreement of results when the same definition of differential expression is used for both studies being compared, using different definitions can result in agreement being lost.

Agreement of differential (co)expression is overall very significant, although low in absolute levels for coexpression. This is true even in cases where one would not expect so, such as when comparing a non-cancer disease study with a cancer study. This suggests that patterns of differential (co)expression is best viewed as being symptomatic of a disease, rather than being specific to it.

Reflections

A confounding factor when dealing with coexpression is that one wouldn't expect it to be informative except for genes that truly are functionally coupled. As such, the coexpression analysis could be complemented by repeat-

ing it and only considering pairs with support for functional coupling in the same manner as was done in paper IV.

When examining the annotation for the studies considered for use in paper V, it became apparent that the definition of what a normal tissue is varies a lot between studies. It is in fact somewhat rare that what is labelled as normal tissue is in fact derived from healthy individuals.

Histologically normal tissue: This is a sample taken adjacent to a tumor, where the normal sample visually seems normal. As that normal tissue very well could have cancer-generating mutations, and altered expression, one would perhaps not expect it to accurately reflect tissue from healthy individuals. However, it does represent tissue that is still functioning mostly normally and genes/pairs identified here might actually be the most interesting ones, seeing as they might represent the final barrier towards pushing the cell into a fully cancerous state.

Tissue from non-cancer patients: Samples are often derived from routine screening for disease or surgery unrelated to disease. In some cases the individuals the samples are drawn from are clearly not healthy, e.g. if someone is so sick that they need a transplant, it would not be unexpected that there are large regulatory differences with tissue from healthy individuals. Same as for histologically normal tissue, such inadequacy might actually be a boon. As shown in paper V, there is often a significantly high agreement between cancer studies and non-cancer studies. Such studies could be suitable to identify differences that are not generally widespread changes simply due to a deviation from a normal state.

Although agreement of differential coexpression is substantially larger than expected by chance, it is nevertheless quite low in absolute levels. This indicates that constructing a network based solely upon coexpression from one study would be treacherous and that meta-analysis would be preferable. At the same time the low agreement is promising for network inference since it suggests that assumptions about independence, used e.g. for naïve bayesian inference, to be sound.

Future Perspectives

The debate around the orthology conjecture highlights a crucial issue for using orthology to transfer information between species: We do not know for certain how reliable orthology based information transfer is. While it reasonably is the best sequence based alternative, at least for 1-to-1 orthologs, the uncertainty is not known.

Current orthology inference algorithms often have parameters set in a more or less ad hoc manner and also do not cover all possible methodologies. For example, it would certainly be possible to run InParanoid using Smith-Waterman sequence comparisons rather than BLAST.

The advances in standardization and benchmarking of orthology inferences could be the solution to resolving the uncertainties of information transfer as well as determining suitable parameters and algorithm components. Extending benchmarking to not only compare databases but also to evaluate the uncertainty of information transfer it would be possible to estimate the effect e.g. post-speciation duplications and evolutionary distance has on the certainty of information transfer. Examining the certainty of information transfer could also be used to improve orthology inferences. Comparing inferences for a larger set of algorithm components than exists in current methodologies, it might be possible to reach an understanding of what impact the possible choices at each step of the inference has and thereby help improving inferences as a whole. Improved inferences and better understanding of uncertainties of orthology transfer would lead to improvements of inferred networks as well as other areas where there is need for transferring information between species.

The results of paper IV show that it is possible to identify subsets of genes for which concordance between mRNA and protein is higher. There are many applications where mRNA abundance are used as proxy for protein abundance and using information about concordance could help obtain more reliable results. E.g. accounting for varying concordance could be applied to the inference of functional coupling networks by separating genes into subsets based on concordance and performing training for coexpression data separately for the subsets. The work in paper IV could be directly extended by calculating descriptors, not only from a single dataset, but through meta-analysis of multiple datasets.

Paper V shows that it would be advantageous to study differential (co)expression across a wide variety of diseases to find common patterns. This could be used to examine what changes diseases have in common and what changes are specific to subsets of diseases or single diseases. Identifying genes and pairs specific to a single disease would enable constructing more specific profiles as well as drawing more reliable conclusions about the mechanisms of the disease.

The conclusions drawn from papers IV and V could be used to augment the Maxlink approach of paper III by utilizing differential (co)expression to gain disease specific context. Including genes with high mRNA-protein concordance that are differentially expressed specifically in the disease examined and weighing links that are differentially coexpressed higher could improve detail and provide better disease gene prioritization.

It would also be possible to adopt a similar approach on a larger scale. The network of functional coupling could first be collapsed down to functional modules, such as pathways. Modules could then be examined for enrichment of disease specific differential expression as well as intra- or inter module enrichment of disease specific differential coexpression in order to facilitate gaining information about the mechanisms of the disease.

Sammanfattning på svenska

Det centrala dogmat i molekylärbiologi säger kortfattat att DNA transkriberas till RNA som sedan translateras till protein. DNA kan sägas utgöra en central informationslagring från vilken kopior (mRNA) av gener görs och används för att konstruera de funktionella komponenterna i cellen, proteiner.

En del sjukdomar har sin orsak i att en gen blir muterad. En konsekvens kan vara att ett protein får felaktig form och därför inte fungerar korrekt. Ett exempel på detta är cystisk fibros där de flesta fall av sjukdomen beror på att en mutation resulterar i ett förkortat protein. I mer komplexa sjukdomar som cancer kan det behövas flera oberoende mutationer för att sjukdomen ska utvecklas. Om flera olika uppsättningar av mutationer kan ge upphov till sjukdomen är det svårare att spåra hur den uppkommer och verkar.

För att t.ex. cancer ska uppstå måste en del viktiga funktioner slås ut. För att få mer kunskap om sådana komplexa sjukdomar kan det vara nödvändigt att undersöka vilka proteiner som är delaktiga i dessa funktioner, vilka proteiner som är funktionellt kopplade. Par av proteiner kan vara funktionellt kopplade på många olika sätt, t.ex. kan de kemiskt modifiera varandra, reglera varandras nivåer, bilda komplex med flera ingående proteiner eller mer indirekt genom att ingå i samma process eller system. Att med experiment avgöra vilka proteiner som är funktionellt kopplade och hur är väldigt svårt. Det kan behövas särskilda förhållanden för att upptäcka att de är kopplade, enskilda experiment av hög kvalité kräver mycket resurser och antalet möjliga par är väldigt stort.

Ett möjligt alternativ i dagsläget är att samla bevis som indikerar funktionell koppling och använda bevisen för att uppskatta om det är troligt att ett visst par av proteiner är funktionellt kopplade eller ej. Även om detta inte ger definitiva svar kan man genom att samla bevis från många olika källor göra trovärdiga förutsägelser om funktionella kopplingar för en stor andel av de möjliga proteinparen. Dessa förutsägelser kan representeras av ett funktionellt nätverk som beskriver vilka proteinpar som har bevis för funktionell koppling och hur stort förtroende man har för kopplingen.

Den här avhandlingen beskriver mitt arbete som har fokuserat dels på att möjliggöra och förbättra uppbyggandet av funktionella nätverk och dels på att använda funktionella nätverk för att förbättra förståelsen av komplexa sjukdomar.

Som tidigare nämnt kan en mutation av en gen ge upphov till ett protein som har förändrad funktion, något som inte är bra om det rör sig om en väldigt nödvändig funktion. Man kan säga att det finns ett evolutionärt tryck för en gen att inte förändras så pass mycket att det resulterande proteinet inte kan uppfylla sin funktion. Det händer att gener dupliceras och när det då finns två kopior av samma gen parallellt i samma art kan den ena kopian förändras medan den andra kopian fortfarande bibehåller den ursprungliga funktionen.

Genom att undersöka geners släktskap kan man dra slutsatser om huruvida de har existerat parallellt inom samma art eller ej och vilken den närmaste släktingen till en gen i en art är i en annan art. När man jämför generna i två arter säger man att två gener är ortologer om de skiljdes åt när en art blev till flera och paraloger om de skiljdes åt när en gen duplicerades till två kopior.

Arbetet i den första och den andra artikeln beskriver den fortlöpande utvecklingen av InParanoid, en metod för att avgöra vilka gener som är ortologer. Då ortologer aldrig existerat parallellt inom samma art är det mer troligt att de har samma funktion än för paraloger. Att veta om två gener är ortologer är därför viktigt om man vill samla bevis om funktionella kopplingar från flera olika arter.

Då vilka system och funktioner som påverkas, snarare än enskilda gener, är av vikt för uppkomsten av en sjukdom är det möjligt att gener som är funktionellt kopplade till kända sjukdomsgener kan vara av intresse. Dessa kopplade gener kan potentiellt ge information om vilka funktioner och system som är viktiga för sjukdomen, användas för diagnostisering eller själva vara sjukdomsgener. Den tredje artikeln beskriver Maxlink, en metod för att hitta tidigare okända sjukdomsgener genom att använda ett funktionellt nätverk för att identifiera gener som är funktionellt kopplade till kända sjukdomsgener. Genom att använda kända cancergener och det funktionella nätverket FunCoup lyckades vi med Maxlink hitta 1800 kandidater med egenskaper karakteristiska för cancergener.

När en sjukdom orsakar stora systemförändringar kan detta ge sig uttryck i att nivåer av mRNA och protein är förändrade hos sjuka jämfört med friska individer. Det finns många studier av framförallt mRNA-nivåer men även av proteinnivåer hos friska och sjuka och data från sådana studier skulle kunna användas för att förbättra Maxlink metoden. Information om hur proteinpar är funktionellt kopplade och under vilka förhållanden är något som generellt saknas i funktionella nätverk och att använda information från uttrycksnivåer om vilka gener eller funktionella kopplingar som är påverkade vore därför av stort värde.

Data för mRNA- och proteinnivåer används även för att bygga funktionella nätverk. Om två proteiner har samma uttrycksmönster över vävnader och

eller olika förhållanden kan detta vara en indikation på att de är funktionellt kopplade.

Även om nivåerna av protein generellt är det som är av intresse använder man istället ofta mRNA nivåer som ett mått på proteinnivåer. Detta görs då det är enklare att mäta nivåer av mRNA än nivåer av protein och mängden protein är beroende av mängden mRNA. Det finns dock andra faktorer som påverkar proteinnivåerna och det exakta förhållandet mellan mRNA och protein är inte känt. Jag undersöker detta förhållande i artikel fyra och visar att det är möjligt att hitta gener för vilka nivåer av mRNA bättre motsvarar nivåer av protein.

Studier av hur uttryck av proteiner och mRNA skiljer sig mellan friska och sjuka kan variera både i sammansättningen av de biologiska prover mätningarna utförs på, hur mätningarna utförs samt hur data från mätningarna bearbetas. Det är av intresse att undersöka om olika studier av samma sjukdom ger jämförbara resultat eller om skillnader i utförandet har för stor inverkan. I den femte artikeln samlar vi ett stort antal studier med prover för friska och sjuka för bröst-, kolon- och lungcancer. Jag undersöker hur bearbetning av data påverkar samstämmigheten mellan studier av samma eller olika cancer och kontrasterar resultaten mot samstämmigheten mellan cancerstudier och andra typer av studier för samma vävnader. Våra resultat visar att samstämmigheten är hög, även mellan studier av olika sjukdomar, men bara när data används på rätt sätt.

Acknowledgements

In a similar manner to this thesis being the sum of my scientific endeavours I am the sum of the people I have met who have influenced who I am today. While I view all people I've met and been influenced by as contributors, some are more contributors and deserve special mention.

The Adams family – My parents for always being supportive despite me being the black sheep of the family. Dear sister, may there never be an as good "torrdämmare" as you. My university years would not have been the same without my cousin Olof introducing me to the glory of "kall glögg och hårdrock" and proving that it indeed is possible to have a negative sense of direction, *in ovo veritas*!

Monica for guiding my first tenative steps towards academia and Doris for believing so fiercely in me that it rubbed off.

My supervisors – Erik for trusting me to eventually get to a point and helping me actually make one once I got there. Cristina for uplifting chats and always giving valuable feedback about both scientific research and scientific life.

Boyzen - Omar, perhaps the most caring and oblivious person I know; Big Mike, Södertäljes most friendly gangsta; Shandy, the master meme creator and my "follower" (I totally set him up with both an education, job and indirectly family, promise); Henke, the man of the brilliant plans; Ram, always true to himself; "Lillen", not as small as one might think. Thanks for daring adventures and escapades.

The dudes - Justin, Kev, Niels and Dave, thanks for relaxing running and for chat.

Past and present members of the Sonnhammer group – Kristoffer for always interesting discussions and shocking revelations; Dave for always being friendly, helpful and patient; Walter for being a total dude, rock on (but only carefully until she gets older)!; Thomas and Oliver for both being the opposite of a scheißkopf, always providing valuable insights such as enlightening me as to my future career as a comedian; Fabian and Matt for not only putting up with my rants but providing valuable feedback too; Andreas for his keen perspectives; Dimitri, the coolest cat in the lab; Mats for giving me a

head start doing a PhD; Emil for work related and unrelated discussions as well as introducing me to dr Bozze.

Maria Sallander and the rest of the administrative staff at DBB for always being friendly and helpful even when having to go an extra mile on account of an absent minded PhD student.

Neighbors throughout the years - “Mia”, Aymeric, Hossein and Joel making me almost miss being back at Roslagstull. Tony and Ida for making the going part of going for coffee more enjoyable than the coffee part. Tove and Anna for shared woes of thesis writing and parenthood. Rauan for always being fun to hang out with and putting “the machine” into words.

And most of all my beloved ladies, the lights of my life, for having made me feel like the good kind of "S-pappa" even when I've been the bad kind of "S-pappa"

References

Abbott MA, Poiesz BJ, Byrne BC, Kwok S, Sninsky JJ, Ehrlich GD. 1988. Enzymatic gene amplification: qualitative and quantitative methods for detecting proviral DNA amplified in vitro. *J Infect Dis* **158**: 1158–1169.

Agaton C, Falk R, Höidén Guthenberg I, Göstring L, Uhlén M, Hober S. 2004. Selective enrichment of monospecific polyclonal antibodies for antibody-based proteomics efforts. *J Chromatogr A* **1043**: 33–40.

Alexeyenko A, Schmitt T, Tjärnberg A, Guala D, Frings O, Sonnhammer ELL. 2012. Comparative interactomics with Funcoup 2.0. *Nucleic Acids Res* **40**: D821–D828.

Alexeyenko A, Sonnhammer ELL. 2009. Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Res* **19**: 1107–1116.

Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, et al. 2005. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **33**: D418–D424.

Allison DB, Cui X, Page GP, Sabripour M. 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7**: 55–65.

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* **5**: e1000262.

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. 2010. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res* **39**: D289–D294.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. *PLoS Comput Biol* **8**: e1002514.

Alwine JC, Kemp DJ, Stark GR. 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A* **74**: 5350–5354.

Andersson JO, Andersson SG. 1999. Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev* **9**: 664–71.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–29.

Bader GD, Cary MP, Sander C. 2006. Pathguide: a pathway resource list. *Nucleic Acids Res* **34**: D504–D506.

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, et al. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**: 246.

Baldi P, Long AD. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**: 509–519.

Baudot A, Gómez-López G, Valencia A. 2009. Translational disease interpretation with molecular networks. *Genome Biol* **10**: 221.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Berglund L, Bjorling E, Oksvold P, Fagerberg L, Asplund A, Szigyarto CA, Persson A, Ottosson J, Wernerus H, Nilsson P, et al. 2008. A genecentric Human Protein Atlas for expression profiles based on antibodies. *Mol Cell Proteomics* **7**: 2019–2027.

Bergmann S, Ihmels J, Barkai N. 2004. Similarities and Differences in Genome-Wide Expression Data of Six Organisms. *PLoS Biol* **2**: E9.

Bjellqvist B, Ek K, Righetti PG, Gianazza E, Görg A, Westermeier R, Postel W. 1982. Isoelectric focusing in immobilized pH gradients: Principle, methodology and some applications. *J Biochem Biophys Methods* **6**: 317-339.

Björkholm P, Sonnhammer ELL. 2009. Comparative analysis and unification of domain-domain interaction networks. *Bioinformatics* **25**: 3020–3025.

Bock JL. 2000. The new era of automated immunoassay. *Am J Clin Pathol* **113**: 628–646.

Boeckmann B, Robinson-Rechavi M, Xenarios I, Dessimoz C. 2011. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief Bioinform* **12**: 423–435.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**: 185–193.

Bonner RF, Emmert-Buck M, Cole K, Pohida T, Chuaqui R, Goldstein S, Liotta LA. 1997. Laser capture microdissection: molecular analysis of tissue. *Science* **278**: 1481-1483.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**: 365–371.

Busch F, Paul W. 1961. Isotopentrennung mit dem elektrischen Massenfilter. *Zeitschrift für Physik* **164**: 581-587.

Carvalho PC, Hewel J, Barbosa VC, Yates JR. 2008. Identifying differences in protein expression levels by spectral counting and feature selection. *Genet Mol Res* **7**: 342-356.

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. 2011. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**: D685–690.

Chan SY, White K, Loscalzo J. 2012. Deciphering the molecular basis of human cardiovascular disease through network biology. *Curr Opin Cardiol* **27**: 202–209.

Chatr-Aryamontri A, Breitkreutz B-J, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, et al. 2013. The BioGRID interaction database: 2013 update. *Nucleic Acids Res* **41**: D816–823.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS, Stoeckert CJ. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**: D363–368.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One* **2**: e383.

Chen Y, Jiang T, Jiang R. 2011. Uncover disease genes by maximizing information flow in the phenome-interactome network. *Bioinformatics* **27**: i167–i176.

Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y. 2012. Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. *PLoS Comput Biol* **8**: e1002503.

Choi JK, Yu U, Yoo OJ, Kim S. 2005. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics* **21**: 4348–4355.

Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**: 613-619.

Date S V, Marcotte EM. 2005. Protein function prediction using the Protein Link EXplorer (PLEX). *Bioinformatic* **21**: 2558–2559.

Dawson J a, Kendziorski C. 2011. An Empirical Bayesian Approach for Identifying Differential Coexpression in High-Throughput Experiments. *Biometrics* **68**: 455–465.

DeLuca TF, Cui J, Jung J-YY, Gabriel KCS, Wall DP, St Gabriel KC. 2012. Roundup 2.0: Enabling comparative genomics for over 1800 genomes. *Bioinformatics* **28**: 715–716.

DeLuca TF, Wu I-H, Pu J, Monaghan T, Peshkin L, Singh S, Wall DP. 2006. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics* **22**: 2044-2046.

Dessimoz C, Gabaldon T, Roos DS, Sonnhammer E, Herrero J, the Quest for Orthologs Consortium. 2012. Toward Community Standards in the Quest for Orthologs. *Bioinformatics* **28**: 900–904.

Eisen JA. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr Opin Genet Dev* **10**: 606–11.

Fenn J, Mann M, Meng C, Wong S, Whitehouse C. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**: 64-71.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* **19**: 99–113.

Forslund K, Pekkari I, Sonnhammer EL. 2011. Domain architecture conservation in orthologs. *BMC Bioinformatics* **12**: 326.

Forslund K, Sonnhammer ELL. 2008. Predicting protein function from domain content. *Bioinformatics* **24**: 1681–1687.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, Von Mering C, et al. 2012. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**: D808–D815.

Freudenberg JM, Sivaganesan S, Wagner M, Medvedovic M. 2010. A semi-parametric Bayesian model for unsupervised differential co-expression analysis. *BMC Bioinformatics* **11**: 234.

Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S. 2007. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm Genome* **18**: 463–472.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. *Nat Rev Cancer* **4**: 177–183.

Gabaldon T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer EL, Lewis S. 2009. Joining forces in the quest for orthologs. *Genome Biol* **10**: 403.

Ge H, Liu Z, Church GM, Vidal M. 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29**: 482–486.

George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters M a. 2006. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* **34**: e130.

Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **100**: 6940–5.

Glazko GV, Mushegian AR. 2004. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol* **5**: R32.

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. 2007. The human disease network. *Proc Natl Acad Sci U S A* **104**: 8685–8690.

Goldsby RA, Kindt TJ, Kuby J, Osborne BA. 2002. *Immunology, Fifth Edition*. W. H. Freeman.

Gonçalves JP, Francisco AP, Moreau Y, Madeira SC. 2012. Interactogeneous: Disease Gene Prioritization Using Heterogeneous Networks and Full Topology Scores. *PloS One* **7**: e49634.

Gray MW. 2012. Mitochondrial evolution. *Cold Spring Harbor Perspectives in Biology* **4**: a011403.

Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, Zhang L, Mei N, Chen T, Herman D, Goodsaid FM, et al. 2006. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotechnol* **24**: 1162–1169.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**: 994–999.

Görg A, Postel W, Günther S. 1988. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **9**: 531–546.

Hebenstreit D, Fang M, Gu M, Charoensawan V, Van Oudenaarden A, Teichmann SA. 2011. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* **7**: 497.

Hegyi H, Gerstein M. 2001. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res* **11**: 1632–1640.

Hegyi H, Gerstein M. 1999. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* **288**: 147–164.

Henricson A, Forslund K, Sonnhammer EL. 2010. Orthology confers intron position conservation. *BMC Genomics* **11**: 412.

Holloway AJ, Oshlack A, Diyagama DS, Bowtell DD, Smyth GK. 2006. Statistical analysis of an RNA titration series evaluates microarray precision and sensitivity on a whole-array basis. *BMC Bioinformatics* **7**: 511.

Hubbell E, Liu WM, Mei R. 2002. Robust estimators for expression analysis. *Bioinformatics* **18**: 1585–1592.

Huber W, Von Heydebreck A, Sültmann H, Poustka A, Vingron M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**: S96–S104.

Hudson NJ, Reverter A, Dalrymple BP. 2009. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput Biol* **5**: e1000382.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldon T, Gabaldón T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Brief Bioinform* **12**: 442–448.

Hulsen T, Huynen MA, De Vlieg J, Groenen PM. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* **7**: R31.

Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, Coller HA, Troyanskaya OG. 2009. Exploring the human genome with functional maps. *Genome Res* **19**: 1093–1106.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**: 249–264.

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

Jensen LJ, Julien P, Kuhn M, Von Mering C, Muller J, Doerks T, Bork P. 2008. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* **36**: D250–D254.

Jonsson PF, Bates PA. 2006. Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**: 2291–7.

Kamburov A, Wierling C, Lehrach H, Herwig R. 2009. ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Res* **37**: D623–D628.

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

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**: D109–D114.

Kasperkovitz P V, Timmer TCG, Smeets TJ, Verbeet NL, Tak PP, Van Baarsen LGM, Baltus B, Huizinga TWJ, Pieterman E, Fero M, et al. 2005. Fibroblast-like synoviocytes derived from patients with rheumatoid arthritis show the imprint of synovial tissue heterogeneity: evidence of a link between an increased myofibroblast-like phenotype and high-inflammation synovitis. *Arthritis and Rheum* **52**: 430–441.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, et al. 2012. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **40**: D841–D846.

Keshava PTS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. 2009. Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**: D767–D772.

Kim WK, Krumpelman C, Marcotte EM. 2008. Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. *Genome Biol* **9**: S5.

Klose J. 1975. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**: 231–243.

Koonin E V. 2005. ORTHOLOGS, PARALOGS, AND EVOLUTIONARY GENOMICS. *Annual Review of Genetics* **39**: 309-338.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. 2011. Computational methods for Gene Orthology inference. *Brief Bioinform* **12**: 379–391.

Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM. 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet* **24**: 539–551.

Köhler G, Milstein C. 1975. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**: 495–497.

Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, et al. 2007. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25**: 309–316.

Lai Y, Wu B, Chen L, Zhao H. 2004. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics* **20**: 3146–3155.

Lassmann T, Frings O, Sonnhammer EL. 2009. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res* **37**: 858–865.

Lassmann T, Sonnhammer EL. 2005. Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* **6**: 298.

Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY. 2010a. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nat Biotechnol* **28**: 149–156.

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. 2011. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* **21**: 1109–1121.

Lee I, Lehner B, Vavouri T, Shin J, Fraser AG, Marcotte EM. 2010b. Predicting genetic modifier loci using functional gene networks. *Genome Res* **20**: 1143–1153.

Lee I, Li Z, Marcotte EM. 2007. An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PloS One* **2**: e988.

Li L, Zhang K, Lee J, Cordes S, Davis DP, Tang Z. 2009. Discovering cancer genes by integrating network and functional properties. *BMC Med Genomics* **2**: 61.

Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, et al. 2012. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* **40**: D857–D861.

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

Lindskog M, Rockberg J, Uhlen M, Sterky F. 2005. Selection of protein epitopes for antibody production. *Biotechniques* **38**: 723–727.

Linghu B, Snitkin ES, Hu Z, Xia Y, DeLisi C. 2009. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol* **10**: R91.

Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. 2012. Revisiting Global Gene Expression Analysis. *Cell* **151**: 476-482.

Makarova KS, Sorokin A V, Novichkov PS, Wolf YI, Koonin EV. 2007. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct* **2**: 33.

Malone JH, Oliver B. 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology* **9**: 34.

Marcotte EM. 2000. Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol* **10**: 359–365.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

May C, Brosseron F, Chartowski P, Schumbrutzki C, Schoenebeck B, Marcus K. 2011. Instruments and methods in proteomics. *Methods Mol Biol* **696**: 3–26.

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

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621–628.

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

Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, Von Mering C, Doerks T, Jensen LJ, et al. 2010. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* **38**: D190–D195.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.

Navlakha S, Kingsford C. 2010. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26**: 1057–1063.

Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol* **7**: e1002073.

Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**: 376–386.

Ortholog Benchmarking Webservice. Ortholog Benchmarking Webservice. http://orthology.benchmarkservice.org/cgi-bin/gateway.pl (Accessed February 8, 2013).

Oshlack A, Wakefield MJ. 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4**: 14.

Ostlund G, Lindskog M, Sonnhammer EL. 2010a. Network-based Identification of novel cancer genes. *Mol Cell Proteomics* **9**: 648–655.

Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL, Östlund G, Köstler T. 2010b. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38**: D196–D203.

O'Farrell PH. 1975. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**: 4007–4021.

Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes H-W, et al. 2005. The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**: 832–834.

Panchenko A, Przytycka TM, eds. 2008. *Protein-protein Interactions and Networks: Identification, Computer Analysis, and Prediction (Computational Biology)*. Springer.

Pappin DJC, Hojrup P, Bleasby AJ. 1993. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* **3**: 327-332.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**: 4285–4288.

Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A. 2009. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci* **18**: 1306–1315.

Powell S, Szklarczyk D, Trachana K, Roth a, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, et al. 2011. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res* **40**: D284–D289.

Rattei T, Tischler P, Gotz S, Jehl MA, Hoser J, Arnold R, Conesa A, Mewes HW. 2010. SIMAP--a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res* **38**: D223–D226.

Razick S, Magklaras G, Donaldson IM. 2008. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**: 405.

Reference Genome Group of the Gene Ontology Consortium. 2009. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol* **5**: e1000431.

Remm M, Storm CEV, Sonnhammer ELL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**: 1041-1052.

Reverter A, Ingham A, Lehnert SA, Tan SH, Wang Y, Ratnakumar A, Dalrymple BP. 2006. Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. *Bioinformatics* **22**: 2396–2404.

Ronaghi M. 1998. DNA SEQUENCING:A Sequencing Method Based on Real-Time Pyrophosphate. *Science* **281**: 363-365.

Ross PL. 2004. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Mol Cell Proteomics* **3**: 1154-1169.

Roth ACJ, Gonnet GH, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**: 518.

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**: D449–D451.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467–470.

Schmidt A, Kellermann J, Lottspeich F. 2005. A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* **5**: 4-15.

Schmitt T, Messina DN, Schreiber F, Sonnhammer EL. 2011. Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief Bioinform* **12**: 485–488.

Schreiber F, Sonnhammer ELL. 2013. Hieranoid: Hierarchical orthology inference. *J Mol Biol* **in press**.

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**: 337–342.

Shen L, Chepelev I, Liu J, Wang W. 2010. Prediction of quantitative phenotypes based on genetic networks: a case study in yeast sporulation. *BMC Syst Biol* **4**: 128.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**: Article3.

Sonnhammer EL, Koonin E V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* **18**: 619–620.

de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. 2009. Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5**: 1512–1526.

Syvänen AC, Bengtström M, Tenhunen J, Söderlund H. 1988. Quantification of polymerase chain reaction products by affinity-based hybrid collection. *Nucleic Acids Res* **16**: 11327–11338.

Szalkowski A, Ledergerber C, Krahenbuhl P, Dessimoz C. 2008. SWPS3 - fast multi-threaded vectorized Smith-Waterman for IBM Cell/B.E. and x86/SSE2. *BMC Res Notes* **1**: 107.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41.

Tatusov RL, Galperin MY, Natale DA, Koonin E V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33–36.

Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. 2009. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* **27**: 199–204.

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* **38**: 615–643.

The UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**: D71–D75.

Thomas PD, Wood V, Mungall CJ, Lewis SE, Blake JA. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS Comput Biol* **8**: e1002386.

Tiffin N, Andrade-Navarro MA, Perez-Iratxeta C. 2009. Linking genes to diseases: it's all in the data. *Genome Med* **1**: 77.

Trachana K, Larsson T a, Powell S, Chen W-H, Doerks T, Muller J, Bork P. 2011. Orthology prediction methods: A quality assessment using curated protein families. *Bioessays* **33**: 769–780.

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**: 5116–5121.

Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, et al. 2010. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* **28**: 1248–1250.

Uhlén M, Oksvold P, Älgenäs C, Hamsten C, Fagerberg L, Klevebring D, Lundberg E, Odeberg J, Pontén F, Kondo T, et al. 2012. Antibody-based protein profiling of the human chromosome 21. *Mol Cell Proteomics* **11**: M111.013458.

VanGuilder HD, Vrana KE, Freeman WM. 2008. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* **44**: 619–626.

Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. 2010. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* **6**: e1000641.

Vogel C, de Sousa Abreu R, Ko D, Le S-Y, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO. 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* **6**: 400.

Wall DP, Fraser HB, Hirsh a E. 2003. Detecting putative orthologs. *Bioinformatics* **19**: 1710–1711.

Wang L, Tu Z, Sun F. 2009. A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. *BMC Genomics* **10**: 220.

Wang PI, Marcotte EM. 2010. It's the machine that matters: Predicting gene function and phenotype from protein networks. *J Proteomics* **73**: 2277–2289.

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, et al. 2010. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* **38**: W214–W220.

Wollnik H. 1993. Time-of-flight mass analyzers. *Mass Spectrom Rev* **12**: 89–114.

Wu X, Jiang R, Zhang MQ, Li S. 2008. Network-based global inference of human disease genes. *Mol Syst Biol* **4**: 189.

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. 2004. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J Am Stat Assoc* **99**: 909-917.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

Yates JR. 1998. SPECIAL FEATURE: Mass Spectrometry and the Age of the Proteome. *Proteome* **33**: 1–19.

Yu H, Liu BH, Ye ZQ, Li C, Li YX, Li YY. 2011. Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. *BMC Bioinformatics* **12**: 315.

Zaluzec EJ, Gage DA, Watson JT. 1995. Matrix-Assisted Laser Desorption Ionization Mass Spectrometry: Applications in Peptide and Protein Characterization. *Protein Expr Purif* **6**: 109-123.

Zheng Y, Roberts RJ, Kasif S. 2002. Genomic functional annotation using coevolution profiles of gene clusters. *Genome Biol* **3**: RESEARCH0060.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol* **18**: 292–298.

Zhu Q, Miecznikowski JC, Halfon MS. 2010. Preferred analysis methods for Affymetrix GeneChips. II. An expanded, balanced, wholly-defined spike-in dataset. *BMC Bioinformatics* **11**: 285.