The interplay between genome and network evolution in eukaryotes

Like Fokkens (2013)
The interplay between genome and network evolution in eukaryotes.
PhD thesis, Utrecht University
Cover design by: Thijs Fokkens and Like Fokkens
ISBN: 978-90-8891-624-3

The interplay between genome and network evolution in eukaryotes

De wisselwerking tussen genoom en netwerk evolutie in eukaryoten

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 28 mei 2013 des middags te 4.15 uur

door

Like Fokkens

geboren op 11 december 1979 te Groningen

Promotor: Prof. dr. P. Hogeweg
Co-promotor: Dr. B. Snel

Contents

Chapter 1

Introduction

1.1 Introduction

Evolution is a balancing act between conservation and variation. Variation within a population is crucial in adaptive evolution as natural selection acts upon differences between individuals. Mutations are necessary to establish diversity. On the other hand, from the perspective of the individual organism, mutations are seldom beneficial and need to be avoided. In general, a certain level of conservation is necessary to secure any adaptive progress. This dilemma manifests itself in the differences in evolutionary dynamics of proteins. Some proteins are very static in evolution as their function does not allow for much variation. Other proteins are very changeable, they are used to adapt to any changing circumstances or to occupy a novel niche. The difference in evolutionary dynamics between proteins corresponds to differences in selection pressure that is related to their role in the cellular machinery. Whether a mutation is fixated in evolution depends on the functional relations of the proteins it affects, how it changes these relations and how this influences the functioning of the machinery as a whole.

In this thesis we describe the results of our study of the interplay between genome evolution and protein function in eukaryotes. How do functional relations of proteins translate into selective constraints? How do changes on the genome affect the organization of proteins in the molecular machinery? We can test possible answers to these questions thanks to two current advances: the advent of completely sequenced eukaryotic genomes and the generation of genome-wide protein interaction data in the fungus *S. cerevisiae*.

Eukaryotic genomes have only relatively recently become available in numbers that permit comparative genomic studies aimed at discovering general and fundamental evolutionary principles. At the same time, experimental techniques to determine protein-protein interactions in eukaryotes have been up-scaled in order to test a large portion of a species' proteome for physical interactions (Uetz et al., 2000; Ito et al., 2001; Walhout and Vidal, 2001; Gavin et al., 2002, 2006; Krogan et al., 2006). Repeated screens for direct physical interactions and co-complex membership in *S. cerevisiae* allow for estimates of quality and completeness of the data (Yu et al., 2008). The sheer size of these datasets does not only provide information on functional relations for a large number of proteins, it offers an unprecedented view of the large-scale architecture of the molecular machinery in this species.

We integrate this new functional data with information from comparative genomics in eukaryotes to provide novel insights on feedback mechanisms between genome and cellular machinery. Does the cellular machinery have a modular architecture, in which proteins cooperate in relatively independent functional modules? How does this influence protein evolutionary dynamics? In the first two chapters we study (disrupted) co-evolution of collaborating proteins. In chapter

4 and 5 we implement certain assumptions on how genomic changes affect the local functional organization of proteins in the cell, in a very basic null model and study the consequences of these assumptions in terms of protein evolution and global architecture. In this chapter we describe some of the evolutionary and functional information that is available, and the strengths and pitfalls of combining these two types of information.

1.2 Comparative genomics in eukaryotes: protein families

Before assembly, a genome sequence consists of a collection of short reads. Partially overlapping reads are joined together to form contigs that are themselves collected into supercontigs or scaffolds that should in the end correspond to chromosomes. In eukaryotes typically only a small fraction of the DNA on these chromosomes actually consists of protein coding genes. These are identified using gene prediction algorithms that compare the newly sequenced genome to the sequences of the known genes and detect patterns in the DNA that are indicative of a gene or a promoter region. Nucleotide triplets (codons) of these genes are translated into amino acids yielding a set of protein sequences: the proteome.

In this study we use protein- rather than gene sequences as this is more suitable for large evolutionary distances and we typically compare distantly related species. We group homologous proteins (i.e. proteins that are related through common descent) into protein families. Homologous proteins that belong to the same species (and thus originated from a duplication event) are known as paralogs. Orthologous groups form a specific case of protein families as they are defined with respect to a certain speciation event. For example, if we compare the proteomes of only eukaryotic species, each orthologous group represents a single protein in the last eukaryotic ancestor, or, when only species from a single clade (e.g. Fungi) belong to this family, in the ancestor of this clade. Orthologous groups thus provide a valuable reference point in time: the ancestral species indicates a maximum period of evolutionary divergence between any two proteins in the same orthologous group. These relatively closely related homologs (how close depends on the reference speciation event) are known as orthologs. Proteins from the same species that belong to the same orthologous group are known as inparalogs. Again this term is used with respect to a speciation event, indicating that the duplication that gave rise to these proteins occurred after this speciation event. Similarly, paralogs that belong to different orthologous groups are known as outparalogs, indicating they originate from a relatively old duplication event (Sonnhammer and Koonin, 2002; Koonin, 2005).

There is a large variety of methods to construct orthologous groups, but they

roughly fall into two different categories: phylogenetic tree annotation and graph clustering. Both start by collecting homologous sequences: pairs of protein sequences are aligned and if the similarity score that is assigned to the aligned regions (Pearson and Lipman, 1988; Altschul et al., 1990; Rognes, 2001) is significantly high, the proteins are assumed to share a common ancestor. Methods that are based on phylogenetic tree annotation, proceed to construct a multiple sequence alignment for each group of homologous proteins and infer a phylogenetic tree based on this alignment. This process can be repeated several times, each time including different sets of aligned positions, to provide bootstrap support for each branch in the tree. This (bootstrapped) tree is then annotated to distinguish branching through gene duplication from branching through speciation. Each node representing a single protein in the ancestral species of choice forms a separate orthologous group: all leaves that can be traced back to this node are collected into this group.

This method is computationally expensive, individual steps are complex and often require multiple iterations and manual intervention. On the other hand, a phylogenetic tree provides detailed information on the evolutionary history of a protein family, although misplacement of a single protein can lead to inferences of a large number of duplications and losses. This can be avoided through tree rearrangement, a process in which the gene tree is compared to the species tree and branches with low bootstrap support are placed elsewhere in the tree if this leads to a more parsimonious evolutionary scenario. Phylogenetic tree annotation is mostly used to study individual protein families (e.g. (van Dam et al., 2009, 2011; Sarikas et al., 2011; Dittmer and Misteli, 2011; Middelbeek et al., 2010)) rather than general evolutionary trends. This may change due to recent heroic efforts to provide phylogenetic trees for a large number of protein families (Huerta-Cepas et al., 2011; Muller et al., 2010; Gabaldon, 2008) that can be combined with automated reconciliation, rearrangement and annotation of gene trees (Chen et al., 2000; Vilella et al., 2009).

A more common and less laborious approach to define orthologous groups is to partition a graph in which homologous proteins are connected, into nonoverlapping clusters (Tatusov et al., 2000, 2003; Remm et al., 2001; Li et al., 2003; Dehal and Boore, 2006; Muller et al., 2010). Proteins are connected in this graph depending on whether they share significant sequence similarity and on additional criteria that may vary between different methods. For example, many methods require that proteins are Bidirectional Best Hits[1] (e.g. (Remm et al., 2001; Li et al., 2003)) and that the aligned region spans at least half of the shortest protein sequence (Remm et al., 2001). This last constraint is applied to avoid connecting proteins based on sequence similarity in only a small part of the protein. Moreover, after clustering the graph of putatively homologous proteins, the res-

[1] Protein A in species a is the Bidirectional Best Hit of protein B in species *b* if B is most similar to A (in terms of sequence) of all proteins in *b* and A is most similar to B of all proteins in *a*. Species *a* and *b* can be the same species.

ulting groups can be subjected several post-processing steps. These are aimed at detecting fusion and fission events, or duplications in or predating the last common ancestor of the species represented in the group (Tatusov et al., 2000, 2003; Muller et al., 2010; Jothi et al., 2006). In case of a fusion or fission event, a protein is split up and in parts assigned to different groups, whereas in case of old duplications the group is split up.

In this thesis we focus on differences in family sizes between species without considering detailed evolutionary reconstructions, hence we use orthologous groups based on graph clustering rather than based on explicit phylogenetic trees. In Chapter 2 and 3 we define special-purpose orthologous groups. The specific methodology that we implement differs between the two chapters. In Chapter 3 we only consider two species, whereas in Chapter 2 we include 34 complete eukaryotic genomes. In Chapter 4 and 5 we use an off-the-shelf set of protein families.

The comparison of complete proteomes of multiple species by grouping proteins of common descent into families, allows us to register differences between species as well as between protein families. In some cases, duplication or loss of proteins can be directly related to adaptive needs. For example, if an organism switches to a parasitic lifestyle and takes some of the metabolites it needs for survival directly from its host, its genome can lose the genes encoding for proteins involved in producing these metabolites (e.g. (Keeling et al., 2010; Morrison et al., 2007)). Similarly, species that turn pathogenic can expand certain gene families, for example those coding for enzymes that are capable of degrading the host's cell wall (Dodds, 2010).

Comparing protein family sizes in different species can thus shed light on how they adapted to environmental changes in the course of evolution. However, not all protein family dynamics is readily explained. One of the reasons is that, because we lack comprehensive, system-wide functional data in most species, we don't know how genomic changes translate into changes in the organization of the molecular machinery. If a protein is lost, is the entire subsystem (e.g. protein complex) lost? Is the protein replaced? If a protein has duplicated, do both daughter proteins perform part of the ancestral function? Do they gain new functionalities, e.g. new interactions with other proteins? Literature supplies examples of almost any imaginable scenario. Can we extend beyond individual examples and provide rules of thumb that can be used when there is no experimental data available? To answer these questions we combine information from protein families with data from large-scale experimental studies in eukaryotes. This data allows us to define the function of a protein in terms of its interactions with other proteins, i.e. in terms of its position in a protein network.

1.3 Large-scale protein networks in eukaryotes

The first system-wide functional screens in eukaryotes measured gene expression levels in different experimental conditions using microarrays, obtaining for each gene an expression profile: the levels of expression of a gene in all conditions measured. The up- or downregulation of a gene in a specific experimental condition provides information on the cellular processes this gene is involved in. Moreover, if two genes are up- and or downregulated in the same conditions, these genes are likely to be involved in the same processes (Hughes et al., 2000; Wolfe et al., 2005). The correlation of expression profiles can be represented in a (weighted) network. Microarray experiments tend to be very sensitive towards minor changes in experimental conditions, replicates of the same experiments may produce very distinct profiles (Bammler et al., 2005; Kuo et al., 2006). Overlaying different datasets, either from the same or from multiple organisms, increases confidence in network connections (Stuart et al., 2003; van Noort et al., 2003). Although coregulation of genes indicates that they're likely to be involved in similar cellular processes, to disentangle how exactly different proteins work together in these processes additional experiments are required. Not all proteins that strongly cooperate are coregulated (Jeffery, 2009; Copley, 2012; Han et al., 2004; de Lichtenberg et al., 2005) and because regulation also occurs post transcription, mRNA levels do not necessarily correspond to protein abundance.

The classic way of determining a protein's function is to inactivate or delete the gene that encodes for this protein and study the mutant's phenotype. Initially, in large scale studies, a phenotype was defined in terms of growth speed (Giaever et al., 2002; Winzeler et al., 1999) or its change in growth speed in response to certain experimental conditions (Dudley et al., 2005; Hillenmeyer et al., 2008). A major leap forward is the more detailed phenotypic characterization in terms of the change in expression level of all genes in a cell, comparing a deletion mutant to the wild type (Featherstone and Broadie, 2002; van Wageningen et al., 2010). Another important advancement is the creation of double mutants (Tong et al., 2001, 2004). By comparing the growth speed of the double mutant to the growth speed that would be expected based on the fitness effect of the two single mutants, these screens reveal functional interdependence as well as redundancy in the molecular machinery (Collins et al., 2006; Baryshnikova et al., 2010). If the fitness effect of a double mutant is not stronger than that of a single mutant, both proteins are mutually dependent: the loss of one protein prevents full functionality of the other protein, hence deleting this other protein does not have a significant additional fitness effect. This relation is known as a positive Genetic Interaction and is observed between for example proteins that belong to the same protein complex (Kelley and Ideker, 2005). In contrast, if the fitness effect of a double mutant is very severe compared to that of a single mutant, one protein was compensating for the lack of function of the other protein. This is known as a negative Genetic Interaction and is observed for example when proteins take part

in parallel pathways (Kelley and Ideker, 2005).

Expression- as well as deletion mutant studies provide valuable information on protein function, but a detailed description of cellular processes requires knowledge of the actual physical interactions between proteins as well. Over time, physical interactions from a range of small-scale experiments have been collected from literature and stored in various databases (e.g. (Stark et al., 2006; Xenarios et al., 2000; Hermjakob et al., 2004)). The up-scaling of Yeast-Two-Hybrid (Y2H) screens in 2000 provided genome-wide information of protein-protein interactions obtained from a single experimental setup (Uetz et al., 2000). Systematic screening for protein complexes using Tandem Affinity Purification followed by Mass Spectrometry (TAP/MS) provided even more information on both direct and indirect physical interactions between proteins. This information on interactions (the interactome) is typically represented in a protein network connecting interacting proteins. The edges in this network can be weighted as repeated screens allow for overlaying different datasets to estimate reliability (Gavin et al., 2006; Krogan et al., 2006; Collins et al., 2007a).

The overlap between protein interaction datasets that have been generated with different techniques, is typically very small, reflecting systematic biases that can be attributed to each specific experimental setup (Ivanic et al., 2009; Fernandes et al., 2010; Hakes et al., 2005; Sambourg and Thierry-Mieg, 2010). For example, TAP/MS is aimed to detect protein complexes and thus returns indirect interactions as well. Moreover, because Mass Spectrometry is used to identify proteins, networks based on TAP/MS data are enriched for abundant proteins. In Y2H screens the interacting proteins activate a reporter gene, hence proteins that can not (easily) enter the nucleus, for example because they're attached to the membrane, are underrepresented in Y2H based networks. Some proteins are capable of auto-activating the reporter gene and may thus introduce a large number of false positive interactions (Walhout and Vidal, 1999). The lack of overlap between different networks is due to both False Positives as well as False Negatives: different techniques sample different portions of the yeast interactome (Yu et al., 2008) and detect different types of functional relations (van Noort et al., 2007).

Metabolic networks represent enzymes and their chemical reactions in a directed graph. In chapter 2, we study metabolic pathways as an example of a functional module, a subgraph in which all components are strongly interdependent. Other examples of directed networks are gene regulatory networks (GRNs) that are based on experiments that identify interactions between proteins and DNA, such as Y1H, ChIP-on-ChIP or ChIP-seq (Grove and Walhout, 2008; Walhout, 2011). Gene Regulatory Networks fall outside the scope of this work.

The number of False Positive connections in networks is decreased by overlaying different datasets, which can be repeated screens in the same experimental

setup (Ito et al., 2001), data from different types of experiments (Myers et al., 2009) or even from different species where a connection is confirmed if it is also found between orthologs in the other species (Walhout et al., 2000). Filtering a network with other datasets increases confidence its edges but this comes at the cost of decreased coverage (von Mering et al., 2002). We aim to connect patterns from networks with patterns from genome sequences. In this context, avoiding False Negatives is as important as avoiding False Positives, which calls for comprehensive functional screens and completely sequences genomes. Recent high-throughput datasets test almost all proteins encoded in the *S. cerevisiae* genome for physical interactions (Gavin et al., 2006; Krogan et al., 2006) whereas for example the labour intensive Genetic Interaction studies mainly sample subsystems (Zheng et al., 2010; Fiedler et al., 2009; Wilmes et al., 2008; Collins et al., 2007b; Schuldiner et al., 2005). Therefore we focus on protein-protein interactions.

1.4 Combining results from comparative genomics with data on protein function.

1.4.1 Function prediction

One of the first applications of comparative genomics is to infer functional information for a protein based on experimental evidence obtained for a homologous protein in a model organism. In addition to comparisons between individual gene- and protein sequences, novel genome sequences are characterized by comparing them to models representing protein domains, signal peptides, binding sites, etc. Protein networks are defined or extended by transferring connections between homologous proteins (Yu et al., 2004; Walhout et al., 2000; von Mering et al., 2003). The use of comparative genomics in function prediction is important as for most species the amount of experimental data that is available, lags far behind the amount of sequence information. Given recent progress in sequencing techniques, this imbalance is likely to increase (Metzker, 2010).

The transfer of functional characteristics between homologous proteins is commonplace, but to what extent and in which conditions this practice is sound, is not well known (Lewis et al., 2012). The confidence in inference of functional properties based on homology depends on the extent of sequence divergence, whether the proteins are orthologs or paralogs and the level of detail of the functional description. The main reason why it is so difficult to apply knowledge from well-studied model organisms to other species is that one-to-one homology relations between proteins in different species are relatively rare. We do not know how evolutionary events translate into changes in the organization of the molecu-

lar system under study hence lineage specific duplications and/or losses hinder the comparison of molecular systems between different species. We address the translation of genomic events into network level changes in chapter 4 and 5.

Another way in which comparative genomics is used in function prediction is by comparing the presence and absence patterns or entire phylogenetic trees of different protein families (Kensche et al., 2008; Pazos et al., 2008; Pellegrini, 2012). The underlying assumption is that proteins that belong to families that have similar phylogenetic trees or taxonomic distributions, are subjected to similar or shared selective constraints and thus likely to be engaged in the same cellular processes. Moreover, the need for dosage balance of components of protein complexes leads to coduplication of complex members (Papp et al., 2003). On the other hand, many proteins that we know are involved in the same system, for example proteins that belong to the same protein complex, exhibit very distinct evolutionary profiles, indicating that strong cooperation does not guarantee similar selective constraints. We specifically address the relation between functional context and sequence similarity and co-evolution in chapters 2 and 3.

1.4.2 Network evolution

Given the common practice of transferring functional information between homologous proteins, it is important to have some estimate of the extent and rate of network rewiring in evolution (Lewis et al., 2012). Network evolution is studied by aligning networks that are defined in different species, by matching homologous proteins. The lack of comprehensive networks in multiple species prohibits exact estimates of rewiring, but general trends can be observed. For example, several studies indicate that the extent of rewiring depends on the type of connection that is represented in the network. For example, stable interactions, such as between members of a protein complex tend to be very conserved whereas rewiring of transient interactions (e.g.phosphorylation) occurs more frequently (van Dam and Snel, 2008; Shou et al., 2011). Similarly, rewiring between different modules is more common than within a module. Both observations are obviously related, as transient interactions are more likely to occur between different modules.

Network evolution is not only studied through network alignments. The timeframe in which a protein family originated can be inferred from the taxonomic distribution of the species that are represented in the family. This information can be used to reconstruct how a network has expanded in the course of evolution (Qin et al., 2003; Eisenberg and Levanon, 2003; Kim and Marcotte, 2008). A caveat in this approach is that the age of a family does not necessarily correspond to the age of a protein because many proteins result from gene duplications that obviously occur after invention of the gene that founded the family. How this affects the age-structure in protein networks is the subject of chapter 4.

1.4.3 Relating a protein's evolutionary properties to it's network properties

In addition to reconstructing network evolution, the age of a protein family is used to compare network properties of 'old' proteins to those of 'young' proteins. For example, Eisenberg et al. claim that older proteins are engaged in more protein-protein interactions (Eisenberg and Levanon, 2003) and suggest that proteins tend to gather more and more interactions over time. Warnefors et al. showed that older proteins are more tightly regulated (Warnefors and Eyre-Walker, 2011) and suggest that old proteins are involved in core cellular processes.

In addition to the problem of gene duplications, a caveat in linking protein age to a network property is that the observed relations depend on how protein age was defined. Proteins that exhibit slow sequence evolution tend to be 'older' as homologs in distant species are more easily recognized. As in some studies proteins that have many interaction partners evolve more slowly (they are more constrained), it is unclear whether high connectivity is associated with older proteins or with slowly evolving proteins.

A common conjecture is that the rate of sequence evolution of a protein corresponds to it's 'fitness density', the fraction of it's amino acids that is important for proper functioning of the protein (Pal et al., 2006; Zhou et al., 2008; Wolf et al., 2010). Because specific sites that are important for the structure of a protein or its interactions with other proteins, tend to more conserved, proteins that have more of these constrained sites will evolve more slowly. The question is whether a protein's fitness density can be predicted from its position in a protein network (Pal et al., 2006). There have been many attempts to answer this question. Different evolutionary properties, such as protein age, the propensity to be lost in evolution and rate of sequence evolution are correlated, not only because they all provide indications of (relaxed) selection pressure, but also because of methodological reasons, mainly the fact that slow sequence evolution aids the recognition of homologs (Wolf et al., 2009; Krylov et al., 2003; Fokkens et al., 2012; Elhaik et al., 2006). Similarly, network properties, such as degree and centrality correlate as well. These confounding correlations spur the discussion on which properties are most suitable to represent the interplay between evolution and cellular organization.

Even if a strong correlation between evolutionary and network properties is observed, this does not need to be caused by selection pressure on the level of a protein's position in the network. For example, the rate of sequence evolution is probably strongly determined by selective pressure on effective translation (Drummond et al., 2006). The rate of sequence evolution of a protein is generally measured by the ratio of nonsynonymous (leading to an amino acid change) versus synonymous nucleotide substitutions Ka/Ks (or in terms of codon substitutions: dN/dS (Hirsh et al., 2005)) derived from an alignment with an ortholog

in a closely related species. This ratio is reflects the number of amino acid substitutions over time: a high Ka/Ks is generally assumed to indicate that a protein is under positive selection. A low Ka/Ks, indicating a protein is under negative selection, correlates with the number of functional constraints of a protein and its dispensability (Krylov et al., 2003). However, for highly expressed proteins, the cost of misfolding constrains codon usage, affecting both Ka and Ks. This indicates that Ka/Ks measures selective pressure not only on protein function, but also on efficiency and reliability of translation (Drummond and Wilke, 2008). Proteins encoded by highly expressed genes are generally more abundant hence they are more likely to have many interaction partners. This can also explain the correlation between the number of interaction partners and the rate of sequence evolution of a protein (Bloom and Adami, 2003).

The long-tailed degree distribution observed in many biological networks is associated with evolutionary robustness as removal of most nodes and most edges will not affect network structure. This suggests that proteins that are pivotal for maintaining network structure, such as the prolific interactors (also known as network hubs), are less likely to be lost in evolution. Several studies indicate that this is indeed the case (Fraser et al., 2002, 2003; Krylov et al., 2003; Jordan et al., 2003). First of all, this may be due to a positive correlation between the propensity to be lost in evolution and the rate of sequence evolution (Krylov et al., 2003). Secondly, a myriad of network growth models that simulate network evolution without explicit natural selection reproduce networks with a long-tailed degree distribution (e.g. (Kim and Marcotte, 2008; Vazquez et al., 2003; Barabasi and Albert, 1999)). This demonstrates that hubs can also arise naturally from small-scale processes and that avoiding disruption of network structure is not necessarily the main mechanism behind the degree distribution that we observe in biological networks. On the other hand, most network growth models that simulate evolution of protein interaction networks do not incorporate gene loss, hence hubs are conserved automatically. We study the effect of incorporating gene loss on network architecture in chapter 5.

1.5 Outline

The assumption that the evolutionary dynamics of a protein family is influenced by the functional relations that it's members engage in, is a central paradigm in this thesis. There are several ways in which this influence may take shape. An extreme case is a functional module in which all components are strongly dependent on each other. If a protein's function is only relevant on the level of the module, selection will act on module level as well, leading to distinct phylogenetic patterns: either the module is completely present or completely absent, the functional module is also an evolutionary module.

Previous studies in prokaryotes demonstrate that evolutionary modularity of potential functional modules is limited and that some types of functional modules evolve more cohesively than others (Snel and Huynen, 2004; Campillos et al., 2006). There are two possible explanations. First of all, the organization of proteins in the cellular machinery need not be very static in evolution. Strong functional ties may be more often broken than previously thought, because the associated decrease in fitness either less severe or more easily overcome. If this would be the case, we would observe a high level of rewiring when we compare networks in different species. Despite the upscaling of experimental techniques, protein networks are rather fragmented, especially in those organisms with large proteomes. The severity of this problem of missing network connections is multiplied when comparing two fragmented networks, hence the level of rewiring in evolution is very difficult to assess. Nevertheless, those studies that estimate rewiring rates, suggest that at least in protein complexes, rewiring does not occur very often (van Dam and Snel, 2008; Shou et al., 2011). As proteins rarely change their complex membership, we explore alternative explanations for disrupted coevolution of members of the same protein complex. For example, our characterizations of the functional ties could be incorrect and/or incomplete, hence proteins that we assume are very much interdependent, are in fact not strongly or not exclusively cooperating.

In chapter 2, we study co-evolution of components of protein complexes and pathways. Do we observe any cohesive evolution of functional modules in eukaryotes? Can we use the protein interaction datasets that recently became available, to explain why some modules evolve cohesively while others don't, or why some components deviate from the rest of the functional module? To what extent can what appears to be evolutionary flexibility, be explained by incorrect module definitions or unreliable protein families? In other words: what are the limitations of our methods to detect evolutionary cohesiveness?

One methodological limitation is our inability to recognize homology between strongly diverged protein sequences. In chapter 3 we investigate whether we can use functional information, in this case co-complex membership, to infer or confirm homology that we normally wouldn't recognize due to extensive sequence divergence. Can we recover subunits that are 'absent' in certain species using conserved functional context and thus restore evolutionary cohesiveness of the protein complex? If so, why have these proteins diverged so much in sequence, whilst keeping a similar functional context?

In the second half of this thesis we switch from top down inference of evolutionary rules from data analyses, to bottom up simulations implementing hypothetical evolutionary rules. We study how specific rules affect the large scale organization of protein networks.

In chapter 4, we investigate the influence of duplication followed by subfunc-

tionalization on how proteins of different ages are distributed in protein-protein interaction networks. We argue that the observed tendency to interact with proteins of a similar age may be due to how paralogs (that belong to the same family and are thus of the same age) are embedded in the network after duplication. What is the influence of interacting paralogs on the age structure of a protein interaction network? Does this pattern arise as a side effect of duplication followed by subfunctionalization?

In chapter 5 we extend our null model of duplication and subfunctionalization and include gene loss. We implement different assumptions on how the dispensability of a protein is reflected in its network properties and study how these assumptions convert into global network architecture. How do networks generated by this more complete null model differ from networks generated by duplication and subfunctionalization without gene loss? How does this depend on our conjectures regarding protein-protein interactions and the dispensability of genes?

In summary, in the first two chapters we focus on how protein evolution depends on its functional context, whereas in the last two chapters we focus on how genomic changes are translated into network changes and how this again depends on functional context.

Chapter 2

Cohesive versus flexible evolution of functional modules in eukaryotes

Like Fokkens and Berend Snel

PLoS Computational Biology, 2009

Abstract

BACKGROUND: Although functionally related proteins can be reliably predicted from phylogenetic profiles, many functional modules do not seem to evolve cohesively according to case studies and systematic analyses in prokaryotes.

METHODOLOGY: In this study we quantify the extent of evolutionary cohesiveness of functional modules in eukaryotes and probe the biological and methodological factors influencing our estimates. We have collected various datasets of protein complexes and pathways in *S. cerevisiae*. We define orthologous groups on 34 eukaryotic genomes and measure the extent of cohesive evolution of sets of orthoogous groups of which members constitute a known complex or pathway. Within this framework it appears that most functional modules evolve flexibly, rather than cohesively.

CONCLUSIONS: Even after correcting for uncertain module definitions and potentially problematic orthologous groups, only 46% of pathways and complexes evolves more cohesively than random modules. This flexibility seems partly coupled to the nature of the functional module as biochemical pathways are generally more cohesively evolving than complexes.

Author's summary

Components of a protein complex or a metabolic pathway strongly cooperate to perform a specific function. Because of this functional interdependence, proteins that form a complex or pathway are expected to be present and absent together in different species. Phylogenetic profiling methods, in which proteins with similar presence and absence patterns are inferred to be functionally linked, are based on this assumption. In this report, we quantify to what extent proteins that together constitute a complex or pathway (a functional module) in yeast are present and absent together (evolve cohesively) in other eukaryotic species. We find that more than half of all complexes and pathways are only partially present in a number of species. It appears that evolution of functional modules is very flexible; components are not indispensable; they can be replaced or reused in a different functional context. This places a limit on how well phylogenetic profiling methods can detect functionally related proteins. Functional modules that evolve cohesively are typically involved in biological processes such as translation and amino acid metabolism.

2.1 Introduction

Phylogenetic profiling is a successful method to predict or confirm functional relations between proteins. If the phylogenetic patterns of two proteins are alike, they are likely to be functionally related (Pellegrini et al., 1999). However, this does not necessarily mean that all functionally related proteins have similar phylogenetic patterns. In depth phylogenetic reconstructions of specific pathways and complexes have yielded a number of examples of complexes and pathways gradually gaining and losing components during evolution (Monahan et al., 2008; Huynen et al., 1999; Gabaldon et al., 2005; Tanaka et al., 2005; Smits et al., 2007; Singh et al., 2008a; Bourbon, 2008). A preponderance of flexible evolution has also been suggested by a number of large scale studies in prokaryotes (Snel and Huynen, 2004; Glazko and Mushegian, 2004; Campillos et al., 2006). Both types of studies thus reveal limited modularity or 'cohesiveness' in evolution of functional modules, showing that the flexibly evolving examples are not an exception.

Recent application of phylogenetic profiling methods on eukaryotes has not been as successful in identifying functional relations as in prokaryotes (Snitkin et al., 2006). This raises the question to what extent, if at all, functional modules evolve cohesively in eukaryotes. The organization of bacterial genomes into operons should facilitate modular evolution of functionally linked proteins. In eukaryotes however, gene order and genome organization are unlikely to play an important role and any modular coevolution would be the result of nongenomic, e.g. system level, properties of the functional module. The study of evolutionary cohesiveness of functional modules in eukaryotes may therefore enable us to shed new light on the way functional organization influences the evolutionary dynamics of the genome and vice versa. The recent availability of a sufficient number of sequenced and assembled genomes across the eukaryotic species tree, as well as the accessibility of high throughput functional data, yield the opportunity to look at possible cohesive evolution in eukaryotes in a systems biological context.

Our aims in this study are twofold: we want to define and quantify evolutionary cohesiveness of functional modules in eukaryotes, and, given this quantification, we want to understand the evolutionary behavior which we observe. In order to meet these goals, we collect a diverse set of functional modules (pathways and complexes). For each module we describe the evolutionary dynamics of its constituents across 34 species from 6 major eukaryotic divisions. We select a measure to determine from the dynamics whether we should consider a module to display cohesive evolution. Once this quantification of the degree of cohesive evolution of functional modules in eukaryotes is established, we are able to compare cohesively with flexibly evolving modules and gain insight in both methodological as well as biological factors which contribute to our result.

Dataset	Number of Modules	Average Module Size
SGD	106	4.56
KEGG	92	14.89
MIPS	199	5.91
Aloy	87	6.95
PE	433	4.37
Socio-affinity	461	11.15
All	1285	8.02
All curated	447	7.51

Table 2.1. Datasets used in this study. The number of modules and the average number of subunits in the modules are listed per dataset, as well as for the nonredundant combination of all datasets ('all') and of all curated datasets ('all curated'). The SGD pathways and KEGG datasets are curated and consist mainly of metabolic pathways. The PE and socio-affinity datasets both result from clustering Tandem Affinity Purification (TAP) data. The differences between these two datasets include the fact that PE clusters are based on raw data from the study by Krogan et al. (Krogan et al., 2006) as well as from Gavin et al. (Gavin et al., 2006), the similarity score (Purification Enrichment versus Socio-affinity) and the algorithm used to cluster the proteins. MIPS and Aloy are two curated complex datasets, the Aloy dataset is a manual selection based on extensive literature curation, information on protein structures and previous TAP derived protein complexes (Aloy et al., 2004). Curated datasets comprise approximately one third of all modules.

2.2 Results

2.2.1 Scoring cohesiveness

We gather 6 datasets containing protein complexes and pathways, defined in *S. cerevisiae*, as our set of functional modules (Table 2.1). In order to measure coevolution of the components of a functional module, we assign all proteins which are part of a module to orthologous groups, based on predefined euKaryotic Orthologous Groups (KOGs) (Tatusov et al., 2003), for all proteins from 34 eukaryotic species (see Materials and Methods), resulting in 214.342 (out of 368.358, >58%) assigned proteins. The (partial) presence or absence of a module in a species depends on whether there are proteins from that species assigned to the orthologous group to which the module components belong (Figure 2.1).

No standard method exists to measure the degree to which a module evolves cohesively. Hence we implement several scoring schemes, both from the literature as well as newly defined. We compare individual modules to a random background in order to decide whether a pattern is the result of evolutionary dynamics or could have been obtained randomly. We adopt the strategy from Campillos et al. (Campillos et al., 2006): for each size N of functional modules, we generate 100.000 random modules by randomly selecting N groups from the set of orthologous groups which are part of at least one functional module. Each

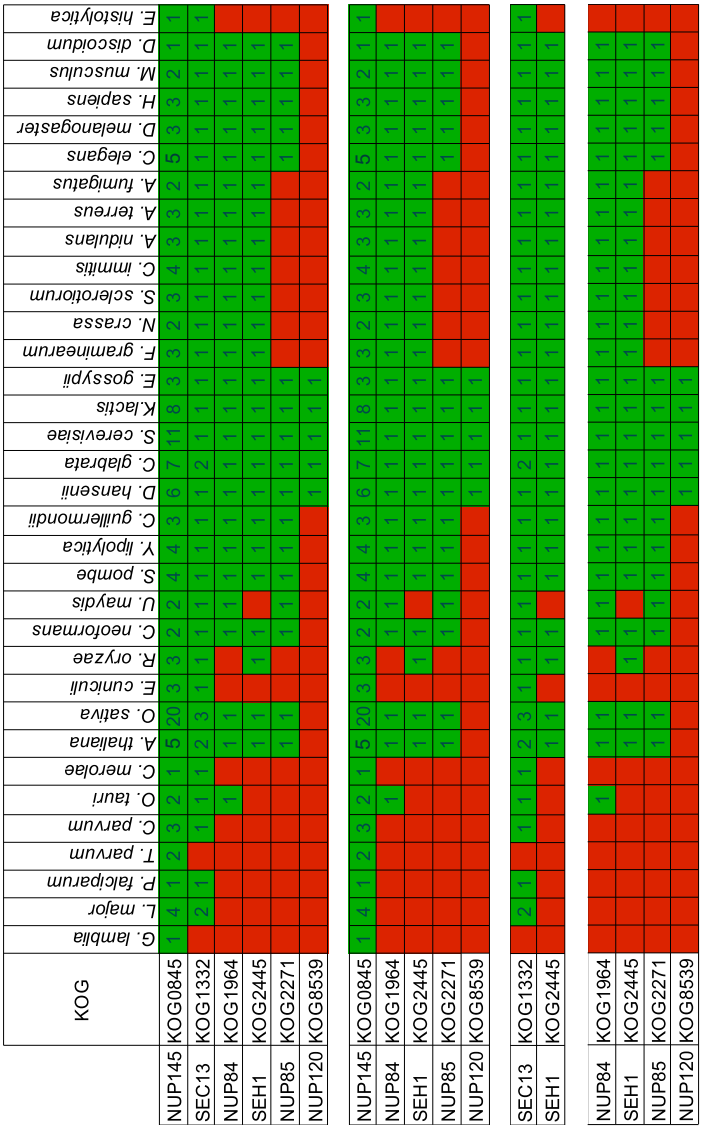

Figure 2.1. Example of a flexibly evolving complex: Nup84 subcomplex of the Nuclear Pore Complex.

Figure 2.1. A) The profile of the Nup84 complex, red indicating absence, green presence (number of paralogs in dark green). The raw score of this complex is (5,0), which means that there are 5 species in which this complex is completely present and none in which this complex is completely absent. The cohesiveness score, which is the fraction of random modules of the same size which score better both in the number of species in which the module is present as well as in the number of species in which the module is absent, is 0.48. This complex from the Aloy dataset occurs also in the MIPS dataset and, with some additional subunits, in the PE and Socio-affinity clusters, so it passes the cross-comparison filter without losing any subunits. B) The profile after cross-comparison with TAP data. SEC13, which is also part of the COPII complex, has the lowest PE score with the other subunits and has a higher propensity to interact with a protein outside the module (namely with SEC31, an other member of the COPII complex) than with any other member of this module. Removal of this protein from the module results in a subcomplex which is not evolving more cohesively than the original module. C) Apart from improving the module definition, we attempt to filter possible noise originating from the use of orthologous groups to describe a modules evolutionary dynamics. KOG0845, KOG1964, KOG2271 and KOG8539 are considered unreliable because they have less than 90% overlap with a orthoMCL derived orthologous group. Removal of those orthologous groups leads to a more cohesively evolving module, with a raw score (24,2) and a cohesiveness score 0.87. D) Removal of orthologous groups which are likely to have functionally differentiated (groups containing many inparalogs, in this example KOG0845 and KOG1332) results in a submodule which we consider evolutionary cohesive: it has a raw score of (5,8) and a cohesiveness score of 0.996. More details on this module and some additional examples can be found in chapters SE1-SE5 of the Supplementary Material†.

functional module is assigned a cohesiveness score defined as the fraction of random modules with a lower 'raw' score. At a cutoff of 0.99, reflecting a probability to obtain a pattern this cohesive by chance of 0.01, we regard a functional module to evolve cohesively.

We observe that regardless of the specific scoring scheme implemented, the majority of functional modules evolve flexibly (Table 2.2). In the remainder of our investigation we use the score which is most successful in separating real from random modules. This turns out to be a two dimensional vector consisting of the number of species in which the module is completely present and the number of species in which the module is completely absent (Figure 2.2). This score identifies 27% of all modules and 37% of all curated modules as cohesively evolving.

An additional merit of this score is that it does not correlate with module size, in contrast to other scores that seem to benefit larger modules (Table 2 in Text S1†). This is linked to a difference between cohesive large and small cohesive modules: manual inspection reveals that large modules typically distinguish themselves from the random background by being completely present in several species, while they're usually never completely absent. Yet small modules distinguish themselves from the random modules by being completely absent in at least a few species.

We carried out the quantification of cohesiveness in eukaryotes and, similarly to what has been observed previously in prokaryotes, we observe that the majority of functional modules evolves flexibly: 27% evolves cohesively on average, ranging from 21%-33% of complexes to 38%-44% of biochemical pathways. There is a host of potential technical and biological reasons for this observation. Are most of our pathways and complexes in fact not functional modules? Is functional modularity defined more appropriately on a different level (domain, protein, network)? Can proteins be functionally related but not co-evolving, because the intrinsic nature of their relationship makes it plastic in evolution? Does the timespan in our orthologous groups allow for so many duplications and subsequent independent losses that the real evolutionary history of the module is obscured?

The effects of these potential causes are difficult to disentangle. Nevertheless, we will attempt to assess the relative importance of module and orthologous group definition in the remainder of this study, in order to get a better estimate of the extent of cohesive evolution of complexes and pathways in eukaryotes. We improve our module definition by cross-comparison of our different datasets and by filtering our modules with data on interactions and cellular locations. Subsequently, we filter out those orthologous groups which are most likely to obfuscate the evolutionary history of a module component. Finally, we will discuss differences in characteristics between cohesively and flexibly evolving modules in order to gain further insights into the why of this observed level of flexibility.

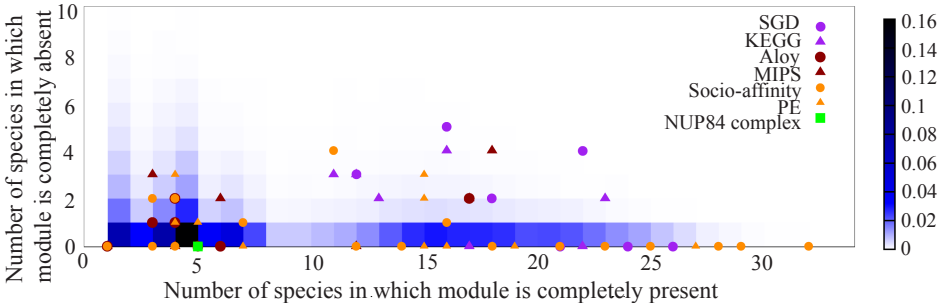

Figure 2.2. Scores and random background. This figure shows the raw scores for modules composed of six subunits from all datasets, with the Nup84 complex from Figure 2.1 highlighted in green. The random background density for all score bins is shown in shades of blue, turning darker as the number of random modules with a score in that particular bin increases.

2.2.2 Effects of module definition

The fractions of cohesive modules per dataset as listed in Table 2.2 reveal a considerable disparity in the degree of evolutionary cohesiveness among datasets when they are different with respect to their underlying concepts. In contrast, results on datasets of the same category ('pathways', 'curated complexes' or 'com-

Dataset	Average Cooccurrence	Average Deviation from Modular	Homogeneous Columns	Species Absent	Species Present	Species Absent, Species Present
SGD	0.14	0.15	0.09	0.06	0.03	0.44
KEGG	0.24	0.24	0.17	0.08	0.16	0.38
MIPS	0.17	0.17	0.15	0.05	0.1	0.33
Aloy	0.21	0.23	0.16	0.02	0.1	0.31
PE	0.08	0.08	0.06	0.03	0.05	0.21
Socio-affinity	0.27	0.3	0.2	0.01	0.19	0.24
All	0.18	0.2	0.14	0.03	0.12	0.27
All curated	0.19	0.19	0.15	0.06	0.1	0.37

Table 2.2. Fraction of cohesive modules for different datasets and different scoring schemes. Average Co-occurrence: for each pair of module subunits we calculate the fraction of species in which both subunits are either present or absent together. We average over all component pairs to obtain a score per module. Average deviation from modular: the sum of the deviation of the number of components of the functional module for each genome to the average number of module components per genome, adopted from Snel et al. (Snel and Huynen, 2004). Homogeneous Columns: the number of species in which a module is either completely present or completely absent, adopted from Gavin et al. (Gavin et al., 2006). Species Present, Species Absent: the number of species in which a module is completely present and the number of species in which the module is completely absent. Those two values together make up the raw score which is used throughout the article.

plexes based on high throughput data') are much more congruent. These results suggest that curated datasets are of better quality compared to high throughput data based module definitions. Hence part of the flexible evolution observed here could be just a matter of poor module definition, as has been suggested previously (Snel and Huynen, 2004). We explicitly test this by applying different filters to enhance our module definition and see whether the level of cohesiveness is increased.

First we find that modules which are defined in multiple datasets tend to evolve more cohesively than modules which are not (P = 8e-06, Table 4a in Text S1†). Second we probed for a functional core that evolves more cohesively by trimming from all functional modules the parts that do not overlap with at least one other module definition. We observe that confirmed submodules evolve more cohesively than the original modules (P = 0.001, Table 4b in Text S1†), especially in those datasets containing large modules. Moreover the fraction of cohesive modules increases from 27% to 36% (Figure 2.3, Table 4c in Text S1†). The primary observation that curated modules seem to evolve more cohesively than modules inferred from high throughput data, is bolstered by an increase in the extent of evolutionary cohesiveness after application of a cross-comparison filter. The combined evidence thus strongly suggests that part of the observed evolutionary flexibility can be attributed to the incorrect definition of functional modules.

A physical interaction often indicates a functional relation, we therefore next combine our module definition with Tandem Affinity Purification (TAP) data, which is the base of our two high-throughput derived module datasets (Gavin et al., 2006; Collins et al., 2007a). We restrict this analysis to complexes, because physical interactions are biologically most relevant in that context and many pathway components (i.e. metabolic enzymes) do not have any interaction partner in our TAP dataset.

Cohesive complexes have a higher mean PE score than flexibly evolving ones, but this observation is biased towards the multitude of complexes that are automatically generated from high throughput interaction data (P = 0.017, Table 4a in Text S1†). If we look at the curated complex datasets separately, results point in a different direction. Much to our own surprise, cohesive modules from curated datasets tend to have a lower mean PE score than flexibly evolving ones from the same dataset (P = 0.001, Table 4a in Text S1†). Similarly, removal of subunits which are most loosely attached to the rest of the complex, has a small and mixed effect on evolutionary cohesiveness (Text S1, Table 4b†).

If we remove subunits which most likely interaction partner is not part of the same module, we observe no significant increase in cohesiveness (Table 4d in Text S1†). A strong interaction of a module component with a protein outside the module apparently does not indicate that this component is not part of the module, or that it has an additional function outside the module. On the contrary:

it may be its function within the module to interact with other parts of the system. These results indicate that a probable physical interaction is neither a necessary nor sufficient condition for a functional relation. Even given the fact that the TAP experiments are not exhaustive with respect to different growth conditions and there are probably many more interactions than those we know about, it is clear that functional relations extend physical ones.

2.2.3 Orthologs and Inparalogs

We have established that reducing the conceptual and technical ambiguity in functional module definition increases the observed evolutionary cohesiveness. Now we test the robustness of our results to the definition of orthologous groups. We run orthoMCL (Li et al., 2003) with default parameters on our set of species. Using this orthology as sole data source, the fraction of cohesive modules and the average cohesiveness scores are qualitatively the same as when we use our KOG-based orthology assignments (Table 6a in Text S1†).

More importantly, we can cross-compare our original KOG-based orthologous groups with the groups defined by the orthoMCL method. If we trust only those orthologous groups with 90% overlap with an orthoMCL group, removing the unreliable orthologous groups results in an increase in cohesiveness, except for the datasets which contain large modules: KEGG and the socio-affinity clusters. The orthologous groups deemed unreliable typically contain more species than the trusted ones ($P = 0.0$). Discarding unreliable orthologous groups means we remove components which are present in many species, which, within our scoring scheme, has more negative impact on the evolutionary cohesiveness of large modules than of small modules. If we compare submodules to original modules we find no significant increase in cohesiveness, except for the datasets derived from high-throughput experiments (Table 6b in Text S1†). However, the overall fraction of cohesiveness increases from 27% to 31% (Figure 2.3, Table 6c in Text S1†), an increase which mainly results from removing modules which consist solely of unreliable KOGs. As was the case with module definitions, we observe that a more conservative definition of orthologous groups results in a higher degree of cohesiveness.

Apart from obvious problems with incorrect assignments, which we tried to tackle by cross filtering with orthoMCL, there are more ways in which the use of orthologous groups to infer presence and absence of module components in different genomes distorts the quantification of cohesive evolution. A module which is completely absent in a certain species could have retained a functionally differentiated recent duplicate of one of its components. In the phylogenetic profile this would correspond to a column of all zeros and a one, while the actual module is completely missing. This phenomenon of functional differentiation is more likely to occur as a family has more duplications and we expect that compon-

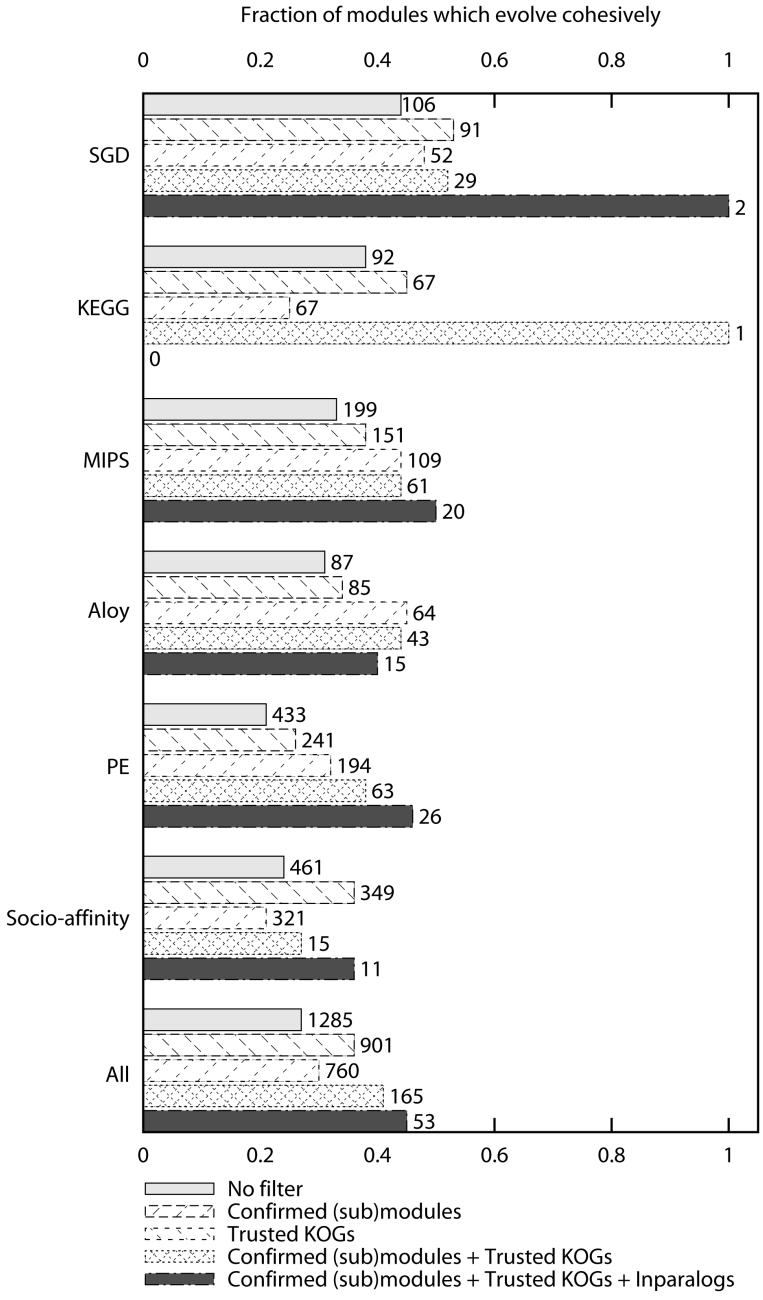

Figure 2.3. (Combined) effect of different filters on the fraction of cohesive modules. On top of each bar we show the number of (sub)modules passing the filter.

ents of cohesive modules are generally assigned to orthologous groups with few inparalogs.

We find that indeed cohesively evolving modules tend to be composed of orthologous groups which contain few inparalogs (P = 5e-07, Table S7a in Text S1†). We adopt the approach described by Snel et al. (Snel and Huynen, 2004) and remove the 50% orthologous groups containing most inparalogs from our datasets. The resulting submodules evolve more cohesively than the original modules (P = 0.02, Table 7b in Text S1†) and the fraction of cohesive modules across all datasets increases from 27% to 33% (Table 7c in Text S1†). Datasets which comprise mainly of large modules do not show an increase in cohesiveness. We can explain this by the fact that large modules are often distinctively cohesive by virtue of being completely present in a large number of species. Removing presence which is possibly but not necessarily spurious, is therefore not likely to increase the measured evolutionary cohesiveness in large modules.

The paralogy filter strongly suggests that on the level of protein families, functional divergence is likely to be one of the factors influencing evolutionary cohesiveness. However, whether this caused by the fact that sometimes a functionally diverged duplicate is present, while a duplicate which retained the original function is lost, or whether it is the case that large families typically are not part of cohesive modules, remains debatable.

We tested multifunctionality on the level of individual proteins by integration of high throughput and literature derived functional information (Text S2†). However, we have not been able to show convincingly that multifunctionality of a protein plays an important role in explaining the observed evolutionary flexibility.

2.2.4 Cohesively versus flexibly evolving functional modules and pathways versus complexes

Given the fact that some modules evolve cohesively and others do not, one of the questions we want to answer is whether, and if so, in what respects cohesively evolving modules are different from flexibly evolving modules. Cohesively evolving modules tend to have a lower rate of sequence evolution (P=0.0009, comparing Dn/Ds rates from (Hirsh et al., 2005) of cohesively versus flexibly evolving modules), reflecting that they're subject to stronger negative selection pressure. As mentioned above, components of cohesively evolving modules tend to duplicate less often than components of flexibly evolving modules. We compared the average propensity of module components to interact with each other between cohesively and flexibly evolving modules. We found to our own surprise, that for the curated complex datasets, components of cohesively evolving complexes actually were less likely to interact among each other than components of

flexibly evolving complexes.

Another interesting question is whether cohesive evolution is more likely to occur in certain biological processes than others. We detect overrepresented Gene Ontology (GO) categories (Ashburner et al., 2000) of proteins in cohesive modules with respect to all proteins in functional modules using the BiNGO plugin in Cytoscape (Shannon et al., 2003). (Figure 8, Tables S9-S11 in Text S1†). Proteins which are part of cohesively evolving modules are involved in core processes: amino acid metabolism, protein ribosome biogenesis, electron transport and generation of precursor metabolites and energy.

It may be the case that modules engaged in these essential processes are not particularly cohesively evolving, but just very conserved. A comparison of the number of species assigned to KOGs containing cohesively evolving module components assigned to these overrepresented GO categories, to a background of all KOGs shows that indeed proteins involved in translation, cytoplasm organization and biogenesis, ribosome biogenesis and assembly are more conserved than the background. In contrast, proteins involved in the other overrepresented core processes such as, for example, amino acid metabolism, are less conserved compared to the background of all module components (Table S9†). This shows that there are in fact modules which do not evolve cohesively only because all components are essential (and therefore conserved). These modules are mainly involved in core metabolic processes.

The overrepresentation of metabolic GO categories among cohesively evolving modules corresponds to a striking difference in cohesiveness observed between datasets containing complexes, and pathway datasets (Table 2.2). Biochemical pathways evolve more cohesively than complexes (P=0.00012 comparing pathways with curated complexes, P<1e-100 comparing pathways to all complexes). In fact, whether a module is a pathway or a complex, is a good predictor for cohesive evolution (Figure 3 and Table 3a in Text S1†). The difference between pathways and complexes is more significant among small modules, which distinguish themselves from the random background by being completely absent in multiple species (Table 3b in Text S1†).

2.3 Discussion

The present study is the first large scale investigation of cohesive evolution of functional modules in eukaryotes. We show similar evolutionary behavior of functional modules in eukaryotes to what is previously observed for prokaryotes: most modules evolve flexibly (Snel and Huynen, 2004; Glazko and Mushegian, 2004; Campillos et al., 2006) and curated modules evolve more cohesively than modules derived from high throughput interaction data (Snel and Huynen, 2004). As

eukaryotes do not contain operons that facilitate the simultaneous loss of module components, all cohesive evolution that we observe is the result of nongenomic properties of the functional module. Hence the system level properties of functional modules are important in the cohesive loss of subunits. Nonetheless a substantial level of flexibility seems resistant to conceptual and technical filtering.

We attempt to estimate the relative importance of mistakes in the definition of functional modules and the use of orthologous groups to determine presence and absence of module components in our set of genomes. We increase reliability, both of our set of functional modules as well as our set of orthologous groups and find that cohesiveness is increased with approximately 30%. Removing orthologous groups which are likely to have functionally differentiated also increases the fraction of cohesive modules with $\sim$30%.

Ideally, we want to overlay all those filters on top of each other, but if we do, we remove so many modules and module components that we are left with less than 13% of our original number of modules and the modules which remain are typically very small (2 or 3 components). Even after application of all these filters we still observe that most functional modules do not evolve more cohesively than random (46% of modules have a cohesiveness score > 0.99)(Figure 2.3).

Naturally, our approach has some limitations in capturing and classifying the diversity in possible evolutionary scenario's illustrated by manually curated examples (Monahan et al., 2008; Huynen et al., 1999; Gabaldon et al., 2005; Tanaka et al., 2005; Smits et al., 2007; Singh et al., 2008a; Bourbon, 2008), (Suppl. Examples SE1-SE5 [†]). The assignment of proteins to orthologous groups is neither exhaustive nor completely correct and not all of the mistakes can be filtered out. Moreover, there are many other ways in which proteins co-evolve (similar rate (Pazos and Valencia, 2001), compensatory mutations (Neher, 1994), coduplication (Cordero et al., 2008)) and our cohesiveness score is restricted to co-occurrence only. These limitations, inherent in a large scale analysis, also apply to the use of phylogenetic profiles to determine functional relations between pairs of proteins. Recent evaluations of phylogenetic profiling methods show that reliable results are obtained at a cost of very low sensitivity, especially in eukaryotes (Jothi et al., 2007). The evaluated methods are more advanced than simply counting co-presence and co-absence. Nevertheless either many functionally related pairs are not detected or many unrelated pairs are being classified as coevolving. Strikingly, the related pairs which can be reliably retrieved belong to functional classes representing those cellular processes, which are fundamental for any cell in any kingdom of life, which corresponds to what we have observed in this study.

Manual reconstructions of the evolution of functional modules, complex purification with missing subunits across different species (Kroiss et al., 2008), as well as previous large scale investigation of evolutionary cohesiveness of functional mod-

ules and the evaluations of phylogenetic profiling methods all point into the same direction. Therefore, even though the exact degree of coevolution is probably underestimated, we conclude that functionally related proteins do not necessarily coevolve, and functional modules do not need to behave as evolutionary modules.

2.4 Methods

2.4.1 Module datasets

We obtained the SGD pathway dataset from the Saccheromyces Genome Database (`ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/biochemical_pathways.tab`), the KEGG pathway datasets from the KEGG website (`ftp://ftp.genome.jp/pub/kegg/pathway/organisms/sce/sce_gene_map.tab` and `ftp://ftp.genome.jp/pub/kegg/pathway/map_title.tab`), the socio-affinity clusters were provided in the Supplementary Information of the publication (Gavin et al., 2006) and the Purification Enrichment clusters were obtained from personal communication. The MIPS dataset was downloaded from `ftp://ftpmips.gsf.de/yeast/catalogues/complexcat` (Mewes et al., 2008) and the Aloy dataset from `www.russell.embl.de/complexes/` (Aloy et al., 2004).

We deleted per dataset the modules of which a submodule was also present in that dataset. We deleted from the pathway datasets those modules which were complexes rather than pathways (SGD pathways: pyruvate dehydrogenase, KEGG pathways: Ribosome, Proteasome, DNA polymerase, RNA polymerase). Modules from which components are assigned to one orthologous group, as well as modules which consist of only one protein or for which we could only map one protein to a systematic ORF name were excluded. Mapping to systematic ORF names was done via the gene registry file from SGD (`ftp://genome-ftp.stanford.edu/pub/yeast/gene_registry/registry.genenames.tab`).

2.4.2 Orthologous groups

Due to dynamics of protein evolution such as protein fusion, protein fission and domain acquisition and loss, defining orthologous groups is a nontrivial task. Therefore we choose a set of well-established, manually curated orthologous groups from the KOG database (Tatusov et al., 2003) as our starting point. A set of 34 eukaryotic species (Figure 2.1), including metazoa, amoebazoa, alveolates, excavata and plantae, is selected based on completeness and quality of annotation of their genomes, yielding a total of 368358 proteins. We perform

all against all Smith Watermann with Paralign (Rognes, 2001) on the protein sequences from the selected species and ran Inparanoid (Remm et al., 2001) with default parameters (except that we used a threshold on the score rather than on the E-value) on this data for each pair of species. Proteins within one Inparanoid cluster which are from different species are connected with an edge, resulting in a graph connecting 237538 proteins. First, we assign 162250 proteins to pre-existing KOGs from the KOG database (Tatusov et al., 2003) with a KOGnitor script. Subsequently, each unassigned protein connected to at least two proteins which are assigned to the same KOG and have an edge between them is assigned to that KOG. This leaves us with 206108 unassigned proteins. In our large graph we identify triangles (trios of interconnected proteins), if a triangle has two components assigned to the same orthologous group, we assign the third component to that group as well. In this way, another 14555 proteins were added to pre-existing KOGs. The remainder of triangles we clustered into 5704 novel orthologous groups using CFinder (Adamcsek et al., 2006), a program which implements the clique percolation method to detect clusters of fully connected subgraphs of different sizes (in this case size three). We have assigned 214342 out of 368358 proteins to a total of 10548 orthologous groups, more than half of which is novel.

We defined an alternative for our orthologous groups by running the orthoMCL program (Li et al., 2003) with default parameters on our set of genomes. We assign a total of 275953 proteins to 40239 orthologous groups.

2.4.3 Module definition filters

For our cross-comparison filter we check for each dataset, for each module, whether there is (complete or partial) overlap with another module in another dataset. If a module is not completely confirmed, we remove unconfirmed subunits such that we keep the largest overlap we have encountered in other datasets.

In order to filter the functional modules with high throughput interaction data, we use the Purification Enrichment (PE) score from (Collins et al., 2007a). This score integrates data from two large scale interaction (TAP/MS) studies ((Gavin et al., 2006) and (Krogan et al., 2006)). Both presence and absence of associations are taken into account to derive a measure denoting the likelihood two proteins directly or indirectly interact (see (Collins et al., 2007a) for further detail). We downloaded these PE scores from `http://interactome-cmp.ucsf.edu/` on February 23, 2008. For our PE score filter, we only consider modules which components have at least one interaction with another protein (within the module or outside) with a confidence score higher than 0.2 (Collins et al., 2007a). We first remove all components with a zero PE score with all other module components and cluster the remaining components with single linkage with -1*PE as distance. We obtain two clusters and remove the smallest cluster.

Acknowledgements

We would like to thank Patrick Kemmeren for providing the PE clusters dataset. We also thank Jos Boekhorst, John van Dam, Otto X. Cordero, Gabino Sanchez-Perez and Radek Szklarczyk for helpful discussions.

Supplementary Material

[†]Supplementary Material for this chapter (Tables S1-S3, text S1-S2) can be found online at `http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371\%2Fjournal.pcbi.1000276#s5`

Chapter 3

Enrichment of homologs in insignificant BLAST hits by co-complex network alignment

Like Fokkens, Sandra MC Botelho, Jos Boekhorst and Berend Snel

BMC Bioinformatics, 2010

Abstract

BACKGROUND: Homology is a crucial concept in comparative genomics. The algorithm probably most widely used for homology detection in comparative genomics, is BLAST. Usually a stringent score cutoff is applied to distinguish putative homologs from possible false positive hits. As a consequence, some BLAST hits are discarded that are in fact homologous.

RESULTS: Analogous to the use of the genomics context in genome alignments, we test whether conserved functional context can be used to select candidate homologs from insignificant BLAST hits. We make a co-complex network alignment between complex subunits in yeast and human and find that proteins with an insignificant BLAST hit that are part of homologous complexes, are likely to be homologous themselves. Further analysis of the distant homologs we recovered using the co-complex network alignment, shows that a large majority of these distant homologs are in fact ancient paralogs.

CONCLUSIONS: Our results show that, even though evolution takes place at the sequence and genome level, co-complex networks can be used as circumstantial evidence to improve confidence in the homology of distantly related sequences.

3.1 Background

Comparative genomics involves large scale investigations to identify which parts of different genomes are of common descent in order to predict function or to study genome evolution. A common first step towards detecting homology between genes or proteins within a genome or between different genomes, is to do a BLAST search with a set of genes or proteins against a database and regard each hit with an E-value below a certain cutoff to be homologous (Altschul et al., 1990). Additionally, several filters and clustering algorithms can be applied to separate sets of homologs into orthologous groups (e.g. (Remm et al., 2001; Li et al., 2003)). Usually, a stringent score cutoff is used to ensure that the hits that are included are indeed homologs. Naturally, homologs whose sequences have diverged strongly, are incorrectly excluded.

On a smaller scale, more sensitive searches based on profiles of groups of related amino acid sequences (such as PSI-BLAST or HMMer) or, if available, protein three dimensional structures are commonly used to avoid False Negatives without losing confidence in the putative homologs returned (Soding, 2005; Altschul et al., 1997). In these searches, instead of using the same scores or probability for each position, a multiple sequence alignment is used to define position specific substitution scores or transition probabilities.

Besides improving sequence based homology searches, one can also use information on the genomic context of sequences to aid detection of a common descent of sequences. Genome alignments can be very useful when there are difficulties in determining homology between sequences, for example between intergenic regions. Boekhorst and Snel showed that genome alignments can be used to select candidates from a set of insignificant BLAST hits in prokaryotes (Boekhorst and Snel, 2007). In eukaryotes, gene order is less conserved across large phylogenomic distances such as between fungi and animals and therefore less likely to make a valuable contribution to the detection of homology at these large evolutionary distances (Koonin, 2009). As a result, conserved synteny is mainly employed in eukaryotes for the detection of orthologs, between closely related species, e.g. within ascomycete fungi or within vertebrates (Byrne and Wolfe, 2005; Wapinski et al., 2007).

The availability of protein interaction networks allows for the comparison of genomes and the functional context simultaneously. Information on the functional context of proteins is already used in comparative genomics of eukaryotes to select from a set of inparalogs, the protein that is functionally similar to the query sequence (the 'functional ortholog') (Espadaler et al., 2008; Singh et al., 2008b; Bandyopadhyay et al., 2006). In the comparative analysis of protein interaction networks, spurious protein interactions can be separated from biologically relevant interactions if the protein-protein interaction occurs in different species. We

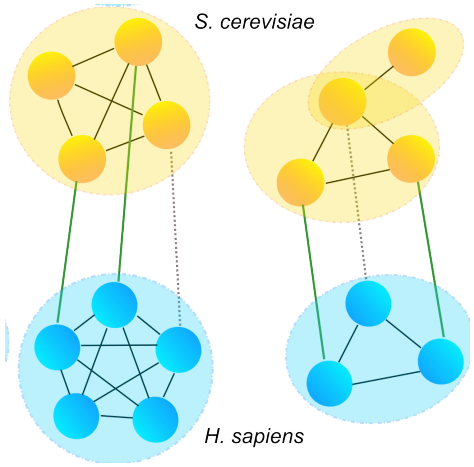

Figure 3.1. Co-complex network alignment and homology inference in insignificant BLAST hits. Green lines: human-yeast unambiguous and readily identifiable orthologs (human and yeast proteins in one Inparanoid cluster), gray dotted line: insignificant BLAST hit. If two proteins with an insignificant BLAST hit are subunits of homologous complexes, are these proteins more likely to be homologous than would follow from the score returned by BLAST?

here test if the reverse is also in principle applicable: can the network alignment help to separate spurious homology links from real ones? Analogously to genome alignment, we test the expectation that the insignificant blast hit between protein a from species A and protein b from species B is more likely to reflect homology if protein a is functionally closely related to proteins which are readily identifiable as orthologous to proteins in species B that are functionally closely related to b (Figure 3.1). To answer this question, we select candidate pairs from BLAST hits between human and yeast proteins based on conserved functional context, in this case homologous complexes, and determine whether this selection contains relatively more homologs than a background of hits with similar BLAST scores.

3.2 Results and discussion

3.2.1 Are hits with conserved functional context more likely to be homologous?

We perform an all-against-all BLAST search between the human and yeast proteomes using a substantially more inclusive threshold than normally is applied to allow a comprehensive survey of insignificant BLAST hits. For each query-hit pair BLAST returns an E-value. We bin the E-values into 8 bins ranging from [E <= 10^{-5}] to [10 < E <= 100]. We define co-complex networks for human

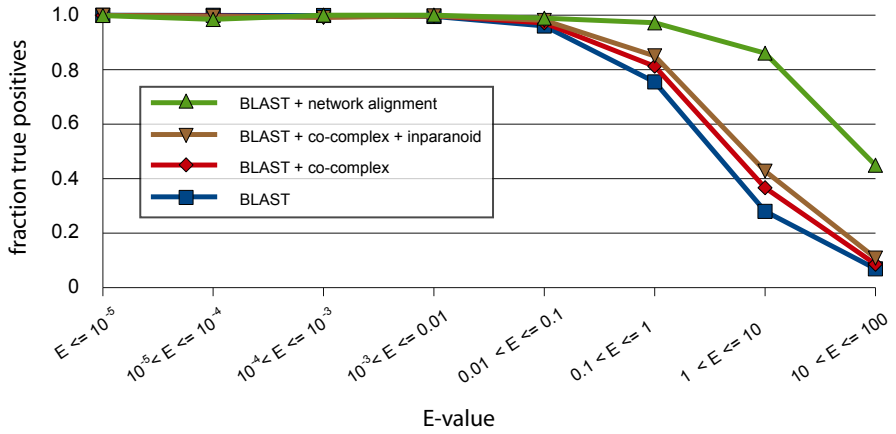

Figure 3.2. Fraction of True Positives for different E-value bins for different subsets of BLAST hits with that E-value. The fraction of True Positives for all BLAST hits ('BLAST', blue line), the BLAST hits for which both the human query as the yeast hit are part of a co-complex network ('BLAST+cocomplex', red line), the BLAST hits for which both the human query as the yeast hit are part of a co-complex network and both have a direct co-complex network neighbour that has a clear ortholog in the other species (is part of a human-yeast Inparanoid cluster) ('BLAST+cocomplex+inparanoid', brown line), the BLAST hits for which both the human query as the yeast hit are part of a co-complex network and both have a direct co-complex network neighbour and these neighbours are clear orthologs of each other (are part of the same human-yeast Inparanoid cluster) ('BLAST+network alignment', green line).

and yeast based on two curated complex datasets per species, and use Inparanoid clusters between human and yeast to align these networks (Figure 3.1 and Methods) (Remm et al., 2001; Matthews et al., 2009; Mewes et al., 2008; Ruepp et al., 2010). If the query and the hit contain a domain which belongs to the same Pfam clan, we consider them to be True Positives. For each of our 8 bins, we calculate the fraction of query-hit pairs which are True Positives, with and without co-complex network alignment.

We find that the use of co-complex information results in a considerable increase in the fraction of true homologs among the returned hits, compared to BLAST without co-complex information (Figure 3.2). This difference is most eminent in bins representing E-values normally considered to be insignificant (the 'gray-zone'). At E-values between 1 and 10 almost 90% of the returned hits share a Pfam clan, which means a substantial, 8 fold increase in the percentage of True Positives. This is not due to a bias resulting from being a member of the co-complex network or being in a conserved region of the co-complex network, as only after alignment of the co-complex network we see a big improvement in the fraction of True Positives (Figure 3.2).

Only a small subset of yeast and human proteins ($\sim$12% of human and $\sim$26% of yeast proteins) is part of a co-complex network within each species. Moreover,

BIN	BLAST		Co-complex network alignment	
	Number of pairs	Number of queries	Number of pairs	Number of queries
$E <= 10^{-5}$	180299	16271	15999	791
$10^{-5} < E \leq 10^{-4}$	19238	6771	102	88
$10^{-4} < E \leq 10^{-3}$	15916	6566	102	86
$10^{-3} < E \leq 0.01$	23882	8626	152	118
$0.01 < E \leq 0.1$	27273	10818	164	122
$0.1 < E \leq 1$	53861	22861	192	155
$1 < E \leq 10$	233649	44138	288	222
$10 < E \leq 100$	1108105	46427	787	495

Table 3.1. Each row contains the number of query-hit pairs and the number of distinct human query proteins in a particular E-value bin, for all BLAST hits and the subset of BLAST hits which falls into the 'co-complex network alignment' category.

many of those are not functionally linked to proteins that have readily identifiable orthologs in the other species. As a consequence, this method is applicable to only a small fraction of query-hit pairs (Table 3.1). If we include high-throughput co-complex datasets for yeast and human, the coverage is increased a little at a cost of a slightly inferior performance (see Additional file 1[†]).

We show that alignment of co-complex networks can facilitate the identification of true homologs among gray zone BLAST hits. In a simple and completely automated procedure, we obtain a subset of hits which, despite very high E-values, is substantially enriched for homologs. This allows us to infer homology for pairs with co-complex network alignment with an E-value ranging between 0.1 and 1 with the similar confidence as for pairs before co-complex network alignment and an E-value of 0.01 (Figure 3.2). Our framework would likely be improved if we could use statistics on (locally) missing connections in both co-complex networks. To date, protein complex datasets are too fragmentary to make any sensible estimates of the number of missing connections.

3.2.2 Detection of missing complex subunits

Previous large scale investigations towards presence and absence of protein complex subunits in prokaryotes and eukaryotes reveal that most complexes are only partially present in other species (Campillos et al., 2006; Fokkens and Snel, 2009; Snel and Huynen, 2004). In these studies, an orthology definition based on BLAST is used to determine presence and absence of subunits in different species and part of the subunits regarded absent may be missing due to detection problems. Hence the disrupted co-evolution of protein complexes might partly be an artefact.

The use of co-complex information is potentially useful in the detection of yeast homologs of subunits of human protein complexes. Especially as the most important disadvantage, the lack of coverage of the co-complex networks, is less urgent because the queries are subunits and hence are all part of the human co-complex network. We take the opportunity to test the applicability of our method to a problem in comparative genomics and assess the added value of co-complex network alignment in detecting homologs in yeast for subunits of human complexes. For the complexes in the CORUM dataset, we initially find homologs for complex components by running a BLAST search with all subunits against the yeast proteome, applying a commonly used E-value cutoff of 0.001. Then, for the subunits we did not find homologs for, we use a less stringent E-value cutoff of 1, in combination with co-complex network alignment, to see how many additional subunits we pick up.

Using BLAST only with an E-value cutoff of 0.001, we find yeast homologs for 1199 out of 1901 (63.07%) subunits. We find that 172 out of 710 complexes

(24.23%) have a homolog in yeast for all subunits, 427 (60.14%) have homologs for some subunits and 111 (15.63%) complexes are completely absent. Even when only comparing two species, we find that for most human complexes, only part of all subunits have a homolog in yeast. However, as we have argued before, some subunits may be called absent due to detection problems.

For the 702 subunits for which we detected no homolog in yeast, we select, using the co-complex network alignment, candidate homologs in yeast for 52 additional subunits, belonging to 62 complexes (some subunits are part of multiple complexes). Using Pfam, CDD and PSI-BLAST, we confirm that the 49 out of 52 candidates recovered with co-complex network alignment are in fact homologous. With the 49 confirmed homologs we retrieved, an additional 19 complexes are completely present in yeast (see Additional file 2[†]).

One striking observation when comparing individual complexes in human to complexes in yeast, is that there is very little congruence between human and yeast complex definitions (see Additional file 3[†]). Factors such as the incompleteness of data in both species, individual decisions on what does belong to a complex and what does not, and proteins belonging to multiple complexes, obscure a one-to-one relation between yeast and human complexes, assuming such a correspondence exists.

Fortunately, because we align human and yeast complexes on a network level rather than as individual complexes, we are able to retrieve homologs with the co-complex network alignment for complexes which do not exist as such in yeast. A good example is the Multisynthetase complex (Figure 3.3). This complex is composed of 8 aminoacyl-tRNA synthetases and 3 auxiliary proteins. The individual tRNA synthetases all have a homolog in yeast found with the initial straightforward BLAST search. The yeast homologs of the tRNA sythetases are not known to be organized in a complex, with one important exception: methionyl and glutamyl synthetases MES1 and GUS1 associate into a complex with ARC1, an auxiliary protein (and homolog of the human auxiliary protein p43, SCYE1) which increases catalytic efficiency and ensures correct localization into the cytoplasm. Via this complex, human JTV1, a scaffold required for the assembly and stability of the multi-tRNA synthetase complex, is linked to a short N-terminal stretch of yeast GUS1, whose C-term is unambiguously homologous to the glutaminyl synthetase in human, QARS (Figure 3.3). When we do a PSI-BLAST with human JTV1 as a query protein, we retrieve GUS1, aligned to the GST_C domain in JTV1 (E-value 1e-05) after three iterations.

We recovered a homolog for another subunit of the Multisynthetase complex via an unrelated complex: the Ribosome. Human EEF1E1 has a hit with yeast EFB1 with an E-value of 0.015. EFB1 is located at the ribosome, as is YHR020W, which is the readily identifiable yeast ortholog of the human bifunctional glutamyl-prolyl tRNA synthetase EPRS. Both EFB1 and EEF1E1 are translation elongation

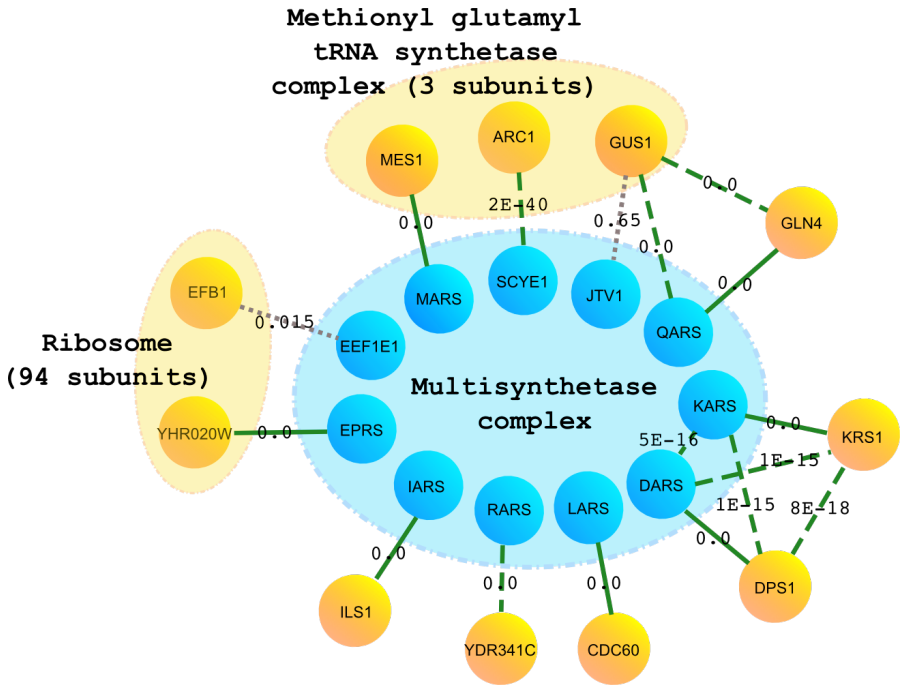

Figure 3.3. The Multisynthetase complex. Yeast homologs were detected for all subunits of the Multisynthetase complex. Green solid lines link proteins which are together in an Inparanoid cluster, green dashed lines indicate a significant BLAST hit between the two proteins linked, gray dashed lines indicate insignificant BLAST hits between proteins for which homology is confirmed by the co-complex network alignment.

factors (EEF1E is a translation elongation factor 1 epsilon and EFB1 a translation elongation factor 1 beta) and both contain a domain which belongs to the GST_C_superfamily. The HSP lies in the regions where the GST_C_superfamily domain lies in both proteins and these regions in the protein sequences are, albeit very distantly, evolutionary related. EFB1 has a much more similar homolog in human (namely EEF1B2, BLAST E-value 1e-33), suggesting that EEF1E1 and EFB are related through a very old duplication event and the translation elongation factor 1 epsilon EEF1E1 ortholog is lost in yeast.

Applying the co-complex network alignment to the set of protein complex subunits in CORUM, we select candidate homologs in yeast for 52 proteins, out of which we could confirm homology with Pfam, CDD or PSI-BLAST for 49 pairs. The observation reported in both large and small scale investigations (Campillos et al., 2006; Fokkens and Snel, 2009; Snel and Huynen, 2004; Gabaldon et al., 2005; Kroiss et al., 2008), that most complexes are 'incomplete' in many species, remains unchallenged because we can only show for a few complexes that their incompleteness is a result of an undetected homology.

3.2.3 Are the recovered distant homologs orthologs?

Exploiting the co-complex network alignment we find yeast homologs for 52 subunits of human complexes that are not revealed by standard BLAST. This is markedly less than the 405 human queries for which co-complex information is applicable (Table 3.1). The likely crucial difference between our initial survey of all BLAST hit pairs and the detection of missing complex subunits is the fact that in the latter, we applied the co-complex alignment only to those query proteins for which we could not find a homolog with BLAST alone. Therefore we expect that many query proteins for which we recover a distant homolog with the co-complex network alignment in the initial survey, have an additional, significant hit in yeast and are therefore not used as a query when looking for additional homologs for complex subunits. Indeed, we find that this is the case for no less than 85% (347 out of 405) of the query-hit pairs with an E-value > 0.01.

There are a few possible evolutionary histories that can explain the fact that for a certain query protein in human, we find a close homolog and a distant homolog with conserved functional context in yeast. First of all, distant homologs recovered by co-complex network alignment could be ancient paralogs (outparalogs with respect to branching of fungi and metazoa), in which the high degree of divergence is due to time rather than rapid sequence evolution (Figure 3.4a). For instance, EEF1E1 and EFB1 in the Multisynthetase example discussed above are ancient paralogs. Another possibility is a more recent duplication in yeast followed by asymmetric divergence in the duplicates, in which case the divergence is caused by accelerated evolution on one branch (Figure 3.4b (Bandyopadhyay et al., 2006; Notebaart et al., 2005). Finally, the two yeast hits may be homo-

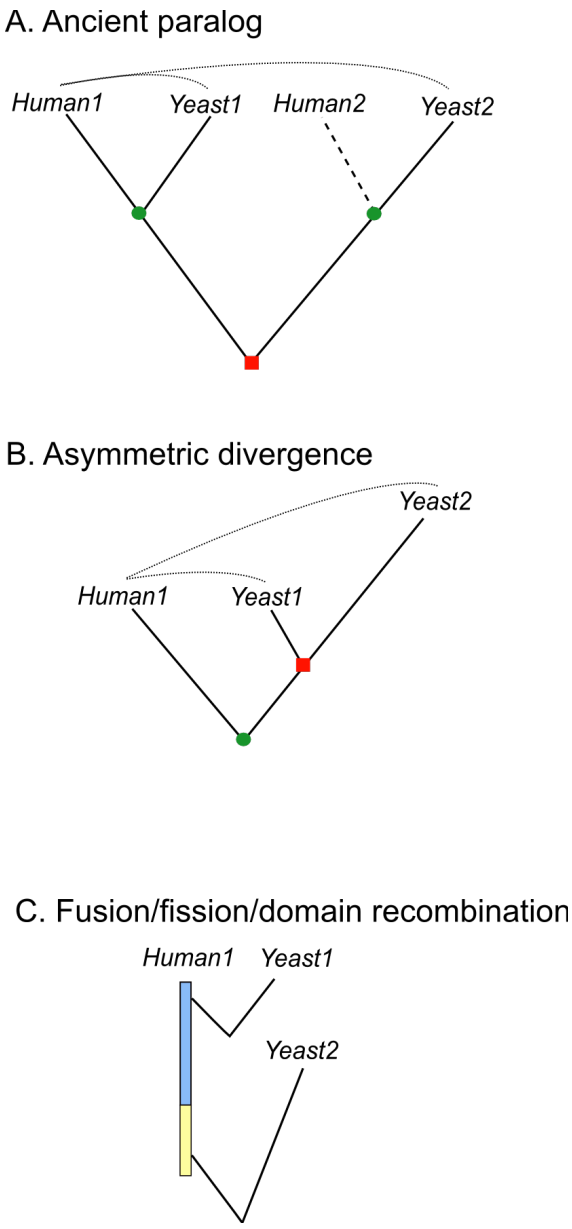

Figure 3.4. Evolutionary histories that explain why for a query protein in human, we find both a close and a distant homolog in yeast. Some proteins for which we recover a distant homolog in yeast with our method, in fact have a better hit (a closer homolog) in yeast. The two scenarios depicted here, 'Ancient paralog' and 'Asymmetric divergence', could both have this effect. We test which scenario occurs more often by looking whether the distant homolog in yeast (Yeast2 in this figure) have a closer homolog in human than Human1. Red square: gene duplication event, green circle: speciation event.

logous to different regions of the query protein due to fusion, fission or domain recombination events (Figure 3.4c), in which one domain/region has a markedly higher rate of sequence evolution than the other.

In the fusion/fission/domain recombination scenario, the two yeast hits of the human query protein in yeast are not homologous. For 29 of our 347 trios the two yeast hits are not a significant hit in BLAST, neither do they share a homologous domain according to Pfam. For 27 of these trios the best scoring BLAST HSP of the two yeast proteins is in a different region in the human protein. In the remaining two pairs, the distant homologs that we retrieve share only a short KOW motif with the query protein, while the best hit shares both the KOW motif (not recognized by Pfam, but part of the HSP) and also the adjacent Ribosomal L27e domain.

If we consider only those 318 trios of proteins in which the two yeast proteins are homologous according to Pfam, we find that in 307 of them, the distant homolog has a significant hit in human, suggesting it is in fact an ancient paralog (Figure 3.4a). A recent study towards the fate of duplicated protein complex subunits showed that 31% of duplicates resides in different complexes, 31% stayed in the same complex and in 38% of the cases one of the duplicates is not known to be part of any complex (Szklarczyk et al., 2008).We investigate the fate of the 307 yeast pairs that are outparalogs according to our analysis. The yeast pairs are not a random sample of ancient yeast duplicates, on the contrary. Because one of the yeast paralogs is a close homolog of a human protein which is part of a complex which is homologous to the complex the other yeast paralog is part of, we expect a bias towards duplicates remaining in the same complex.

We find that for 139 pairs (45.3%), both duplicates are in the same complex, for 25 pairs (8.14%) the duplicates are in overlapping complexes (sharing more than half of their subunits), for 108 pairs (35.2%) they are in a different complexes and for 35 pairs one of the duplicates (the one most closely related to the human query protein) is not known to be part of any complex in yeast. We expect that the human homologs of the 108 pairs in which the yeast ancient duplicates belong to distinct complexes, are more often part of multiple complexes, indicating that the yeast duplicates subfunctionalized. We do observe a significant overrepresentation of proteins that are part of multiple complexes in ancient paralogs when compared to all subunits in the CORUM complex dataset ($P=0.007$), but not in ancient paralogs which are part of the same complex when compared to those which have ended up in distinct complexes ($P=0.58$).

The lion's share of distant homologs we recover using the co-complex network alignment consists of ancient paralogs rather than orthologs. Duplications in general are very important in the evolution of protein complexes (Szklarczyk et al., 2008; Pereira-Leal et al., 2006) and many structures are known to consist of subunits resulting from very old duplications (e.g. the proteasome). We find that

in most cases both duplicates are part of the same or overlapping complexes. This suggests that the duplicates we detect have sub- or neofunctionalized within one complex, although some might be the result of outparalogs that have been independently recruited to a biological process.

3.3 Conclusions

We test whether contextual information from the functional network, in this case conserved co-complex relations, can aid homology detection. Functional context information has been used before to help in choosing functional orthologs from a set of inparalogs, but to our best knowledge, this is the first time functional networks are used to aid distant homology detection. Using an aligned co-complex network, we can identify a subset highly enriched for homologs of BLAST hits with an E-value which would normally be regarded as insignificant. This shows that, even though evolution takes place at the sequence level, one can use co-complex networks as circumstantial evidence to improve confidence in the homology of distantly related sequences.

The interspecies co-complex network includes only a small fraction of all proteins, which impedes applicability. As more high-throughput datasets become available in more species, we expect that the proof of principle we established here, can be applied and tested on a larger scale, between more distantly related species and with other types of functional relations. We apply our co-complex network alignment to a dataset of human complexes in order to determine how many homologous subunits we can detect that we missed in an initial BLAST search. We thereby recovered homologs for only a few additional subunits, despite the fact that coverage is less a limiting factor in this context. We find that one reason we retrieve less additional subunits than expected, is that with the co-complex alignment, we mainly detect outparalogs rather than orthologs.

It has been shown that subunits of a protein complex diverge at similar rates, presumably because subunits of a protein complex are functionally strongly interdependent and subject to very similar evolutionary constraints (Chen and Dokholyan, 2006). In contrast, the co-complex network alignment method is based on the fact that some subunits diverged between human and yeast to such an extent that they are not picked up in a regular BLAST search and other subunits are conserved such that the human and yeast orthologs are still detected by Inparanoid. In this light it is not surprising that most homologs we recover with our method are ancient paralogs rather than orthologs: the difference in the extent of divergence is due to difference in time, as opposed to difference in evolutionary rates between subunits of the same protein complex.

Researchers studying the evolution of individual protein complexes have used

functional information to find diverged homologs successfully despite absence of proof from a large scale study (Boube et al., 2002; Smits et al., 2007). Our results provide this proof. Numerous predictions made in these small scale studies were subsequently confirmed by profile vs profile alignments or the comparison of protein three dimensional structures upon availability. Interestingly, many of these predictions represent initial BLAST hits with E-values even higher than 100 (the cutoff used in this study). Hence it is possible that in our study still many homologs have gone undetected, and the formal integration of functional context with more sensitive homology detection methods might help in the development of automatic bioinformatic methods to uncover these distant homologs and improve our insights into the ancient evolution of the protein interaction network.

3.4 Methods

3.4.1 Co-complex network

To construct a human co-complex network, we download the set of CORUM Core complexes from `http://mips.gsf.de/genre/proj/corum` and stored the complexes as sets of co-complex pairs (Ruepp et al., 2010). We added 'direct complex' pairs downloaded from Reactome `http://www.reactome.org/download/current/homo_sapiens.interactions.txt.gz` (Matthews et al., 2009), which, in combination with pairs from the CORUM dataset, results in a co-complex network containing 32415 unique pairs in total. For the yeast co-complex network, we stored MIPS complexes from `ftp://ftpmips.gsf.de/yeast/catalogues/complexcat` as binary co-complex relationships (Mewes et al., 2008) and complexes from SGD GO cellular component annotation (Ashburner et al., 2000) as in (Szklarczyk et al., 2008). This resulted in 20075 unique pairs in total.

3.4.2 BLAST and Pfam

We downloaded 46704 human protein sequences from Ensembl (Hubbard et al., 2005), (Homo_sapiens.NCBI36.50.pep.all.fa) and yeast protein sequences from the Saccharomyces Genome Database (orf_trans_all.fasta) in July 2008. We run BLAST between human and yeast with the maximum returned E-value set to 100, maximum number of hits and alignments set such that it is no limiting factor (Altschul et al., 1990). We did not adjust the database size. If two proteins have multiple HSPs (regions aligned by BLAST), we keep only the HSP (High Scoring Sequence Pair) with the lowest E-value. We downloaded Pfam HMMs (version 23, July 2008) and data on homologous Pfam families (Pfam clans) from the Pfam website (`ftp://ftp.sanger.ac.uk/pub/databases/Pfam`), searched

for domains in human and yeast proteins using hmmpfam in the HMMer package (Bateman et al., 2004) with default cutoffs.

For each BLAST hit, for both the human query protein and the yeast hit, we determine the overlap between the HSP and each Pfam domain and divide the number of amino acids in the overlap with the length of the shortest region (either Pfam domain or HSP) to get a percentage of overlap. If a query and a hit have a Pfam domain that belongs to the same clan and the overlap of the domain and the HSP is greater than 50%, we call this BLAST hit a True Positive. BLAST hits for which this overlap is less than or equal to 50% in either the human query or the yeast target protein are ignored as the gold standard for homology (Pfam clans) can't be fully applied to these proteins.

3.4.3 Co-complex network alignment

To align the co-complex networks of yeast and human we look for yeast orthologs for all proteins in the human co-complex network using Inparanoid. We run Inparanoid 3.0 with default parameters, so for each bidirectional best hit which forms a seed pair for an Inparanoid cluster, it is required that the minimum BLAST bitscore is 50 and the overlap of the alignment relative to the shortest of the two proteins is at least 50% (Remm et al., 2001).

For each BLAST query-hit pair, if the human query protein has at least one direct neighbour in the human co-complex network that is orthologous (in one Inparanoid cluster) to the direct neighbour of the yeast protein in the yeast co-complex network (Figure 3.1), we assign this pair to the 'co-complex network alignment' category (Figure 3.2). We bin E-values in 8 bins ranging from [$E \leq 10^{-5}$] to [$10 < E \leq 100$] and calculate for each bin the percentage of True Positives (hits that each have a Pfam domain belonging to the same clan), also known as the Positive Predictive Value. Each bin contains at least 60 pairs and at least 50 query proteins (Table 3.1). Normalizing for family size gives similar results (see Additional file 4[†]).

3.4.4 Detection of missing complex subunits

To avoid biases due to overlapping complexes as much as possible, we removed 803 complexes which are a subcomplex of another complex from the set of CORUM Core complexes. If we remove all supercomplexes instead of all subcomplexes, we get qualitatively the same results. We first attempt to find a yeast homolog with BLAST and an E-value cutoff of 0.001 for all subunits. Subsequently, on those subunits we did not find a homolog for, we applied the co-complex network alignment with an adjusted E-value of 1 (expected percentage

of False Positives < 3% (Figure 3.2)).

Acknowledgements

This work was cofinanced by the Netherlands BioInformatics Centre (NBIC) which is part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research. Sandra Botelho is supported by a grant from the Fundação para a Ciência e Tecnologia, SFRH\BD\33212\2007.

Authors' contributions

BS conceived the study and assisted in writing the manuscript. LF, SB and JB performed the analysis and LF wrote the manuscript. All authors have read and approved the manuscript.

Supplementary Material

†Supplementary Material for this chapter (Additional files 1-3) can be found online at `http://www.biomedcentral.com/1471-2105/11/86/additional`

Chapter 4

Gene duplications contribute to the overrepresentation of interactions between proteins of a similar age

Like Fokkens, Paulien Hogeweg and Berend Snel

BMC Evolutionary Biology, 2012

Abstract

BACKGROUND: The study of biological networks and how they have evolved is fundamental to our understanding of the cell. By investigating how proteins of different ages are connected in the protein interaction network, one can infer how that network has expanded in evolution, without the need for explicit reconstruction of ancestral networks. Studies that implement this approach show that proteins are often connected to proteins of a similar age, suggesting a simultaneous emergence of interacting proteins. There are several theories explaining this phenomenon, but despite the importance of gene duplication in genome evolution, none consider protein family dynamics as a contributing factor.

RESULTS: In an *S. cerevisiae* protein interaction network we investigate to what extent edges that arise from duplication events contribute to the observed tendency to interact with proteins of a similar age. We find that part of this tendency is explained by interactions between paralogs. Age is usually defined on the level of protein families, rather than individual proteins, hence paralogs have the same age. The major contribution however, is from interaction partners that are shared between paralogs. These interactions have most likely been conserved after a duplication event. To investigate to what extent a nearly neutral process of network growth can explain these results, we adjust a well-studied network growth model to incorporate protein families. Our model shows that the number of edges between paralogs can be amplified by subsequent duplication events, thus explaining the overrepresentation of interparalog edges in the data. The fact that interaction partners shared by paralogs are often of the same age as the paralogs does not arise naturally from our model and needs further investigation.

CONCLUSION: We amend previous theories that explain why proteins of a similar age prefer to interact by demonstrating that this observation can be partially explained by gene duplication events. There is an ongoing debate on whether the protein interaction network is predominantly shaped by duplication and subfunctionalization or whether network rewiring is most important. Our analyses of *S. cerevisiae* protein interaction networks demonstrate that duplications have influenced at least one property of the protein interaction network: how proteins of different ages are connected.

4.1 Background

The wealth of sequence data from a wide range of species, has allowed for large-scale studies of genome evolution and detailed reconstruction of the parts lists of our earliest ancestors (Koonin, 2010; Glansdorff et al., 2008). The study of network evolution not only requires these detailed 'parts lists', but also information on how these parts are assembled into a molecular machinery in different organisms. Despite the progress in both the generation of large-scale functional data in multiple organisms, as well as the inference of functional relations from sequence data, the overlap in functional networks in different species is typically very small. The reconstruction of ancestral networks on a scale that would allow for general statements on network evolution is not yet possible (Ali and Deane, 2010).

Previous studies attempt to circumvent this problem by assigning an age to proteins in an *S. cerevisiae* protein interaction network, assuming that patterns of connectivity between proteins of different ages offer a glimpse on how the network changed over time. A recurrent observation in these studies is the simultaneous emergence of interacting proteins (Qin et al., 2003; Kim and Marcotte, 2008; Capra et al., 2010). To date, two distinct theories have been put forward to explain this phenomenon. Multiple interacting proteins, added to the network at the same time, may be more likely to be functional and therefore under positive selection (Qin et al., 2003). Alternatively, a tendency to interact with proteins of similar age can arise as a side effect of a neutral network expansion process in which new proteins are added to network peripheries while old proteins are mainly located at network cores (Kim and Marcotte, 2008). In this work, we amend both these explanations by demonstrating that gene duplication events contribute to the overrepresentation of interactions between proteins of similar age.

Protein age, as defined by the taxonomic distribution of the family it belongs to, is assumed to correspond to a time frame in which the protein was 'added' to the network (Qin et al., 2003; Kim and Marcotte, 2008; Capra et al., 2010). However, few genes in *S. cerevisiae*'s genome and thus in its protein interaction network have emerged absolutely de novo. Most genes are the result of either small scale or whole genome duplications, replacing an ancestral gene by two daughter genes.

In the classical view of functional divergence after gene duplication, one of the daughters keeps the ancestral function while the other is free to evolve an entirely new function (neofunctionalization) (Stoltzfus, 1999). On the network level, this would indeed correspond to a node being 'added' to the network (namely the node evolving a new function), but the protein evolving a new function, unless it is not recognized as a homolog, belongs to the same family as its paralog and thus has the same 'age' by definition. Thus, even if network evolution can be

considered as a process in which new nodes are simply 'added' to the network, the age of a protein does not correspond to the time frame of emergence in the network.

Moreover, neofunctionalization is not the only possible scenario of divergence after duplication (Bergthorsson et al., 2007; Francino, 2005; Innan and Kondrashov, 2010; Force et al., 1999; Bershtein and Tawfik, 2008; Levasseur and Pontarotti, 2011). For example, duplicate genes are preserved in the genome to achieve a dosage increase (Szklarczyk et al., 2008) or daughter genes both perform part of the ancestral function (Lynch and Force, 2000) (subfunctionalization). These processes cannot be modeled by 'adding' proteins to a network. If, due to network rewiring, genome and network evolution would be completely independent, we would expect paralogs to behave like random pairs in the network. On the other hand, if gene duplication events leave an imprint on the network, we would expect paralogs to share more interactions partners than non-paralogs, reminiscent of their initial complete redundancy. Indeed, even if the vast majority of paralog pairs does not share any interaction partners, the relative overlap in interaction partners of paralogs is higher than of pairs belonging to different families (Musso et al., 2007; AIMC, 2011).

Here, we investigate the influence of gene duplication events on the age structure of *S. cerevisiae* protein interaction networks. We find that interparalog interactions account for a small part of the overrepresentation of interactions between proteins of a similar age. Intriguingly, we find another, unexpected effect of gene duplications on the age structure of the network. It turns out that the major contribution to the observation that proteins interact with proteins of a similar age is from interaction partners that are shared by paralogs, mostly likely an ancestral interaction that is preserved after duplication. We investigate whether this result can occur as a side effect of neutral network growth by duplication and divergence, and find that our simple model can only replicate an overrepresentation of interparalog edges, not the conservation of edges with proteins of the same age after duplication.

4.2 Results and discussion

We perform an in depth analysis on the effect of gene duplications on age structure in an *S. cerevisiae* literature curated protein-protein interaction network (PIN) (Reguly et al., 2006), consisting of 3268 nodes and 12058 edges. We assign an age to the 2476 nodes that belong to a known protein family (Muller et al., 2010), based on the taxonomic distribution of this family. We use the work by Kim and Marcotte as an anchor point and group proteins into the same 4 age categories they use, ranging from families that have members from all three kingdoms (Archaea, Bacteria and Eukaryotes, named ABE), those with members from

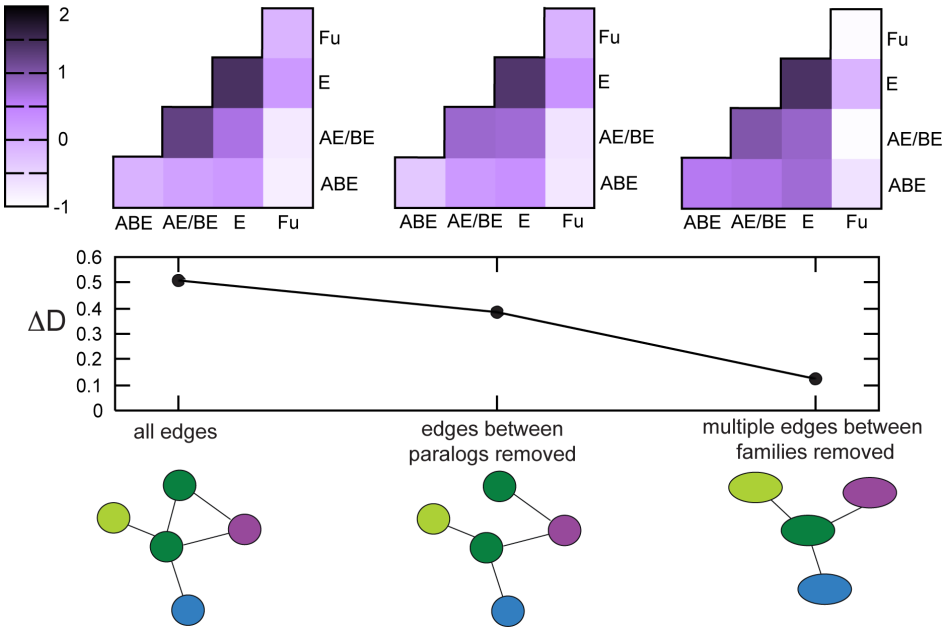

Figure 4.1. ΔD and interaction densities between age groups in the original and collapsed protein interaction network. We calculate the ΔD value and the normalized interaction densities for an *S. cerevisiae* literature curated network, using the taxonomic distribution of EggNOG orthologous groups to determine protein age. Round network nodes correspond to proteins whereas oval network nodes correspond to protein families. Different colors indicate different families. If we remove all edges between paralogs, the ΔD value decreases and the normalized interaction densities for this network show that the strongest effect is in interactions between proteins of age AE/BE. It turns out that mainly interactions between homologous components of the Spliceosome have been removed in this age category. We then continue to remove all edges that are redundant on a protein family level, thus collapsing the network into a network where nodes are no longer proteins, but protein families. We show that ΔD is decreased dramatically in this network. There is neither a specific family nor age group overrepresented among the nodes that have edges removed.

only two kingdoms (AE/BE) to Eukaryote- (E) and Fungal- (Fu) specific families (E) (Kim and Marcotte, 2008). Moreover, we use their method to calculate normalized interaction densities between different age groups and implement the statistic they propose, ΔD, to measure age-dependence among these interaction densities (see Methods and (Kim and Marcotte, 2008) for further detail). A positive value of ΔD indicates a higher than expected connectivity between proteins of similar age categories. The literature curated PIN has a ΔD value of 0.51 (Figure 4.1). In addition to ΔD we define a new measure to quantify interaction densities between age groups and the potential gradient in these densities. Results using this alternative measure ΔD_{new} are discussed in the last section of the results and discussion.

4.2.1 No evidence for artifacts in the data causing the observed interaction preference among proteins.

The tendency to interact with proteins of a similar age has been reported by several independent studies, each using a different PIN, different families to infer age and different levels of granularity in age categories. However, we need to be as sure as possible that this phenomenon is not caused by any artifacts in the data. To correct for possible biases in the literature curated PIN, we do the same analyses on 3 other networks, one based on Y2H (Yu et al., 2008), one on TAP/MS (Collins et al., 2007a; Gavin et al., 2006; Krogan et al., 2006) data and a combination of both techniques (HTP network from (Kim and Marcotte, 2008), see Methods for more detail), and find ΔD values ranging from 0.48 to 0.63 (Table S1†). The relatively small overlap in interactions of these networks (Figure S2†) indicates they sample different portions of the underlying real PIN (Yu et al., 2008), though of course some of the interactions that occur in only a single network are False Positives. Interactions between abundant proteins are likely to be overrepresented in all of these networks (Ivanic et al., 2009). We compare the abundance of proteins in the different PINs to a background distribution of all proteins for which abundance was measured (Table S3†, data obtained from (Ghaemmaghami et al., 2003)). We find that only networks including interactions based on TAP/MS data differ significantly from the background. To ensure the interaction preference among proteins of a similar age is not limited to abundant proteins and thus not representative of the underlying complete interaction network, we remove the 10, 50, 100, 500 and 1000 most abundant proteins and recalculate ΔD. We find that removal of the most abundant proteins does not lead to a decrease in ΔD (Table S4†) and conclude that interaction preference among proteins of a similar age is not limited to abundant proteins. Similarly, we determine which functional categories assigned to protein families are overrepresented in the different networks, remove all proteins from the categories and find again that ΔD does not decrease (Table S5†).

We experiment using different age groups representing various other intervals on the species tree and find that ΔD does not depend on the specific age categories ABE, AE/BE, E and Fu (Table S6†). Because our definition of age is dependent on the taxonomic distribution of a protein family, we expect that slowly evolving protein families, as their members are recognized across more distant species, tend to be older (Wolf et al., 2009). Indeed, if we compare the distribution of Dn/Ds ratios (Wall et al., 2005) among the different age categories, we find faster sequence evolution for young proteins (Figure S7†). Interacting proteins are under similar evolutionary constraints and tend to have similar rates of evolution (Hakes et al., 2007; Lovell and Robertson, 2010; Pazos and Valencia, 2008), thus the overrepresentation of interactions between proteins of a similar age could be a side-effect of the correlation of protein age with evolutionary rate. If the observed ΔD value would depend on similar rates of sequence evolution rather than on similar age, we would expect that if we bin proteins according to their Dn/Ds ratio (as if this was their age), ΔD for these categories would exceed ΔD based on age groups. In contrast, we find that if we calculate ΔD based on evolutionary rate, it is -0.05 while for this network (a subnetwork of the original network as not all proteins in the network are assigned a Dn/Ds ratio), ΔD based on age groups is 0.54 (Table S8†). Even though protein age correlates with the rate of sequence evolution, the latter is not the determining factor in the interaction preference among proteins of a similar age. In conclusion, we have found no evidence that a positive ΔD value is caused by certain biases in the data.

4.2.2 Interactions between paralogs play a minor role in the interaction preference among proteins of a similar age

Several studies investigating the evolution of protein complexes revealed that they often originate from duplications of genes encoding self-interacting proteins (Pereira-Leal et al., 2007, 2006; Pereira-Leal and Teichmann, 2005; Veretnik et al., 2009; Liu et al., 2009). On a network level, this would result in clusters of interacting proteins of the same age (Figure S9†). Interparalog interactions are thus a possible explanation for the 'simultaneous emergence' of interacting proteins. Of the 7210 interactions between proteins that belong to a known family, 430 are interactions between paralogs ($\simeq$ 6%), belonging to 107 different families (Table S10†). Of these 430 interparalog edges, 258 are interactions between members of the same protein complex ($\simeq$ 60%).

Even though they comprise only a small fraction of all the edges in the PIN, interactions between paralogs are more abundant than one would expect given the size distribution of protein families and even the age structure of the network ($P<10^{-4}$, 100000 random redistributions of family labels over nodes). Family labels are only shuffled within the same age category to preserve ΔD. If we remove all interparalog edges from the network, we reduce the network to 3228 nodes

and 11628 edges (in the original network, 40 proteins interact only with family members. In this reduced network they have no edges, and therefore they are removed) and ΔD decreases with approximately 24% to a value of 0.38 (Figure 4.1). This value of ΔD is still significantly higher than random ($P<10^{-4}$, randomization by redistributing family-labels over the network without interparalog edges 100000 times), indicating that growth of functional modules (e.g. protein complexes) by duplication of subunits only accounts for part of the overrepresentation of interactions between proteins of a similar age.

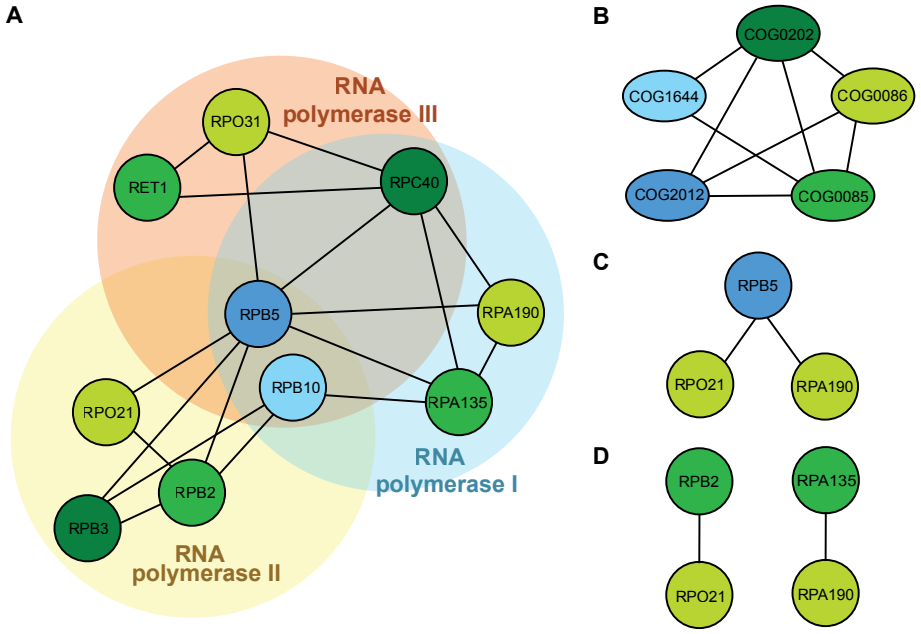

Figure 4.2. Example of network collapse into protein families. We show here part of the Literature Curated *S. cerevisiae* protein interaction network, involving interactions between certain components of RNA polymerases. a. The sub network in its original state. Nodes are colored according to their age (green: ABE, blue: AE/BE), individual families are in different shades of the same color. b. The sub network collapsed into protein families. c. Duplication in a single family: two edges that share a node (RBP5) d. Duplications in both families: non-overlapping edges (RPB2-RPO21, RPA135-RPA190, RET1-RPO31)

Interaction partners *shared by* paralogs play a major role in the interaction preference among proteins of a similar age

Gene duplications do not only influence the PIN by generating interactions between paralogs. Ancestral interactions with other proteins, if conserved in both paralogs after duplication, can also alter network topology. In this specific liter-

ature curated PIN, the relative overlap of interaction partners between paralogs is significantly higher than of pairs belonging to different families ($P \simeq 0.0$, Table S11†). This overlap does not necessarily affect the age structure of the network. Interestingly, we find that the interaction partners shared by paralogs are more often of the same age as the paralogs, than the interaction partners they don't share ($P < 4.6e\text{-}17$, Table S12†), indicating that duplication of protein interactions can also contribute to a positive ΔD.

In order to reduce the effect of interaction conservation after duplication events on ΔD, we collapse the network into connected families (see Figure 4.2 for an example of collapsing network of interacting proteins into a network of interacting families). Interestingly, ΔD decreases to 0.12, a value that is not significantly different from random ($P \simeq 0.1$, randomization by redistributing family-labels over the collapsed network 100000 times, Figure 4.1). In addition to intrafamily edges, we have removed all edges that occur multiple times between family pairs (Table S13†). The decrease in ΔD shows that families of a similar age often have multiple edges connecting their members.

The most likely scenario (requiring the smallest number of evolutionary events) in which gene duplication generates additional edges between two families, is when a member A of one family duplicates and both daughters A' and A" keep the ancestral interaction with the protein B from the other family. The two edges representing these interactions overlap as both contain the protein B. For example, RPB5 (COG2012) and RPB10 (COG1644) are RNA polymerase subunits common to all three polymerases and are connected to different members of COG0085 and COG0086 (Figure 4.2c). On the other hand, if proteins from both families duplicate, the edges representing the interactions do not necessarily overlap: e.g. if A' interacts with B' and A" interacts with B" and A' does not interact with B" and A" does not interact with B'. This scenario occurs in the RNA polymerases as well: proteins RPA135, RPB2 and RET1, members of COG0085, are connected to RPA190, RPO23 and RPO31 respectively, members of COG0086, and part of RNA Polymerase I, II and III (Young, 1991; Cornelissen et al., 1988; Gabrielsen and Sentenac, 1991) (Figure 4.2d).

For each family-pair that occurs multiple times in the network (i.e. multiple edges exist between members of these families), we calculate the fraction of protein-pairs that is overlapping. We find that for 80% of the families, all protein-pairs overlap (A'-B and A"-B, Table S14), suggesting that the amplification of the number of interactions between proteins of a similar age occurs mainly through asymmetric expansion rather than duplication and reuse of small functional modules. Interestingly though, if both families are of the same age, this fraction is much lower (65%). However, there is a strong bias towards pairs of old families, suggesting gradual duplication of functional modules (given more time, duplication of additional subunits is more likely), rather than duplication of entire functional modules at a time (Pereira-Leal and Teichmann, 2005).

4.2.3 Age-dependent interaction densities in an extended Duplication-Divergence model

The results described above demonstrate that gene duplications contribute strongly to the observed interaction preference among proteins of a similar age. First of all, interparalog edges explain part of the overrepresentation of edges among proteins of a similar age in the network, suggesting a role for functional module growth by duplication of subunits. The major contribution in most networks however, is from the conservation of ancestral interactions with proteins of a similar age. Are these interactions preferentially conserved? In other words, if the ancestral protein interacted with some proteins that are older and some proteins that are of the same age, do the daughter genes after duplication typically lose the interactions with the older proteins and do they tend to keep those interactions with the proteins of the same age? Or is a small bias in the number of interactions with proteins of a similar age of the ancestral protein, amplified by subsequent duplication events? In other words, does natural selection play a role or does this phenomenon arise as a side effect of network growth by duplication and divergence?

Due to limited availability in protein interaction data in different species, the direct inference of ancestral protein interaction networks and subsequent evolutionary events is primarily anecdotal (Ali and Deane, 2010). Therefore we prefer to use a network growth model to directly test some of our assumptions on network evolution. First and foremost, we want to establish whether conservation of interactions with proteins of a similar age after duplication, arises as a trivial side effect of neutral network growth. We adjust a well-studied and simple model of network growth by node duplication (Vazquez et al., 2003), which we will refer to as the Duplication-Divergence (DD) model, to accommodate family relations between nodes. In the model, we use families to define the age of a node and to calculate statistics on paralogs in order to compare them to those obtained from the data.

The model is initialized with a fully connected graph of 4 nodes, which are of 4 different families but have the same age. When a randomly selected node is duplicated, both copies are connected to the same nodes to which the ancestral node was connected. Duplication is followed directly by a rapid subfuctionalization process: for each ancestral neighbor, we delete its edge with one of the two daughter nodes with a probability q. During the subfunctionalization steps, it is possible to favor one of the two daughter nodes when deleting an edge (parameter s), leading to systematic asymmetric divergence (Wagner, 2002; Kim and Yi, 2006; Evlampiev and Isambert, 2007). In our extension of the original model, with a probability a, this subfunctionalization process is accompanied by drastic changes in sequence, leading to one of the paralogs founding a new family (i.e. not recognizable as a paralog). Otherwise, both paralogs belong to the same family and thus have the same age. With a probability p a new connection is formed

between the duplicates, analogous to e.g. a homodimer becoming a heterodimer (see Methods and Figure S15† for more detail).

Previously published results show that a DD model without the implementation of protein families can only yield networks with a negative ΔD value, i.e. networks in which nodes mostly interact with nodes of a different age (Kim and Marcotte, 2008). This is because in the DD model, nodes with a high degree have a larger probability to connect to a new node, by duplication of one of their neighbors. Since in network growth models in general, old nodes have more neighbors than young nodes (simply because they have had more time to gain edges) this results in old nodes preferably gaining a new edge. In a model without family relations between nodes, one of the twin nodes will always be assigned a new age after duplication and therefore the edges gained will be mainly connecting an old node to a young node.

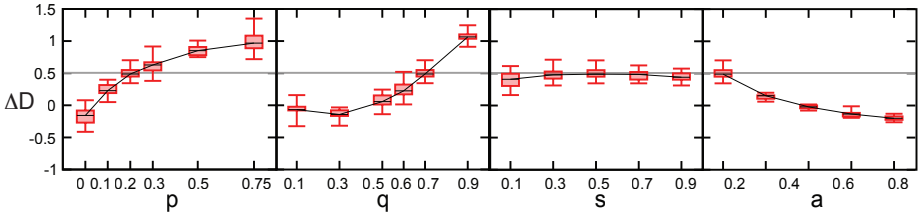

Figure 4.3. ΔD values for the extended DD model under different parameter conditions. Increase in values of p and q and decrease of a lead to higher ΔD values in the network, divergence symmetry s has very little effect. Default parameter conditions are p=0.2, q=0.7, s=0.5, a=0.2, each plot shows ΔD values when one of these parameters is varied while the others are kept at default values (for full parameter sweeps we refer to Figure S16†). The gray line is the ΔD value of the yeast LC PIN. Boxes show the.25 and.75 percentile of 20 runs, the error bars show the extreme values and the black line is the mean of 20 runs.

Using our implementation of protein age, that is more congruent with the bioinformatic data analysis, we systematically study the DD model by running it under many different parameter conditions. We find that in our extended model a positive ΔD value is possible under parameter conditions that have been shown to yield networks that are topologically similar to yeast protein interaction networks (Kim and Marcotte, 2008). Given a low probability of founding a new family (parameter a), a high level of divergence after duplication (parameter q) and a relatively high probability of a connection between twin nodes after duplication (parameter p) our model yields networks with a ΔD value that is comparable to that of a yeast PIN (Figure 4.3, Figure S16†).

These specific parameter conditions lead a high number of interparalog interactions in the network: due to high divergence after duplication the relative contribution of novel edges between twin nodes is higher (Figure 4.4, Figure S17†). In the DD model the positive ΔD value hinges on interactions between members of the same family, as is also illustrated by the high interaction densities between

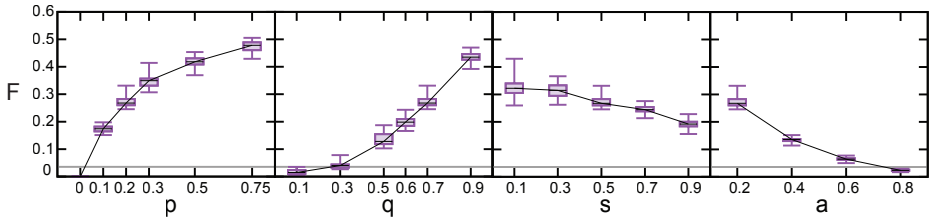

Figure 4.4. The fraction of interparalog edges for the extended DD model under different parameter conditions. As is the case for ΔD values, an increase in values of p and q and decrease of a lead to higher fraction of edges that connect paralogs in the network, divergence symmetry s has very little effect. Note that low values of p already have a large impact on the fraction of edges connecting paralogs: with p=0.2 and a=0.2, the probability of gaining an edge between daughter nodes after a duplication event in the network is $p^*(1\text{-}a)$ = 0.16, yet this small probability of an edge gain between daughter nodes after duplication leads to ±30% of all edges connecting paralogs. Default parameter conditions are p=0.2, q=0.7, s=0.5, a=0.2, each plot shows the fraction of interparalog edges when one of these parameters is varied while the others are kept at default values (for full parameter sweeps we refer to Figure S18†). The gray line is the fraction of edges that connect paralogs of the yeast LC PIN. Boxes show the .25 and .75 percentile of 20 runs, the error bars show the extreme values and the black line is the mean of 20 runs.

proteins of the exact same age (Figure S18†). If we remove interparalog edges from a model network, ΔD decreases below zero and if we collapse the model networks into networks of protein families, ΔD decreases even further (Figure S19†). In the data we do not observe such a preference for young families to interact with old families.

We gain two important insights from the extended DD model. First of all, we find that the number of interparalog edges in networks produced by the model is much higher than one might expect based on the values of p and a alone. It turns out that only a small fraction (0-2%, depending on parameter conditions) of interparalog edges in the model stem directly from the gain of an interaction between two daughter nodes immediately after a duplication event. After a duplication event in which an edge is gained between daughter nodes, this new edge can be propagated in the network through subsequent duplication of these daughter nodes. Importantly, this demonstrates that the effect of relatively rare events on network topology can be amplified in networks that grow by duplication and subfunctionalization of nodes. Moreover, this mechanism indicates that previous estimates of the degree of neofunctionalization after duplication that are based on the overrepresentation of interactions between paralogs are likely to be too high (Wagner, 2003; Gibson and Goldberg, 2009). Secondly, despite the fact that conservation of ancestral interactions is more likely to occur under these parameter conditions (Figure S20†), we find that low levels of functional divergence alone do not lead to a higher ΔD value (Figure 4.3, Figure S21†). This indicates that an overrepresentation of edges between proteins of a similar age, due to conservation of ancestral interactions in both duplicates, does not arise

automatically from a process of network growth by node duplication such as we modeled here.

4.2.4 Alternative measures for age-dependence in interaction densities.

In (Kim and Marcotte, 2008) the number of interactions between members of two age groups is normalized with respect to the number of nodes in each age group (representing the maximum number of edges that is possible between these age groups (see Methods)). As a consequence, age groups with low connectivity in general have lower interaction densities (Figure 4.5). Moreover, ΔD is sensitive to random removal of nodes or edges from the network: it declines as more nodes or edges are removed (Figure S22†), while random removal of nodes and edges should not affect the overall tendency to interact with nodes of a similar age in the network.

We define an alternative measure for the tendency to interact with proteins of a similar age, ΔD_{new}, based on interaction densities normalized by the age groups connectivity (see Methods for more detail). This new measure neither reflects differences in connectivity for different age groups (Figure 4.5) nor does it scale with the number of nodes or edges in the network (Figure S22†). We reperform all of our analyses using ΔD_{new} instead of ΔD. We find ΔD_{new} values ranging from 0.35 to 0.56 (Table S1) indicating that the interaction preference among proteins of similar age is neither due to artifacts in the measure of interaction density nor to the measure of the gradient in interaction densities. We test how ΔD_{new} depends on possible biases in the data, such as protein abundance, overrepresented functional categories, evolutionary rate and the choice of age groups and find that, like ΔD, these biases do not affect the positive value of ΔD_{new} (Table S4, Table S5, Table S6 and Table S8 †).

If we remove interparalog edges we find that ΔD_{new} is decreased for all networks (Table S1, Figure S23 †). If we collapse our networks into networks of protein families (Figure 4.1), we find that ΔD_{new} decreases in 3 out of 4 networks (Table S1, Figure S23 †). If we compare the ΔD_{new} values to those of randomized networks (randomization by redistributing family labels over the network), we find that ΔD_{new} is not significantly different from random networks for two out of 4 networks: the Y2H and the TAP network (Table S1†). In the model networks, there is little difference between ΔD_{new} and ΔD (Figure S24†). In conclusion, we have factored out age group properties confounding the previous definition of interaction densities, namely size and connectivity, and find that our results remain largely unchanged.

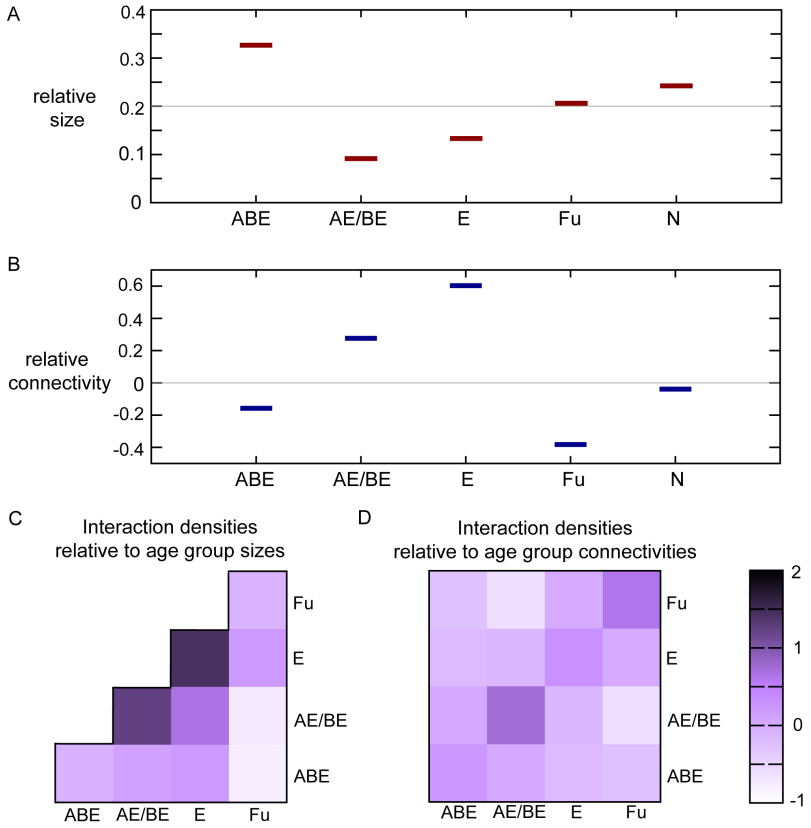

Figure 4.5. Interaction densities depend on size and connectivity of age groups. Interaction densities depend on age group sizes: the connectivity between age groups is normalized by the maximum possible connectivity: the number of connections if all members from the age groups would be interacting. This density is then again normalized by the interaction density of the entire network to allow for comparison between networks of different edges densities (see Methods for more detail). Age groups that are small and/or have low general connectivity will have low interaction densities with any age group. Interaction densities thus do not only represent a relative connectivity between a pair of age groups (or within one age group), it also reflects the overall connectivity as well as the size for each individual age group. A. Relative size for each age group in the LC network: the proportion of nodes that belong to this age group. The grey line denotes the relative size if all age groups would be of the same size. ABE and Fu are the largest groups. N is the group of proteins that are not assigned to an eggNOG family. B. Relative connectivity for each age group in the LC network, calculated as follows: $\log_2$ (avg(degree age group) / avg(degree network)). The grey line denotes the relative connectivity if it would have been the same for all age groups. Age groups ABE and Fu have a relatively low degree. N is the group of proteins that are not assigned to an eggNOG family. C. Interaction densities between age groups in the LC network: age groups ABE and Fu have in general low interaction densities with each age group reflecting their large size and low connectivity rather than a specific relation between two age groups. D. An improved method of calculating interaction densities: connectivity normalized by expected connectivity (see Methods for more detail). These densities are independent of the age groups sizes or degree and represent only a specific property of a pair of age groups. Our alternative measure ΔD_{new} is calculated based on these interaction densities.

4.3 Conclusion

Several studies relate the age of a protein to how that protein is embedded in the molecular machinery (Dalmolin et al., 2011; Waterhouse et al., 2011; Kunin et al., 2004; Capra et al., 2010; Warnefors and Eyre-Walker, 2011; Eisenberg and Levanon, 2003). In order to use this information to understand the evolution of the molecular machinery, one needs a clear conception of what 'age' actually is. Concerns regarding potential biases in protein age defined through the taxonomic distribution of detected homologs, have been raised before (Wolf et al., 2009; Saeed and Deane, 2006). This is important because an incorrect understanding of protein age can lead to premature conclusions on network evolution. For example, the observation that old proteins tend to have more interactions has been proposed as evidence supporting the Preferential Attachment model of network evolution (Barabasi and Albert, 1999; Eisenberg and Levanon, 2003), but a slow rate of sequence evolution as well as a low propensity for gene loss have been associated with increased connectivity (Koonin and Wolf, 2006; Wolf et al., 2008; Park and Choi, 2009; Wall et al., 2005), which would be an alternative explanation.

We test whether the overrepresentation of interactions between proteins of a similar age can be explained by biases in both the genomic as well as the functional data and find that this is not the case. In contrast, interactions between paralogs as well as interaction partners shared by paralogs account for part of the tendency to interact with proteins of a similar age. The fact that interaction partners that are shared by paralogs are more often of the same age has not been previously reported. The most parsimonious evolutionary scenario explaining the fact that two paralogs share an interaction partner, is one in which the pre-duplication ancestor of the two paralogs had an interaction with another protein and this ancestral interaction was conserved in both daughter nodes after duplication. An initial small bias to interact with proteins of a similar age could have been amplified by duplication events.

We test this hypothesis in a network growth model, which is initialized with a fully connected network of 4 nodes of the same age. We find evidence that amplification through duplication is possible in the case of interparalog edges: novel edges between paralogs are created at a low rate but because of subsequent duplications of these interactions, creating these novel edges can have a profound effect on network topology. If duplication and the conservation of ancestral interactions with proteins of a similar age would be sufficient to generate an interaction preference among proteins of a similar age we expect it to emerge from this model. The fact that our model can only explain the part of the interaction preference among proteins of a similar age that is caused by interacting paralogs, suggests that future work should be directed at identification of additional important factors. For example, our model neither implements de novo gene in-

vention or interaction gain, network rewiring, nor gene loss. Moreover, protein interaction networks tend to include several types of interactions, ranging from phosphorylation to possibly indirect interactions of proteins that belong to the same complex. In summary: our analyses of protein interaction data suggest an important role for gene duplications in the preference to interact with proteins of a similar age. Yet results from our model indicate that a process of duplication and subfunctionalization alone does not explain the preference to interact with proteins of a similar age we observe in *S. cerevisiae* protein interaction networks.

4.4 Methods

4.4.1 Protein families and protein age in protein interaction networks.

The literature curated (LC) network and the network based on Y2H data combined with TAP/MS data (HTP network) were taken from the Supplementary Material of the paper by Kim and Marcotte (Kim and Marcotte, 2008). The literature curated network was based on data from BioGRID (Reguly et al., 2006). Interactions that were only supported by high throughput data were removed, as well as all protein-RNA interactions, interactions supported only by co-localization or co-fractionation or data from (Collins et al., 2007a; Ptacek et al., 2005; Grandi et al., 2002) were excluded. The HTP network was created by compiling data from (Ito et al., 2001; Uetz et al., 2000; Gavin et al., 2006, 2002; Krogan et al., 2006; Ho et al., 2002), including only those interactions that have been supported by more than one study (studies (Gavin et al., 2006) and (Gavin et al., 2002) were counted as one). From both the LC and the HTP network ribosomal proteins were excluded (see original paper (Kim and Marcotte, 2008) for more detail). The network based on Y2H data (Yu et al., 2008) was downloaded from `interactome.dfci.harvard.edu/S_cerevisiae/download/Y2H_union.txt`. We construct a binary TAP/MS network by using PE scores calculated by (Collins et al., 2007a) based on data from (Gavin et al., 2006; Krogan et al., 2006) and a PE score cutoff of 0.2. PE scores were downloaded from `http://interactome-cmp.ucsf.edu`.

We want to investigate the effect of gene duplications on the tendency to interact with proteins of a similar age and to avoid unnecessary complications we use protein families (as in (Qin et al., 2003; Capra et al., 2010)) rather than domain families (as in (Kim and Marcotte, 2008)) to define the age of a protein. We download EggNOG orthologous groups (Muller et al., 2010) (COG.NOG) from `ftp://eggnog.embl.de/eggNOG/2.0/` and assign an age to each group based on the species distribution in this group. EggNOG uses NCBI taxonomy identifiers for its species, we use NCBI taxonomy (nodes.dmp in taxdump.tar.gz, downloaded from `ftp://ftp.ncbi.nih.gov/pub/taxonomy/`) to determine which in-

ternal node in the species tree provided by EggNOG corresponds to the ancestor of e.g. all Bacteria, gathered all leaves that were under that internal node and scanned the EggNOG families for the presence of species with these NCBI taxonomy identifiers. Two identifiers have changed in NCBI: 5058 in EggNOG is 746128 in NCBI and 382253 in EggNOG is 434922 in NCBI. If a group only contains fungal proteins, it was assumed to have been invented in Fungi and was assigned the age Fu, if it consists of Eukaryotes only it was assigned the age E. If a group contains at least one protein from either Bacteria or Archaea it was assumed to have emerged in the ancestor shared by Archaea and Eukaryotes or in the First Eukaryotic Common Ancestor (assuming that proteins that are present in Bacteria and Eukaryotes only result from an endosymbiosis event leading to the mitochondrion) and was assigned the age AE/BE. If a group contains at least one protein from Bacteria and at least one from Archaea it was assumed to have been present in Last Universal Common Ancestor and was assigned the age ABE. If we implement age categories based on other intervals on the species tree, ΔD range between 0.6 and 0.79 (Table S6†).

All data (e.g. abundance, age, complex membership, etc.) on individual proteins used in this study is provided in Table S25†. For complex membership, we use the list of yeast proteins assigned to different GO macromolecular complexes obtained from `http://www.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl`.

Interaction preference with proteins of a similar age: ΔD and ΔD_{new}. We use the metric described in, (Kim and Marcotte, 2008) for each pair of age groups m and n, the normalized interaction density is calculated as follows:

$$D_{m,n} = log_2 \frac{I_{m,n}/E_{m,n}}{2L/(N(N-1))}$$

where
$I_{m,n}$ = the number of edges observed between age groups m and n
$E_{m,n}$ = the maximum number of edges that is possible between age groups m and n, and is calculated as follows:
within an age group: $E_{m,n} = N_n(N_n - 1))/2$
between different age groups: $E_{m,n} = N_n \times N_m$
N_m, N_n = the number of nodes with age m, n respectively
N = the total number of nodes
L = the total number of edges.

To compare these interaction densities, we calculate the average interaction density gradient ΔD.

$$\Delta D = \frac{\sum_{n=2}^{G} \sum_{m<n} D_{m+1,n} - D_{m,n}}{G(G-1)/2} \qquad (1 \leq m < n \leq G)$$

where G is the number of age groups (4 in this study). These equations are equal to those in the original paper by Kim and Marcotte (Kim and Marcotte, 2008).

The interaction densities, as calculated by Kim and Marcotte, are normalized with respect to the number of nodes in the age groups. However, the connectivity differs quite strongly per age group. For example, the fungal specific proteins are not as densely connected as older proteins hence interaction densities between the Fu age group and all other age groups are typically low (Figure 4.1). If we want a low density to correspond to an underrepresentation of interactions between two specific age groups we should consider normalizing by the number of interactions we would expect based on the connectivity of the two age groups rather than their size. We therefore define alternative interaction densities in which we divide the frequency of observing an interaction between proteins of age group X and age group Y by the expected frequency of observing this interaction. For example, the LC network has 12058 edges, 7210 of which are between proteins that are assigned to a protein family and thus have an age. This corresponds to 2*7210 = 14420 'edge ends', of which 5561 are occupied by a protein of age 'ABE': there are 1256 edges between two proteins that both have age 'ABE' and 3049 edges between a protein of age 'ABE' and a protein of a different age (2*1256 + 3049 = 5561). The observed frequency of edges between proteins that both belong to age group 'ABE' equals 2512/14420 $\simeq$ 0.174, while the expected frequency equals (5561/14420)2 $\simeq$ 0.149. The normalized interaction density between 'ABE' and 'ABE' is $\log_2$(0.174/0.149) = 0.23.

The original measure ΔD was calculated based on only part of the differences in densities between pairs of age groups. For our measure ΔD_{new} we use all the differences between pairs of age groups to quantify the gradient in our new set of interaction densities:

$$\Delta D_{new} = \frac{\sum_{n=2}^{G}\sum_{m<n}^{G-1} D_{m+1,n} - D_{m,n} + \sum_{n=1}^{G-1}\sum_{m>n}^{G} D_{n,m-1} - D_{n,m}}{G^2}$$

where G is the number of age groups (4 in this study) and $D_{m,n}$ is the interaction density between age groups m and n normalized by the expected interaction density as described above.

4.4.2 Network growth model

We implement the extended Duplication Divergence model using the Igraph package to represent graphs. We initialize the model with a fully connected graph consisting of 4 nodes. The seed graph does affect network topology in the DD model (Hormozdiari et al., 2007), we choose a seed graph similar to the one used in (Kim and Marcotte, 2008). We want to focus on the effect of implement-

ing protein families rather than other topological characteristics such as the shape of the degree distribution, etcetera. If the DD model is initialized with this graph it can produce networks with topological characteristics similar to *S. cerevisiae* PINs (Kim and Marcotte, 2008). These nodes in thise seed graph all belong to different families, but these families do have the same age. We initialize the families and ages in the model as such because we want to test whether an initial interaction preference for proteins of a similar age will be amplified through a process of duplication and subfunctionalization.

At the end of a model run, when the network reached its target size of 3000 nodes, we group the different ages into 4 different groups, trying to keep the 4 groups of approximately the same size, if possible, to avoid large variance in ΔD due to sparsely populated age groups. Keeping the sizes of the age groups similar to those observed in the data has little effect on either ΔD or ΔD_{new} (Figure S26†). In the data $\sim$20% of proteins have no age, leading to a lower fraction of edges that connect paralogs. We randomly select 600 nodes from the model network and designate them as nodes without an age in order to rule this out as the major contributor to the difference in the percentage of connected paralogs. We find that ΔD remains very similar and that the fraction of edges that connect paralogs is decreased but still a lot higher than in the data (Figure S27†).

Each iteration a random node X is selected and duplicated with all of its edges, resulting in nodes A and B. If a random number between 0 and 1, is lower than a, daughter node A (identical to B) is assigned a new age, while B still has the ancestral age (we assume one node needs to perform part of the ancestral function). Then, for each interaction partner Y of A and B, if a random number between 0 and 1 is lower than q, the interaction between either A and Y or B and Y is deleted. If a random number is lower than s, we delete the interaction between A and Y, otherwise we delete the interaction between B and Y. This means that if $s > 0.5$, the interaction with the daughter node that can be assigned a new age is more likely to be deleted (the node that diverges faster in sequence, also loses more interactions). Finally, we draw a random number and if this number is lower than p, we create a new edge between A and B.

During the subfunctionalization process, it is possible for a node to lose all of its edges. In this case the node will be deleted and the network remains unchanged except for the fact that the node that 'duplicated' may have a new age.

Acknowledgements

The authors would like to thank Wouter van Veldhoven for his help in starting up the project, prof. dr. Wan Kyu Kim for sharing his code on the DD model and calculating ΔD, Kirsten ten Tusscher and Michael Seidl for their comments on the manuscript and Jan Kees van Amerongen for technical support.

Author's contributions

PH and BS conceived of the study and assisted in writing the manuscript, LF participated in the design of the study, performed the analyses and wrote the manuscript. All authors read and approved the manuscript.

Supplementary Material

[†]Supplementary Material for this chapter can be found online at `http://www.biomedcentral.com/1471-2148/12/99/additional`

Chapter 5

Consequences of gene loss demonstrate the importance of gain of novel interactions in network evolution

Like Fokkens, Paulien Hogeweg and Berend Snel

in preparation

Abstract

BACKGROUND:Most network growth models assume that networks expand in evolution while comparative genomics studies indicate that this is unlikely. Here, we adapt a well-studied model of network growth by duplication and subfunctionalization, called the Duplication-Divergence model. We add rules regarding gene loss and network rewiring, independent of duplication events, and study how each of these processes affect the global structure of the simulated networks.

RESULTS:We find that unless networks expand, duplication followed by subfunctionalization jeopardizes maintaining network integrity on longer timescales. We incorporate gain of novel interactions to compensate for decrease in connectivity due to gene loss. We find that, of all the different models researched here, the model implementing duplication followed by subfunctionalization, random gene loss and preferential interaction gain with a node with a high degree most accurately reproduces topology observed in biological protein interaction networks. Under biologically relevant conditions, duplication followed by subfunctionalization alone is not sufficient to explain large-scale architecture of protein-protein interaction networks and rewiring is needed to prevent network breakup.

CONCLUSION: We use network growth models to study how processes on the level of individual proteins affect global network architecture. This approach enhances our understanding of the topological consequences of our assumptions regarding the retention and loss of genes.

5.1 Introduction

Large-scale protein-protein interaction networks (PINs) offer an unprecedented view into the organization of the cellular machinery in *S. cerevisiae* (Gavin et al., 2006; Batada et al., 2006; Yu et al., 2008). "Nothing in biology makes sense except in the light of evolution" (Dobzhansky, 1964) hence the pending question is whether a PIN has certain characteristics because they offer a selective advantage or because they arise as a side-effect of small-scale evolutionary processes. The long-tailed degree distribution in PINs is associated with increased evolutionary robustness and the modular organization of the network may improve it's adaptability. On the other hand, simple models simulating network growth without applying selection on the network as a whole, are able to generate networks whose architecture resembles that of real PINs (e.g. (Barabasi and Albert, 1999; Watts and Strogatz, 1998; Vazquez et al., 2003; Kim and Marcotte, 2008). This demonstrates that PINs do not necessarily have a particular architecture because of a selective advantage: network properties could also arise from small-scale, local evolutionary processes.

There is a large variety among network growth models: they simulate processes ranging from gene duplication (Vazquez et al., 2003) and network rewiring (Watts and Strogatz, 1998; Berg et al., 2004) to the invention of novel genes (Kim and Marcotte, 2008). The common denominator in all these models is network expansion. Individual gene families expand, while other families decrease in size or are lost completely from a genome. The number of genes in *S. cerevisiae* is comparable to the estimated number of genes in the reconstructed Last Eukaryotic Common Ancestor (Koonin, 2010). This calls into question whether the assumption that gene repertoires and protein networks expand, is valid. Gene loss is obviously important in shaping a genome (Snel et al., 2002; Krylov et al., 2003; Wapinski et al., 2007). A model that simulates network evolution should include this process.

We incorporate gene loss into the Duplication Divergence (DD) model, in which gene duplication is followed by rapid subfunctionalization. We select this particular model for various reasons. First of all, the DD model is well studied and produces networks that share multiple topological properties with real PINs (Vazquez et al., 2003; Sole and Valverde, 2008; Hormozdiari et al., 2007). Secondly, gene duplications play an important role in the evolution of gene content (Snel et al., 2002; Krylov et al., 2003; Wapinski et al., 2007). This model allows us to investigate a specific theory on the retention of duplicate genes in the genome, namely the hypothesis that both copies subfunctionalize by gathering complementary, degenerative mutations (Force et al., 1999). Finally, paralogs share more interactions partners than random protein pairs, even if they result from an ancient duplication event (Musso et al., 2007; Pereira-Leal et al., 2007; Navlakha and Kingsford, 2011; AIMC, 2011; Fokkens et al., 2012). This demon-

strates that gene duplication events leave a trace in the network that models that do not incorporate gene duplication, will not reproduce.

We balance duplication of network nodes with removal of network nodes, thus maintaining a stable gene repertoire size throughout a large number of iterations. We test two different hypotheses regarding gene loss. As a null model, we select nodes to be removed from the network randomly. We compare the resulting networks with those generated by a model in which we select a node for deletion with a probability that is inversely proportional to it's degree, assuming that a protein's connectivity correlates with it's dispensability. We seek to answer the following questions: (i) whether a more complete null model, including gene loss, can reproduce characteristics of PIN topology, (ii) which network properties are most strongly affected by gene loss and (iii) how different implementations of gene loss affect network architecture.

We monitor how networks change over time via a wide range of network statistics, including the degree distribution, clustering coefficient, modularity, average path length and the size distribution of network components. The use of multiple, complementary observables provides insight into how the evolutionary processes that we simulate, affect network topology. Moreover, certain network properties, such as a long-tailed degree distribution, have proven to be rather generic: many distinct network growth models produce networks with this characteristic. The use of multiple observables allows us to distinguish between different models and pinpoint their specific strengths and weaknesses. We analyze model networks against a background of four *S. cerevisiae* PINs, representing different types of experimental data, as well as different methods to compile this data into a network. Comparing multiple PINs allows us to identify any network-specific biases and estimate which network properties are more variable than others, providing for each observable a range of 'correct' values rather than just a single number.

We find that some network properties that are reproduced by the original DD model, strongly depend on network expansion. For example, maintaining network integrity and retaining highly connected hubs is nearly impossible unless networks are growing. After duplication both daughter nodes inherit complementary parts of their ancestor's interactions, hence the number of times that a protein can duplicate and subfunctionalize, depends on its number of connections. This limits the long-term viability of networks that only gain new interactions through duplication.

We extend the model by including the gain of entirely novel interactions, independent of duplication events. Again, we implement two different hypotheses: we compare a null model in which interactions are gained by randomly selected nodes, with a model in which nodes with a high degree preferably gain new interactions. We find that the latter model generates networks that are most similar to real PINs.

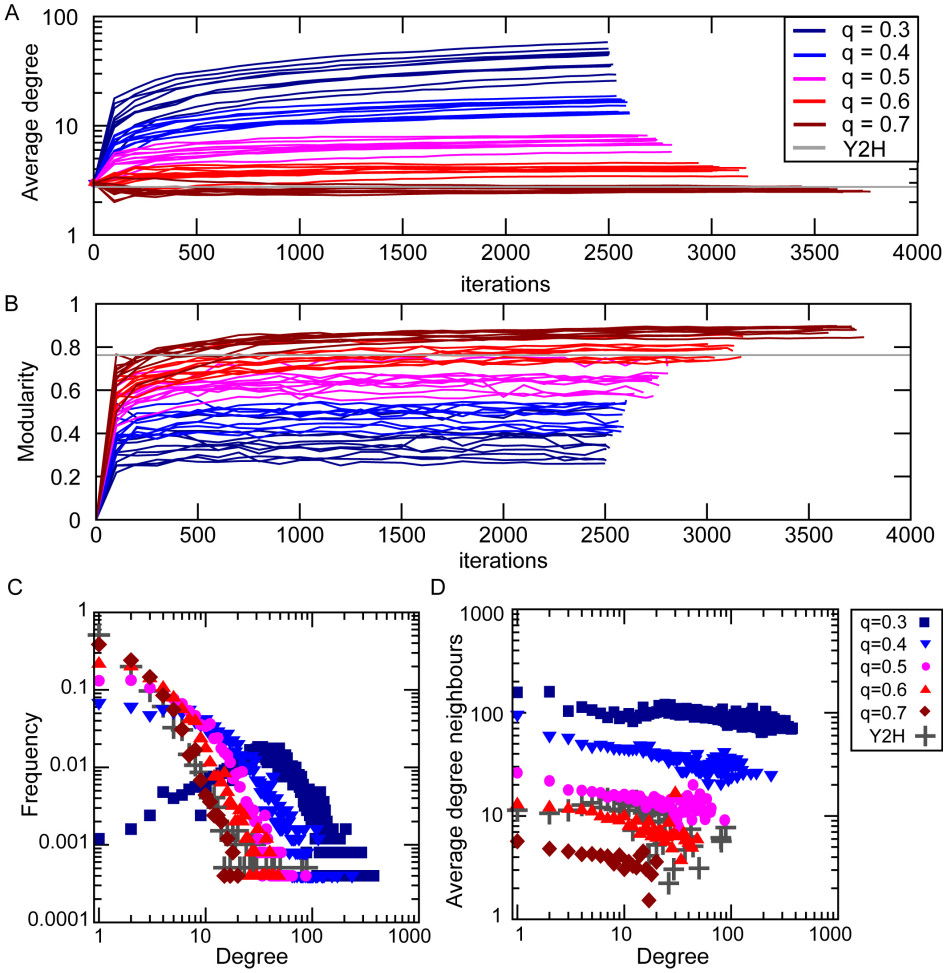

Figure 5.1. Network statistics for the original DD model, for different q. A,B) The average degree (i.e. number of interaction partners) resp. modularity per iteration for the DD model for different values of the divergence parameter q. For each parameter setting we plot 10 runs of the model. The grey line is the average degree resp. modularity in the Y2H network. For higher values of q networks become more and more sparse and modular over time. Each model is run until the network has reached a size of 2500 nodes, for higher values of q this takes more iterations as nodes are more likely to become a singleton (because one of the daughter nodes loses all connections) after duplication. C) The degree distribution after the network has reached its final size of 2500 nodes, for different values of q. The degree distribution of the Y2H network is depicted in grey for reference. D) Per degree <k>, the average degree of the interaction partners of all nodes with degree <k>,for different values of q and for the Y2H network (in grey). The degree anticorrelation is less strong for lower values of q.

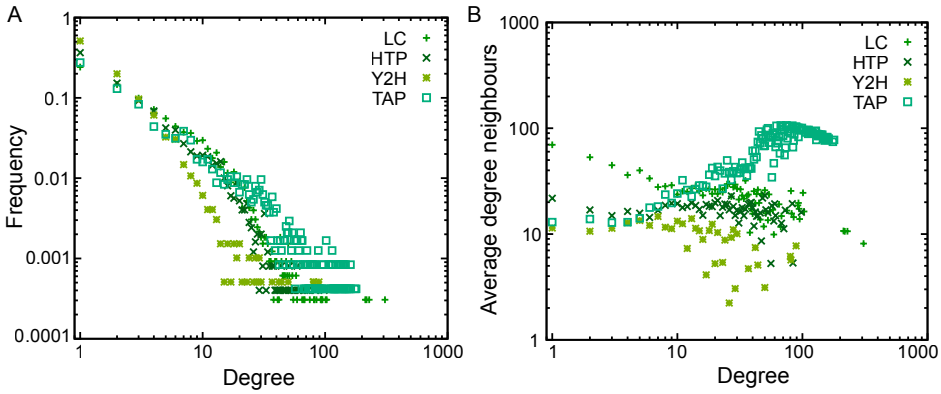

Figure 5.2. Degree distribution and degree anticorrelation in 4 different *S. cerevisiae* PINs. A) The frequency distribution of node degrees in 4 different *S. cerevisiae* PINs. The networks differ in overall connectivity, the Y2H network being relatively sparse and the TAP (co-complex) being relatively densely connected, but, compared to the degree anticorrelation depicted in panel B, the shape of the distributions is similar. B) Per degree <k>, the average degree of interaction partners (network neighbours) for all nodes with <k> neighbours. In the Y2H and the LC networks, the node degree correlates negatively with the average degree of node neighbours, indicating that hubs are usually connected to nodes with a low degree. In contrast, in the TAP network, hubs are located in densely connected clusters (large complexes) and are typically connected to other hubs. In the HTP network there is no relation between a node's degree and the average degree of its neighbours.

5.2 Results

5.2.1 Network growth by duplication and subfunctionalization

The original Duplication-Divergence model, as described in (Vazquez et al., 2003), is initialized with a small network. In each iteration, a randomly selected node is duplicated along with all of its edges. With a probability p a new edge is created between the twin nodes. In the subsequent subfunctionalization step, for each of the interaction partners of the twin nodes, one of the two edges is deleted with a probability q (see Methods and Figure 5.3). The model assumes a rapid initial divergence between daughter nodes after duplication (He and Zhang, 2005), thus the subfunctionalization step is completed before a new node is selected for duplication. Subsequent duplication of shared neighbours of the daughter nodes leads to further functional divergence on a longer timescale.

We fix p at 0.1, as higher rates are unlikely given the number of interparalog interactions observed in *S. cerevisiae* PINs (Fokkens et al., 2012), and vary the level of subfunctionalization after duplication q. We initialize the network with a small clique of 4 nodes and study how the network properties change as the

network grows. Because the network grows by duplicating parts of the existing network, the network with which the model is initialized does affect its outcome, but mainly in terms of maximum clique size (Hormozdiari et al., 2007). For low values of q (<0.5), the relative overlap of interaction partners of paralogs in the model is high (Fokkens et al., 2012) and the network becomes increasingly dense, whereas for higher values of q, the network becomes sparse and highly modular (Sole and Valverde, 2008) (Figure 5.1A, B).

As a node gains connections by duplication of its neighbours, nodes with a high degree are more likely to gain a new interaction than nodes with a low degree. In sparse networks ($q>0.5$) this rich-get-richer dynamics results in a degree distribution where most nodes have a low degree and few nodes have a very high degree, resembling the long-tailed degree distributions of real PINs (Figure 5.1C and Figure 5.2A) (Vazquez et al., 2003). Moreover, in sparse networks, nodes with a high degree that gained edges through duplication of their neighbours are typically connected to nodes with a low degree that lost connections due to subfunctionalization (Figure 5.1D) (Kim and Marcotte, 2008). This phenomenon, degree anticorrelation, is also observed in some, but not all, PINs (Figure 5.2B) (Maslov and Sneppen, 2002; Hakes et al., 2005). Most nodes belong to a giant component but, in contrast to many other network growth models, it is possible in the DD model to have more than one component in the network.

Duplication followed by subfunctionalization is able to reproduce a large number of network characteristics under conditions of network growth. On long timescales there is a limit to how often nodes can duplicate. Most nodes only have one interaction partner (Figure 5.1C), which means that after duplication of these nodes, daughter nodes can not afford to lose any interactions in the subfunctionalization step without turning into a singleton and being removed from the network (unless they gain an interaction with their twin, which occurs with a small probability p). The parameter q thus does not only affect the loss of interactions after duplication, it also influences the number of interactions needed to grow a network of a certain size: for high q more nodes turn into singletons after duplication (Figure 5.1A, B). Nodes with a low degree have a higher probability of one of their daughter nodes being a singleton (and thus being removed immediately) after duplication, especially for high values of q.

5.2.2 Stable gene repertoire size by incorporating gene loss in the DD model

We switch from a model of network expansion, to a model in which we maintain a more or less stable gene repertoire size. We define the probabilities for duplication or deletion events such that they explicitly depend on the number of nodes in the network, relative to a target size of 2500 nodes (see Methods and Figure

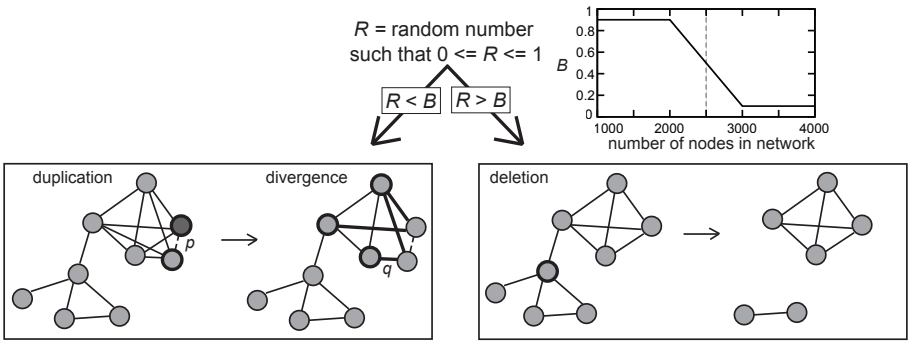

Figure 5.3. The Duplication-Divergence model with node deletion. The number B ranges between 0.1 and 0.9 and depends on the number of nodes in the network compared to a target number of nodes (2500, dotted line), a minimum (2000 nodes) and a maximum (3000 nodes). In case of a duplication event (left panel), a node is randomly selected and duplicated with all of its edges. With a probability 0.1 the two daughter nodes are connected (dotted line). Subsequently, for each of the ancestral interaction partners (thick lines), the connection with one of the two daughter nodes is removed with a probability q. In case of a deletion event (right panel), a random node is selected (thick line) and removed from the network along with all of its edges. At the end of each iteration, nodes without any connections are removed from the network.

5.3). In addition to deleting nodes that are randomly selected in case of a gene loss event, we remove all singletons at the end of each iteration, assuming fast pseudogenization when a protein is no longer involved in the cellular machinery. We initialize the model with a clique of 6 nodes. For the first $\sim$2000 iterations the network will grow fast as duplication events greatly outnumber deletion events. When the target size is approached, the number of nodes in the network stabilizes. We let the model run for 100000 iterations. This enables us to study long-term effects. We record over 10 different topological statistics in order to monitor how the networks change over time.

The DD model that incorporates gene loss, produces sparse and modular networks. This is independent of q (figure 5.4A, B). The degree distribution is long-tailed, just as the degree distribution of real PINs, but the tail is shorter due to a lack of highly connected hubs (Figure 5.4C). In networks generated with low levels of subfunctionalization ($q < 0.5$), there is no degree anticorrelation (Figure 5.4D): nodes with a high degree typically interact with nodes that also have a high degree. This indicates that in these networks, nodes with a high degree are located in relatively large and densely connected clusters, resembling co-complex networks rather than protein-protein interaction networks (Figure 5.2B).

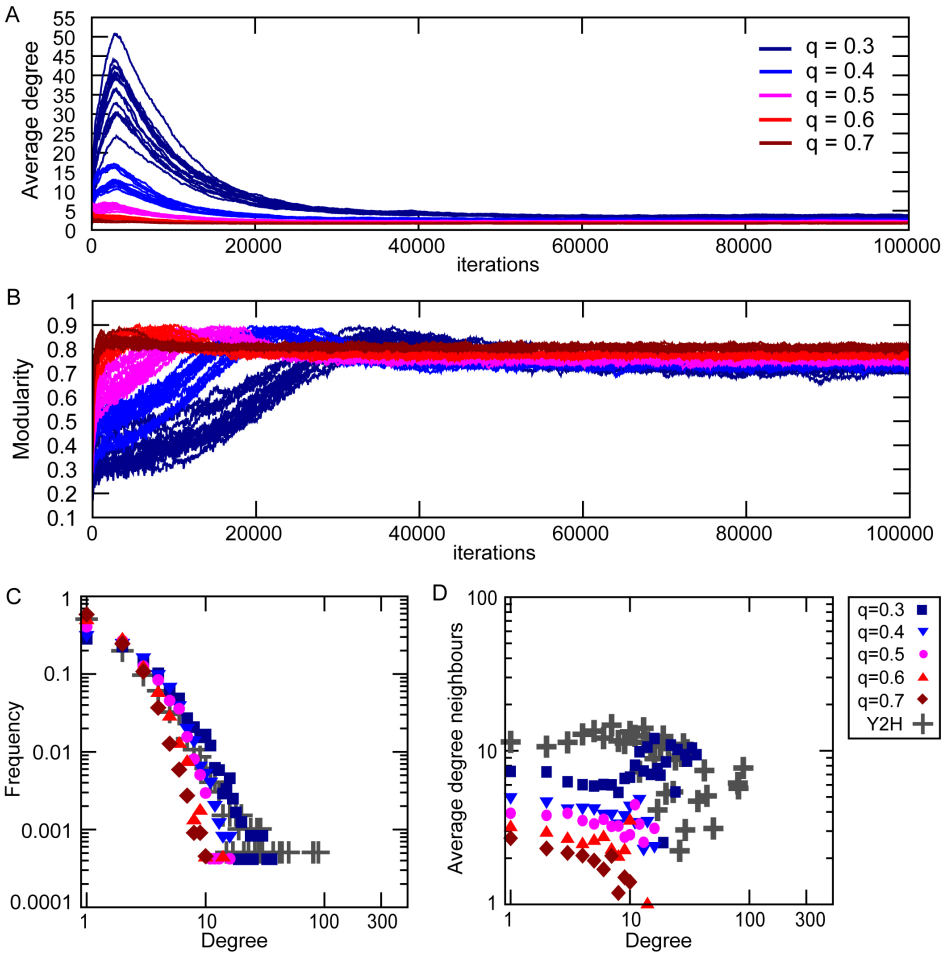

Figure 5.4. Network statistics for the DD model with random node deletion, for different q. A,B) The average degree (i.e. number of interaction partners) resp. modularity per iteration for the DD model for different values of the divergence parameter q. For each parameter setting we plot 10 runs of the model. The grey line is the average degree resp. modularity in the Y2H network. Each model is run untill the network has reached a size of 2500 nodes, for higher values of q this takes more iterations as nodes are more likely to become a singleton (because one of the daughter nodes loses all connections) after duplication. C) The degree distribution after 100000 iterations, for different values of q. The degree distribution of the Y2H network is depicted in grey for reference. D) Per degree <k>, the average degree of the interaction partners of all nodes with degree <k>, after 100000 iterations for different values of q and for the Y2H network (in grey). For low values of q there is no degree anticorrelation, indicating that nodes with many connections reside in large, densely connected clusters, resembling a co-complex network such as the TAP network depicted in Figure 5.2.

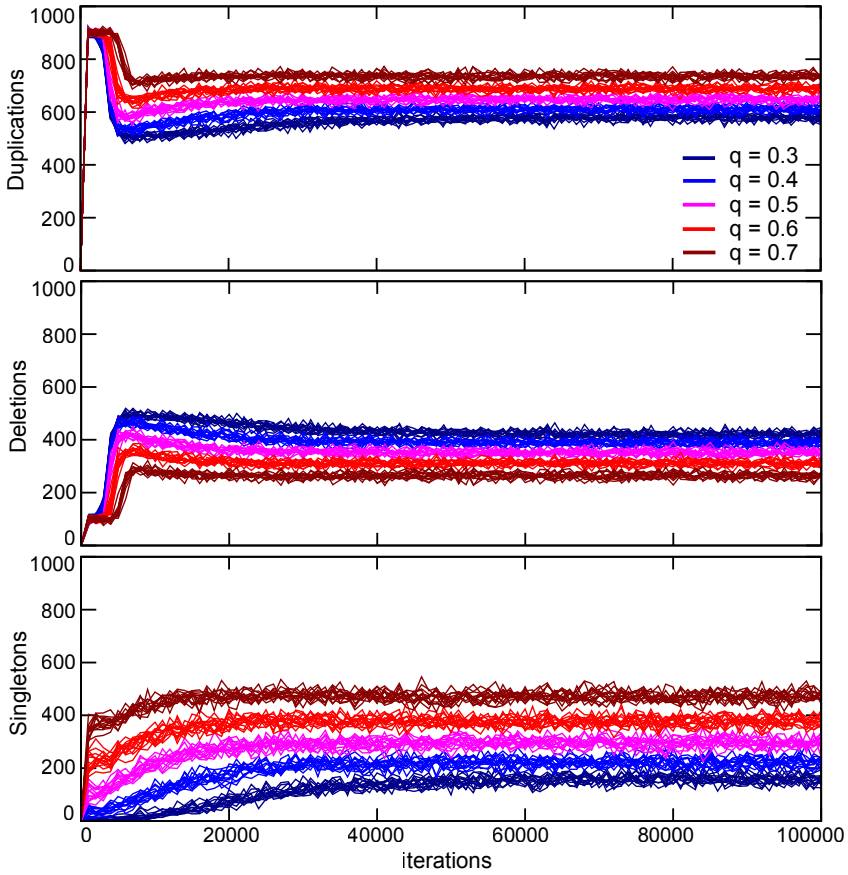

Figure 5.5. The number of duplication- and deletion events and the number of singletons for the DD model with random node deletion, for different values of q. The probability for duplication resp. deletion event depends on the number of nodes int the network compared to a target size of 2500 nodes (Figure 5.3 and Methods). In addition to randomly selected nodes, nodes that are completely disconnected from the network a.k.a. singletons, are removed as well. This occurs more frequently in the sparse networks generated under higher values of q, hence there are more duplication events to compensate and keep the number of nodes in the network stable.

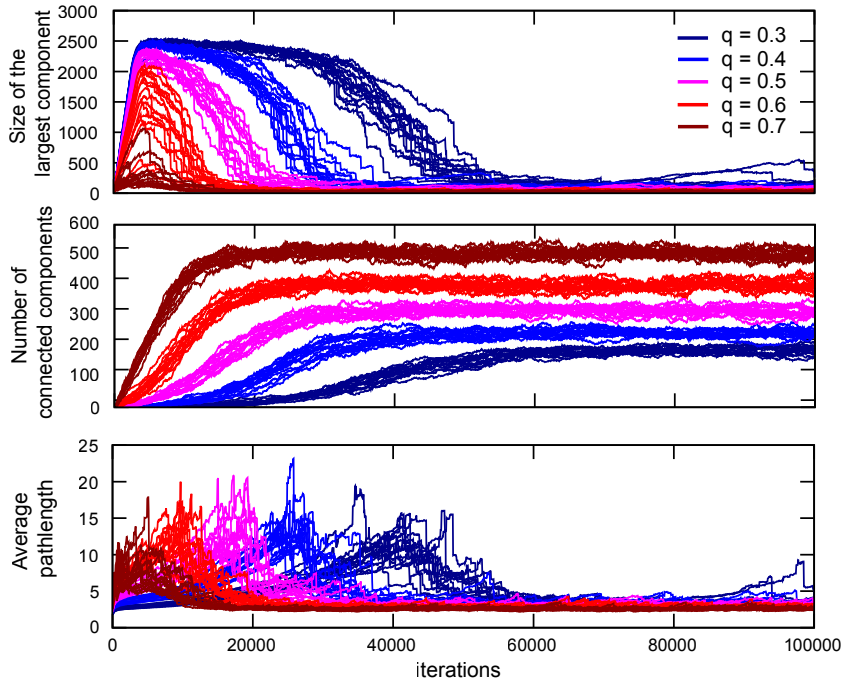

Figure 5.6. Phase transition in the DD model with random node deletion. In the initial growth phase, lasting for approximately 2500 iterations, the size of the largest component increases. As networks approach their target size, more and more deletion events occur (Figure 5.5B) and networks become more sparse (Figure 5.4A) and fall apart into smaller components.

5.2.3 From small world to large world to many tiny worlds

Gene loss decreases network connectivity: given an average degree of <k>, a network on average loses <k> edges when a node is deleted, while it only gains p + <k>(1-q) edges when a node duplicates. It is not possible in this model to balance duplication and deletion events such that both the number of network nodes as well as the number of network edges are stable. Although the average degree increases with increased q, the difference in connectivity between networks generated using different values of q is less striking than under growth conditions. In the model, the number of duplication and deletion events is balanced to maintain a certain number of nodes. As singletons arise less frequently in networks generated using low q values, these networks are subjected to deletion of randomly selected nodes more often (Figure 5.5). Hence networks generated using low values of q lose more interactions through explicit gene loss.

The loss of network connectivity greatly affects network cohesion. Networks undergo a phase transition: after an initial growth phase, gene loss events cause the network to become more and more sparse (Figure 5.6A). This results in the network breaking up into disconnected components (Figure 5.6B and 5.6C). The average path length increases at first: the network is slowly 'pulled apart' as shortcuts are deleted. The average path length reaches its maximum just before the network breaks up. When the phase transition takes place, depends on the connectivity in the network: in networks that are generated under lower values of q (and thus are more dense) the phase transition occurs later.

q	1 edge	2 edges	3 or 4 edges	5 or 6 edges	> 6 edges
0.3	1.52	0.95	0.85	0.83	0.81
0.4	1.46	0.89	0.77	0.72	0.75
0.5	1.37	0.82	0.69	0.66	0.68
0.6	1.31	0.77	0.63	0.58	0.57
0.7	1.24	0.73	0.58	0.54	0.73

Table 5.1. Degree dependent node loss in models with random node deletion For the networks from iteration 5000, 6000, 7000, ..., 99000 we determine which nodes are deleted after 1000 iterations. Each column in this table represents a degree bin. For each network we determine for each bin [% of nodes with this degree that is deleted after a 1000 iterations] / [% of all nodes that was deleted after a 1000 iterations in the whole network]. The table contains averages over 91 networks (from iteration 5000 - 99000, with steps of a 1000 iterations). Ratios > 1 indicate that nodes that have this degree are deleted relatively often.

5.2.4 Degree-dependent gene loss

If the number of interaction partners of a protein reflects it's importance in the functioning of the cell, we would expect it to correlate with evolutionary rate, essentiality and the propensity that a protein is lost in evolution. Indeed, several

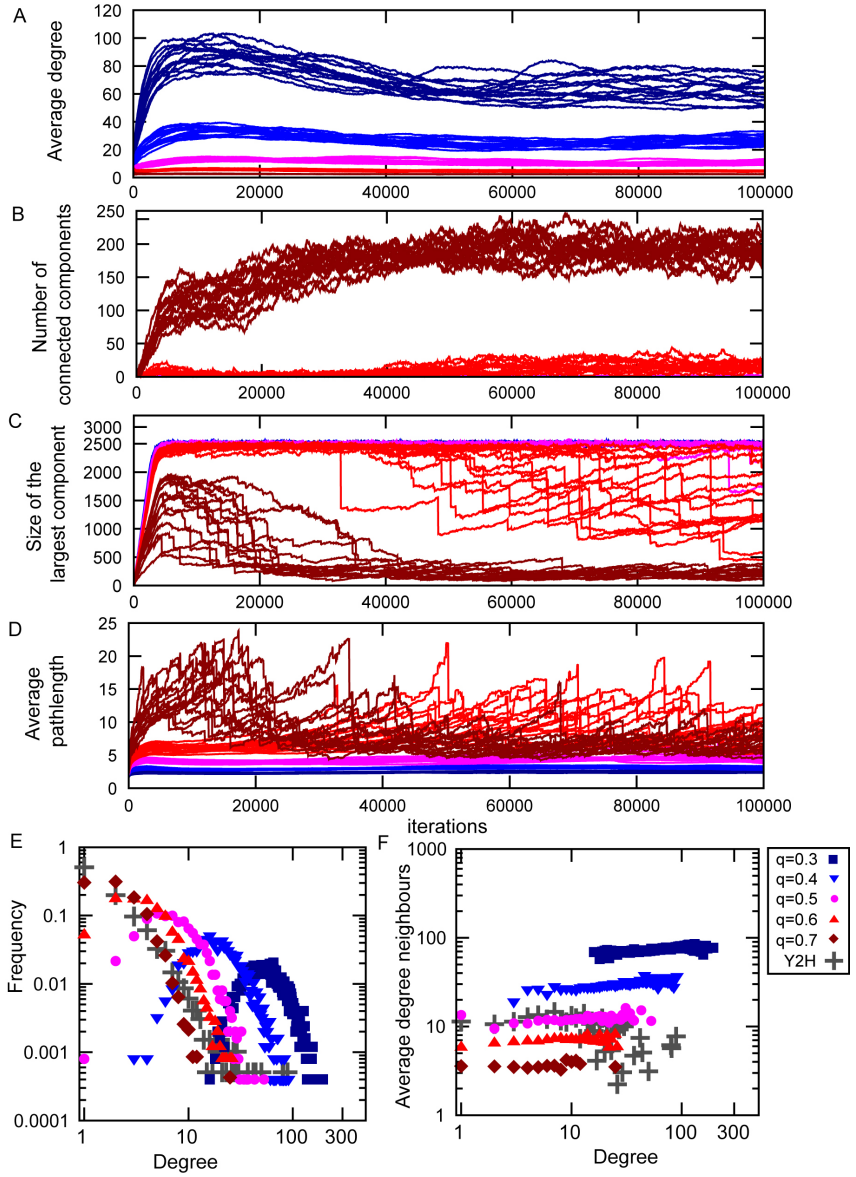

Figure 5.7. Network statistics for the DD model with explicit degree-dependent deletion, for different values of q. A,B,C,D) The average degree, number of connected components, size of the largest component and the average path length in time for the DD model with degree-dependent node deletion. For each parameter setting q, we plot 10 model runs. Panel E and F depict the degree distribution and the average degree of neighbours per degree for an arbitrary run, for different values of q. Preferential selection of nodes with a low degree for deletion results in densely connected networks (A) in which most nodes have more than one connection (E), network integrity is maintained except for very high levels of divergence after duplication ($q>=0.6$) (B and C) and the average path length is low, except for high levels of divergence after duplication ($q>=0.6$) (D).

studies report a correlation between these properties (Fraser et al., 2003; Krylov et al., 2003). On the other hand, several alternative explanations have been proposed for the observed correlations. For example, essentiality could be a property of protein complexes rather than of individual proteins. Hence the overrepresentation of essential proteins among well-connected hubs in PINs can be explained by the fact that these large, essential complexes consist of a number of these hubs (Hart et al., 2007; Zotenko et al., 2008). Moreover, the number of interaction partners correlates with other network properties, such as for example a node's centrality or betweenness, that may better reflect whether a protein is pivotal in the cellular machinery or not (Jeong et al., 2001).

Even though in our model we select nodes to be deleted independently of their network properties, the probability that a node is lost during a simulation, depends on its degree. This is because nodes that have few interactions are more likely to become a singleton (Table 5.1). If we assume that more 'important' nodes are less likely to be lost per se and explicitly strengthen this relation between a node's connectivity and it's propensity to be deleted during a simulation, do we still observe network breakup? How does this assumption affect network architecture? We adjust the model and explicitly select nodes to be deleted with a probability that is inversely proportional to their degree (see Methods for more detail). We find that networks that are subjected to this explicit degree-dependent deletion, do maintain a giant component, but only given low values of q (Figure 5.7B and C). However, these networks are extremely dense (Figure 5.7A) and their degree distributions are very distinct from those observed in real PINs. The majority of nodes has a relatively high degree (»1), as nodes with a few interactions are often deleted (Figure 5.7E). In summary, the architecturse of networks generated by a model implementing degree-dependent deletion differ from those of real PINs.

The only mechanism to gain interactions in our models is through duplication of neighbours, thus assuming that gain of completely novel interactions does not occur frequently enough to influence the large-scale network structure. The extent of de novo interaction gain is still under debate and this assumption may not be valid. Obviously, proteins sometimes gain completely novel interactions (other than through duplication of interaction partners) (Berg et al., 2004; Kim et al., 2006), but without comprehensive interaction data in multiple species it is hard to assess how often this occurs (Ali and Deane, 2010; Gibson and Goldberg, 2009). Whether gain of de novo interactions strongly influences PIN architecture, and which rules govern this process, are still open questions.

We incorporate de novo interaction gain into our model. We investigate whether gain of novel interactions will prevent network breakup, without losing a modular network structure. Moreover, we want to know whether the expected increase in network connectivity allows for more highly connected hubs.

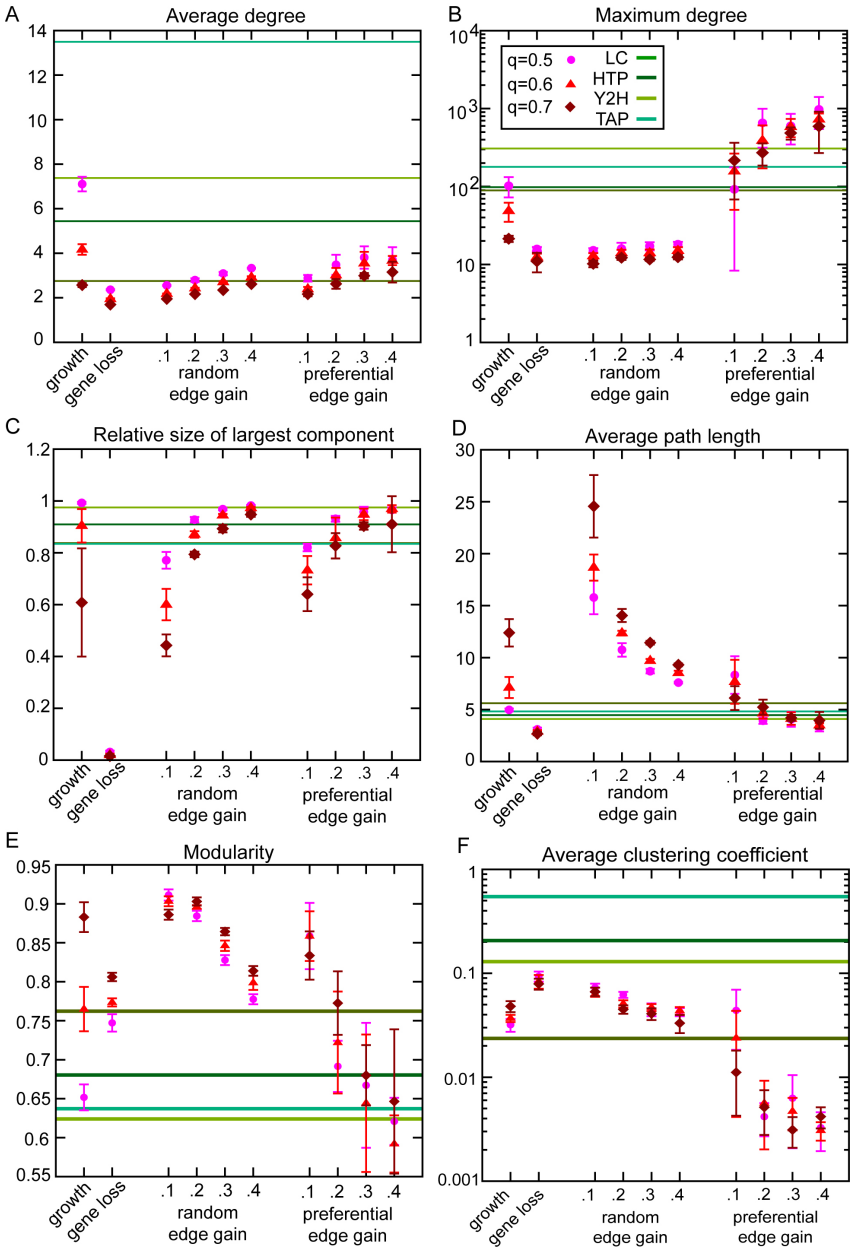

Figure 5.8. Network statistics for the original DD model, the DD model with random node deletion and the DD model with random node deletion and different implementations of de novo interaction gain, for different values of q.

Figure 5.8. The mean (symbol), maximum and minimum (errorbars) values are shown for 5 independent runs of each model, after 100000 iterations. The statistics for the 4 different *S. cerevisiae* PINs used in this study are depicted in different shades of green for reference. A) The average degree in the network. In the original DD model, without node deletion, network connectivity strongly depends on q. In networks with node deletion this is less important as all networks are sparse. Introducing de novo edge leads to only a small increase in network connectivity. Values for real PINs range from 2.75 (Y2H) to 13.5 (TAP). B) The maximum node degree in the network. Among models that implement random node deletion, only models incorporating preferential gain of de novo interactions contain very well connected hubs, i.e. hubs with more than 50 interaction partners. In real PINs the maximum degree ranges from 89 (Y2H) to 308 (LC). C) The relative size of the largest component i.e. the fraction of nodes that belong to the giant component. In real networks most nodes (from 83% in the TAP network to 97% in the LC network) reside in a giant component. We find that in networks generated with random node deletion this giant component desintegrates after switching from growth to maintaining a stable gene repertoire size. In networks with de novo interaction gain, even if the average degree is only slightly increased, network integrity is maintained. D) The average path length is very short in models with random node deletion without de novo edge gain due to network breakup into small disconnected clusters. With random de novo edge gain network integrity ismaintained (see panel C) but the average path length in model networks is still higher than in real PINs. With preferential de novo edge gain the average path length is comparable to that of real PINs, indicating that highly connected hubs contribute to shortcuts in the network (see Table 5.2) E) Modularity. Increasing the probability of de novo edge gain reduces modularity in the network. F) Global clustering coefficient. Networks generated with preferential rewiring have a low global clustering coeffcient: the well-connected hubs bridge relatively small clusters.

5.2.5 Gain of de novo interactions results in a giant component, but for highly connected hubs preferential attachment is required.

Little is known about which rules govern network rewiring, or even if such rules exist. We first study a null model in which each iteration, with a probability e, we randomly select a pair of nodes and connect them. Notably, in contrast to the model described in (Pastor-Satorras et al., 2003),interaction gain is not coupled to a duplication event in our model, hence we assume that subfunctionalization and neofunctionalization occur on different timescales (He and Zhang, 2005).

We find that if $e>=0.2$, networks generated by the model combine a giant component that contains > 80% of the nodes, with a high level of modularity (Figure 5.8 C,E). Interestingly, the maximum node degree is comparable to that of networks generated without edge gain (Figure 5.8B). Even if networks maintain a stable average degree during a simulation, the rich-get-richer dynamics in DD models that incorporate gene loss, does not lead to highly connected hubs. This is in accordance with the observation in an *S. cerevisiae* PIN that hubs do not preferably interact with members of the same protein family, which is what you

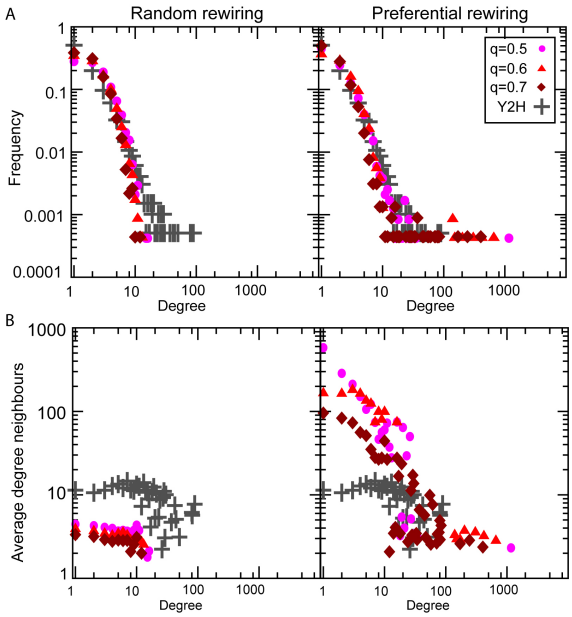

Figure 5.9. Degree distribution and degree anticorrelation in networks generated by models that incorporate rewiring. A) Degree distribution of networks generated with $e = 0.2$. Networks generated with random rewiring (left panel) do not contain well-connected hubs (i.e. nodes with more than 20 interactions). The degree distribution of networks generated with preferential rewiring resembles that of real PINs. B) Degree anticorrelation is very strong in networks generated with preferential rewiring, indicating that most low degree nodes are connected to well-connected hubs.

would expect if interactions are primarily gained through duplication of network neighbours (Berg et al., 2004).

As hubs do not arise automatically from a process of duplication and random rewiring, the pending question is whether hubs evolve to become hubs. There are two observations that suggest that they do. First of all, hubs have longer amino acid sequences, often have multiple and repeated domains and longer disordered regions than proteins with a few interaction partners (Haynes et al., 2006). This indicates that hubs have been well adapted to handle many different interactions. We assume that proteins that already have many interaction partners can acquire novel interactions relatively easily, for example because they have a long disordered region to which many different proteins can bind. We implement this assumption by connecting a randomly selected node to a node that is selected with a probability proportional to its degree (as in (Berg et al., 2004)), reminiscent of the Preferential Attachment model (Barabasi and Albert, 1999) (see Methods).

We find that the networks generated by the model implementing preferential edge gain do contain well-connected hubs, even though the average degree is comparable to that of networks generated with random edge gain. Moreover, the degree distribution and the average path length is more similar to what we observe in *S. cerevisiae* PINs than the average path length and degree distribution in networks produced using random rewiring (Figure 5.8 and 5.9). We conclude that preferential edge gain better explains network topology than random random edge gain.

Nevertheless, the model incorporating preferential edge gain does not reproduce all features that are shared by different PINs. For example, the short average path length in networks generated by our model hinges on the presence of highly connected hubs, which is not the case for real PINs. If we reduce the degree of those hubs that have more than 20 interactions, by randomly removing 'excess' interactions of these hubs from real PINs, the average pathlength in real PINs hardly changes. In contrast, in our model networks, the average path length almost doubles after removing connections from hubs (Table 5.2). This effect is not due to removal of edges per se, because removing the same number of randomly selected edges from the network does not affect the average pathlength.

The clustering coefficient for nodes with a high degree (>20) is much higher in real PINs than it is our model networks, indicating that in real PINs there are paths around hubs, whereas in our model networks alternative, hub-avoiding routes do not occur as often. One possible explanation is that, in addition to preferential interaction gain by hubs, novel interactions are more likely to arise between proteins that are relatively close in the network. They are more likely to be in each other's close proximity: an obvious first requirement for a physical interaction.

networks	no edges removed		X edges removed from hubs		X edges removed from random nodes	
	In largest component	Average path length	In largest component	Average path length	In largest component	Average path length
LC	0.97	4.09	0.86	5.12	0.88	4.43
HTP	0.91	4.47	0.82	5.19	0.82	4.69
Y2H	0.84	5.61	0.78	6.59	0.75	5.81
TAP	0.83	4.82	0.79	5.86	0.62	5.28
models						
1	0.9	4.02	0.25	9.35	0.5	4.36
2	0.85	4.4	4.46	8.79	0.63	4.6
3	0.93	4.45	0.6	6.7	0.66	4.89
4	0.7	5.56	0.48	7.33	0.5	5.75
5	0.9	4.77	0.48	9.07	0.66	5.22

Table 5.2. Effect of decreasing degree of hubs by removing edges We compare 5 model networks (numbered 1-5, all generated with q=0.6 and e=0.2, random node deletion and preferential edge gain) with real PINs. The fraction of nodes that is connected in the giant component as well as the average path length is comparable between model networks and real PINs. We investigate how important the highly connective hubs are for maintaining network integrity and remaining a small world. We randomly select excess edges from hubs that have a degree > 20, such that the maximum degree in the network equals 20. We then remove these edges from the network (X in total) and study whether the largest component breaks up and/or the average path length increases. In order to exclude that any effect is just due to the removal of edges and is not specific for highly connected hubs, we also remove X randomly selected edges from the network and compare the efefct on the size of the largest component and the average path length. We perform these steps 10 times, the table contains averages over these 10 repeats. We find that in real PINs, removing edges from the highly connected hubs has very little effect on network integrity and the average path length. In contrast, in model networks, the largest component decreases in size and the average path length is much longer compared to when we remove the same number of randomly selected edges. This indicates that hubs are important for maintaining network integrity and for remaining a small world in model networks but not in real PINs.

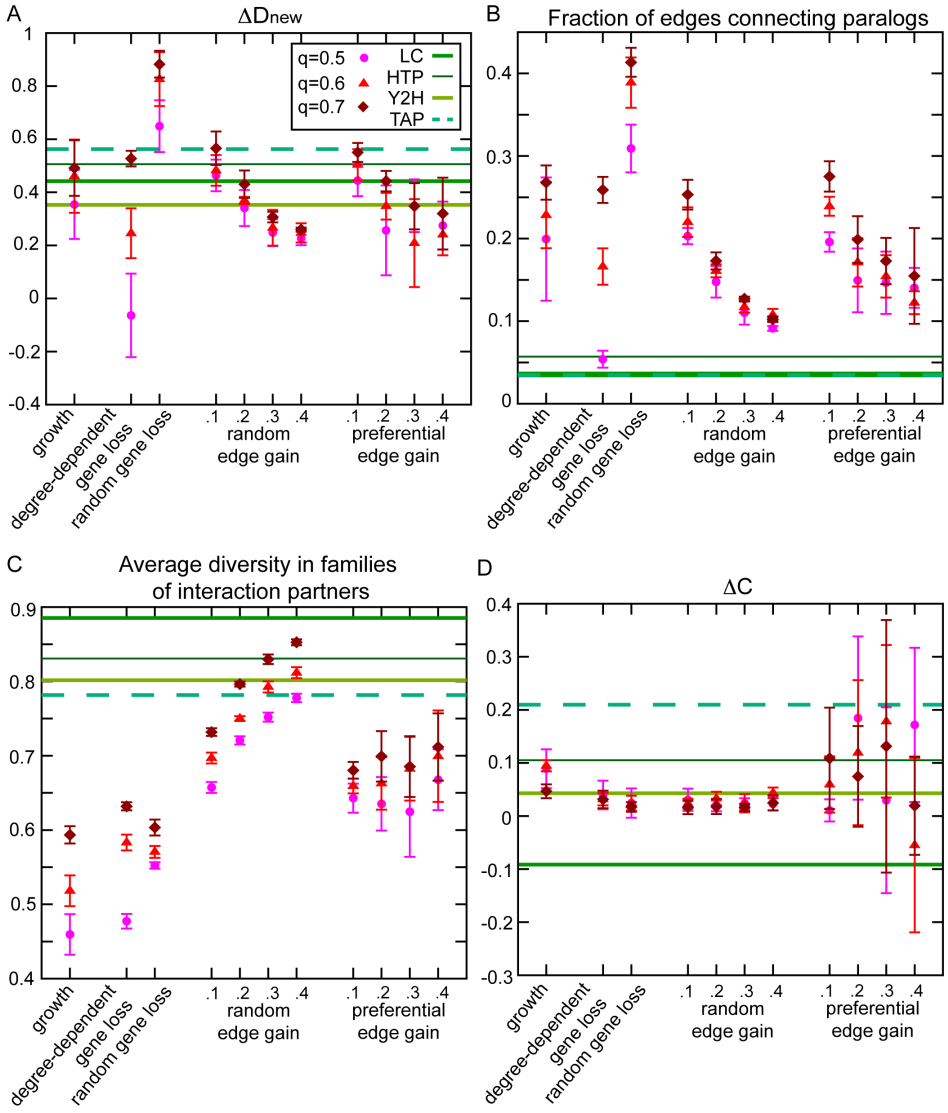

Figure 5.10. Age structure in real and model networks. A,B. The positive ΔD_{new} can be attributed to a relatively high number of edges that connect paralogs. C. Members of a protein family cluster in the network, especially in networks generated without rewiring. D. A strong correlation between protein age and connectivity is observed in neither real PINs nor in model networks.

5.2.6 Protein age

The taxonomic distribution of species in which we can detect homologs for a protein, provides a rough estimate of when this protein was invented and thus of its age. Including the age of proteins as an additional feature adds a realm of fascinating observations regarding the evolution of biological networks. For example, the observation that old proteins tend to have more interactions than younger proteins suggests that proteins gradually accumulate interactions over time (Eisenberg and Levanon, 2003; Capra et al., 2010). However, the relationship between protein age and -connectivity is confounded by the fact that proteins that experience slower sequence evolution tend to be estimated as being older. Slowly evolving proteins have more interactions in some PINs, e.g. due to an experimental bias towards more abundant proteins (such as the TAP/MS network (Ivanic et al., 2009), Figure 5.10D). Whether this relationship is observed thus depends on the exact definition of protein age, the different age categories that are compared (e.g. Capra et al. only distinguish proteins originating from before or after the Whole Genome Duplication (WGD) event) and the PIN that was used. In 3 out of 4 PINs we find no tendency for older proteins to have more interactions, congruent with (Kunin et al., 2004). We quantify the gradient in connectivity per age in a measure ΔC (see materials and methods for more detail). In neither of our DD models we find a relation between protein age and -degree (Figure 5.10C). The fact that older proteins have had more time to accumulate novel edges is balanced by the fact that older proteins also have more time to lose interactions, either by duplicating or because interaction partners are removed from the network through gene loss. In networks that grow through Preferential Attachment of novel genes, old nodes tend to have more interactions than younger nodes (Eisenberg and Levanon, 2003). In contrast, in networks generated by the model that incorporates preferential edge gain, this is not the case (Figure 5.10C).

Patterns of connectivity between proteins of different ages may offer a glimpse on how a network changed over time. Particularly interesting is the trend that proteins preferentially interact with proteins of a similar age (Qin et al., 2003; Kim and Marcotte, 2008; Capra et al., 2010). We extensively studied the influence of gene duplications on the age structure of PINs in (Fokkens et al., 2012) and found that both interparalog interactions as well as with interactions shared by paralogs contribute to the observed tendency to interact with proteins of a similar age. We could reproduce the effect of interparalog interactions with simulations of network evolution using the DD growth model, but not the effect of interactions shared by paralogs.

We study the age structure of networks generated using DD models that include gene loss and thus operate on a more realistic timescale than growth models. We implement protein age and quantify the overrepresentation of proteins of a similar age using the ΔD_{new} measure described in (Fokkens et al., 2012) (see Methods

for more detail). We find high values of ΔD_{new} in models that do not incorporate edge gain. In these networks nodes tend to interact with nodes of a similar age (Figure 5.10A). A node with one interaction can only duplicate if either both copies keep the ancestral interaction or an interaction is gained between daughter nodes after duplication (cf. a homodimer that becomes a heterodimer, that occurs with a probability p=0.1, see Figure 5.3 and Methods). Hence, in these sparse networks, such as those generated by models without edge gain, interactions between paralogs are even more strongly overrepresented than in networks generated without node deletion (Figure 5.10B). In models without de novo edge gain, interactions are only gained by duplication of network neighbours, hence most proteins interact with proteins that belong to a single family that is usually the same family as the protein itself (Figure 5.10C). De novo edge gain tapers the overrepresentation of interparalog interactions in the model networks, leading to a reduced ΔD_{new}. In conclusion, these results are similar to what we observed in the DD model without node deletion: the tendency to interact with proteins of a similar age can be attributed to an overrepresentation of interparalog edges. This is not observed to this extent in real PINs (Fokkens et al., 2012).

5.3 Discussion

Comparative genomics studies show that some gene families are more changeable in evolution than others (Snel et al., 2002; Krylov et al., 2003; Wapinski et al., 2007). We expect the volatility of a gene family to be related to functional properties of its members and vice versa. One way to describe the function of a protein is through its position in a network, for example a protein interaction network. By representing proteins as nodes in a network and converting genomic changes into network changes, we can model how gene duplication or -loss affects the functional properties of proteins in the cell. Moreover, we can experiment with different ways in which a network property (here: node degree) influences the likelihood of fixation of gene loss. This approach allows us to study how certain conjectures on the mutual influence of genome evolution and cellular organization translate into global network architecture, thus improving our understanding of their more farreaching consequences.

Capra et al. found that proteins that result from a very recent (post WGD) duplication tend to have a lower degree in *S. cerevisiae* PINs, suggesting that duplication comes at a cost of initially losing interactions, e.g. via subfunctionalization (Capra et al., 2010). We find that in our models with constant network size, it is vital that after duplication both daughter nodes gain new interactions over time to compensate for those that were lost in the subfunctionalization step. If not, the number of times that a node can duplicate is very limited. On a global level, more edges are removed from the network by node deletion than are added to the network by node duplication. Unless node duplications greatly outnumber node

losses (i.e. network expansion), subfunctionalization combined with gene loss thus leads to a gradual loss in connectivity that results in a catastrophic network breakup. In our models, networks, despite their long-tailed degree distributions, are not robust towards the cumulative loss of interactions.

In our simulations, we keep the size of our protein repertoire stable. In real life it is more likely that periods of genome expansion alternate with periods of streamlining (Scannell et al., 2007; Koonin, 2007; Mittenthal et al., 2012; Cuypers and Hogeweg, 2012). As long as on average more interactions are removed from the network through gene loss than are added through gene duplication, which is the case when nodes only gain interactions through duplication of neighbours, we do not expect that this affects our general conclusions regarding network viability without de novo interaction gain.

There are several potential mechanisms that may prevent a catastrophic breakup of the network. One example is loss of network modules, instead of single genes: localized loss may save other parts of the network. However, several studies of co-evolution of components of network modules (protein complexes) indicate that loss of an entire module does not often occur. Moreover, in our model, loss of a gene that is part of a module, increases the probability to be removed from the network for the remaining module members because loss of a member reduces their degree.

Explicit conservation of edges that connect different clusters in the network may also contribute to maintaining network integrity. Essentiality of protein-protein interactions (PPIs), rather than proteins, can explain why hubs tend to be essential (proteins with many interactions are more likely to be engaged in an essential interaction) as well as why some proteins are essential yet have very few interactions (He and Zhang, 2006). On the other hand, studies investigating conservation of interactions indicate that complex membership tends to be conserved (van Dam and Snel, 2008; Zinman et al., 2011; Shou et al., 2011; Hart et al., 2007; Zotenko et al., 2008) whereas interactions that bridge different modules are much more changeable (Zinman et al., 2011). Hence, even if essential PPIs exist, those are unlikely to hold the PIN together.

Another potential mechanism is the gain of completely novel interactions (hence not through duplication) to compensate for any decrease in network connectivity due to duplication- and loss events. De novo interaction gain can occur between existing proteins or with entirely novel genes. As novel genes tend to preferably connect to the network's periphery (Capra et al., 2010), we do not expect invention of novel genes to be pivotal in maintaining network integrity. We find that gain of novel interactions between existing proteins at a relatively low rate (compared to gene duplication) is sufficient to maintain network integrity.

Changes in expression or localization as well as domain recombination may lead to gain of a novel interaction. We propose a model in which some proteins, due

to their structure (disordered regions) and/or abundance, are more likely to gain a novel interaction, for example with a protein that acquires a new domain. We implement this by selecting a random node (cf. a protein that acquires a new domain) and connect this node to a node selected with a probability that is proportional to its degree (cf. a hub by nature: e.g. a protein with large disordered regions). Networks that are generated by this model, a hybrid of the DD model with gene loss and the Preferential Attachment model (Barabasi and Albert, 1999) (or the model of (Berg et al., 2004)), contain well-connected hubs such as those found in real PINs.

Analyses of *S. cerevisiae* PINs described here and in (Berg et al., 2004) reveal that proteins interact with members of multiple, different families. This is not congruent with duplication of neighbours as the main motor driving gain of interactions. Our network models demonstrate that gain of interactions through duplication of interaction partners alone is not adequate to maintain sufficient network connectivity and that network rewiring is crucial to prevent catastrophic breakups of PINs.

5.4 Methods

5.4.1 Protein interaction networks and topological statistics

We use the same *S. cerevisiae* PINs, including the age of nodes, as in (Fokkens et al., 2012). In short, the LC is based on data from BioGRID (Reguly et al., 2006). Some interactions were removed, namely protein-RNA interactions, interactions that were only supported by high throughput data, co-localization, co-fractionation or data from (Collins et al., 2007a; Ptacek et al., 2005; Grandi et al., 2002). The HTP network is based on high-throughput Y2H and TAP/MS data from (Ito et al., 2001; Uetz et al., 2000; Gavin et al., 2006, 2002; Krogan et al., 2006; Ho et al., 2002) and includes only those interactions that have been supported by more than one study, where studies (Gavin et al., 2006) and (Gavin et al., 2002) were counted as one. The LC and the HTP network were taken from the Supplementary Material from (Kim and Marcotte, 2008). The Y2H network from (Yu et al., 2008) was downloaded from `interactome.dfci.harvard.edu/S_cerevisiae/download/Y2H_union.txt`. The TAP network is based on PE scores downloaded from `http://interactome-cmp.ucsf.edu`. Proteins are connected in the TAP network if their PE score exceeds 0.2.

We use the Igraph package to represent real as well as model networks. This package was used to calculate the level of modularity in the network (Newman and Girvan, 2004) and the global clustering coefficient (defined as the number of cliques of size 3 divided by the number of triplets in the network (Wasserman,

1994)).

5.4.2 Protein families, -age and family/age-related statistics

Protein age is defined based on EggNOG orthologous groups, we use the 4 age categories as in (Fokkens et al., 2012) and (Kim and Marcotte, 2008). We assign the age Fu to groups that only contain fungal proteins, assuming that the founder gene was invented in Fungi. Similarly we assign the age E to groups that consist of eukaryotic proteins only. If a group contains at least one protein from either Bacteria or Archaea it was assigned the age AE/BE. If a group contains proteins from both Bacteria as well as from Archaea it was assigned the age ABE (see (Fokkens et al., 2012) for more detail).

We calculate ΔC from the relative connectivity C_g per age group (the average degree of proteins in an age group divided by the average degree of proteins that have an age, see also Table 3) as follows:

$$\Delta C = \frac{\sum_{g=1}^{g=G-1} C_g - C_{g+1}}{G-1}$$

where G is the number of age groups and g=1 corresponds to the oldest age group (ABE) and g=4 corresponds to the youngest (Fu).

We calculate ΔD_{new} as in (Fokkens et al., 2012) :

$$\Delta D_{new} = \frac{\sum_{n=2}^{G} \sum_{m<n}^{G-1} D_{m+1,n} - D_{m,n} + \sum_{n=1}^{G-1} \sum_{m>n}^{G} D_{n,m-1} - D_{n,m}}{G^2}$$

where $D_{m,n}$ is the relative interaction density between age groups m and n, calculated by dividing the frequency of observing an interaction between proteins of age group m and age group n by the expected frequency of observing this interaction based on the overall connectivity of m and n (see (Fokkens et al., 2012) for more detail).

The fraction of interactions between paralogs is simply calculated by dividing the number of edges that connect two members of the same family by the total number of edges in the network. The family diversity among interaction partners for node N is calculated by counting the number of different families to which this node is connected, by the total number of its interaction partners. If all its interaction partners belong to different families, this number is 1. If all its interaction partners belong to the same family this number is 1/<k> where <k> is the degree of node N. We average over all nodes with two or more interactions. Nodes with a high degree that only interact with members of a single family will lower this average more than nodes with a low degree.

5.4.3 Network growth model

We initialize the model with a fully connected graph consisting of 6 nodes (for the simulations without gene loss the network was initialized with a clique of 4 nodes). Each iteration we determine whether there will be a duplication or deletion event. The probability of either event depends on the number of nodes in the network (see Figure 5.3). In a duplication event a random node X is selected and duplicated with all of its edges. One of the daughter nodes is assigned a new age with a probability a=0.1, corresponding to a daughter node that diverges beyond recognition (Wolfe, 2004). For each interaction partner Y of X, if a random number between 0 and 1 is lower than q the interaction between one of the daughter nodes and Y is deleted. Finally, with a probability p, we create a new edge between the daughter nodes.

In a deletion event a node is selected and removed along with all of its edges. In models with explicit degree-dependent selection nodes are selected with a probability inversely proportional to $<k>^2$, where $<k>$ is the degree of the node. In models with de novo edge gain, we draw a random number between 0 and 1, if this number is below the probability of a new interaction e, we select two nodes and connect them in the network. In case of preferential edge gain, first one node is selected with a probability that is proportional to $<k>^2$, the a second node is randomly selected from the remaining nodes. At the end of each iteration all singletons, i.e. nodes that have a degree $<k> = 0$ are removed, as well as any double edges that may occur because an interaction was gained between two nodes that already were connected in the network.

Acknowledgements

The authors would like to thank Wouter van Veldhoven for his help in starting up the project and Jan Kees van Amerongen for technical support.

Chapter 6

Discussion

Summarizing discussion

We study the complex interplay between the functional organization of proteins in the cell and the evolutionary dynamics of protein families. The organization of the cellular machinery can be viewed as a hierarchy of modules (Ravasz et al., 2002; Gavin et al., 2006; Wagner et al., 2007). A sequence of amino acids folds into a domain, multiple domains are combined in a single protein, several proteins bind to form a protein complex, and protein complexes cooperate in biological processes. Hierarchical modularity implies hierarchical functionality. This is reflected in the structure of protein function classification schemes (Ashburner et al., 2000; Ruepp et al., 2004). How does this hierarchical modular organization affect evolution? How does natural selection act on different levels of modularity?

Strong selection on the level of protein modules -such as complexes or metabolic pathways- leads to distinct phylogenetic patterns: modules tend to be either completely present or completely absent in different species. We can estimate the extent and frequency of module level selection by scoring the cohesiveness of the combined presence-absence patterns of components of functional modules. Previous studies that use this approach are largely based on prokaryotes and report limited evolutionary cohesiveness for functional modules when compared to a random background (Snel and Huynen, 2004; Campillos et al., 2006). Components of cohesively evolving modules tend to be encoded in the same operon (Campillos et al., 2006) and proteins encoded in the same operon tend to evolve more cohesively than coregulated proteins (Snel and Huynen, 2004).

The genomic organization in prokaryotes -encoding cooperating proteins in one operon- provides a feedback mechanism that may enhance evolutionary cohesiveness of functional modules. We investigate evolutionary modularity in eukaryotes that lack such an operon structure in their genomes (chapter 2). We find that, as in prokaryotes, the majority of functional modules evolves flexibly. Unfortunately, differences in the specific module definitions, protein families and modularity scores prohibit a direct comparison of our results to those obtained in prokaryotes. We do observe very similar trends. For example, metabolic pathways evolve more cohesively than macromolecular complexes. Removing potentially confounding protein families or less involved module members, slightly increases observed evolutionary cohesiveness.

The advent of large-scale protein interaction datasets allow us to filter our set of protein complexes with this data and remove proteins that are less strongly attached to the complex or more attached to an other complex. Interestingly, filtering our functional modules with TAP/MS data (Collins et al., 2007a) has a small and mixed effect. This indicates that erroneous clustering and/or false positive physical interactions are not important contributors to the observed flexibility. Unexpectedly, in curated datasets, cohesive complexes are less well-connected

than non cohesive complexes. Apparently a physical interaction does not necessarily indicate strong functional interdependence.

A direct measure of functional interdependence is the quantification of genetic interactions (GIs) (Collins et al., 2006; Baryshnikova et al., 2010). This data could therefore prove to be a valuable intermediate step, bridging the gap between protein interactions and evolutionary profiles. Large scale (ğenome wide) quantitative GI data has recently become available (Costanzo et al., 2010; Ryan et al., 2012). To further increase coverage, this data can be combined with previous studies focusing on specific cellular subsystems (Zheng et al., 2010; Fiedler et al., 2009; Wilmes et al., 2008; Collins et al., 2007b; Schuldiner et al., 2005), by integrating the different scores associated with distinct platforms (Linden et al., 2011). A lot of effort is going into linking genetic- to physical and metabolic interactions (Kelley and Ideker, 2005; Michaut et al., 2011; Szappanos et al., 2011; Baryshnikova et al., 2010), but how genetic interactions within or between functional modules relate to cohesive evolution is yet to be explored.

The hierarchical modular organization of the cellular machinery is often explained in terms of selection for increased evolvability (Wagner et al., 2007; Rorick and Wagner, 2011; Tusscher and Hogeweg, 2011), especially when adapting to frequently changing environmental conditions (Samal et al., 2011; Kashtan and Alon, 2005). The paradigm that modules can be removed or tinkered with without severely disrupting the rest of the system, implies that modularity leads to reduced pleiotropy. A modular organization can thus increase a system's robustness (Wagner et al., 2007). Moreover, if modules can easily be copied, modified and reused elsewhere, they can function as building blocks in the cellular machinery.

The potential use of modules as building blocks is facilitated by colocating module components on the genome. This genomic reinforcement of module structure does occur at the bottom of the module hierarchy (domains consisting of amino acids and proteins comprised of multiple domains) but as we move upwards, feedback from genomic organization is less obvious. In chapter 2, we assume that in eukaryotes, genomic feedback on evolutionary cohesiveness of protein modules is negligible compared to the feedback in prokaryotes (Fischer et al., 2006; Seoighe et al., 2000; Langkjaer et al., 2000). On the other hand, eukaryotic gene order is far from random: genes that are colocated on the genome can be coregulated via chromatin remodeling or, like genes in an operon, simply by sharing transcription factor binding sites. Genes that belong to the same protein complex are often located within relatively short distance of each other. This can be explained in terms of dosage balance, both on a short timescale (coregulation) as well as on a longer timescale (coduplication), hence genomic colocalization kills two birds with one stone (Teichmann and Veitia, 2004; Papp et al., 2003). Both in fungi and in plants, secondary metabolic pathways are encoded in operon-like clusters (Osbourn, 2010; Rep and Kistler, 2010). A recent study indicates that

even if specific gene order has changed between species, genomic organization may still be conserved on the level of biological processes (Al-Shahrour et al., 2010). More research is needed to determine whether this conservation persists over longer phylogenetic distances and whether this correlates with modular presence/absence patterns or coduplication of the protein families in question.

Network growth models offer a contrasting perspective on the evolution of modularity in biological networks. Neutral models of network- or genome evolution replicate a number of generic properties of cellular systems, such as a scale-free degree distribution (Vazquez et al., 2003; Barabasi and Albert, 1999; van Noort et al., 2004; Pastor-Satorras et al., 2003) and overrepresentation of certain network motifs (Cordero and Hogeweg, 2006) in biological networks, as well as the size distribution of protein- and domain families in genomes (Karev et al., 2002, 2003). Hallinan and Solé use a network growth model to demonstrate that a (hierarchical) modular organization can arise as a side-effect of genome expansion through gene duplication and subfunctionalization (Hallinan, 2004; Sole and Valverde, 2008). Models of network growth via node duplication provide biological ground for growth through Preferential Attachment (Barabasi and Albert, 1999).

The patterns of how proteins of different ages are connected in a network provides information on how the network has expanded over time (Qin et al., 2003; Eisenberg and Levanon, 2003; Kim and Marcotte, 2008; Capra et al., 2010; Warnefors and Eyre-Walker, 2011). Proteins tend to physically interact with proteins of a similar age (Qin et al., 2003; Kim and Marcotte, 2008) and origin (Capra et al., 2010). Kim and Marcotte implement protein age in a number of different network growth models and find that a model based on node duplication does not produce networks in which nodes of a similar age prefer to interact. On that basis they conclude that gene duplication does not strongly influence the architecture (including hierarchical modularity) of the protein interaction network (Kim and Marcotte, 2008).

On the other hand, gene duplications leave their traces in the interaction network, e.g. paralogs share more interactions than random pairs of proteins (Musso et al., 2007; AIMC, 2011). Moreover, studies of the evolution of protein complexes as well as of metabolic pathways demonstrate that small scale duplications play an important role (Pereira-Leal et al., 2006, 2007). In chapter 3 we also find that protein complexes contain ancient paralogs. This suggests that duplications contribute to the network's modularity. Moreover, protein age is defined on the level of protein families and paralogs thus have the same age. We expect that gene duplications influence network properties in terms of protein age. In chapter 4 we demonstrate that indeed interactions between paralogs as well as interactions shared by paralogs contribute to the overrepresentation of interactions between proteins of a similar age.

We've extended the Duplication-Divergence model originally proposed by Vazquez et al., incorporating protein families. In this model, daughter nodes subfunctionalize after duplication, losing complementary sets of connections, but typically sharing part of the ancestral interactions as well. We find that this model can replicate the overrepresentation of interparalog interactions even when the probability of connecting daughter genes after duplication is low. In contrast, paralogs share relatively few interaction partners in networks generated by our model. We find that in networks produced by our model, the overrepresentation of interactions between proteins of a similar age very strongly depends on interparalog edges and we do not observe this in most real protein interaction networks.

We further extend the model and include node loss. We compare network topology for networks generated while keeping a stable network size instead of continuously growing. Without gain of novel interactions, networks break up into a large number of small disconnected components. In addition to randomly selected nodes, we remove all nodes that are completely disconnected from the network. This leads to degree-dependent node loss: the probability that a node will be deleted at some point in time, depends on the number of its interactions. If a node is lost, the degree of all other nodes in the same module decreases, thus increasing the probability they will be removed. When observing the network at coarse grained time intervals, this process will resemble concerted loss of module members.

Network growth models allow us to investigate the consequences of basic assumptions on how protein family dynamics translate into network level changes, thus providing a new angle from which we can study the interplay between genomic- and network evolution. A modular organization of the cellular machinery is a very robust outcome of our models, indicating that although the mechanism of duplication followed by subfunctionalization alone is not sufficient to explain all aspects of protein interaction network topology, it may be pivotal in generating the modular organization we observe in the cell. Hence a useful extension to the model would be to generate multiple species by restarting different simulations from the same ancestral network. This will enable us to investigate how network modules change in evolution, whether rewiring between modules is more extensive than rewiring within modules (as reported in the literature (Zinman et al., 2011; Roguev et al., 2008)) and to what extent network modules evolve cohesively without direct, explicit module-level selection.

Bibliography

Adamcsek, B., G. Palla, I. J. Farkas, I. Derenyi, and T. Vicsek, 2006: Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics (Oxford, England)*, **22 (8)**, 1021–1023. (Cited on page 30.)

AIMC, 2011: Evidence for network evolution in an arabidopsis interactome map. *Science (New York, N.Y.)*, **333 (6042)**, 601–607. (Cited on pages 52, 71, and 98.)

Al-Shahrour, F., P. Minguez, T. Marques-Bonet, E. Gazave, A. Navarro, and J. Dopazo, 2010: Selection upon genome architecture: conservation of functional neighborhoods with changing genes. *PLoS computational biology*, **6 (10)**, e1000 953. (Cited on page 98.)

Ali, W. and C. M. Deane, 2010: Evolutionary analysis reveals low coverage as the major challenge for protein interaction network alignment. *Molecular bioSystems*, **6 (11)**, 2296–2304. (Cited on pages 51, 58, and 82.)

Aloy, P., et al., 2004: Structure-based assembly of protein complexes in yeast. *Science (New York, N.Y.)*, **303 (5666)**, 2026–2029. (Cited on pages 18 and 29.)

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990: Basic local alignment search tool. *Journal of Molecular Biology*, **215 (3)**, 403–410. (Cited on pages 4, 35, and 46.)

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, 1997: Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, **25 (17)**, 3389–3402. (Cited on page 35.)

Ashburner, M., et al., 2000: Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, **25 (1)**, 25–29. (Cited on pages 27, 46, and 96.)

Bammler, T., et al., 2005: Standardizing global gene expression analysis between laboratories and across platforms. *Nature methods*, **2 (5)**, 351–356. (Cited on page 6.)

Bandyopadhyay, S., R. Sharan, and T. Ideker, 2006: Systematic identification of

functional orthologs based on protein network comparison. *Genome research*, **16 (3)**, 428–435. (Cited on pages 35 and 42.)

Barabasi, A. L. and R. Albert, 1999: Emergence of scaling in random networks. *Science (New York, N.Y.)*, **286 (5439)**, 509–512. (Cited on pages 11, 63, 71, 86, 92, and 98.)

Baryshnikova, A., et al., 2010: Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature methods*, **7 (12)**, 1017–1024. (Cited on pages 6 and 97.)

Batada, N. N., T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, L. D. Hurst, and M. Tyers, 2006: Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS biology*, **4 (10)**, e317. (Cited on page 71.)

Bateman, A., et al., 2004: The pfam protein families database. *Nucleic acids research*, **32 (Database issue)**, D138–41. (Cited on page 47.)

Berg, J., M. Lassig, and A. Wagner, 2004: Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC evolutionary biology*, **4**, 51. (Cited on pages 71, 82, 86, and 92.)

Bergthorsson, U., D. I. Andersson, and J. R. Roth, 2007: Ohno's dilemma: evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences of the United States of America*, **104 (43)**, 17 004–17 009. (Cited on page 52.)

Bershtein, S. and D. S. Tawfik, 2008: Ohno's model revisited: measuring the frequency of potentially adaptive mutations under various mutational drifts. *Molecular biology and evolution*, **25 (11)**, 2311–2318. (Cited on page 52.)

Bloom, J. D. and C. Adami, 2003: Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC evolutionary biology*, **3**, 21. (Cited on page 11.)

Boekhorst, J. and B. Snel, 2007: Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. *BMC bioinformatics*, **8**, 356. (Cited on page 35.)

Boube, M., L. Joulia, D. L. Cribbs, and H. M. Bourbon, 2002: Evidence for a mediator of rna polymerase ii transcriptional regulation conserved from yeast to man. *Cell*, **110 (2)**, 143–151. (Cited on page 46.)

Bourbon, H. M., 2008: Comparative genomics supports a deep evolutionary origin for the large, four-module transcriptional mediator complex. *Nucleic acids research*, **36 (12)**, 3993–4008. (Cited on pages 17 and 28.)

Byrne, K. P. and K. H. Wolfe, 2005: The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome research*, **15 (10)**, 1456–1461. (Cited on page 35.)

Campillos, M., C. von Mering, L. J. Jensen, and P. Bork, 2006: Identification

and analysis of evolutionarily cohesive functional modules in protein networks. *Genome research*, **16 (3)**, 374–382. (Cited on pages 12, 17, 18, 27, 39, 42, and 96.)

Capra, J. A., K. S. Pollard, and M. Singh, 2010: Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome biology*, **11 (12)**, R127. (Cited on pages 51, 63, 64, 89, 90, 91, and 98.)

Chen, K., D. Durand, and M. Farach-Colton, 2000: Notung: a program for dating gene duplications and optimizing gene family trees. *Journal of computational biology : a journal of computational molecular cell biology*, **7 (3-4)**, 429–447. (Cited on page 4.)

Chen, Y. and N. V. Dokholyan, 2006: The coordinated evolution of yeast proteins is constrained by functional modularity. *Trends in genetics : TIG*, **22 (8)**, 416–419. (Cited on page 45.)

Collins, S. R., P. Kemmeren, X. C. Zhao, J. F. Greenblatt, F. Spencer, F. C. Holstege, J. S. Weissman, and N. J. Krogan, 2007a: Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Molecular & cellular proteomics : MCP*, **6 (3)**, 439–450. (Cited on pages 7, 23, 30, 54, 64, 92, and 96.)

Collins, S. R., M. Schuldiner, N. J. Krogan, and J. S. Weissman, 2006: A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome biology*, **7 (7)**, R63. (Cited on pages 6 and 97.)

Collins, S. R., et al., 2007b: Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, **446 (7137)**, 806–810. (Cited on pages 8 and 97.)

Copley, S. D., 2012: Moonlighting is mainstream: paradigm adjustment required. *BioEssays : news and reviews in molecular, cellular and developmental biology*, **34 (7)**, 578–588. (Cited on page 6.)

Cordero, O. X. and P. Hogeweg, 2006: Feed-forward loop circuits as a side effect of genome evolution. *Molecular biology and evolution*, **23 (10)**, 1931–1936. (Cited on page 98.)

Cordero, O. X., B. Snel, and P. Hogeweg, 2008: Coevolution of gene families in prokaryotes. *Genome research*, **18 (3)**, 462–468. (Cited on page 28.)

Cornelissen, A. W., R. Evers, and J. Kock, 1988: Structure and sequence of genes encoding subunits of eukaryotic rna polymerases. *Oxford surveys on eukaryotic genes*, **5**, 91–131. (Cited on page 57.)

Costanzo, M., et al., 2010: The genetic landscape of a cell. *Science (New York, N.Y.)*, **327 (5964)**, 425–431. (Cited on page 97.)

Cuypers, T. D. and P. Hogeweg, 2012: Virtual genomes in flux: an interplay of neutrality and adaptability explains genome expansion and streamlining. *Genome biology and evolution*, **4 (3)**, 212–229. (Cited on page 91.)

Dalmolin, R. J., M. A. Castro, J. L. R. Filho, L. H. Souza, R. M. de Almeida, and J. C. Moreira, 2011: Evolutionary plasticity determination by orthologous groups distribution. *Biology direct*, **6**, 22. (Cited on page 63.)

de Lichtenberg, U., L. J. Jensen, S. Brunak, and P. Bork, 2005: Dynamic complex formation during the yeast cell cycle. *Science (New York, N.Y.)*, **307 (5710)**, 724–727. (Cited on page 6.)

Dehal, P. S. and J. L. Boore, 2006: A phylogenomic gene cluster resource: the phylogenetically inferred groups (phigs) database. *BMC bioinformatics*, **7**, 201. (Cited on page 4.)

Dittmer, T. A. and T. Misteli, 2011: The lamin protein family. *Genome biology*, **12 (5)**, 222–2011–12–5–222. Epub 2011 May 31. (Cited on page 4.)

Dobzhansky, T., 1964: Biology, molecular and organismic. *American Zoologist*, **4**, 443–452. (Cited on page 71.)

Dodds, P. N., 2010: Plant science. genome evolution in plant pathogens. *Science (New York, N.Y.)*, **330 (6010)**, 1486–1487. (Cited on page 5.)

Drummond, D. A., A. Raval, and C. O. Wilke, 2006: A single determinant dominates the rate of yeast protein evolution. *Molecular biology and evolution*, **23 (2)**, 327–337. (Cited on page 10.)

Drummond, D. A. and C. O. Wilke, 2008: Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134 (2)**, 341–352. (Cited on page 11.)

Dudley, A. M., D. M. Janse, A. Tanay, R. Shamir, and G. M. Church, 2005: A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular systems biology*, **1**, 2005.0001. (Cited on page 6.)

Eisenberg, E. and E. Y. Levanon, 2003: Preferential attachment in the protein network evolution. *Physical Review Letters*, **91 (13)**, 138 701. (Cited on pages 9, 10, 63, 89, and 98.)

Elhaik, E., N. Sabath, and D. Graur, 2006: The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Molecular biology and evolution*, **23 (1)**, 1–3. (Cited on page 10.)

Espadaler, J., N. Eswar, E. Querol, F. X. Aviles, A. Sali, M. A. Marti-Renom, and B. Oliva, 2008: Prediction of enzyme function by combining sequence similarity and protein interactions. *BMC bioinformatics*, **9**, 249–2105–9–249. (Cited on page 35.)

Evlampiev, K. and H. Isambert, 2007: Modeling protein network evolution under genome duplication and domain shuffling. *BMC systems biology*, **1**, 49. (Cited on page 58.)

Featherstone, D. E. and K. Broadie, 2002: Wrestling with pleiotropy: genomic

and topological analysis of the yeast gene expression network. *BioEssays : news and reviews in molecular, cellular and developmental biology*, **24 (3)**, 267–274. (Cited on page 6.)

Fernandes, L. P., A. Annibale, J. Kleinjung, A. C. Coolen, and F. Fraternali, 2010: Protein networks reveal detection bias and species consistency when analysed by information-theoretic methods. *PloS one*, **5 (8)**, e12 083. (Cited on page 7.)

Fiedler, D., et al., 2009: Functional organization of the s. cerevisiae phosphorylation network. *Cell*, **136 (5)**, 952–963. (Cited on pages 8 and 97.)

Fischer, G., E. P. Rocha, F. Brunet, M. Vergassola, and B. Dujon, 2006: Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS genetics*, **2 (3)**, e32. (Cited on page 97.)

Fokkens, L., P. Hogeweg, and B. Snel, 2012: Gene duplications contribute to the overrepresentation of interactions between proteins of a similar age. *BMC evolutionary biology*, **12 (1)**, 99. (Cited on pages 10, 71, 74, 75, 89, 90, 92, and 93.)

Fokkens, L. and B. Snel, 2009: Cohesive versus flexible evolution of functional modules in eukaryotes. *PLoS computational biology*, **5 (1)**, e1000 276. (Cited on pages 39 and 42.)

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait, 1999: Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151 (4)**, 1531–1545. (Cited on pages 52 and 71.)

Francino, M. P., 2005: An adaptive radiation model for the origin of new gene functions. *Nature genetics*, **37 (6)**, 573–577. (Cited on page 52.)

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman, 2002: Evolutionary rate in the protein interaction network. *Science (New York, N.Y.)*, **296 (5568)**, 750–752. (Cited on page 11.)

Fraser, H. B., D. P. Wall, and A. E. Hirsh, 2003: A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC evolutionary biology*, **3**, 11. (Cited on pages 11 and 82.)

Gabaldon, T., 2008: Large-scale assignment of orthology: back to phylogenetics? *Genome biology*, **9 (10)**, 235–2008–9–10–235. (Cited on page 4.)

Gabaldon, T., D. Rainey, and M. A. Huynen, 2005: Tracing the evolution of a large protein complex in the eukaryotes, nadh:ubiquinone oxidoreductase (complex i). *Journal of Molecular Biology*, **348 (4)**, 857–870. (Cited on pages 17, 28, and 42.)

Gabrielsen, O. S. and A. Sentenac, 1991: Rna polymerase iii (c) and its transcription factors. *Trends in biochemical sciences*, **16 (11)**, 412–416. (Cited on page 57.)

Gavin, A. C., et al., 2002: Functional organization of the yeast proteome by sys-

tematic analysis of protein complexes. *Nature*, **415 (6868)**, 141–147. (Cited on pages 2, 64, and 92.)

Gavin, A. C., et al., 2006: Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440 (7084)**, 631–636. (Cited on pages 2, 7, 8, 18, 22, 23, 29, 30, 54, 64, 71, 92, and 96.)

Ghaemmaghami, S., W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, 2003: Global analysis of protein expression in yeast. *Nature*, **425 (6959)**, 737–741. (Cited on page 54.)

Giaever, G., et al., 2002: Functional profiling of the saccharomyces cerevisiae genome. *Nature*, **418 (6896)**, 387–391. (Cited on page 6.)

Gibson, T. A. and D. S. Goldberg, 2009: Questioning the ubiquity of neofunctionalization. *PLoS computational biology*, **5 (1)**, e1000 252. (Cited on pages 60 and 82.)

Glansdorff, N., Y. Xu, and B. Labedan, 2008: The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biology direct*, **3**, 29. (Cited on page 51.)

Glazko, G. V. and A. R. Mushegian, 2004: Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome biology*, **5 (5)**, R32. (Cited on pages 17 and 27.)

Grandi, P., et al., 2002: 90s pre-ribosomes include the 35s pre-rrna, the u3 snornp, and 40s subunit processing factors but predominantly lack 60s synthesis factors. *Molecular cell*, **10 (1)**, 105–115. (Cited on pages 64 and 92.)

Grove, C. A. and A. J. Walhout, 2008: Transcription factor functionality and transcription regulatory networks. *Molecular bioSystems*, **4 (4)**, 309–314. (Cited on page 7.)

Hakes, L., S. C. Lovell, S. G. Oliver, and D. L. Robertson, 2007: Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences of the United States of America*, **104 (19)**, 7999–8004. (Cited on page 55.)

Hakes, L., D. L. Robertson, and S. G. Oliver, 2005: Effect of dataset selection on the topological interpretation of protein interaction networks. *BMC genomics*, **6**, 131. (Cited on pages 7 and 75.)

Hallinan, J., 2004: Gene duplication and hierarchical modularity in intracellular interaction networks. *Bio Systems*, **74 (1-3)**, 51–62. (Cited on page 98.)

Han, J. D., et al., 2004: Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430 (6995)**, 88–93. (Cited on page 6.)

Hart, G. T., I. Lee, and E. R. Marcotte, 2007: A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC*

bioinformatics, **8**, 236. (Cited on pages 82 and 91.)

Haynes, C., et al., 2006: Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS computational biology*, **2 (8)**, e100. (Cited on page 86.)

He, X. and J. Zhang, 2005: Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, **169 (2)**, 1157–1164. (Cited on pages 74 and 84.)

He, X. and J. Zhang, 2006: Why do hubs tend to be essential in protein networks? *PLoS genetics*, **2 (6)**, e88. (Cited on page 91.)

Hermjakob, H., et al., 2004: Intact: an open source molecular interaction database. *Nucleic acids research*, **32 (Database issue)**, D452–5. (Cited on page 7.)

Hillenmeyer, M. E., et al., 2008: The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (New York, N.Y.)*, **320 (5874)**, 362–365. (Cited on page 6.)

Hirsh, A. E., H. B. Fraser, and D. P. Wall, 2005: Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Molecular biology and evolution*, **22 (1)**, 174–177. (Cited on pages 10 and 26.)

Ho, Y., et al., 2002: Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, **415 (6868)**, 180–183. (Cited on pages 64 and 92.)

Hormozdiari, F., P. Berenbrink, N. Przulj, and S. C. Sahinalp, 2007: Not all scale-free networks are born equal: the role of the seed graph in ppi network evolution. *PLoS computational biology*, **3 (7)**, e118. (Cited on pages 66, 71, and 75.)

Hubbard, T., et al., 2005: Ensembl 2005. *Nucleic acids research*, **33 (Database issue)**, D447–53. (Cited on page 46.)

Huerta-Cepas, J., S. Capella-Gutierrez, L. P. Pryszcz, I. Denisov, D. Kormes, M. Marcet-Houben, and T. Gabaldon, 2011: Phylomedb v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic acids research*, **39 (Database issue)**, D556–60. (Cited on page 4.)

Hughes, T. R., et al., 2000: Functional discovery via a compendium of expression profiles. *Cell*, **102 (1)**, 109–126. (Cited on page 6.)

Huynen, M. A., T. Dandekar, and P. Bork, 1999: Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends in microbiology*, **7 (7)**, 281–291. (Cited on pages 17 and 28.)

Innan, H. and F. Kondrashov, 2010: The evolution of gene duplications: classifying and distinguishing between models. *Nature reviews.Genetics*, **11 (2)**, 97–108. (Cited on page 52.)

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, 2001: A com-

prehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **98 (8)**, 4569–4574. (Cited on pages 2, 8, 64, and 92.)

Ivanic, J., X. Yu, A. Wallqvist, and J. Reifman, 2009: Influence of protein abundance on high-throughput protein-protein interaction detection. *PloS one*, **4 (6)**, e5815. (Cited on pages 7, 54, and 89.)

Jeffery, C. J., 2009: Moonlighting proteins–an update. *Molecular bioSystems*, **5 (4)**, 345–350. (Cited on page 6.)

Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, 2001: Lethality and centrality in protein networks. *Nature*, **411 (6833)**, 41–42. (Cited on page 82.)

Jordan, I. K., Y. I. Wolf, and E. V. Koonin, 2003: No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC evolutionary biology*, **3**, 1. (Cited on page 11.)

Jothi, R., T. M. Przytycka, and L. Aravind, 2007: Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC bioinformatics*, **8**, 173. (Cited on page 28.)

Jothi, R., E. Zotenko, A. Tasneem, and T. M. Przytycka, 2006: Coco-cl: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics (Oxford, England)*, **22 (7)**, 779–788. (Cited on page 5.)

Karev, G. P., Y. I. Wolf, and E. V. Koonin, 2003: Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics (Oxford, England)*, **19 (15)**, 1889–1900. (Cited on page 98.)

Karev, G. P., Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya, and E. V. Koonin, 2002: Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC evolutionary biology*, **2**, 18. (Cited on page 98.)

Kashtan, N. and U. Alon, 2005: Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, **102 (39)**, 13 773–13 778. (Cited on page 97.)

Keeling, P. J., N. Corradi, H. G. Morrison, K. L. Haag, D. Ebert, L. M. Weiss, D. E. Akiyoshi, and S. Tzipori, 2010: The reduced genome of the parasitic microsporidian enterocytozoon bieneusi lacks genes for core carbon metabolism. *Genome biology and evolution*, **2**, 304–309. (Cited on page 5.)

Kelley, R. and T. Ideker, 2005: Systematic interpretation of genetic interactions using protein networks. *Nature biotechnology*, **23 (5)**, 561–566. (Cited on pages 6, 7, and 97.)

Kensche, P. R., V. van Noort, B. E. Dutilh, and M. A. Huynen, 2008: Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *Journal of the Royal Society, Interface / the Royal Society*, **5 (19)**, 151–170. (Cited on page 9.)

Kim, P. M., L. J. Lu, Y. Xia, and M. B. Gerstein, 2006: Relating three-dimensional structures to protein networks provides evolutionary insights. *Science (New York, N.Y.)*, **314 (5807)**, 1938–1941. (Cited on page 82.)

Kim, S. H. and S. V. Yi, 2006: Correlated asymmetry of sequence and functional divergence between duplicate proteins of saccharomyces cerevisiae. *Molecular biology and evolution*, **23 (5)**, 1068–1075. (Cited on page 58.)

Kim, W. K. and E. M. Marcotte, 2008: Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS computational biology*, **4 (11)**, e1000 232. (Cited on pages 9, 11, 51, 54, 59, 61, 64, 65, 66, 67, 71, 75, 89, 92, 93, and 98.)

Koonin, E. V., 2005: Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, **39**, 309–338. (Cited on page 3.)

Koonin, E. V., 2007: The biological big bang model for the major transitions in evolution. *Biology direct*, **2**, 21. (Cited on page 91.)

Koonin, E. V., 2009: Evolution of genome architecture. *The international journal of biochemistry & cell biology*, **41 (2)**, 298–306. (Cited on page 35.)

Koonin, E. V., 2010: Preview. the incredible expanding ancestor of eukaryotes. *Cell*, **140 (5)**, 606–608. (Cited on pages 51 and 71.)

Koonin, E. V. and Y. I. Wolf, 2006: Evolutionary systems biology: links between gene evolution and function. *Current opinion in biotechnology*, **17 (5)**, 481–487. (Cited on page 63.)

Krogan, N. J., et al., 2006: Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440 (7084)**, 637–643. (Cited on pages 2, 7, 8, 18, 30, 54, 64, and 92.)

Kroiss, M., J. Schultz, J. Wiesner, A. Chari, A. Sickmann, and U. Fischer, 2008: Evolution of an rnp assembly system: a minimal smn complex facilitates formation of usnrnps in drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, **105 (29)**, 10 045–10 050. (Cited on pages 28 and 42.)

Krylov, D. M., Y. I. Wolf, I. B. Rogozin, and E. V. Koonin, 2003: Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome research*, **13 (10)**, 2229–2235. (Cited on pages 10, 11, 71, 82, and 90.)

Kunin, V., J. B. Pereira-Leal, and C. A. Ouzounis, 2004: Functional evolution of the yeast protein interaction network. *Molecular biology and evolution*, **21 (7)**, 1171–1176. (Cited on pages 63 and 89.)

Kuo, W. P., et al., 2006: A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nature biotechnology*, **24 (7)**, 832–840. (Cited on page 6.)

Langkjaer, R. B., M. L. Nielsen, P. R. Daugaard, W. Liu, and J. Piskur, 2000: Yeast chromosomes have been significantly reshaped during their evolutionary history. *Journal of Molecular Biology*, **304 (3)**, 271–288. (Cited on page 97.)

Levasseur, A. and P. Pontarotti, 2011: The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biology direct*, **6**, 11. (Cited on page 52.)

Lewis, A. C., N. S. Jones, M. A. Porter, and C. M. Deane, 2012: What evidence is there for the homology of protein-protein interactions? *PLoS computational biology*, **8 (9)**, e1002 645. (Cited on pages 8 and 9.)

Li, L., C. J. S. Jr, and D. S. Roos, 2003: Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, **13 (9)**, 2178–2189. (Cited on pages 4, 24, 30, and 35.)

Linden, R. O., V. P. Eronen, and T. Aittokallio, 2011: Quantitative maps of genetic interactions in yeast - comparative evaluation and integrative analysis. *BMC systems biology*, **5**, 45–0509–5–45. (Cited on page 97.)

Liu, Y., T. A. Richards, and S. J. Aves, 2009: Ancient diversification of eukaryotic mcm dna replication proteins. *BMC evolutionary biology*, **9**, 60. (Cited on page 55.)

Lovell, S. C. and D. L. Robertson, 2010: An integrated view of molecular coevolution in protein-protein interactions. *Molecular biology and evolution*, **27 (11)**, 2567–2575. (Cited on page 55.)

Lynch, M. and A. Force, 2000: The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154 (1)**, 459–473. (Cited on page 52.)

Maslov, S. and K. Sneppen, 2002: Specificity and stability in topology of protein networks. *Science (New York, N.Y.)*, **296 (5569)**, 910–913. (Cited on page 75.)

Matthews, L., et al., 2009: Reactome knowledgebase of human biological pathways and processes. *Nucleic acids research*, **37 (Database issue)**, D619–22. (Cited on pages 37 and 46.)

Metzker, M. L., 2010: Sequencing technologies - the next generation. *Nature reviews.Genetics*, **11 (1)**, 31–46. (Cited on page 8.)

Mewes, H. W., et al., 2008: Mips: analysis and annotation of genome information in 2007. *Nucleic acids research*, **36 (Database issue)**, D196–201. (Cited on pages 29, 37, and 46.)

Michaut, M., A. Baryshnikova, M. Costanzo, C. L. Myers, B. J. Andrews, C. Boone, and G. D. Bader, 2011: Protein complexes are central in the yeast genetic landscape. *PLoS computational biology*, **7 (2)**, e1001 092. (Cited on page 97.)

Middelbeek, J., K. Clark, H. Venselaar, M. A. Huynen, and F. N. van Leeuwen, 2010: The alpha-kinase family: an exceptional branch on the protein kinase tree. *Cellular and molecular life sciences : CMLS*, **67 (6)**, 875–890. (Cited on

page 4.)

Mittenthal, J., D. Caetano-Anolles, and G. Caetano-Anolles, 2012: Biphasic patterns of diversification and the emergence of modules. *Frontiers in genetics*, **3**, 147. (Cited on page 91.)

Monahan, B. J., J. Villen, S. Marguerat, J. Bahler, S. P. Gygi, and F. Winston, 2008: Fission yeast swi/snf and rsc complexes show compositional and functional differences from budding yeast. *Nature structural & molecular biology*, **15 (8)**, 873–880. (Cited on pages 17 and 28.)

Morrison, H. G., et al., 2007: Genomic minimalism in the early diverging intestinal parasite giardia lamblia. *Science (New York, N.Y.)*, **317 (5846)**, 1921–1926. (Cited on page 5.)

Muller, J., et al., 2010: eggnog v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic acids research*, **38 (Database issue)**, D190–5. (Cited on pages 4, 5, 52, and 64.)

Musso, G., Z. Zhang, and A. Emili, 2007: Retention of protein complex membership by ancient duplicated gene products in budding yeast. *Trends in genetics : TIG*, **23 (6)**, 266–269. (Cited on pages 52, 71, and 98.)

Myers, C. L., C. Chiriac, and O. G. Troyanskaya, 2009: Discovering biological networks from diverse functional genomic data. *Methods in molecular biology (Clifton, N.J.)*, **563**, 157–175. (Cited on page 8.)

Navlakha, S. and C. Kingsford, 2011: Network archaeology: uncovering ancient networks from present-day interactions. *PLoS computational biology*, **7 (4)**, e1001 119. (Cited on page 71.)

Neher, E., 1994: How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the United States of America*, **91 (1)**, 98–102. (Cited on page 28.)

Newman, M. E. and M. Girvan, 2004: Finding and evaluating community structure in networks. *Physical review.E, Statistical, nonlinear, and soft matter physics*, **69 (2 Pt 2)**, 026 113. (Cited on page 92.)

Notebaart, R. A., M. A. Huynen, B. Teusink, R. J. Siezen, and B. Snel, 2005: Correlation between sequence conservation and the genomic context after gene duplication. *Nucleic acids research*, **33 (19)**, 6164–6171. (Cited on page 42.)

Osbourn, A., 2010: Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. *Plant Physiology*, **154 (2)**, 531–535. (Cited on page 97.)

Pal, C., B. Papp, and M. J. Lercher, 2006: An integrated view of protein evolution. *Nature reviews.Genetics*, **7 (5)**, 337–348. (Cited on page 10.)

Papp, B., C. Pal, and L. D. Hurst, 2003: Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424 (6945)**, 194–197. (Cited on pages 9

and 97.)

Park, D. and S. S. Choi, 2009: Why proteins evolve at different rates: the functional hypothesis versus the mistranslation-induced protein misfolding hypothesis. *FEBS letters*, **583 (7)**, 1053–1059. (Cited on page 63.)

Pastor-Satorras, R., E. Smith, and R. V. Sole, 2003: Evolving protein interaction networks through gene duplication. *Journal of theoretical biology*, **222 (2)**, 199–210. (Cited on pages 84 and 98.)

Pazos, F., D. Juan, J. M. Izarzugaza, E. Leon, and A. Valencia, 2008: Prediction of protein interaction based on similarity of phylogenetic trees. *Methods in molecular biology (Clifton, N.J.)*, **484**, 523–535. (Cited on page 9.)

Pazos, F. and A. Valencia, 2001: Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein engineering*, **14 (9)**, 609–614. (Cited on page 28.)

Pazos, F. and A. Valencia, 2008: Protein co-evolution, co-adaptation and interactions. *The EMBO journal*, **27 (20)**, 2648–2655. (Cited on page 55.)

Pearson, W. R. and D. J. Lipman, 1988: Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85 (8)**, 2444–2448. (Cited on page 4.)

Pellegrini, M., 2012: Using phylogenetic profiles to predict functional relationships. *Methods in molecular biology (Clifton, N.J.)*, **804**, 167–177. (Cited on page 9.)

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, 1999: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96 (8)**, 4285–4288. (Cited on page 17.)

Pereira-Leal, J. B., E. D. Levy, C. Kamp, and S. A. Teichmann, 2007: Evolution of protein complexes by duplication of homomeric interactions. *Genome biology*, **8 (4)**, R51. (Cited on pages 55, 71, and 98.)

Pereira-Leal, J. B., E. D. Levy, and S. A. Teichmann, 2006: The origins and evolution of functional modules: lessons from protein complexes. *Philosophical transactions of the Royal Society of London.Series B, Biological sciences*, **361 (1467)**, 507–517. (Cited on pages 44, 55, and 98.)

Pereira-Leal, J. B. and S. A. Teichmann, 2005: Novel specificities emerge by stepwise duplication of functional modules. *Genome research*, **15 (4)**, 552–559. (Cited on pages 55 and 57.)

Ptacek, J., et al., 2005: Global analysis of protein phosphorylation in yeast. *Nature*, **438 (7068)**, 679–684. (Cited on pages 64 and 92.)

Qin, H., H. H. Lu, W. B. Wu, and W. H. Li, 2003: Evolution of the yeast protein interaction network. *Proceedings of the National Academy of Sciences of the United*

States of America, **100 (22)**, 12 820–12 824. (Cited on pages 9, 51, 64, 89, and 98.)

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi, 2002: Hierarchical organization of modularity in metabolic networks. *Science (New York, N.Y.)*, **297 (5586)**, 1551–1555. (Cited on page 96.)

Reguly, T., et al., 2006: Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae. *Journal of biology*, **5 (4)**, 11. (Cited on pages 52, 64, and 92.)

Remm, M., C. E. Storm, and E. L. Sonnhammer, 2001: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, **314 (5)**, 1041–1052. (Cited on pages 4, 30, 35, 37, and 47.)

Rep, M. and H. C. Kistler, 2010: The genomic organization of plant pathogenicity in fusarium species. *Current opinion in plant biology*, **13 (4)**, 420–426. (Cited on page 97.)

Rognes, T., 2001: Paralign: a parallel sequence alignment algorithm for rapid and sensitive database searches. *Nucleic acids research*, **29 (7)**, 1647–1652. (Cited on pages 4 and 30.)

Roguev, A., et al., 2008: Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science (New York, N.Y.)*, **322 (5900)**, 405–410. (Cited on page 99.)

Rorick, M. M. and G. P. Wagner, 2011: Protein structural modularity and robustness are associated with evolvability. *Genome biology and evolution*, **3**, 456–475. (Cited on page 97.)

Ruepp, A., et al., 2004: The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic acids research*, **32 (18)**, 5539–5545. (Cited on page 96.)

Ruepp, A., et al., 2010: Corum: the comprehensive resource of mammalian protein complexes–2009. *Nucleic acids research*, **38 (Database issue)**, D497–501. (Cited on pages 37 and 46.)

Ryan, C. J., et al., 2012: Hierarchical modularity and the evolution of genetic interactomes across species. *Molecular cell*, **46 (5)**, 691–704. (Cited on page 97.)

Saeed, R. and C. M. Deane, 2006: Protein protein interactions, evolutionary rate, abundance and age. *BMC bioinformatics*, **7**, 128. (Cited on page 63.)

Samal, A., A. Wagner, and O. C. Martin, 2011: Environmental versatility promotes modularity in genome-scale metabolic networks. *BMC systems biology*, **5**, 135–0509–5–135. (Cited on page 97.)

Sambourg, L. and N. Thierry-Mieg, 2010: New insights into protein-protein interaction data lead to increased estimates of the s. cerevisiae interactome size. *BMC bioinformatics*, **11**, 605–2105–11–605. (Cited on page 7.)

Sarikas, A., T. Hartmann, and Z. Q. Pan, 2011: The cullin protein family. *Genome biology*, **12 (4)**, 220–2011–12–4–220. Epub 2011 Apr 28. (Cited on page 4.)

Scannell, D. R., A. C. Frank, G. C. Conant, K. P. Byrne, M. Woolfit, and K. H. Wolfe, 2007: Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proceedings of the National Academy of Sciences of the United States of America*, **104 (20)**, 8397–8402. (Cited on page 91.)

Schuldiner, M., et al., 2005: Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, **123 (3)**, 507–519. (Cited on pages 8 and 97.)

Seoighe, C., et al., 2000: Prevalence of small inversions in yeast gene order evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **97 (26)**, 14 433–14 437. (Cited on page 97.)

Shannon, P., et al., 2003: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13 (11)**, 2498–2504. (Cited on page 27.)

Shou, C., N. Bhardwaj, H. Y. Lam, K. K. Yan, P. M. Kim, M. Snyder, and M. B. Gerstein, 2011: Measuring the evolutionary rewiring of biological networks. *PLoS computational biology*, **7 (1)**, e1001 050. (Cited on pages 9, 12, and 91.)

Singh, A. H., D. M. Wolf, P. Wang, and A. P. Arkin, 2008a: Modularity of stress response evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **105 (21)**, 7500–7505. (Cited on pages 17 and 28.)

Singh, R., J. Xu, and B. Berger, 2008b: Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences of the United States of America*, **105 (35)**, 12 763–12 768. (Cited on page 35.)

Smits, P., J. A. Smeitink, L. P. van den Heuvel, M. A. Huynen, and T. J. Ettema, 2007: Reconstructing the evolution of the mitochondrial ribosomal proteome. *Nucleic acids research*, **35 (14)**, 4686–4703. (Cited on pages 17, 28, and 46.)

Snel, B., P. Bork, and M. A. Huynen, 2002: Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome research*, **12 (1)**, 17–25. (Cited on pages 71 and 90.)

Snel, B. and M. A. Huynen, 2004: Quantifying modularity in the evolution of biomolecular systems. *Genome research*, **14 (3)**, 391–397. (Cited on pages 12, 17, 22, 23, 26, 27, 39, 42, and 96.)

Snitkin, E. S., A. M. Gustafson, J. Mellor, J. Wu, and C. DeLisi, 2006: Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC bioinformatics*, **7**, 420. (Cited on page 17.)

Soding, J., 2005: Protein homology detection by hmm-hmm comparison. *Bioinformatics (Oxford, England)*, **21 (7)**, 951–960. (Cited on page 35.)

Sole, R. V. and S. Valverde, 2008: Spontaneous emergence of modularity in cellular networks. *Journal of the Royal Society, Interface / the Royal Society*, **5 (18)**, 129–133. (Cited on pages 71, 75, and 98.)

Sonnhammer, E. L. and E. V. Koonin, 2002: Orthology, paralogy and proposed classification for paralog subtypes. *Trends in genetics : TIG*, **18 (12)**, 619–620. (Cited on page 3.)

Stark, C., B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, 2006: Biogrid: a general repository for interaction datasets. *Nucleic acids research*, **34 (Database issue)**, D535–9. (Cited on page 7.)

Stoltzfus, A., 1999: On the possibility of constructive neutral evolution. *Journal of Molecular Evolution*, **49 (2)**, 169–181. (Cited on page 51.)

Stuart, J. M., E. Segal, D. Koller, and S. K. Kim, 2003: A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, **302 (5643)**, 249–255. (Cited on page 6.)

Szappanos, B., et al., 2011: An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature genetics*, **43 (7)**, 656–662. (Cited on page 97.)

Szklarczyk, R., M. A. Huynen, and B. Snel, 2008: Complex fate of paralogs. *BMC evolutionary biology*, **8**, 337. (Cited on pages 44, 46, and 52.)

Tanaka, T., Y. Tateno, and T. Gojobori, 2005: Evolution of vitamin b6 (pyridoxine) metabolism by gain and loss of genes. *Molecular biology and evolution*, **22 (2)**, 243–250. (Cited on pages 17 and 28.)

Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin, 2000: The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, **28 (1)**, 33–36. (Cited on pages 4 and 5.)

Tatusov, R. L., et al., 2003: The cog database: an updated version includes eukaryotes. *BMC bioinformatics*, **4**, 41. (Cited on pages 4, 5, 18, 29, and 30.)

Teichmann, S. A. and R. A. Veitia, 2004: Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics*, **167 (4)**, 2121–2125. (Cited on page 97.)

Tong, A. H., et al., 2001: Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science (New York, N.Y.)*, **294 (5550)**, 2364–2368. (Cited on page 6.)

Tong, A. H., et al., 2004: Global mapping of the yeast genetic interaction network. *Science (New York, N.Y.)*, **303 (5659)**, 808–813. (Cited on page 6.)

Tusscher, K. H. T. and P. Hogeweg, 2011: Evolution of networks for body plan patterning; interplay of modularity, robustness and evolvability. *PLoS computational biology*, **7 (10)**, e1002 208. (Cited on page 97.)

Uetz, P., et al., 2000: A comprehensive analysis of protein-protein interactions in

saccharomyces cerevisiae. *Nature*, **403 (6770)**, 623–627. (Cited on pages 2, 7, 64, and 92.)

van Dam, T. J., J. L. Bos, and B. Snel, 2011: Evolution of the ras-like small gtpases and their regulators. *Small GTPases*, **2 (1)**, 4–16. (Cited on page 4.)

van Dam, T. J., H. Rehmann, J. L. Bos, and B. Snel, 2009: Phylogeny of the cdc25 homology domain reveals rapid differentiation of ras pathways between early animals and fungi. *Cellular signalling*, **21 (11)**, 1579–1585. (Cited on page 4.)

van Dam, T. J. and B. Snel, 2008: Protein complex evolution does not involve extensive network rewiring. *PLoS computational biology*, **4 (7)**, e1000 132. (Cited on pages 9, 12, and 91.)

van Noort, V., B. Snel, and M. A. Huynen, 2003: Predicting gene function by conserved co-expression. *Trends in genetics : TIG*, **19 (5)**, 238–242. (Cited on page 6.)

van Noort, V., B. Snel, and M. A. Huynen, 2004: The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO reports*, **5 (3)**, 280–284. (Cited on page 98.)

van Noort, V., B. Snel, and M. A. Huynen, 2007: Exploration of the omics evidence landscape: adding qualitative labels to predicted protein-protein interactions. *Genome biology*, **8 (9)**, R197. (Cited on page 7.)

van Wageningen, S., et al., 2010: Functional overlap and regulatory links shape genetic interactions between signaling pathways. *Cell*, **143 (6)**, 991–1004. (Cited on page 6.)

Vazquez, A., A. Flammini, A. Maritan, and A. Vespignani, 2003: Modeling of protein interaction networks. *ComplexUs*, **1 (38)**. (Cited on pages 11, 58, 71, 74, 75, and 98.)

Veretnik, S., C. Wills, P. Youkharibache, R. E. Valas, and P. E. Bourne, 2009: Sm/lsm genes provide a glimpse into the early evolution of the spliceosome. *PLoS computational biology*, **5 (3)**, e1000 315. (Cited on page 55.)

Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney, 2009: Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, **19 (2)**, 327–335. (Cited on page 4.)

von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, 2003: String: a database of predicted functional associations between proteins. *Nucleic acids research*, **31 (1)**, 258–261. (Cited on page 8.)

von Mering, C., R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, 2002: Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417 (6887)**, 399–403. (Cited on page 8.)

Wagner, A., 2002: Asymmetric functional divergence of duplicate genes in yeast. *Molecular biology and evolution*, **19 (10)**, 1760–1768. (Cited on page 58.)

Wagner, A., 2003: How the global structure of protein interaction networks evolves. *Proceedings.Biological sciences / The Royal Society*, **270 (1514)**, 457–466. (Cited on page 60.)

Wagner, G. P., M. Pavlicev, and J. M. Cheverud, 2007: The road to modularity. *Nature reviews.Genetics*, **8 (12)**, 921–931. (Cited on pages 96 and 97.)

Walhout, A. J., 2011: What does biologically meaningful mean? a perspective on gene regulatory network validation. *Genome biology*, **12 (4)**, 109–2011–12–4–109. Epub 2011 Apr 11. (Cited on page 7.)

Walhout, A. J., R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg, and M. Vidal, 2000: Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science (New York, N.Y.)*, **287 (5450)**, 116–122. (Cited on page 8.)

Walhout, A. J. and M. Vidal, 1999: A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome research*, **9 (11)**, 1128–1134. (Cited on page 7.)

Walhout, A. J. and M. Vidal, 2001: High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods (San Diego, Calif.)*, **24 (3)**, 297–306. (Cited on page 2.)

Wall, D. P., A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen, and M. W. Feldman, 2005: Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **102 (15)**, 5483–5488. (Cited on pages 55 and 63.)

Wapinski, I., A. Pfeffer, N. Friedman, and A. Regev, 2007: Natural history and evolutionary principles of gene duplication in fungi. *Nature*, **449 (7158)**, 54–61. (Cited on pages 35, 71, and 90.)

Warnefors, M. and A. Eyre-Walker, 2011: The accumulation of gene regulation through time. *Genome biology and evolution*, **3**, 667–673. (Cited on pages 10, 63, and 98.)

Wasserman, S., 1994: *Social network analysis : methods and applications*. Cambridge University Press, Cambridge. (Cited on page 92.)

Waterhouse, R. M., E. M. Zdobnov, and E. V. Kriventseva, 2011: Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. *Genome biology and evolution*, **3**, 75–86. (Cited on page 63.)

Watts, D. J. and S. H. Strogatz, 1998: Collective dynamics of 'small-world' networks. *Nature*, **393 (6684)**, 440–442. (Cited on page 71.)

Wilmes, G. M., et al., 2008: A genetic interaction map of rna-processing factors reveals links between sem1/dss1-containing complexes and mrna export and splicing. *Molecular cell*, **32 (5)**, 735–746. (Cited on pages 8 and 97.)

Winzeler, E. A., et al., 1999: Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science (New York, N.Y.)*, **285 (5429)**, 901–906. (Cited on page 6.)

Wolf, M. Y., Y. I. Wolf, and E. V. Koonin, 2008: Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biology direct*, **3**, 40. (Cited on page 63.)

Wolf, Y. I., I. V. Gopich, D. J. Lipman, and E. V. Koonin, 2010: Relative contributions of intrinsic structural-functional constraints and translation rate to the evolution of protein-coding genes. *Genome biology and evolution*, **2**, 190–199. (Cited on page 10.)

Wolf, Y. I., P. S. Novichkov, G. P. Karev, E. V. Koonin, and D. J. Lipman, 2009: The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, **106 (18)**, 7273–7280. (Cited on pages 10, 55, and 63.)

Wolfe, C. J., I. S. Kohane, and A. J. Butte, 2005: Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC bioinformatics*, **6**, 227. (Cited on page 6.)

Wolfe, K., 2004: Evolutionary genomics: yeasts accelerate beyond blast. *Current biology : CB*, **14 (10)**, R392–4. (Cited on page 94.)

Xenarios, I., D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, 2000: Dip: the database of interacting proteins. *Nucleic acids research*, **28 (1)**, 289–291. (Cited on page 7.)

Young, R. A., 1991: Rna polymerase ii. *Annual Review of Biochemistry*, **60**, 689–715. (Cited on page 57.)

Yu, H., et al., 2004: Annotation transfer between genomes: protein-protein interologs and protein-dna regulogs. *Genome research*, **14 (6)**, 1107–1118. (Cited on page 8.)

Yu, H., et al., 2008: High-quality binary protein interaction map of the yeast interactome network. *Science (New York, N.Y.)*, **322 (5898)**, 104–110. (Cited on pages 2, 7, 54, 64, 71, and 92.)

Zheng, J., et al., 2010: Epistatic relationships reveal the functional organization of yeast transcription factors. *Molecular systems biology*, **6**, 420. (Cited on pages 8 and 97.)

Zhou, T., D. A. Drummond, and C. O. Wilke, 2008: Contact density affects protein evolutionary rate from bacteria to animals. *Journal of Molecular Evolution*, **66 (4)**, 395–404. (Cited on page 10.)

Zinman, G. E., S. Zhong, and Z. Bar-Joseph, 2011: Biological interaction networks are conserved at the module level. *BMC systems biology*, **5 (1)**, 134. (Cited on pages 91 and 99.)

Zotenko, E., J. Mestre, D. P. O'Leary, and T. M. Przytycka, 2008: Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS computational biology*, **4 (8)**, e1000 140. (Cited on pages 82 and 91.)

7

List of Publications

Fokkens L, Hogeweg P, Snel B. Gene duplications contribute to the overrepresentation of interactions between proteins of a similar age. *BMC Evolutionary Biology*, 2012 (Chapter 4 in this thesis).

van Wageningen S, Kemmeren P, Lijnzaad P, Margaritis T, Benschop JJ, de Castro IJ, van Leenen D, Groot Koerkamp MJ, Ko CW, Miles AJ, Brabers N, Brok MO, Lenstra TL, Fiedler D, Fokkens L, Aldecoa R, Apweiler E, Taliadouros V, Sameith K, van de Pasch LA, van Hooff SR, Bakker LV, Krogan NJ, Snel B, Holstege FC. Functional overlap and regulatory links shape genetic interactions between signaling pathways. *Cell*, 2010.

Fokkens L, Botelho SM, Boekhorst J, Snel B. Enrichment of homologs in insignificant BLAST hits by co-complex network alignment. *BMC Bioinformatics*, 2010 (Chapter 3 in this thesis).

Fokkens L, Snel B. Cohesive versus flexible evolution of functional modules in eukaryotes. *PLoS Computational Biology*, 2009 (Chapter 2 in this thesis).

Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *The American Journal of Human Genetics*, 2006.

8

curriculum vitæ

The author was born in Groningen, the Netherlands, on December 11, 1979. From 1990 to 1997 she attended the Willem Blaeu College in Alkmaar, where she obtained her VWO diploma. In September 1997 she started her studies Cognitive Artificial Intelligence (CKI) at Utrecht University (UU) and she obtained both her bachelor's degree in CKI as well as her master's degree in Theoretical Biology and Bioinformatics in 2006.

In September 2006 she started her postgraduate research under supervision of dr. Berend Snel and Prof. dr. P. Hogeweg. The results of this research are described in this thesis.

9

Nederlandse samenvatting

Een veelgebruikte metafoor voor een levende cel is dat van een fabriek. Een aanzienlijk deel van de taken in deze fabriek wordt uitgevoerd door grote moleculen genaamd eiwitten. Deze eiwitten staan gecodeerd op het genoom. Een verandering op het genoom kan dus de werking van de fabriek beïnvloeden. Omgekeerd beïnvloedt de organisatie van eiwitten in de fabriek welke mutaties voordelig of in ieder geval niet te nadelig zijn en geconserveerd zouden kunnen worden. Er is dus een duidelijke wisselwerking tussen het genoom en de fabriek die zij codeert.

Recente onderzoeken, onder andere in gist, hebben grootschalige netwerken van samenwerkingsverbanden tussen eiwitten in de cel blootgelegd. Deze netwerken hebben een zogenaamde modulaire organisatie. Dat wil zeggen dat eiwitten vaak nauw samenwerken in groepjes die zelf relatief onafhankelijk opereren van de rest van het systeem, vergelijkbaar met een machine in een fabriek. Deze modules zijn zelf ook weer georganiseerd in grotere modules. In dit proefschrift onderzoeken we welke processen belangrijk zijn voor het ontstaan en het in stand houden van een dergelijke organisatie.

Als een module eenmaal is ontstaan, is de onderlinge functionele afhankelijkheid van de deelnemende eiwitten in deze module zo groot, dat zij zonder de andere leden van deze module geen bestaansreden hebben? In dat geval verwachten we dat functionele modules ook modules zijn in de evolutie: als één geheel aan- of afwezig zijn in verschillende soorten. Dit is al eens onderzocht, maar die onderzoeken zijn helemaal of grotendeels gebaseerd op prokaryote genomen, die een andere genoomorganisatie kennen dan eukaryote genomen. Pas sinds kort zijn er genoeg eukaryote genoomsequenties beschikbaar om een dergelijk onderzoek in dit domein mogelijk te maken.

Wij nemen aan dat eiwitcomplexen en metabole paden die bekend zijn in gist

kunnen worden beschouwd als functionele modules. Door de aminozuurvolgorde van de eiwitten in die modules uit gist te vergelijken met de aminozuurvolgorde in van alle eiwitten die bekend zijn in 33 andere eukaryote soorten, bepalen we hun aan- en afwezigheid in die soorten. We scoren de 'cohesie' of 'modulariteit' van de aan- en afwezigheidspatronen van functionele modules. Om te bepalen hoe significant de geobserveerde evolutionaire cohesie is, vergelijken we de score van elke module met een achtergrond van 100000 willekeurig samengestelde modules van dezelfde grootte.

We vinden voor het merendeel van de functionele modules dat hun evolutionaire cohesie niet significant hoger is dan we op basis van onze willekeurig samengestelde modules zouden mogen verwachten. Komt dit omdat sommige deelnemers niet zo nauw betrokken zijn bij de module als we aan hebben genomen? In recente grootschalige experimenten is systematisch uitgezocht welke eiwitten in gist hoe vaak samen in één complex zitten. Deze gegevens gebruiken we om eiwitten die potentieel onterecht tot de module gerekend worden, eruit te filteren. Ook vergelijken we verschillende module-definities en filteren we de eiwitten die in maar één definitie tot de module behoren eruit. Vervolgens kunnen we de cohesie van de aan- en afwezigheidspatronen van gefilterde modules vergelijken met die van de originele modules. Na het toepassen van bovengenoemde filters is het aantal modules met een significante cohesie in aan- en afwezigheidspatronen met ongeveer een derde vermeerderd. Ook onze definities van aan- en afwezigheid nemen we onder de loep. We hanteren striktere criteria voor aanwezigheid en we filteren eiwitten die vaak gedupliceerd zijn uit modules, met wederom een klein effect op de evolutionaire cohesie van de modules.

We concluderen dat de precieze samenstelling van modules niet geconserveerd is. Aanpassingen van de cellulaire fabriek gebeuren dus voornamelijk door het geleidelijk veranderen van machines door onderdelen weg te halen en toe te voegen, en niet door het introduceren of weghalen van hele machines tegelijk.

In hoofdstuk 3 onderzoeken we of familierelaties (homologie) kunnen vaststellen tussen eiwitten die wat betreft hun aminozuurvolgorde niet erg op elkaar lijken, maar wel in dezelfde module (in verschillende soorten) zitten. We vinden dat dit inderdaad het geval is. We identificeren welke complexen homoloog zijn tussen mens en gist aan de hand van complex-deelnemers waarvan de sequentie wel geconserveerd is. Helaas, ondanks dat beide organismen al sinds lange tijd grondig bestudeerd worden, zijn de gegevens over welke eiwitten bij elkaar in één complex zitten verre van volledig en blijkt deze methode niet op grote schaal praktisch toepasbaar voor het voorspellen van familierelaties.

We vragen ons af hoe het kan dat binnen één complex sommige eiwitten in hun sequentie zo gedivergeerd zijn dat de familierelatie nauwelijks vast te stellen is, terwijl dat voor andere deelnemers niet geldt. Evolueren sommige eiwitten binnen een complex sneller dan andere? We vinden dat in er meestal sprake is

van oude duplicaties (voordat mens en gist zich van elkaar afsplitsten) en dat de divergentie verklaart kan worden uit het feit dat er meer tijd is verstreken. Dit komt overeen met wat er al bekend is over hoe eiwitcomplexen veelal zijn ontstaan, namelijk vaak door duplicatie van eiwitten die aan zichzelf binden. Aan veel eiwitcomplexen liggen dus heel oude duplicaties ten grondslag.

In hoofdstuk 4 en 5 benaderen we de vraag over het ontstaan en behoud van een modulaire organisatie van een andere kant. We gaan uit van een netwerkmodel dat bepaalde evolutionaire processen waarvan we weten dat die plaatsvinden op het genoom, expliciet koppelt aan veranderingen in de samenwerkingsverbanden in de cel. Deze samenwerkingsverbanden worden samengevat in een netwerk waarin knopen staan voor eiwitten en de connecties tussen de knopen fysieke interacties tussen eiwitten representeren. In het model groeit een netwerk door duplicatie gevolgd door subfunctionalisatie van de knopen in het netwerk. Dat wil zeggen dat het model een aantal keren itereert en elke iteratie wordt een willekeurige eiwit geselecteerd, die wordt gedupliceerd en de twee 'dochters' nemen elk een deel van de ouderlijke taken op zich, wat inhoudt dat ze elk een complementair deel van de ouderlijke connecties behouden en verliezen. Het model wordt ook wel het Duplicatie-Divergentie model genoemd. De netwerken die voortkomen uit dit model lijken wat betreft hun architectuur heel erg op de netwerken die zijn verkregen uit de grootschalige experimenten waarin naar fysieke interacties tussen eiwitten is gezocht. Zij zijn bijvoorbeeld ook modulair georganiseerd. Dit is opmerkelijk omdat een modulaire organisatie wordt geassocieerd met het vermogen om zich makkelijk aan nieuwe omstandigheden aan te passen. Nu blijkt echter dat een dergelijke organisatie ook kan ontstaan in modellen zonder expliciete natuurlijke selectie.

De vraag is nu of dit proces van duplicatie en subfunctionalisatie ook daadwerkelijk verantwoordelijk is voor de modulaire architectuur van het eiwitinteractienetwerk. In een artikel uit in PloS Computational Biology uit 2008 betogen Kim en Marcotte dat dit niet het geval kan zijn. Aan de hand van de taxonomische distributie van de soorten waarin het eiwit aanwezig is, delen zij eiwitten in in verschillende leeftijdscategorieen. Als een eiwit bijvoorbeeld wordt gevonden in bacteriën wordt het ouder geschat dan als het alleen terug te vinden is in schimmels. Zij vinden dat eiwitten vaak interacteren met eiwitten van gelijkende leeftijd. Zij simuleren netwerkevolutie met een aantal verschillende modellen, waaronder het Duplicatie-Divergentie model. Zij vinden dat het feit dat eiwitten vaak interacteren met eiwitten van gelijkende leeftijd, niet verklaard kan worden door het Duplicatie-Divergentie model. Ze concluderen dat duplicatie van genen en subfunctionalisatie van de eiwitten waarvoor ze coderen, niet belangrijk is in de evolutie van eiwitinteractienetwerken.

Aan de andere kant is diverse keren vastgesteld dat genduplicatie juist een cruciale rol speelt, bijvoorbeeld bij het ontstaan van modules zoals eiwit complexen. Bovendien kan het Duplicatie-Divergentie model veel andere kenmerken van ei-

witinteractienetwerken, zoals hun modulaire organisatie, wèl goed verklaren. De auteurs hebben in hun implementatie van leeftijd van eiwitten in het model, geen rekening gehouden met het feit dat eiwitten na duplicatie een gelijkende aminozuurvolgorde hebben en dus even oud geschat zullen worden. Immers, beiden zullen in dezelfde soorten teruggevonden worden. Het is dus mogelijk dat interacties tussen eiwitten die uit een duplicatie zijn ontstaan (paralogen) de observatie dat eiwitten vaak interacteren met eiwitten van dezelfde leeftijd, tenminste ten dele verklaren. Wij vinden dat dit inderdaad het geval is. Echter, fysieke interacties die gedeeld worden door dochter eiwitten na duplicatie van het ouderlijk gen, drukken een nog grotere stempel op de leeftijdsstructuur van het netwerk.

Wij incorporeren het feit dat er familierelaties tussen eiwitten bestaan in het Duplicatie-Divergentie model en onderzoeken of het nu wel netwerken genereert waarin eiwitten relatief vaak interacteren met eiwitten van dezelfde leeftijd. We vinden dat dit het geval is. Echter, de verantwoordelijke mechanismen in het model blijken niet dezelfde te zijn als in de werkelijkheid. In het model zijn interacties tussen eiwitten van dezelfde leeftijd voornamelijk interacties tussen paralogen terwijl dit in interactie-netwerken in gist niet het geval is. Desalniettemin hebben wij met dit onderzoek weerlegd dat genduplicatie geen belangrijke rol speelt in netwerk evolutie.

In de meeste netwerk modellen, inclusief het Duplicatie-Divergentie model, wordt alleen netwerk groei gemodelleerd. Vergelijkende genoomstudies hebben echter uitgewezen dat het aantal verschillende genen niet per se toeneemt in evolutie en dat grote verschillen tussen soorten net zo goed ontstaan door gendeletie als door genduplicatie. Wij zijn benieuwd hoe het Duplicatie-Divergentie model zich zal gedragen als we gendeletieprocessen in het model inlijven.

We nemen aan dat een cel een bepaald aantal eiwitten nodig heeft om redelijk te kunnen functioneren. Daarom laten we de kans dat een duplicatie of deletie plaatsvindt afhangen van het aantal knopen in het netwerk, naarmate het netwerk minder dan 2500 knopen heeft is de kans op duplicatie groter en naarmate het netwerk meer dan 2500 knopen heeft is de kans op een deletie weer groter (2500 is overigens een vrij arbitrair getal). In het geval van een deletie kiezen we een knoop uit en verwijderen die. Als een knoop helemaal geen enkele connectie meer heeft, wordt die sowieso ook verwijderd. Wat betreft genduplicatie volgen we dezelfde routine als in hoofdstuk 4. We draaien het model onder verschillende aannames. Wat betreft duplicatie testen we verschillende maten van divergentie direct na duplicatie. Wat betreft deletie selecteren we in sommige simulaties knopen willekeurig, en in andere simulaties nemen we aan dat knopen die minder connecties hebben minder belangrijk zijn voor het functioneren van de cel, en selecteren we de knopen expliciet op basis van hun (gebrek aan) connectiviteit.

Iedere 100 iteraties meten we een groot aantal netwerkkenmerken. Zo kunnen

we vaststellen hoe de netwerk architectuur verandert gedurende de simulatie. Een voorbeeld van een netwerk kenmerk is zijn modulariteit, maar we kijken ook naar gemiddelde connectiviteit van knopen, en bijvoorbeeld of oudere knopen meer connecties hebben dan jonge. We vergelijken wat we zien in netwerken die gegenereerd zijn door het model met 4 verschillende eiwitinteractienetwerken in gist.

We zien dat gedurende een simulatie de connectiviteit binnen het netwerk geleidelijk afneemt en het uiteindelijk in verschillende stukken breekt. Als we knopen voor deletie selecteren met een kans die omgekeerd evenredig is aan hun aantal interacties, blijft de connectiviteit binnen het netwerk in sommige omstandigheden redelijk op peil. Echter, het netwerk dat gegenereerd wordt door dit model verschilt weer in andere opzichten van echte interactie netwerken in gist. Bovendien is het in netwerken waarin knopen willekeurig worden geselecteerd ook zo dat knopen met minder interacties een grotere kans lopen om uiteindelijk uit het netwerk verwijderd te worden, simpelweg omdat ze een grotere kans hebben om al hun interacties te verliezen.

We concluderen dat duplicatie van interacties na het dupliceren van knopen alleen niet voldoende is om de integriteit van het netwerk te waarborgen. We breiden het model verder uit zodat knopen ook geheel nieuwe interacties kunnen krijgen. Ook hier vergelijken we twee verschillende mechanismen. In het eenvoudigste model selecteren we in het geval van een compleet nieuwe interactie, twee knopen willekeurig. Uit recente onderzoeken blijkt echter dat eiwitten met veel verschillende eiwitten interacteren, bepaalde eigenschappen hebben die dat mogelijk maken. Als we aannemen dat sommige eiwitten makkelijker nieuwe interacties aangaan dan andere omdat ze bijvoorbeeld vanwege hun structuur, veel verschillende eiwitten kunnen binden, zouden eiwitten die al veel interacties hebben makkelijker in staat moeten zijn om een nieuwe interactie (met een willekeurig ander eiwit) aan te gaan. We vinden dat modellen waarin het verkrijgen van nieuwe interacties tussen knopen op deze laatste manier is geimplementeerd, netwerken produceren waarvan de architectuur dat van echte eiwitinteractienetwerken het dichtst benadert.

Samenvattend hebben we in dit proefschrift de relatie tussen de organisatie van het stelsel van samenwerkingsverbanden tussen eiwitten en evolutionaire gebeurtenissen die op het genoom voorkomen vanuit twee hoeken bestudeert. Uit onze analyses in de eerste twee hoofdstukken blijkt onder meer dat de precieze samenstelling van eiwitmodules geen vaststaand gegeven is in evolutie, maar dat dit frequent wordt aangepast. Verder onderzoek, waarin nieuwe gegevens over zogenaamde 'genetische interacties' tussen eiwitten worden gebruikt kan de relatie tussen functionele afhankelijkheid en co-evolutie verder verduidelijken. Uit het onderzoek dat we beschrijven in de laatste twee hoofdstukken concluderen we dat genduplicatie mogelijk toch een zeer belangrijke drijvende kracht achter een modulaire organisatie is. Verder hebben we gemerkt dat we, door naar veel ver-

schillende netwerk eigenschappen tegelijk te kijken, zowel in modellen als in de data, een beter begrip hebben gekregen van hoe genoom en netwerkevolutie met elkaar verweven zijn.

10

Acknowledgements/Dankwoord

Eén ding wist ik zeker na mijn studie: ik zou niet in Utrecht blijven. Ik zou bevrijd de wijde wereld intrekken. Dus toen Paulien het idee opperde om een promotieonderzoek te doen bij Berend, in Utrecht, had ik daar in eerste instantie helemaal geen oren naar. Maar Paulien heeft vaker goede ideeën. In de eerste ontmoeting met Berend werd me de aard van zowel het onderzoeksproject als de begeleider heel duidelijk en ik besloot om de wijde wereld toch nog maar even te laten voor wat hij was.

Afgezien van misschien een enkel moment in december 2012, heb ik dit besluit nooit betreurd. Dit komt natuurlijk door de aard van mijn werk, maar ook door de mensen met wie ik dit werk heb gedaan, die mij hebben geholpen en/of waarvan het gewoon fijn is dat ze bestaan.

Berend, als ik hakkel over iets waarvan ik eigenlijk alleen de contouren nog kan zien, kan jij in drie woorden precies samenvatten wat ik blijk te bedoelen. Ik heb heel erg veel van jou geleerd, over evolutie, over fylogenie, over vergelijkend genoomonderzoek en over hoe je je resultaten presenteert. Dank daarvoor. Bedankt dat je altijd zo recht door zee bent en onomwonden zegt wat je wel en niet goed vindt. Als ik echt zo'n complete idioot ben als ik me soms voel, zal jij me dat heus wel zal laten weten. Dat is een hele geruststelling. Bedankt voor je vertrouwen in mij en je complimenten. Bedankt voor je betrokkenheid, je begrip en je steun. Voor dat je me met rust liet als het water me aan mijn lippen stond, en dat je een kraan dichtdraaide als je dat kon. Heel, heel erg bedankt voor alles. Ik had me echt geen betere begeleider kunnen wensen.

Paulien, ik weet dat mensen hun leven veranderen voor jou ongeveer dagelijkse kost is, maar dat neemt niet weg dat dank daarvoor op zijn plaats is. Ik heb me een te groot deel van mijn leven verveeld. Je vervelen is niet leuk. Je hebt me de

schoonheid in biologie laten zien en dit heeft me mede geholpen om uit mijn sluimertoestand te ontwaken en mijn hersens eens wat beter te gebruiken. Jij hebt me zo ongelofelijk veel geleerd, over wat biologie is en over wat en waar informatie is. Ik ben je echt heel dankbaar. Ook dat je zo principieel, onverstoorbaar en betrokken bent, waardeer ik heel erg. De wetenschap dat een krachtig persoon als jij achter me stond tijdens mijn promotieonderzoek, heb ik als een echte steun ervaren. Heel hartelijk bedankt.

Michael, you came in with a bang. You were young , you were ambitious, you gave a great talk and you were very nice company during diner. You stirred up our comfortable but slightly dusty group of roommates and I think we all enjoyed that. Your unrestrained laughter whenever I shared my moments of misery still resonates in my head. It is an absolute pleasure to know you and I am truly honoured to be your paranimf.

Lydia, you are one of the Dutchest people I know. Both your feet are firmly on the ground and I really enjoyed hearing your down-to-earth comments in that amazing voice. Thank you for showing how much you care. Thanks for caring so much. You and Michael really deserve eachother.

Jos, ik ben heel blij dat ik jou ken. Je bent zo ongelofelijk gezellig en zo ontzettend aardig. Bovendien weet je alles. Je plezier in je werk was erg aanstekelijk. Ik heb heel veel aan je gehad, door directe hulp en geduldige uitleg maar vooral gewoon door je aanwezigheid en je kijk op de wereld. 'Zie je nou wel, Jos vindt het ook', dacht ik vaak. Bedankt Jos, voor het funderen van mijn mening, maar vooral voor je zeer prettige gezelschap. Ik hoop heel erg dat we ooit weer een werkkamer delen. De kamer ernaast is ook goed (maar dan wel graag met de deur open).

John, jij en ik hebben heel lang naast elkaar gezeten maar daar houdt verder ook elke gelijkenis wel op. Ons onderzoek, onze werktijden en onze karakters zijn heel verschillend. Ik heb er van genoten om met een mengeling van verbazing en jaloezie naar rechts te kijken. Hopelijk maak ik ook ooit zulke mooie plaatjes van mijn genbomen, dat lijkt me voor mij haalbaarder dan een regelmatig dagritme. Het was heel fijn om naast en met je te werken en een grote eer om je paranimf te zijn. Dank je wel.

Dear Gabino, thank you for your warmth, your feedback and your interesting stories. Nothing can say 'You worry too much' like your smile nested in your beard. Thank you. And please finish your rates-of-sequence-evolution-manuscript: it deserves to be published.

Ik heb werkelijk buitengewoon geboft met de masterstudenten die mij geholpen bij mijn onderzoek. Laura, Wouter en Esther: het was een plezier en een voorrecht om jullie te begeleiden. Omdat we er dan tenminste samen niks van begrepen, omdat ik mezelf kon horen nablaten wat Berend ooit tegen me heeft gezegd, en vooral omdat het geweldig is als mensen dingen ontdekken die je zelf niet zou

hebben gezien. Laura, thank you so much for your indestructable optimism and cheerfullness. The ladies room has not been the same since you left. Wouter, heel hartelijk bedankt voor je doorzettingsvermogen en het fundament dat je hebt gelegd voor maar liefst twee hoofdstukken in dit proefschrift. Esther, heel erg bedankt voor je rust, je intelligentie en je onverstoorbaarheid.

Een groot voordeel van het werken op deze vakgroep is dat je aardige, bijzondere en slimme collega's hebt. En daar ook nog eens heel veel van. Ben, Rob, Sacha, Anton, Otto, Nobuto, Jan Kees, Kirsten, Levien, Eva, Marian, Milan, Rikkert, Daniel van der Post, Daniel Weise, Stan, Veronica, Can, Thomas, Folkert, Ronald, Sandro, Chris, Jorg, Ioana, Christian, Ilka, Joost, Boris, Tjibbe, Henk-Jan, Hanneke, Paola, Sai, Tendai, Rao, Klaartje, Renske, Adrian, Alessia, Rianne, Erik, Nijuscha en Harmen: thank you for your pleasant company and the insightful presentations of your research. Otto, bedankt voor het mede-assisteren van Bioinformatische Processen, het is een intensieve bezigheid en ik ben heel blij dat juist met jou mocht doen. Boris, Marian, Hanneke en Milan: extra dank voor extra gezelligheid, bij jullie thuis, of in een moestuin. Stan en Veronica, bedankt voor jullie gastvrijheid en enthousiasme. Radek, ook als was je er maar een maand, ook jouw enthousiasme was heel aanstekelijk. Levien, Eva en Tjibbe bedankt voor het meefietsen naar Schoorl. Paola, please keep reminding me I should attend your parties.

Jan Kees, zonder jou had ik maar een fractie van dit werk kunnen doen. Jij vond me in het begin heel boos en dat was ik ook, want eigenlijk haat ik computers. Jij hebt mijn werk meer dan dragelijk gemaakt. Dank je wel!

Ik wil Rob, Sacha, Stan, Can, Kirsten, Nobuto, Marian en Milan graag apart bedanken voor hun uitleg, geduld en hulp tijdens colleges of werkgroepen. Zonder deze fundering, gelegd in het laatste deel van mijn studie, had ik dit proefschrift niet kunnen schrijven. Lude en Ciska wil ik om ongeveer dezelfde reden bedanken. Ze me hebben laten zien hoe leuk het is om onderzoek te doen en hoe fijn het is als iemand meteen begrijpt waar je het over hebt.

En als ik het dan toch over fundering heb: papa en mama, bedankt voor het leven dat jullie me hebben gegeven. Dank voor jullie genen, voor jullie niet-repressieve opvoeding en jullie onvoorwaardelijke liefde. Ik heb nooit last van faalangst. Ik hoef niet te voldoen aan enig verwachtingspatroon maar mag het wel. Ik hoef niets te bereiken maar mag het wel. Dank jullie wel voor die vrijheid. Dank jullie wel voor alle praktische hulp. Dank jullie wel voor mijn lieve broertje Thijs.

Lief broertje Thijs, dank je wel voor je oprechte interesse in mijn werk dat zo dichtbij en toch zo ver van het jouwe ligt. Ik ben zo blij dat je er bent. Dank je wel voor je ontwerp voor de kaft.

Lieve Rosa, bedankt voor je steun in mijn donkerste uren. Te weten dat mijn jonge baby in veilige handen was gaf me rust, anders had ik die allerlaatste, allertaaiste

stukken niet kunnen schrijven. Heel erg bedankt. Annelies en Joop, Jacques, Fleur en Renee, dank jullie wel voor al jullie mentale en praktische steun.

Esther, heel erg bedankt voor al je kaartjes, die zo maar uit het niets op de mat lagen. Zo lief en attent. Het heeft me vaak echt opgebeurd en goed gedaan dat er in dat verre Friesland iemand was die aan me dacht, die wist dat ik het moeilijk had en met me meeleefde. Dank je wel.

Mijn vrienden, met name Corine, Maaike, Julia, Lars, Ronald, Martin, Edwin, Tijn en Sanne wil ik graag bedanken omdat ze nog steeds mijn vrienden zijn. De combinatie van proefschrift en kinderen laat verder weinig tijd en energie voor een sociaal leven. Daar komt nu dus eindelijk verandering in, hoera!

Corine is niet voor niks mijn paranimf. Zij geeft je wat je nodig hebt, ongeacht wat dat is. Dat is bijzonder. Dank je wel, lieve Corine, dat ik mijn afstudeerscriptie bij jou in de tuin mocht afschrijven en dat je me ook nog verwende met lekkere broodjes, dagen achter elkaar. En ook nu weer, toen ik he-le-maal vastzat met mijn conclusie, probeerde je met een soort interview mijn gedachten te ordenen. Eigenlijk wilde ik zeggen dat je op moest lazeren en dat ik nu juist even niet aan die rotconclusie wou denken, maar je bent zo aardig dus deed ik dat niet, en warempel: het hielp. Niets van die ordening heeft uiteindelijk het papier gehaald maar het heeft me wel uit de mist getrokken. Ik ben gezegend met zo'n supervriendin.

Lieve Rikkert, heel hartelijk bedankt voor je onvoorwaardelijke steun in de laatste maanden. Bedankt dat je nooit zei 'Is het nou nog niet af?' terwijl je dat natuurlijk wel dacht. Bedankt dat je mijn proefschrift voorrang boven alles hebt gegeven. Dat is zwaar, zeker met twee kinderen, maar het was nodig om het binnen redelijke tijd (bij mij ook nog eens een zeer rekbaar begrip) af te krijgen. Het is vreselijk om samen te leven met iemand die een proefschrift afmaakt en je hebt nooit geklaagd. Dank je wel voor zowel je genadeloosheid als je empathie. Ik weet dat het proefschrift zelf je geen fluit kan schelen, dat het je niks doet of ik nu wel of niet gepromoveerd ben, dat je helemaal niet begrijpt waarom ik hier überhaupt aan ben begonnen en vooral niet waarom ik het per se af moest maken. De enige reden waarom je me zo steunt is omdat ik het ben. En dat is echt heel erg lief. Dank je wel.

Kinderen kosten een hoop tijd en nachtrust en ze houden je ontzettend van je werk. Maar als Boris opeens 'wawawawawa' zegt in plaats van 'waaaaaaaaaa' en Oscar kan fietsen op een echte fiets, ben ik Zo Ongelofelijk Trots. Daar kan een proefschrift niet aan tippen. Wat een rendement. Allerliefste Oscar, allerliefste Boris. Dank jullie wel dat jullie gewoon doorgroeien als ik stilsta. Ik ben dolblij met jullie!