Function and evolution of genes in the human protein interaction network

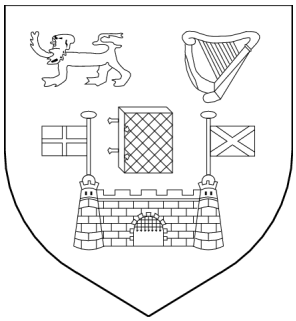
by

Åsa Pérez-Bercoff

A thesis submitted to
The University of Dublin
for the degree of

Doctor of Philosophy

Department of Genetics
Trinity College
University of Dublin



February, 2012

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. This thesis may be deposited in the University's open access institutional repository, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement. The copyright belongs jointly to the University of Dublin and Åsa Pérez-Bercoff.

Signature of Author .....

Åsa Pérez-Bercoff

28 October 2011

Acknowledgements

There are many to thank in this thesis. Obviously I will start with my supervisor Aoife McLysaght for unless she had let me join her group I would not be where I am today. So thank you for letting me join your group in my endeavour to understand molecular evolution and for supervision, and also for giving me the freedom to explore biological networks even when it meant collaborating and visiting the laboratory of another researcher. Therefore also my gratitude to collaborator Gavin Conant for letting me work together with you on such interesting topics, and also for letting me visiting your laboratory at the University of Missouri. Naturally I also wish to thank the past and present lab members in Aoife's group: David, Kirsten, Takashi, Dan, Fergal, Daniele and Aideen, and lab members in Gavin's group: Michaël and Corey. Also to Karsten at the Genetics department at TCD for being so helpful with computational matters, and even updating your paralogon data so that I could use it for the research in chapter 4. I enjoyed the company from all of you both in a scientific and a social capacity. I am grateful to have known and worked with all of you!

I also wish to thank participants of the Wednesday Lunch-time seminars at the Genetics department at TCD for interesting lectures and discussions, and at the University of Missouri, Chris Pires and members from his laboratory who joined Gavin Conant's group for interesting scientific talks and discussions at the Wednesday network meetings.

A big thank you to Dermot from PAS. Your help and advice was

paramount for my PhD! Thanks to my bosses Bernhard, Antonio, Andrea and Thierry at Adamans Ltd. for allowing me take some time off work so I could attend meetings with my supervisor once a month while I was working for you. I am immensely grateful to Gålöstiftelsen för högre utlandsstudier for giving me a very generous scholarship. With the money I received from you I not only paid the university fees of my final year, but my whole financial situation was sorted for the entire year. With this peace of mind as I did not have to worry about supporting myself anymore with a full-time job, and could therefore whole-heartedly focus on my research.

Thanks are also due to my former project supervisors Jens, Bengt and Lasse at Stockholm Bioinformatics Centre for supporting and encouraging me when I went to Ireland to undertake my PhD.

On a social note I would like to thank my friends. I have met many friendly and nice people while in Dublin, and in particular I wish to thank: my good house mates and friends Lilian (you are an incredible generous person, so thank you for so very much!) and Naisha (the three of us shared much laughter in and out of the house!). Yonas (for always *hearing* me out, no matter the topic of discussion, for caring and being supportive), Jerry, Kasia and Haile - together with Yonas you became my family during my exile in Ireland. Thanks also to Linda, your house mate that is always so cheerful and nice to me. Warren (with you also numerous dinners, cycling and swimming, support over coffee and lots of fun, your positive attitude to and confidence in life is inspiring) and Lisa (thanks for friendship and housing me when I have needed it) and Gordon (you are so very knowledgable, and tell things in fun and interesting ways, I really enjoyed attending talks at The Irish Skeptics Society and many other social events in yours and Kirsten's company). Life in Ireland for a student has been, to say the least, interesting and I could not have survived without the help of many friends. Thanks particularly to those friends that housed me when I was between homes. Some have already

been mentioned above. A very special thank you to Joe and Shirley who after meeting my parents only once while diving in Spain housed me for free for two weeks upon my arrival in Dublin until I had my housing situation sorted. A thanks to Lilian's aunt Janice and Lilian's cousin Serena for letting me stay at your house when I have needed a place to stay, and to Naisha's boyfriend Andrew for driving some of my belongings when I moved to my flat in Rathmines. Thanks to Kirsten's (then boyfriend, now husband) Dan for together with Kirsten you two offered me a place to stay if I needed it. Thanks also to fellow cyclists Tim from the Genetics department who took me to Dublin 3DTri triahtlon club, and Nicholas captain of the TCD cycling club for cycling with and waiting for me so that I would not get lost around Meath, Kildare or up in the beautiful Wicklow mountains!

Thanks also to the people I got to know in Columbia Missouri that were so friendly to me. Especially my landlords Corey, Katy and your precious and sweet daughter Lulu (the memory of you still makes me smile), and also your friendly and protective dog Maggie. You all made me feel very welcome at your home, while I lived there during my time in Missouri. Thanks for inviting me to Sunday family dinners and other social events with your close friends. I am also glad to have met friends Ana, Armelle (thanks for offering to take me to see a doctor when I was home ill over Thanksgiving) and Damien from the French conversation group, and also Michela, and my Swedish next-door neighbour Rigmor and her husband Ronnie for their hospitality while I was in Missouri.

Thanks to friends I already knew back in Sweden (although not all of your are still there); Gabriel (thanks for bringing my racer bike over to Ireland for me!), Muna, Gabi and Kirsten; I am so glad you managed to visit me in Dublin. Thanks also to other friends I knew already before moving from Sweden especially Emelie, Marija and Nikola (I am honoured to have been your wedding witness Marija), Jörg, Johannes, Maria and Patrick.

Finally, but certainly not least thanks to my cousin Danielle in Paris, and Pat, Thomas and Lena and my parents in Sweden, and my grandmother in Argentina for your constant encouragement, love and support. Without you there would be no thesis!

Abstract

The research conducted for this thesis aims to elucidate how the human protein interaction network has evolved, and how protein interactions influence the spatial organisation of the metabolic network. The thesis presents compelling results suggesting that indirect protein-protein interactions between metabolic enzyme proteins, and non-metabolic proteins is a feature conserved from prokaryotes through to mammals. Indirect protein-protein interactions are shown to influence the structure of metabolic networks, very likely through a mechanism called metabolic channelling. This mechanism brings metabolic pathway enzymes into close proximity of each other so that reactions between the enzymes can occur.

In *E. coli* and yeast, reactions that are linked to each other through mediator proteins have a much higher flux compared to reactions lacking them. The thesis presents indirect evidence that indirect protein-protein interactions are very important for the spatial organisation of metabolic protein networks. This thesis has also inferred the likely time of origin of the human protein-protein interactions that were examined, and suggests that most of the protein interactions in human are likely to be conserved in chimpanzee, macaque, mouse, rat, horse, dog and cow. This suggests they date from at least the point in evolution when placental mammals radiated out. Analyses of the function of proteins analyses in the thesis also suggests that the protein interactions have been conserved because they perform fundamental cellular functions. Finally, the duplicability of protein self-interactions (interactions between two proteins encoded by the same gene i.e., homodimers) are investigated. These genes have higher duplicability in comparison to other protein encoding genes, and also contain more interactions in general, which can perhaps be explained by the greater evolutionary age of these genes.

Summary

The increasing availability of high quality genomic sequences is expanding the scope of the biological questions researchers can tackle. Although peptide data and genomic data are available for very many organisms there are not as much data on the interactions between these proteins, and how they form intricate protein interaction networks. Protein interaction data is only available for a restricted number of organisms, and it is estimated that for these networks only ∼10 - 40% (depending on organism) of all protein-protein interactions have been identified so far. Yet, these interactions are intrinsic to all life forms, since proteins are the building blocks of all cells. Proteins not only interact with each other to sustain cell structures, they also perform functions such as metabolism. This thesis considers the different roles protein interactions have in the cell, and how these protein interactions change as the genes encoding them evolve through mechanisms such as, substitution, gene duplication and loss is an integral part in the understanding of cell, molecular and evolutionary biology.

In chapter 3 of this thesis, indirect evidence was presented indicating that proteins, which are not directly involved in metabolic reactions, and are not part of the metabolic network, can act as mediator proteins connecting proteins from different metabolic reactions. These interactions between non-enzymatic mediator proteins and enzyme-proteins were named *indirect protein-protein interactions*. The mediator proteins appear to 'channel' metabolism by bringing enzymes into close proximity of each other, thereby enabling catalytic reactions to occur through a mechanism called metabolic channelling. The evolutionary change of individual proteins affects their protein interactions, and involvement in protein interaction networks. This can affect the evolution of metabolic networks, since mediator proteins can bring different metabolic modules into proximity (through indirect protein-protein interactions), altering existing and/or creating new

metabolic pathways. This spatial organisation of indirect protein-protein interactions is evident in several organisms, from prokaryotes and single-cell eukaryotes to multicellular eukaryotes, such as humans. Spatial organisation is clearly an important property. However, the fact that indirect protein-protein interactions are involved in spatially organising the metabolism of several, distantly related species does not automatically imply that specific interactions have been conserved. The interactions in the protein interaction network can be rewired because interactions between proteins change, which may create new functions, and bring different metabolic modules into close proximity with each, and permit new cell functions.

Nevertheless, as presented in chapter 5, the great majority of nonself protein-protein interactions (or heterodimers) in human are likely to be highly conserved with other mammalian species, including chimpanzee, macaque, mouse, rat, horse, dog and cow. This suggests they date from at least the point in evolution when placental mammals radiated out.

The studies in chapters 3 and 5 investigate the function and evolution of nonself protein-protein interactions. Chapter 4 investigates the evolution of genes encoding homodimers (called self-interacting genes). It is believed that heterodimers evolved from genes that originally encoded homodimeric proteins (i.e., from duplicated self-interacting genes where the duplicate genes kept the interaction between them although the genes were free to diverge, thus forming heterodimers). Therefore in chapter 4 the duplicability of genes encoding proteins forming homodimers was also investigated. It was found that these self-interacting genes tend to duplicate more often than other genes, and that these genes have arisen through whole-genome duplication more often than through small-scale duplication. Furthermore, self-interacting genes that have arisen through whole-genome duplication tend to have more protein-protein interactions compared to other genes. However, this can be explained by the generally greater age of whole genome duplicated genes, which have simply had more time to accumulate interactions.

Contents

1 Introduction **1**

1.1 Brief introduction to networks 2

1.2 Biological networks 6

1.2.1 From metabolic pathways to metabolic networks . . . 10

1.2.2 Protein networks 13

1.3 Gene gains and losses 14

1.3.1 Gene duplication 14

1.3.2 Whole Genome Duplication 16

1.3.3 Comparisons of Whole Genome Duplication vs Smaller Scale Duplications 18

1.3.4 Gene dosage and gene duplicability 19

1.3.5 How gene duplication affects the protein interaction network 20

1.4 Constraints on protein sequences 23

1.4.1 Selection vs genetic drift 23

1.4.2 Molecular co-evolution and protein interactions 24

1.5 Network evolution 26

1.5.1 Evolution of protein interaction networks 26

1.5.2 How metabolic pathways and networks evolve 29

1.6 Aim 31

2 Materials and Methods **33**

2.1 Constructing biological networks and assessing their quality . 33
2.1.1 Protein interactions - detection and assessment 33
2.1.2 Metabolic network data and assessment 35
2.2 Flux Balance Analysis - Assessing robustness and fitness of metabolic networks 39
2.3 Permutation tests 40
2.4 Maximum likelihood 42
2.4.1 *branch models* - a codon substitiution model for measuring evolutionary constraint 44

3 iPPIs and metabolic channelling 45
3.1 Introduction 45
3.2 Methods 47
3.2.1 Protein interaction data 47
3.2.2 Gene and protein identifiers used 47
3.2.3 Mapping of metabolic and protein interaction networks 49
3.2.4 Robustness of iPPI enrichment to method of PPI detection 49
3.2.5 Pathway analysis 49
3.2.6 Gene Ontology analyses 50
3.2.7 Flux balance analysis 50
3.2.8 Comparison of metabolic and protein interaction networks by randomisation 51
3.3 Results 54
3.3.1 Indirect interactions between neighboring enzymes 58
3.3.2 Analysis of yeast pathways for dPPIs and iPPIs 61
3.3.3 Functional annotation of mediator proteins 61
3.3.4 Association of iPPI presence and high metabolic flux . 62
3.4 Discussion 64

3.5 Conclusions 66

4 Duplicability of self-int human genes 67

4.1 Introduction 67

4.2 Methods 69

4.2.1 Filtering of human protein-protein interaction data . . 69

4.2.2 Definition of singleton and duplicate genes 69

4.2.3 Definition of singleton and duplicate genes for the comparative study 70

4.2.4 Comparison of WGD duplicate genes vs SSD duplicate genes 71

4.2.5 Controlling for age and protein connectivity 72

4.2.6 Gene Ontology analyses 73

4.3 Results and Discussion 73

4.3.1 Higher duplicability of self-interacting genes 73

4.3.2 WGD genes are enriched for self-interactions by comparison with SSD genes 75

4.3.3 WGD genes have more interaction partners on average than SSD genes 77

4.3.4 Self-interacting genes are enriched for developmental and essential biological processes, and WGD self-interacting genes are involved in metabolism 77

4.4 Conclusions 85

5 History of the human PIN 87

5.1 Introduction 87

5.2 Methods 92

5.2.1 A parsimony-based reconstruction of the ancestral states of the human protein interaction network 92

5.2.2 Assessing co-evolution between interacting proteins using mirrortrees 93
5.2.3 Estimating co-evolution using the correlation of ω values in the phylogeny 94
5.2.4 Shared signals of adaptive and co-evolution 96
5.2.5 Association of the degree of selective constraint and protein interaction network position 96
5.2.6 Statistical analysis of the network weights 97
5.2.7 Gene Ontology analysis 97
5.2.8 Protein degree and position in the protein interaction network 98
5.3 Results 100
5.3.1 Reconstructing the ancestral states of the human protein interaction network 100
5.3.2 Functional annotation of primate-specific PPI genes . 100
5.3.3 Protein degree of primate-specific PPI genes 101
5.3.4 Patterns of co-evolution support the ancient nature of most human PPIs 106
5.3.5 Weak evidence for shared instances of adaptive evolution between PPI partners 107
5.3.6 Proteins interact with other proteins of similar constraint more often than expected 110
5.4 Discussion 110
5.5 Conclusions 112

6 Conclusions **115**

List of Tables

3.1 Statistics regarding the networks employed 48
3.2 Metabolic network structure and dPPI prevalence 55
3.3 More iPPIs in real metabolic networks than in randomised ones 60

4.1 Proportion of self-interacting duplicated genes generated by different mechanisms 72
4.2 Frequency of self-interaction of singleton and duplicate (E-value threshold) genes under alternative parameters for paralog definition 72
4.3 Frequency of self-interactions amongst WGD and SSD duplicate genes under alternative WGD identification methods . . . 72
4.4 Duplicate genes are enriched for self-interactions disregarding E-value cutoff used to define duplicate genes 74
75table.4.5
4.6 WGD genes are enriched for self-interactions disregarding of WGD detection method 76
4.7 Over-represented GO terms when self- and non-self-interacting genes are compared against each other 82
4.8 Over- and under-represented (italics) GO terms when duplicated genes are compared against singleton genes 83
4.9 Over- and under-representation (italics) of GO terms in self-interacting WGD with respect to SSD genes 84

5.1 Over- and under-represented GO terms of genes present at least once in a primate-specific PPI 102
5.2 Over- and under-represented GO terms of genes in exclusively present in primate-specific PPIs 103
5.3 Connectivity statistics for primate and nonprimate genes . . . 104
5.4 Absolute connectivity difference between proteins in primate and nonprimate PPIs 105
5.5 Co-evolution signal detection using Spearman's Rank Sums correlation coefficient ρ comparison of real and pseudo datasets 108
5.6 Over- and under-represented GO terms of genes present in PPIs where proteins in the protein pair have $\omega > 0.5$ 109

List of Figures

1.1 Network parameters 3
1.2 Regular, small-world and random network topologies 5
1.3 Power-law distribution 7
1.4 Network distributions 8
1.5 Path length and robustness in scale-free vs random networks . 9
1.6 From metabolic pathway to network depiction 11
1.7 Enzyme and compound centric networks 12
1.8 Network rewiring after WGD 28
1.9 Evolution of metabolic pathways 30

2.1 Different types of gene-protein-reaction associations in yeast metabolic network model iND750 37
2.2 Cartoon of a metabolic network and how FBA works 41

3.1 iPPIs provide structure to the yeast metabolic network 56
3.2 iPPIs produce more distinct metabolic subnetworks than expected by chance 59
3.3 Biomass flux and knockdown effects 63

4.1 Data collection 70
4.2 Definition of mouse-specific duplicated genes in human 71
4.3 Increased proportion of self-interacting genes with increased protein connectivity 76

4.4 Relationship of duplication type and number of interactions . 78
4.5 Number of interactions controlled for age 79
4.6 Distribution of synonymous subsitution rate of WGD and SSD duplicate pairs 80

5.1 PPI presence and absence 91
5.2 Mirrortree with protein-protein interactions 92
5.3 Distribution of correlation coefficients 95
5.4 Datasets used in the GO analyses 99
5.5 Distribution of absolute degree difference between PPI pairs . 105
5.6 Branch-pairs where both branches have $\omega > 0.5$ 107

Abbreviations

AP-MS affinity purification mass spectrometry.

ATP adenosinetriphosphate.

BLAST Basic Local Alignment Search Tool.

BLASTP protein BLAST.

DDC model duplication-degeneration-complementation model.

DIP Database of Interacting Proteins.

dPPI direct protein-protein interaction.

EBI European Bioinformatics Institute.

EC Enzyme Commission.

EHMN Edinburgh Human Metabolic Network.

ER Erdös and Rényi.

FBA flux balance analysis.

FRET fluorescence resonance energy transfer.

FSGD fish-specific genome duplication.

GO Gene Ontology.

GPR gene to protein to reaction associations.

GSM genome-scale metabolic.

HIV human immunodeficiency virus.

HPRD Human Protein Reference Database.

iPPI indirect protein-protein interaction.

KEGG Kyoto Encyclopedia of Genes and Genomes.

MIPS Munich Information Center for Protein Sequences.

ML maximum likelihood.

MLE maximum likelihood estimation.

MS mass spectrometry.

MSA multiple sequence alignment.

NAD+/NADH nicotinamide adenine dinucleotide.

ORF open reading frame.

PAML Phylogenetic Analysis by Maximum Likelihood.

PGK phosphoglycerate kinase.

PIN protein interaction network.

PPI protein-protein interaction.

SGD *Saccharomyces* Genome Database.

SSD small-scale duplication.

TAP tandem affinity purification.

WGD whole genome duplication.

WS Watts and Strogatz.

Y2H yeast two-hybrid.

'When were you born?'
Momo pondered this. 'As far as I can remember,' she said at length, 'I've always been around.'

Momo, Michael Ende

Chapter 1

Introduction

Once upon a time, scientists who studied proteins spent years studying a single protein, or a single gene. Their relentless and painstaking work established the basis of our genetic and biochemical understanding of cell and organism function, organisation and evolution. Constant technological development of the methods used on single proteins and genes led to the advent of high-throughput techniques in biotechnology, such as genome sequencing, proteomics and protein interaction identification. The coupling of these methods to computing has enabled the rise of richly populated biological information databases, and the developing science of bioinformatics. A key factor in enabling the development of this field has been to ensure that most data is free and publicly available in databases for anyone to access. This has enabled the possibility for any researcher to move from small-scale to genome-wide studies, and most researchers can now to varying extent use internet based tools to see how their gene or protein of interest, or indeed organism of interest fits into nature's scheme. Our ability to seek general trends or distinct results in a mass of biological data, rather than being restricted to making extrapolations based on the limited data of a few proteins or a distinct region or system in the cell is helping us learn new things about biology, and discovering new areas of interest that can be further explored. That said, bioinformatics data will probably always need to be qualified by high quality data collected through the patient toil of laboratory researchers.

This thesis is, as far as current data allows, a genome-wide study of the evolution of the human protein interaction network, and endeavours to

explore this network's role in metabolic function. Given that this thesis is about the evolution and function of the genes involved in human protein and metabolic networks, the introductory chapter introduces general features of networks before introducing the specific biological features of biological networks. The same mathematical definition of a network applies just as well to the human protein interaction network as it does to a mobile phone communications or air transport network. Molecular evolution analyses will be used to discuss unique aspects of biological network evolution in terms of the forces acting upon and molecular rates by which genes evolve, and the regulation of gene gain and loss, through gene and genome-wide duplication.

1.1 Brief introduction to networks

A network is represented mathematically as a *graph* consisting of *vertices*, also called *nodes* (which is how they will be referred to from here on in this thesis) that are connected to each other through *edges* (Figure 1.1). The number of connections each node has to any other node is called the node's *connectivity* or node *degree* (k) (Albert, 2005). Nodes that are highly connected in the network (i.e. that have a high k) are referred to as *hubs*. Edges in a graph can be directed. In such graphs any edges towards a node are called the node's in-degree, while any edge leaving the node is called the node's out-degree. *Clustering coefficient* (c) is another network parameter. Rather than measuring the direct connectivity of a node it measures the interconnectivity of a node's neighbours. In other words, it measures how well connected the neighbouring nodes of a node are to each other. Other examples of parameters to quantify a network are *path length* (e.g. shortest and longest) between two nodes in the network. Another parameter is *betweenness centrality*, the ratio of the shortest path between two nodes, and the shortest path between the same two nodes passing through a third, predefined, node (Zhu *et al.*, 2007). There is yet another measure called *closeness centrality*, which is the shortest path between a node and all other nodes reachable from it. It can be seen as a measure of how quickly information spreads from one node in the network to all its connecting nodes (Newman, 2008).

How the nodes and edges arrange within a network decides the overall topology, and we can learn a great deal about the network by looking at

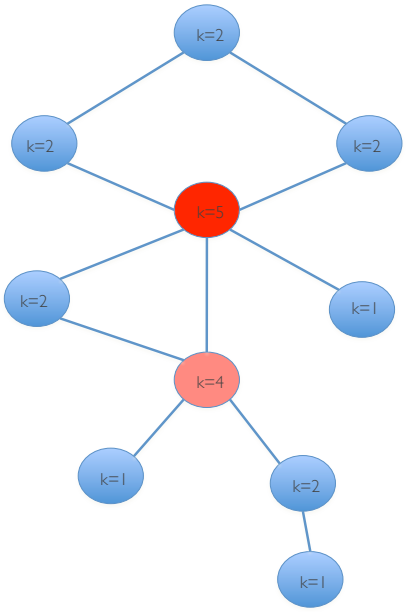


Figure 1.1: Definition of a network and its different parameters. A network (the mathematical term is a *graph*) consists of *nodes* (depicted as circles), or *vertices* as they are also called, and are connected to each other by *edges*. The number of connections (connectivity or degree (k)) of each node is written inside of the respective node. Nodes with many connections i.e. with a high k are called hubs. Here the hub with most connections is depicted in red ($k = 5$). The pink node, although it does not have as many connections as the red node, is also a hub ($k = 4$) as it has more connections than most nodes in the network that are only connected to only one or two other nodes ($k = 1$ or $k = 2$). However, how many connections a node needs to classify as a hub is not defined by a hard-fast rule.

the parameters listed above such e.g., if the network is highly clustered or not. At first it was thought that networks were either regular or random. The regularity of a network is assigned a number between 0 and 1 where $p = 0$ stands for complete regularity while $p = 1$ means the opposite i.e. complete disorder. In a regular network every edge (i.e. connection) in the network is evenly distributed causing the path between any two randomly chosen nodes to be very long (as one has to go through many other nodes to get from the first to the second chosen node). Thus making the path length long, and highly clustered. It is said to be a large-world network. In a random network however all the edges are randomly distributed between the nodes. Thus the random network is a badly clustered, small-world network. However, Watts and Strogatz (1998) realised that many networks, stretching from technological (such as the power grid in the U.S.A.) and social networks amongst movie actors to the neural network of *C. elegans*, were neither completely regular or completely random, but somewhere in-between (i.e. with $0 < p < 1$). The topology of these networks was to be called *small-world* networks because although most connections between nodes were regular, every now and then a node was highly connected (i.e. a hub), which shortened the path length between two randomly chosen nodes substantially (Figure 1.2). Thus, a network could have mostly regular connections, and only a few 'short-cuts', and p increased substantially giving the network small-world characteristics of a random network.

Following the discovery of small-world networks in 1998 Barabási and Albert (1999) decided to study the topology and distribution of several large, real (as existing in nature and society rather than generated specifically and artificially for a computational experiment) networks ranging from the World Wide Web to neural networks and protein networks. Previously it was thought that complex networks were random according to Erdös and Rényi (ER) graph theory. However, as more network data became available Barabási and Albert could test it, they discovered that none of the data they tested arranged itself in a way that the ER random graph theory predicted. Further they discovered that real network data did not really conform to the Watts and Strogatz (WS) model describing small-world networks either. Real networks, they realised, obey two very important rules. Firstly, they grew in size as new nodes were added to the network (something neither the ER nor WS model took into account). Moreover, nodes that were already highly connected were more likely to acquire the new connections, or as Barabási

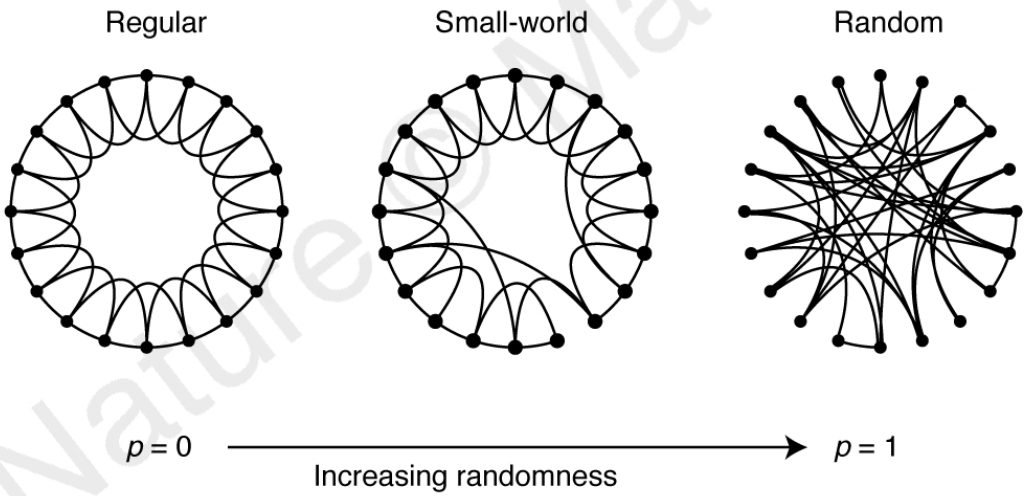


Figure 1.2: Topology of regular, small-world and random neworks. In the regular network $p = 0$ the edges between the nodes are evenly distributed. The network is highly clustered and large-world as it would take long to walk through two randomly chosen nodes. In the random network $p = 1$ the edges are randomly dispersed between the nodes. Thus, causing little clustering, and small-world network as moving between two randomly chosen nodes can be very quick. The small-world network lies in-between the two extremes with most edges randomly distributed. However, every now and then edges connect nodes that would otherwise be far apart. Thereby creating 'short-cuts' in the network, which gives it the small-world characteristics of the random network. Figure from Watts and Strogatz (1998).

and Albert put it the "rich-get-richer". It was as though the network kept re-organising itself into a scale-free state. The *degree distribution* $P(k)$ of a node N is the probability that particular node has k connections, and is calculated by taking the number of nodes with a particular degree $N(k)$ divided by the total number of nodes N. Barabasi and Albert realised that the interactions of the nodes in these networks (i.e., the degree distribution) follow a power-law (Figure 1.3), which can be mathematically expressed as $P(k) \propto k^{-\gamma}$ where $\propto$ stands for 'proportional to' and γ indicates how important hubs are to the network. For example, $\gamma > 3$ indicates hubs are not important, and the network behaves like a random network, while $2 > \gamma > 3$ means there is a hierarchy of hubs i.e. the smaller γ the more important are hubs to the network (Barabási and Oltvai, 2004). Thus, rather than having the node degree follow the Poisson distribution (Figure 1.4Aa-Ac), which is the case for random networks, the connections of the nodes in these networks follow a power-law distribution (Figure 1.4Ba-Bc) consistent with the fact that most nodes have very few connections, while a few nodes have very many connections. These networks were called *scale-free* networks, and today we know that there is a range of these network in the real world (which is why they are also sometimes referred to as 'real-world' networks) e.g. the World Wide Web, how authors of scientific papers are cited in literature, mobile communication networks, several biological networks, such as the protein-protein interactions (PPIs) networks and the neural network in the brain of several animals (including mammals) (Park and Barabási, 2007; Barabási, 2009). Compared to regular and random networks scale-free networks are robust (Figure 1.5). If a node is lost (as long as it is not a highly connected hub) the overall topology will remain intact. In a regular or random network however the loss of any node could disrupt the entire network structure. Also obvious from the same figure is how the longest path length can be substantially shortened in a scale-free network compared to a random network. By passing via a few hubs information can spread very quickly in a scale-free network.

1.2 Biological networks

Molecular cell components directly or indirectly interact with other molecular structures in the cell, often in complex networks, to carry out specific

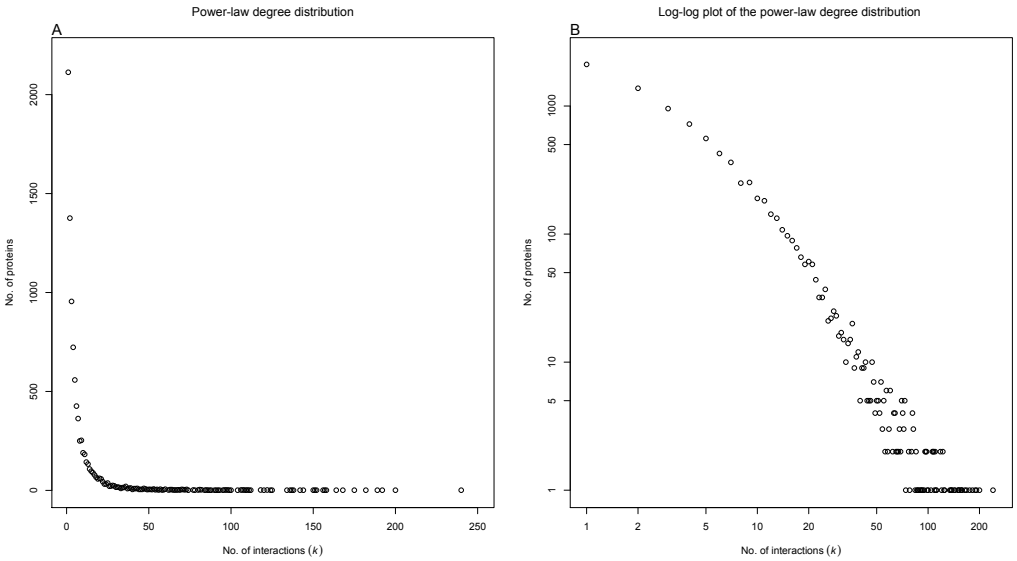


Figure 1.3: Power-law degree distribution and log-log plot of the same distribution. An example of a real-world, scale-free network is the human protein interaction network investigated in the research chapters of this thesis. Here it is clear that very many proteins have few interactions while a few proteins have very many interactions. (In this particular network most proteins only have one interaction, while one protein has 240 interactions.) This relationship of the interactions (*degree*) a) follows the power-law distribution, and b) declines in a straight line when the logarithm of the values are plotted instead. (The data used to plot the distributions in a and b come from the research in chapter 4, section 4.3.1.)

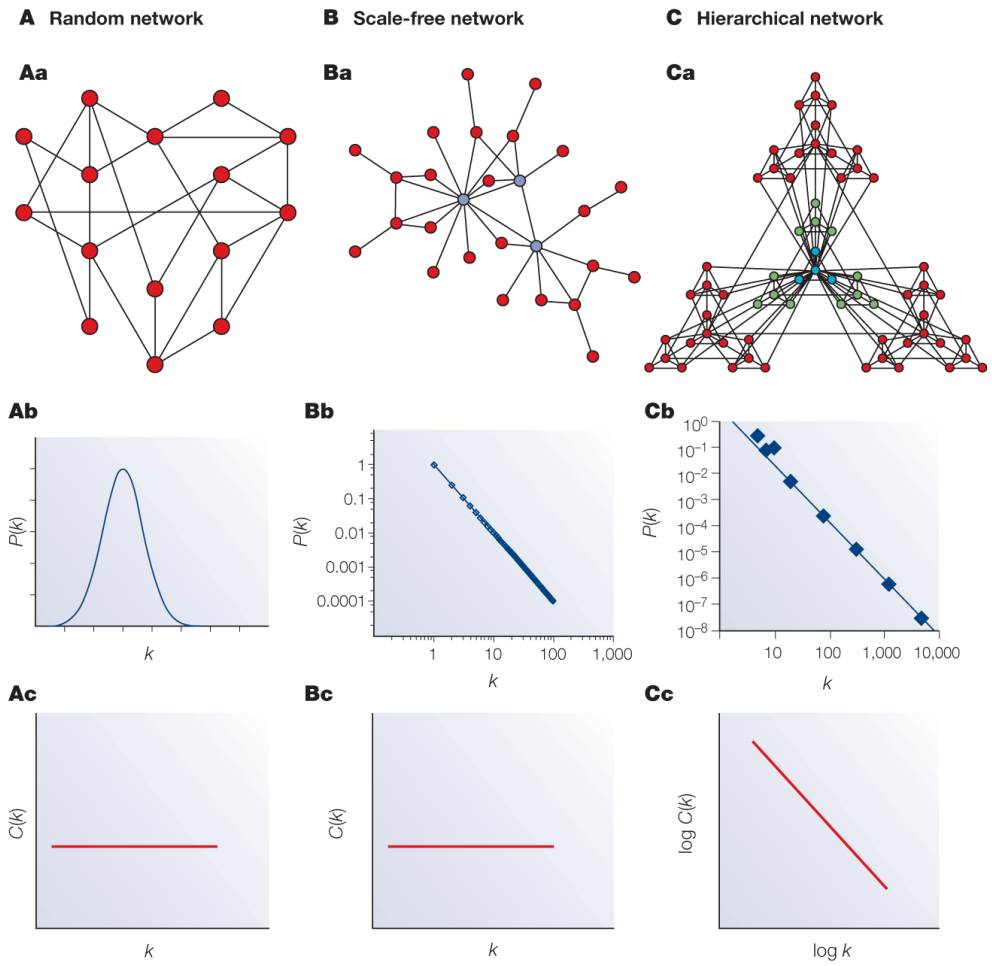


Figure 1.4: Distributions of random, scale-free and hierarchical networks. Figure of networks with their respective node degree distributions and clustering coefficient distributions plotted below. Aa) A random network. Ab) The node degree distribution of a random network follows the Poisson distribution. Ac) The clustering coefficient is independent of how many connections a node has. Hence the horizontal line as these two values are plotted against each other onto a graph. Ba) A scale-free network. Bb) Scale-free networks follow a power-law distribution. Bc) Just as for the random network the clustering coefficient is independent from the number of connections of each node. Ca) A hierarchical network. Cb) Hierarchical networks are a type of scale-free networks, which becomes obvious when the node degree distribution is plotted as it follows the power-law. Cc) Hierarchical networks are highly clustered, and thus the clustering coefficient is not independent of the node degree. The plot shows how nodes with few connections are highly clustered into separate sections in the network. The different parts of the network are connected to each other by a few hubs. Figure from Barabási and Oltvai (2004).

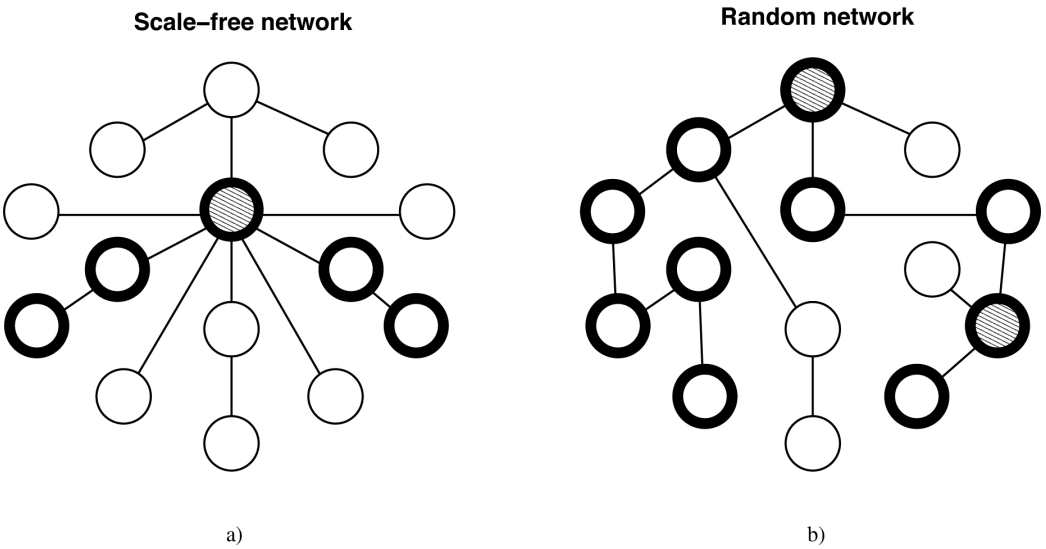


Figure 1.5: Path length and robustness in scale-free vs random networks. Both networks contain the same number of nodes and edges. a) However, the scale-free network has a hub (grey node) with $k = 8$ connections, and the longest path between any two nodes in the network is 4 steps (see highlighted nodes). b) In the random network two nodes have more connections than the rest, but not by much as their connectivity is $k = 3$, compared to the connectivity $k = 1$ or $k = 2$ of the remainder of the nodes. Moreover, the longest path between any two nodes is far longer than in the scale-free network as it is 9 steps. Another thing that becomes very obvious when comparing the two networks is that if a random node is lost the topology of the scale-free network (as long as the hub is not the node to be lost) is far more likely to remain intact i.e. scale-free networks are robust. Meanwhile the random network is very sensitive to disruption. The entire network structure can dissolve from the loss of one node. Figure from Light (2006).

functions. It is possible to distinguish between five different types of biological networks: i) transcription factor binding, ii) protein-protein interaction (PPI), iii) protein phosphorylation, iv) metabolic interactions and v) genetic interaction networks (Zhu *et al.*, 2007). This thesis research focuses on two of these, namely metabolic and PPI networks.

1.2.1 From metabolic pathways to metabolic networks

Metabolism is the biochemical processes whereby the cell's machinery extracts energy in order to function. The cell builds up complex compounds from simpler ones through anabolism, and degrades complex compounds into simple sugars through catabolism. One type of protein, enzymes, catalyse chemical reactions in the cell by lowering the activation energy of a reaction. Thus helping substrates turning into products, which can be consumed by the cell, and/or perform other necessary tasks in the cell's metabolism. A reaction where a substrate S is turned into a product P with the help of an enzyme E can be depicted in a metabolic pathway as follows: $S \xrightarrow{E} P$, or if the reaction occurs in through several reactions before the end product is produced as $A \xrightarrow{E1} B \xrightarrow{E2} C$. Obviously some reactions can also be reversible, depicted by double-edged arrows. When all the metabolic pathways in the cell are put together they form a network. A pathway (Figure 1.6a) can also be depicted as a graph where the nodes are reactions and the edges connecting between the nodes are the substrates (Figure 1.6b). Since some components, such as the co-factors adenosinetriphosphate (ATP) and nicotinamide adenine dinucleotide (NAD+/NADH), are part of very many reactions they are sometimes removed from the network (Figure 1.6c), and thus a more informative pattern in the network may emerge. Such components are often referred to as 'currency metabolites' (Huss and Holme, 2007). Originally, as the metabolic network depiction evolved from how metabolic pathways were written, the substrates and products were the nodes while the reactions (i.e. enzymes) where the edges of the network (Figure 1.7a). However, it is now more common to depict the enzymes as the nodes while the reactions are the edges of a metabolic network (Figure 1.7b).

Metabolism is modular as many components and enzymes interact only with a few other components thus forming an isolated group i.e., a *module* that in turn interacts with other modules via one or a few connections. This

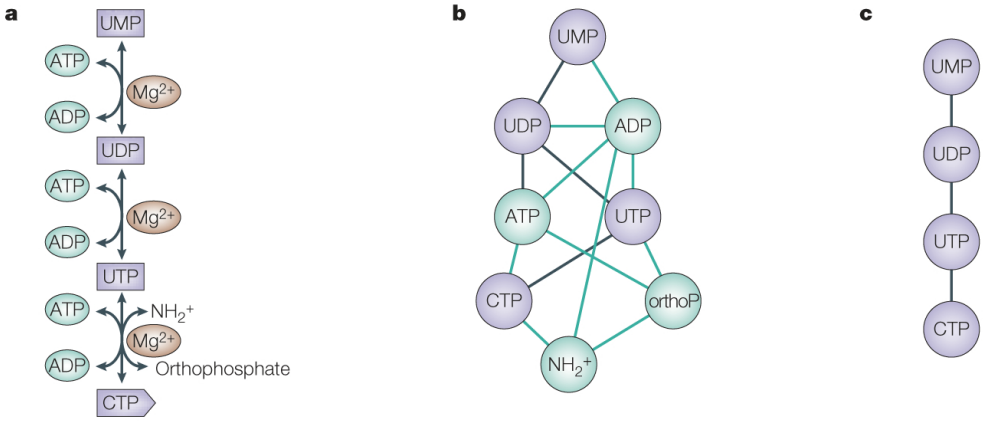


Figure 1.6: From metabolic pathway to metabolic network depiction. Chain of metabolic reactions in a cell depicted: a) In a metabolic pathway b) As a metabolic network where the substrates are the nodes and the reactions connecting these nodes are the edges c) As a simplified metabolic network where co-factors and other promiscuous substrates (often referred to as 'currency metabolites') have been removed creating a metabolic network with less nodes. Thus allowing a more informative pattern emerge in the network. Figure from Barabási and Oltvai (2004).

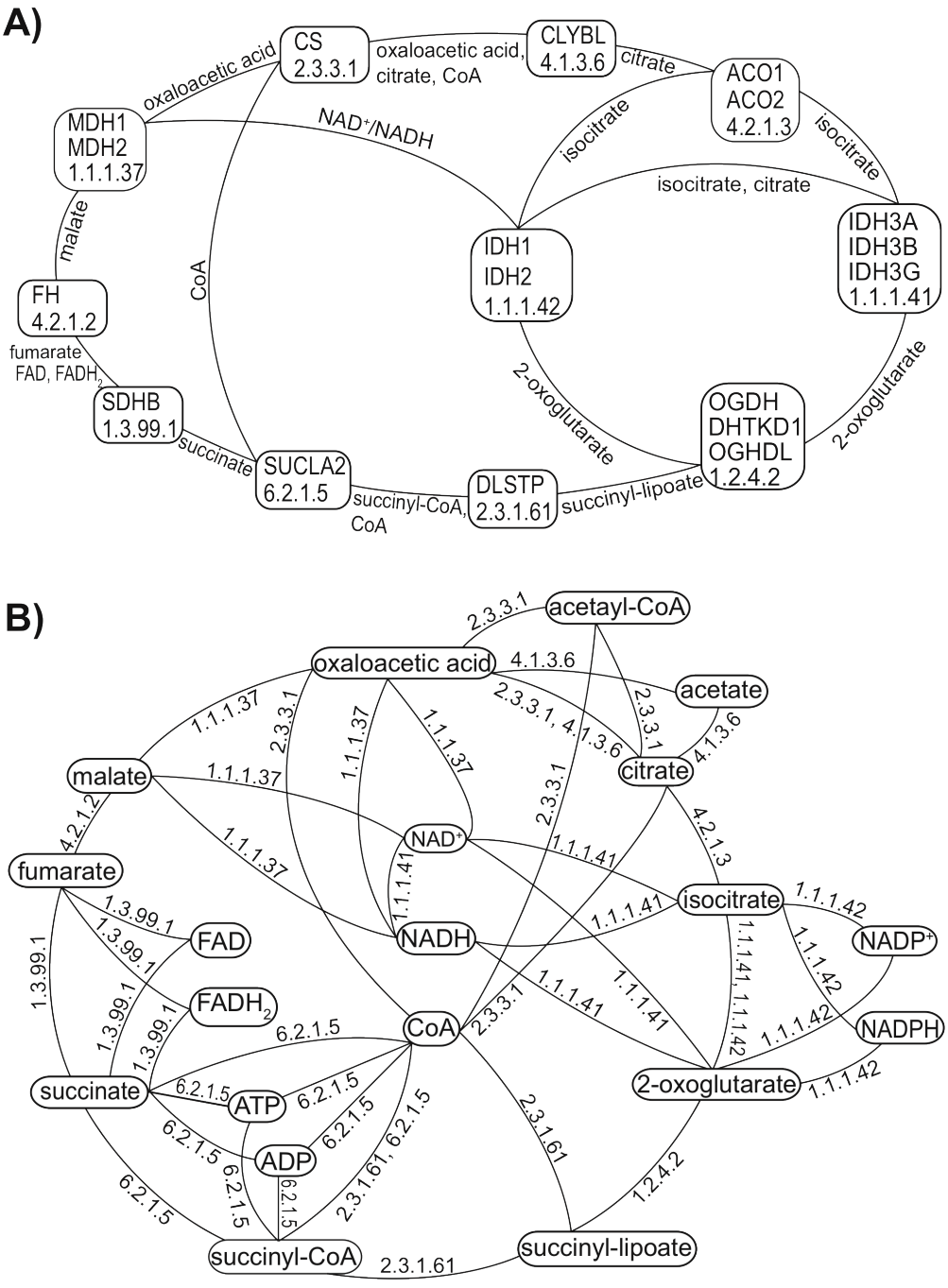


Figure 1.7: Enzyme and compound centric networks. Two representations of metabolic networks a) Enzyme centric network were the enzymes are the nodes, while the compounds are the edges. b) Compound centric network were the metabolites (substrates and products) are the nodes and the enzymes (or reactions) are the edges. Figure by Gavin C. Conant.

allows for the reuse of modules through reconnection of interactions *between* modules while conserving the interactions *within* modules. Obviously then modularity is very important for the evolution of metabolic networks. In fact, metabolic networks are likely to be some of the oldest biological networks (Caetano-Anolles *et al.*, 2009), as will be further elaborated in subsection 1.5.2. The modularity makes the metabolic networks hierarchical. As shown in figure 1.4Ca certain networks (e.g., metabolic networks) subdivide into separate modules, which are connected to each other through a few hubs. Just as for other types of networks the node degree $P(k)$ follows the power-law distribution (Figure 1.4Cb), indicating hiearchical networks are scale-free networks. However, the distribution of the clustering coefficient $C(k)$ (Figure 1.4Cc) also indicate that the network is highly clustered, and not independent of the node degree i.e. the number of connections each node in the network has (Barabási and Oltvai, 2004). Ravasz *et al.* (2002) studied the properties of 43 different metabolic networks, and concluded that they were all hierarchical, and that this hierarchical modularity corresponded to distinct metabolic functions of the cell. Since Ravasz *et al.* (2002) studied metabolic networks in 43 different organisms it has been proposed that metabolic networks are indeed hierarchically organised (as no metabolic network so far has been found to contradict this statement) i.e. they are modular. However, they also have a scale-free topology, and thus $C(k)$ is independent of the network size, and the degree distribution follows the power-law (Ravasz *et al.*, 2002; Barabási and Oltvai, 2004).

1.2.2 Protein networks

The expression "no man is an island" applies to proteins as well. Proteins interact with DNA and RNA, are part of metabolic pathways as enzymes and (as shown in research chapter 3) help to spatially organise metabolism by ensuring the right substrates are in place for a metabolic reactions to occur. Obviously proteins also interact with each other. Transient or permanent, in pairs or as part of a protein complex. If all known protein interactions of an organism are analysed together as an entity its protein interaction network (PIN) can be studied, and conclusions can be drawn from patterns that emerge, which we would not discover if we studied proteins one by one. Not solely because that would be a subjective choice of proteins because even if we could work through all proteins present in a species one by one unless we

know how they interact with each other in a network we cannot really know all there is to know about each and every protein. As mentioned above the PIN has a scale-free topology, meaning that most proteins have a few interactions, while a few proteins have very many protein interaction partners. They are hub proteins. As illustrated by figure 1.5a if a random node in a scale-free network is removed the main network topology remains intact. However, if the hub is removed the entire network can disintegrate. This might seem fragile, but it actually makes the network much more robust to random loss of nodes. In PINs this is interesting as it has been observed that removal of highly connected proteins is correlated with lethality (Jeong *et al.*, 2001). This was further confirmed in a study by Liang and Li (2007) where gene essentiality and protein connectivity was studied in yeast, mouse and human, and where a positive correlation was found between gene essentiality and protein connectivity in both the PIN of yeast and human. As removal of a gene might lead to lethality they realised that whether the gene existed in single-copy or had been duplicated might also correlate to whether or not the gene was essential. They found that highly connected proteins usually exist in a single-copy in yeast, while they have often been duplicated in human. Thus there was a negative correlation between gene essentiality and gene duplicability in yeast, while no such correlation was found in human. Jonsson and Bates (2006) observed that human proteins involved in cancer have many interactions and are central hub proteins in the human PIN. When Rambaldi *et al.* (2008) studied this further they further discovered that not only were genes involved in cancer hub proteins, but that they often only existed in single-copy, and that it was this, rather than their gene function or tissue location, that the cancer genes had in common. Hence, if something happened to a highly connected, single-copy protein this had detrimental effects on the individual as the entire protein network could collapse, affecting far more proteins that those that interacted directly with the affected hub protein.

1.3 Gene gains and losses

1.3.1 Gene duplication

Following a gene duplication event, duplicate genes are most often lost again in the course of evolution. Neutral genes are more likely to be retained,

and become fixed through genetic drift in small populations than in large ones. Duplicated genes can also undergo positive selection through neo- or subfunctionalisation. Therefore, certain genes are more likely to become fixed in a species (thus remaining in the genome) than others. In the study presented in chapter 4 we investigated whether a 'self-interaction' (the word refers to when two or more proteins encoded by the same gene interact with each other to form homomers i.e., it is not as in the phosphoglycerate kinase (PGK) example in section 1.4.2 one protein interacting with two parts of itself) plays a role in gene retention after duplication, and whether there is a difference in retention of self-interacting genes that duplicate through small-scale duplication (SSD) compared to genome wide duplication i.e. whole genome duplication (WGD).

1.3.1.1 Factors governing duplicate gene retention

After a gene duplicates the original gene and its newly formed copy can undergo several different fates. It can gain new, beneficial mutations, *neofunctionalisation*, all while the other copy retains the original function of the gene, or both copies can become complementary, and together cover the function of the single ancestral gene - a mechanism called *subfunctionalisation*. This last process is explained in the duplication-degeneration-complementation model (DDC model). Originally this model suggested that the two copies of a duplicate pair would subdivide the expression pattern and share the function of the ancestral, single copy gene. According to this model, loss of function will be complementary in the two copies so that both copies are needed to perform the function of the original gene, and therefore the subfunctions must also be independent so that mutations in one gene does not affect functions of the other gene (Lynch *et al.*, 2001; Papp *et al.*, 2003b; Prince and Pickett, 2002).

Subfunctionalised copies will be under different selective constraints to the original, ancestral gene. Moreover, the DDC process can occur after neofunctionalisation (Prince and Pickett, 2002). Whether a newly duplicated gene is retained or not depends on the population size. In large populations it is more likely that an allele has sub- or neofunctionalised before duplication, and this of course, does not have to occur in the wild type allele. Moreover, the ancestral locus can become neofunctionalised without the need for beneficial mutations if the new locus becomes established with the wild-type

allele. Furthermore, it takes longer for a gene to become fixed in a large population compared to a small one. As a consequence a subfunctionalised allele will have more time to gather mutations, which may silence the allele if the mutations are degenerate, but may also save the allele if the new mutations are benefitial (Lynch *et al.*, 2001).

A recent study suggests that rather than the DDC model after gene duplication the gene and its duplicate copy will first subfunctionalise, then neofunctionalise, i.e., the duplicate genes will *subneofunctionalise*. It is important to consider the effects on the gene network, and not just on the duplicated gene and its newly formed copy. It is more important to stabilise selection, to preserve the ancestral function, than to drive positive selection. The need for stability may explain the high occurence of neofunctionalisation. Moreover, subfunctionalisation in itself is insufficient to explain the high retention of duplicated genes. Rather, it seems as if gene redundancy and neofunctionalisation in combination with subfunctionalisation are all important for this retention(MacCarthy and Bergman, 2007).

1.3.2 Whole Genome Duplication

When the entire genome of an organism is duplicated, WGD can occur by two different mechanisms; autopolyploidy or allopolyploidy. In the first case, when the number of chromosomes is doubled within one individual, while in the latter, the fusion of the chromosomes of two nuclei from the parental species where the offspring contains the sum of the parental chromosomes (Wolfe, 2001; Freeling and Thomas, 2006; Sémon and Wolfe, 2007; Van de Peer *et al.*, 2009). This leaves three big questions. Firstly, how does an individual that has undergone WGD successfully reproduce unless other individuals of the same species have had their whole genomes duplicated as well? Secondly under what circumstances does WGD occur, and finally what are the biological and evolutionary consequences of WGD?

In plants polyploidy (i.e., WGD) is far more common than in animals. It is estimated that it occurs in 30 - 80 % of all flowering plant species (Otto and Whitton, 2000; Xiong and Pires, 2011). Moreover, WGD is far more common amongst plants and animals that can undergo selfing (i.e., asexual reproduction), and that are present all year round, as a long lifespan increases the chances of rare events (such as WGD) occuring to an individual. While many polyploid plants and insects will reproduce asexually this is not the

case in polyploid amphibia and fish i.e., the latter can reproduce sexually also as polyploids. When a species has undergone WGD e.g., gone from having its diploid genome duplicated and thus become a tetraploid it can (for many species) successfully reproduce triploid offspring with a diploid individual. The triploid offspring will in turn produce haploid, diploid or triploid offspring (Otto and Whitton, 2000). The problem with polyploid species arose with the sex chromosomes (Wolfe, 2001). Although this problem can be overcome when it is not the X-chromosome (in species with XX / XY sex chromosomes), but rather the presence or absence of an Y chromosome that determines the sex i.e., all of XY, XXY, XYYY, XXYY and XXXY will be male. In birds however (where the female has the sex chromosomes ZW, while the male have two identical sex chromosomes ZZ), ZZW individuals will be of unclear, intersexual phenotype, and mortality will be high. In fact, in a study conducted on chicken none of the eggs with haploid, triploid or tetraploid chromosomes hatched. Human triploid or tetraploid individuals rarely survive until birth. When they do triploid humans are sterile while tetraploid babies do not survive beyond infancy. It is believed that imprinting in mammals is hindering proper growth, which is why there are no adult, polyploid humans (Otto and Whitton, 2000).

Many extant diploid species have polyploid ancestry. However, polyploidization has only prevailed a few times in the evolution of eukaryotic life; Four times in angiosperm, once in fungi, two times in chordate/vertebrates (called 1R and 2R after the first and second rounds of genome duplication) (McLysaght *et al.*, 2002; Van de Peer *et al.*, 2010) and a third time in teleost fish i.e., ray-finned fish (fish-specific genome duplication (FSGD), sometimes also referred to as 3R) (Meyer and Van de Peer, 2005). That is, after a WGD event most genomes will slowly, over millions of years, lose genes until most genes have reverted to the number of loci they had before the WGD event (Conant and Wolfe, 2007). However, these WGD descendants will give rise to great species variety e.g., over 25,000 fish species and over 350,000 flowering plant species. Thus, although many polyploids seem to die before birth, reaching a reproductive age or being sterile, hence unable to reproduce, the few times reproduction is successful it seems to generate evolutionary superior species (Meyer and Van de Peer, 2005; Sémon and Wolfe, 2007; Van de Peer *et al.*, 2009). Van de Peer *et al.* (2009) propose that if a WGD occurs in a species at the approximate time of a dramatic ecological event, such as mass extinction, the newly formed polyploid species will probably have enough

genetic variation in order to gain a phenotypic advantage over other diploid species. Hence being able to occupy an ecological niche that has become available due to the dramatic, even cataclysmic ecological changes, while these same polyploid species would not have the same competitive advantage over their diploid sister species in an ecologically stable environment. Several findings support their theory, for example, the FSGD in teleost fish occurred 226 - 316 mya, which overlaps with the Permian-Triassic mass extinction from 250 mya (Van de Peer *et al.*, 2009). However, this is not the only way species undergoing polyplodization (i.e. WGD) could gain an advantage over species that had not undergone WGD. Conant and Wolfe (2007) showed that in *S. cerevisiae* after WGD had occurred all genes in the glycolytic pathway were retained, which could have given *S. cerevisiae* an advantage over other microbes when sugar-rich fruit appeared on earth, which was around the time of the WGD event in fungi.

1.3.3 Comparisons of Whole Genome Duplication vs Smaller Scale Duplications

Gene duplication is proposed to occur through SSD or WGD. The latter model of duplication was championed in 1970 by Susumu Ohno (Ohno, 1970). He stated that it is easier to derive new genes through gene duplication than *de novo*, and suggested that two or even three rounds of WGD had occurred in vertebrates. As stated in the previous subsection for vertebrates there is now evidence that two rounds of genome duplication (2R) occurred in chordates, just before and after the divergence of the lamprey lineage (Van de Peer *et al.*, 2010; Dehal and Boore, 2005), and a third round of genome duplication (3R) in ray-finned fish (Meyer and Van de Peer, 2005; Van de Peer *et al.*, 2010). Davis and Petrov (2005) studied gene duplication in the yeast *S. cerevisiae*. They found that highly expressed genes are more likely to duplicate, and that orthologues of both SSD and WGD duplicates evolve slower than orthologues of the singleton genes they studied (Davis and Petrov, 2005). Finally they found that SSD and WGD genes differed in their function. WGD genes typically encode catalytic proteins, while SSD genes tend to encode binding proteins and enzymatic regulators, which are rarely found amongst WGD duplicates. This is consistent with other studies in yeast (Guan *et al.*, 2007; Hakes *et al.*, 2007a). However, Hakes and colleagues found that although WGD genes are more likely to form a protein complex than SSD genes, the

latter are more likely to be essential genes in a protein complex.

Most genes are lost after duplication, regardless of whether the mechanism was small-scale, or genome-wide. A gene duplicated by SSD needs to become fixed in the entire population to be retained. In order for this to happen, the gene usually needs to undergo selection as described above, or the population size must be small to facilitate fixation through random genetic drift. Most duplicates from SSD events never pass this stage. Genes that are duplicated in a WGD event, on the other hand, are duplicated together with all other genes in that individual, which will then undergo immediate speciation, as occurred at the fish-tetrapod split (Dehal and Boore, 2005). Therefore the duplicated genes do not need to become fixed in the genome of the entire population. However, after a WGD event, the genome will go through massive genome rearrangement and loss, and many of the duplicated genes will be lost at this stage (Davis and Petrov, 2005; Dehal and Boore, 2005).

1.3.4 Gene dosage and gene duplicability

What forces make a newly duplicated gene prevail in the genome, and become fixed in the population? This question is particularly interesting considering the high loss rate of duplicated genes that follows small-scale or genome-wide duplication events. The answer seems to be due to the gene dosage balance effect. This should not be confused with the gene dosage effect, which occurs when there is not enough gene product in a cell due to the deletion of one of the genes alleles (haploinsufficiency). A gene dosage balance effect refers to the relative dosage of genes whose products interact or function in the same pathway. Qian and Zhang (2008) showed that in human and yeast that there is no statistically significant difference between the duplicability of haploinsufficient and haplosufficient genes i.e. that haploinsufficient genes do not tend to duplicate more often than haplosufficient genes. In yeast they showed that haploinsufficient genes that encode subunits in protein complexes tend to be kept in the genome after a WGD event, supporting the theory of gene dosage balance effect theory (Papp *et al.*, 2003a; Qian and Zhang, 2008). Hughes *et al.* (2007) mention in their review that dosage-compensations may act as a force, which *initially* retain duplicate genes after a WGD event, but as time goes by there is no preferential retention of genes involved in protein complexes or different pathways. Hughes and colleagues speculate that this initial retention allows sufficient time for neo-

and subfunctionalisation to occur and thereby ensure selection for their subsequent retention (Hughes *et al.*, 2007; Rastogi and Liberles, 2005). Liang *et al.* (2008) studied the protein structure of more than 12 000 proteins, in several species, including human, and showed that proteins that are sensitive to their surroundings for proper folding, called 'under-wrapped' proteins (meaning that they are dependent on stabilising interactions from other proteins for their protein structure), are dosage sensitive, and hence more vulnerable to dosage imbalance (Liang *et al.*, 2008). This in turn makes duplicate copies of such genes more likely to be lost again after a duplication event. Unless they arose through a WGD event, in which case there will be no gene dosage imbalance since all genes in the organisms genome are duplicated. They also showed that more complex the organisms are less sensitive to gene dosage imbalance, with *E. coli* followed by yeast (*S. cerevisiae*) being the more sensitive organisms whereas human and thale cress (*A. thaliana*) were least dosage sensitive. They found that under-wrapped genes in yeast have a tendency to exist as single-copy genes and not duplicates regardless of their functional category. The correlation between under-wrapped proteins and their decreased gene duplicability appears to be independent of the function of the protein they encode.

1.3.5 How gene duplication affects the protein interaction network

Following gene duplication, the proteins of the two newly formed genes should interact with the same set of proteins as the ancestor gene. Asymmetry in the number of PPI partners will arise after the gene pair paralogues have had time to diverge. The two paralogues will each have a different set of interacting partners, and only share a subset of these with their paralogous copy. Makino and Gojobori studied duplicated gene pairs in yeast in order to elucidate whether the dominant evolutionary force is PPI gain or loss. They found that the slower evolving gene copy would have more interaction partners than the faster evolving copy. Moreover when comparing the two copies, they showed that the evolutionary rate difference is smaller in cases where more protein-protein interactions are still shared between the duplicate pair (Makino *et al.*, 2006; Makino and Gojobori, 2007). Thus while Hahn and Kern (2005) showed that central proteins in the PIN of yeast, worm and fly evolve at slower rates compared to proteins in the periphery of the

protein network regardless of the number of PPIs, Makino and Gojobori (2006) showed that interacting paralogous proteins with different functions and in the sparse part of the PIN evolve slower than paralogous proteins in the central part of the PIN and than interacting paralogues with the same function. Liang and Li (2007) investigated the correlation of gene essentiality, gene duplicability and protein connectivity (how well connected a protein is i.e. measure of interacting partners). They found that essential genes in yeast (defined as having a lethal or sterile knockout phenotype) have high protein connectivity and tend to exist in single-copy, and that there is a negative correlation between protein connectivity and gene duplicability. This can be summarized in the 'central-lethality rule' meaning that an organism will most likely die if highly connected proteins are deleted. They also investigated these properties in human and mouse, and concluded that highly connected proteins in general have high gene duplicability (that is, the opposite trend), and that despite the presence of duplicates these highly-connected genes are usually essential for the organisms ability to survive or reproduce. No correlation was detected between gene essentiality and gene duplicability (Liang and Li, 2007).

Pereira-Leal and colleagues investigated the evolution of protein complexes and concluded that they evolve through the duplication of homodimers. They showed that protein interactions amongst paralogous proteins occur more frequently than can be expected purely by chance in yeast, worm and fruitfly (Pereira-Leal *et al.*, 2007). They also show that protein-protein interactions between homodimers and paralogous dimers in these species are not independent, and conclude that the latter dimer type evolved from the former, something that had been suggested previously (Wagner, 2001). They argue that gene duplication and divergence are important forces driving the expansion of the eukaryotic proteome, and that multiple copies of identical subunits are an economical way of forming larger functional structures. The subunits will be co-localized in the cell, and transcribed by the same mRNA, and as a consequence self-interactions should be overrepresented in protein interaction networks. When these genes duplicate the ancestral gene and a recently created paralogue will be identical, and can interact in a paralogous heteromer (from a genetic point of view this is a heteromer, though from a biochemical standpoint it is a homomer until the protein sequences begin to diverge), which will be as stable as the ancestral homomer, however asymmetry is introduced as the paralogues start diverging (Levy and Pereira-

Leal, 2008). Another study shows that 50-70% of proteins with a known quaternary state (i.e., proteins consisting of several polypeptide chains or subunits (Lenhinger *et al.*, 1993)) form homomers (Levy *et al.*, 2008), which is not surprising since approximately half of all crystallographic structures are homo- or heteromeric protein complexes (Levy *et al.*, 2006). Levy *et al.* (2008) found that certain quaternary structures, such as dihedral (where the same end of the subunits interact) and cyclic (where the opposite ends of the subunits interact), are more prevalent than other types of structures in homomers. The prevalence of dihedral cases can probably be explained in terms of random mutations not causing disruption to binding between two identical subunits to form a dihedral (since they interact with each other by binding to the same end). In contrast, mutation in a cyclic quarternary structure could disrupt the interaction if a compensatory mutation does not appear in the other end of the interacting subunit. Also, disregarding the forces driving molecular evolution, simply by studying iso-chemical properties of proteins one can conclude that proteins will have a higher affinity for interacting with other proteins which are identical to themselves because the lowest energy distribution of a homodimer is generally lower than that of any random heterodimer (Lukatsky *et al.*, 2007).

As stated above, it has previously been shown for yeast, worm and fly that paralogous proteins interact more frequently than would be expected by chance (Pereira-Leal *et al.*, 2007). The same study also showed that PPIs between homodimers and paralogous dimers in these species are not independent, and that the latter dimer type evolved from the former. Ispolatov and colleagues also showed that duplicated genes tend to self-interact more often than can be expected by chance alone in yeast, worm, fly and human. They also found that the likelihood of a protein self-interacting is proportional to its number of interacting partners, and that proteins forming heterodimers with their paralogues tend to have twice as many interaction partners as the average protein in the network. This lead them to conclude that interactions between paralogous proteins forming heterodimers are inherited from self-interacting proteins which form homodimers and from which interactions between heterodimers evolve (Ispolatov *et al.*, 2005).

1.4 Constraints on protein sequences

1.4.1 Selection vs genetic drift

When a mutation occurs in a DNA sequence it contributes to the genetic variation in the population. However, it is not until the substitution is fixed in the population that the mutation becomes a substitution (Bromham, 2008). Mutations at any given allele do not occur at a high rate in a single individual. However, if the entire population, of size N, is considered and if it is a diploid population then there are $2N$ alleles that can mutate at rate μ each generation i.e., there are $2N\mu$ mutations per generation in the population. Thus allowing for great variation within a population (Hartl and Clark, 1997). If the mutation in an individual is passed on to its offspring then the mutation becomes part of the variation in the population (and we should really be talking about the effective population size N_e i.e., the individuals in a population that reproduce and hence pass on their mutation to their offspring). Whether the allele is fixed in the population depends on the selective constraints acting upon the allele and on the size of the population 'harbouring' the allele in question. If we have the DNA sequence of the allele in question we can detect the selective constraints acting on the allele (and driving it either to fixation or out of the gene pool) estimating the nonsynonymous K_a (also denoted d_N) and synonymous K_s (also denoted d_S) substitutions per site in the protein coding sequence. As the name implies nonsynonymous substitutions changes the DNA sequence so that the codon encodes a different amino acid, while a synonymous substitution is silent since it does not change what amino acid is encoded by the codon in the DNA sequence. The ratio of the nonsynonymous over synonymous substitution rates $\omega = K_a/K_s$ is called omega, and is a measure of the selective constraint acting on the protein encoding DNA sequence. If the nonsynonymous mutation gives neither an advantage or disadvantage the nonsynonymous and synonymous mutations will be fixed at the same rate i.e., $K_a = K_s$ and thus $\omega = (K_a/K_s) = 1$. If however the nonsynonymous mutation is deleterious purifying selection will drive the mutation out of the population, and $K_a < K_s$ and consequently $\omega < 1$. If however the nonsynonymous mutation is beneficial natural selection will drive the mutation towards fixation (also referred to as adaptive evolution) and hence $K_a > K_s$ and $\omega > 1$ (Yang, 2001). If a mutation is neutral its frequency

in the population will fluctuate under random genetic drift until it is either fixated in or lost to the population. At the beginning of this section we saw that there are $2N\mu$ mutations in a diploid population of size N. If we consider that the fixation rate of any one mutation is $1/2N$ we realise that for neutral mutations (that are free to fluctuate under random genetic drift since they are not under any selective constraint) the fixation rate is

$$\frac{2N\mu}{2N} = \mu \tag{1.1}$$

i.e., the same rate as the mutation rate itself. However, in reality very many mutations are only nearly neutral. Meaning that they have a small effect on fitness, and thus while slightly deleterious mutations would be removed from large populations they could become fixed due to genetic drift in species with small population sizes e.g., mammals (Ohta, 1995), endosymbiotic microorganisms (Woolfit and Bromham, 2003) or species isolated on islands (Woolfit and Bromham, 2005).

1.4.2 Molecular co-evolution and protein interactions

Lovell and Robertson (2010) define molecular co-evolution as "reciprocal evolutionary change in evolutionary interacting loci". This comes from the original meaning of co-evolution between two species. However, here it is used to refer to the molecular co-evolution between residues. For example, if certain residues in a peptide sequence bind to each other for proper protein folding then purifying selection will constrain these sites from mutating if mutation prevents these sites from binding, and ultimately prevents proper folding of the peptide sequence. However, if mutation at one site is compensated by a complementary mutation at the second site then residues can still bind to each other, the peptide sequence can still fold properly and the two sites are co-evolving. This reciprocal change between interacting residues has been studied to identify site-specific co-evolution e.g., for the study of the V3 loop of the human immunodeficiency virus (HIV) type 1 envelope protein gp120 (encoded by the *env* gene) using an entropy approach derived from Shannon's entropy to identify covarying site pairs within the loop peptide sequence (Korber *et al.*, 1993). In another study the entire *env* gene was scanned for co-evolving residues pairs. 24 such pairs (consisting of 14 residues) were found to be co-evolving within the

V3 loop, and for the entire gene 848 pairs (consisting of 263 residues) were found to co-evolve (Travers *et al.*, 2007). However, although 8 residues in the V3 domain and 5 residues in the co-receptor binding domain of gp120 bind either directly to or within the binding pocket of the host cell's CD4 receptor, and the method applied (Fares and Travers, 2006) detects co-evolution between pairwise sites detection of co-evolution does not necessarily imply an interaction between proteins since also hydrophobicity, molecular weight and the two combined contribute to the co-evolution signal that is detected (Travers *et al.*, 2007).

Goh *et al.* (2000) developed a method to predict PPIs using phosphoglycerate kinase (PGK), a two-domain enzyme which is active when the N- and C-terminal bind each other. In order to remain a functional enzyme the two domains must be able to bind to each other, and consequently they must co-evolve since a change in the amino acid sequence of one terminus must be matched by an equivalent change in the other terminus for the binding to be preserved. Therefore PGK was ideal for developing a measure of co-evolution, which could then be used as a standard reference to compare measures from other amino acid sequences thought to bind to each other. Thus, by constructing one multiple sequence alignment (MSA) per terminus, and one phylogenetic tree per terminus of the PGK the correlation coefficient of all pairwise distances between the two trees could be calculated. Then the overall Pearson's correlation coefficient, r, was calculated from these values, and used as a reference when scanning for co-evolution of binding specificity between chemokine ligands and receptors (Goh *et al.*, 2000).

The method described above was refined in another study where ~67,000 pairs of *E. coli* proteins were scanned for co-evolution signal. This was achieved by constructing MSAs from Basic Local Alignment Search Tool (BLAST) hits and from databases with MSAs. Then for each pair of proteins phylogenetic trees with at least 11 matching branches were constructed i.e. 'mirrortrees'. Then the correlation coefficient values from known (i.e. real) protein interactions were compared against correlation coefficient values from the remaining paired proteins. If the correlation coefficient value of two paired proteins was equal to or higher than that of the protein pairs known to interact it was a good indication that the paired proteins lacking interaction information were indeed interacting as well. Of the 67,209 possible pairs of protein interactions in *E. coli* 2742 were predicted to interact using this 'mirrortree method'. Moreover, as a control the known protein interactions

were also included in the list of possible protein interactions, and these real PPIs were also predicted to interact using this method. (Pazos and Valencia, 2001).

The mirrortree method has also been applied to detect protein interactions in the eukaryote *S. cerevisiae* Hakes *et al.* (2007b). Both Pazos and Valencia (2001) and Hakes *et al.* (2007b) report an average of ∼20% false positive protein interaction predictions with the mirrortree method. It should also be noted that this method cannot really distinguish between interactions between two proteins, and proteins that are simply members of the same protein complex (Lovell and Robertson, 2010). However, Kann *et al.* (2009) successfully identified that although interacting proteins have a co-evolutionary signal along their entire sequence this signal is higher at the binding domain and surrounding binding neighbourhood. Moreover, Tillier and Charlebois (2009) developed a method that measures the strength of co-evolution between any two eukaryotic protein families by finding their largest common distance submatrix. Thus resolving the issue of comparing families of different size and the problem of including paralogous genes into the analysis. Using this method they successfully identified human proteins known to physically interact with each other. Hence distinguishing them from proteins simply belonging to the same biochemical pathway.

The problem with correlated distance methods such as those described above is that nucleotide mutation rates can vary between species. Therefore if the nonsynonymous substitution rates K_a are normalised against the synonymous substitution rates K_s by calculating the rate $\omega = (K_a/K_s)$ per branch (i.e., species), adjustment for the phylogeny is no longer necessary since a species-specific measure was obtained from the rate per branch (Clark and Aquadro, 2010). In fact, Clark and Aquadro found this novel mirrortree method (applied in chapter 5) superior to traditional mirrortree methods (described in paragraph above) that apply pairwise distances to calculate the correlated evolution between pairs of proteins.

1.5 Network evolution

1.5.1 Evolution of protein interaction networks

As mentioned in section 1.1, there are two general characteristics for real networks. Firstly they grow, and secondly they do so by preferential

attachment (Barabási and Albert, 1999), meaning that nodes with many interactions tend to acquire even more interactions. In a social context, this can be explained by the fact that a person who already has many established contacts will make further contacts through them. On the Internet it makes sense that well established websites are hyperlinked more by other pages, and the more credible sites attract more hyperlinks. In biological networks such as the protein interaction network gene duplication is the explanation for network growth. A duplicate and the original gene are identical immediately after gene duplication. Therefore the two will also share interactions. When a gene encoding a highly connected protein is duplicated the duplicate will also have all the same connections i.e., the number of connections will double. Over time though the original gene and its duplicated copy will start diverging, and consequently their interaction partners might change (see section 1.3.5) (Barabási and Oltvai, 2004; Makino *et al.*, 2006; Makino and Gojobori, 2007). However, this also explains how the protein network can be rewired and evolve (see section 1.3 and research chapter 4). Several studies have shown that self-interacting genes seem to be a driving force in the evolution of protein complexes (Wagner, 2001; Ispolatov *et al.*, 2005; Pereira-Leal *et al.*, 2007; Levy *et al.*, 2008). Indeed Presser *et al.* (2008) found that self-interacting genes existed in higher frequency in the ancestral PIN of *S. cerevisiae*. Kunin *et al.* (2004) also studied the *S. cerevisiae* protein network. They discovered that if they mapped the proteins against an evolutionary tree, and then grouped the proteins according to their connectivity in the network it became clear that proteins involved in similar functions appeared around the same time and had similar number of connections in the network. They observed that the most connected proteins in the network appeared at the time of the eukaryotic lineage suggesting their involvement in the organisation of the eukaryotic cell. Moreover they found that the majority of the oldest, pre-eukaryotic proteins had few connections and were metabolic enzymes, consistent with the fact that metabolism is a very old mechanism indeed.

Finally, as is clear from section 1.3 on gene gains and losses WGD plays a major role for successful duplication of dosage sensitive genes (see section 1.3.4), that are involved in protein complexes and/or metabolic pathways. Subsequently WGD is a very important force in the rewiring of protein networks as the duplicated gene pairs can acquire or modify their functions, and gain or lose connections in the network (Figure 1.8).

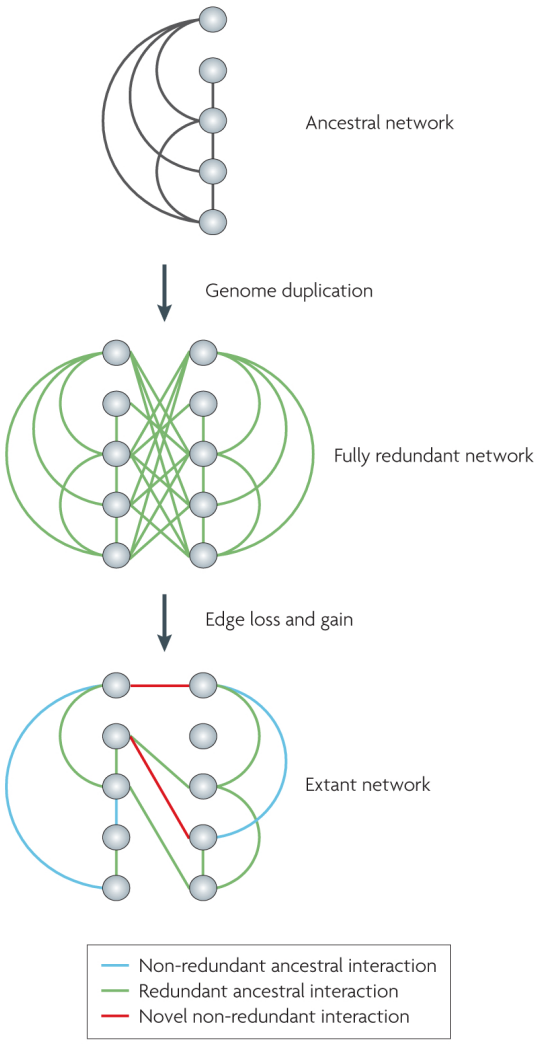


Figure 1.8: Rewiring of a PIN after WGD. Rewiring of a protein network after WGD: All proteins in a PIN are duplicated through WGD. Hence, initially all the edges (PPIs) are also duplicated. However, with time some interactions between proteins are lost, while others are gained (more unusual). Thus rewiring the network. Figure from (Conant and Wolfe, 2008).

1.5.2 How metabolic pathways and networks evolve

There are several theories on how metabolism and metabolic pathways evolved. The first one is called *retrograde evolution* (Horowitz, 1945). As the name suggests the hypothesis proposes that metabolic pathways evolved backwards. That is, at first there is plenty of a particular metabolite A, which is necessary for an organism to survive. With time, however, this metabolite is depleted, giving an advantages to organisms with an enzyme $E1$ capable of producing metabolite A from another metabolite B : $B \xrightarrow{E1} A$ (Figure 1.9; retrograde model). Then as metabolite B is also depleted an organism where enzyme $E1$ has duplicated and diverged to $E2$ will be able to recruit metabolite C in order to produce metabolite B, and finally produce the essential metabolite A :C $\xrightarrow{E2} B \xrightarrow{E1} A$. However, the retrograde evolution model requires an abundance of substrates, ready to be used, which is not a very likely scenario considering that many intermediate metabolites are unstable. Therefore, in 1974 Ycas proposed, and two years later Jensen expanded on, another hypothesis called *patchwork evolution* (Ycas, 1974; Jensen, 1976). This model suggests that in the early days of metabolism organisms with broad-spectrum enzymes could gain an advantage over other organisms as their enzymes could produce a wide number of new metabolites, and new metabolic functions could thus evolve. However, through gene duplication and subsequent sub- and/or neofunctionalisation the enzymes became specialised. Although the new specialised enzymes are not capable of catalysing many different substrates they are very efficient in catalysing one substrate into a desired metabolite (Figure 1.9; patchwork model). A trait that could generate a selective advantage over other organisms. Evidence suggests that both the retrograde and patchwork evolution models have occurred, although the latter for more reactions (Light, 2006; Caetano-Anolles *et al.*, 2009; Yamada and Bork, 2009). A third hypothesis for how metabolism evolved, called the *shell hypothesis*, suggests that the oldest metabolic pathway was the citric acid (or Krebs) circle because it was autotrophic, and capable of producing nutrients from inorganic sources. The theory is that more complex metabolic pathways were added to the existing ones as shells, hence the name of the hypothesis, over evolutionary time (Morowitz, 1999; Caetano-Anolles *et al.*, 2009).

What is known about metabolic networks is that they are modular i.e., the network consists of distinct, isolated groups that interact amongst

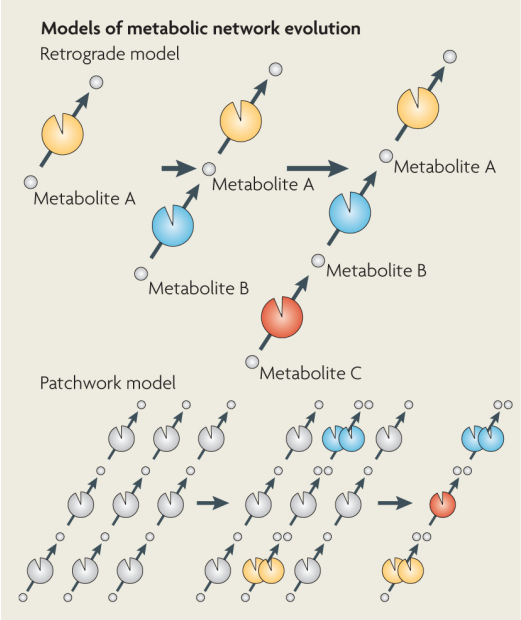


Figure 1.9: Evolution of metabolic pathways. According to the *retrograde model* metabolite A is essential to the organism. As metabolite A is depleted from the environment, organisms with an enzyme capable of producing metabolite A from metabolite B will have an advantage over other organisms. The pathway is expanded (in backwards fashion to the modern pathways) as the metabolites are depleted one by one. According to the *patchwork model* new metabolic functions evolved from broad-spectrum enzymes. However, although these enzymes were able to catalyse many different types of substrates in many different reactions over time duplication and mutation of genes caused enzymes to become more specialised and efficient, which gave organisms with the new specialised function an advantage over organisms with broad-spectrum enzymes. Figure from (Yamada and Bork, 2009).

themselves, but only with other modules via a few connections. This allows for reuse of modules, which can be recruited for new functions, and as a consequence the metabolic network will evolve (Caetano-Anolles *et al.*, 2009). In chapter 3 of this thesis we discovered that proteins, not themselves part of metabolic reactions, and thus not part of the metabolic network, acted as mediator proteins connecting proteins from different metabolic reactions (i.e., proteins that are part of the metabolic network). Thus the mediator proteins channel metabolism by bringing different enzymes into close proximity of each other enabling catalytic reactions to occur. We thus realise that as proteins evolve and as a consequence their protein interactions change this not only affects the protein interaction network and its evolution, but also the evolution of metabolic network since mediator proteins can bring different metabolic modules into proximity thus creating new metabolic pathways.

1.6 Aim

This thesis sets out to analyse the function and evolution of the human protein interaction network. The influence of protein interactions on the spatial organisation of metabolic networks is assessed in chapter 3. The ability of self-interacting human proteins to duplicate and how they evolve is analysed in chapter 4. Finally, the ancestral state of the human protein interaction network is reconstructed, and genes encoding protein-protein interactions specific to the primate branch are scrutinised for function and connectivity in chapter 5. Establishing whether a protein interaction is specific to humans, or if it also exists in a sister lineage, if it has duplicated or exists in single-copy, and whether it interacts with other proteins encoded by the same gene are some of the issues investigated in this thesis. The answers can help us understand the roles, importance and evolutionary legacy of protein interactions, and how these affect our biology.

Chapter 2

Materials and Methods

This chapter describes, in greater detail than in research chapters 3- 5, some of the datasets and the theory to some of the methods used for this thesis.

2.1 Constructing biological networks and assessing their quality

The research of this thesis use publicly available data as 'raw' laboratory material. Therefore this section is dedicated to describing how that data have been generated, assembled and curated by other researchers before it was curated for the purposes of, and used for the research projects described in this thesis methods sections of chapters 3- 5.

2.1.1 Protein interactions - detection and assessment

A protein interaction network (PIN) is a description of the relationship of protein-protein interactions (PPIs), and these datasets are considered the largest, and most diverse of all datasets used to construct biological networks (Zhu *et al.*, 2007). The dynamics of PINs are often measured in seconds or minutes, as protein complexes assemble and disassemble, but also over millions of years as genes are gained or lost, and hence the protein interactions, are gained and lost as well (Levy and Pereira-Leal, 2008). PPIs and PINs can be experimentally characterized and studied with techniques such as yeast two-hybrid (Y2H) and other high-throughput affinity studies. PPIs can be detected either individually between pairs of proteins, or *en-*

masse in high-throughput studies where the protein interactions of entire genomes are detected.

Yeast two-hybrid assays. Y2H systems can be used to establish whether two particular proteins interact. This method can also be used for large-scale studies, and has been used for whole proteome studies of yeast, worm and fly. Another advantage is that Y2H can detect both transient and stable protein interactions. A drawback is that only two proteins can be tested for interaction at a time (therefore protein complexes cannot be detected directly). Y2H is an *in vivo* method that takes place in the nucleus of the cell, which might not be the natural environment of the protein being investigated, and since posttranslational modifications can differ between yeast and other organisms it can be problematic to detect the interactions of cytoplasmic and membrane proteins as well as proteins in prokaryotes and mammals. Detection rates for false positive interactions, as high as 50-90 % have been suggested for Y2H, as well as false-negatives, where interactions known to occur are missed (von Mering *et al.*, 2002; Light, 2006; Shoemaker and Panchenko, 2007).

Mass spectrometry. A newer approach used for high-throughput studies is affinity purification mass spectrometry (AP-MS) or tandem affinity purification (TAP) followed by MS. This technique has been used for protein complex identification, and for large-scale proteome studies. However, affinity purification cannot be used for detecting individual protein interactions. Other drawbacks with this method are that the purification steps will wash out transient interactions, and hence only stable protein interactions will be detected i.e., the method might fail to detect transient interactions e.g., interactions involved in metabolism and signalling. Furthermore, this method suffers from protein contamination, and as a consequence protein interactions can fail to be detected.

Other experimental protein interaction detection methods. Other methods for detecting protein interactions are protein microarrays (detect proteins, rather than genes, using a fluroscent probe against the protein of interest that has been immobilised onto a surface), gene co-expression (interacting proteins have to be expressed at the same time), synthetic lethality (where the protein products of two mutated genes become lethal to the host cell when they interact) and fluorescence resonance energy transfer (FRET) (when two molecules in electronic excited states are within 10100 Å, and excitation is transferred from one molecule to the other without emitting

a photon)). Of these, Y2H, synthetic lethality and FRET are *in vivo* methods i.e., occurring in the cell, while the rest are *in vitro* methods (Shoemaker and Panchenko, 2007). Finally, although not surprising, it should still be noted that these different methods for detecting protein interactions only partially overlap in terms of the protein interactions they manage to detect (von Mering *et al.*, 2002; Templin *et al.*, 2002; Light, 2006; Zhu *et al.*, 2007; Shoemaker and Panchenko, 2007).

There are several different protein interaction databases, and the most complete proteome is that of the yeast *S. cerevisiae* (AP-MS; (Gavin *et al.*, 2002, 2006)). The Database of Interacting Proteins (DIP) (Eisenberg, 2010) contains experimentally validated, high quality protein interaction data for yeast (from Y2H, AP-MS, DNA and protein microarrays) as well as from literature assessment (Shoemaker and Panchenko, 2007; Xenarios *et al.*, 2000, 2001; Salwinski *et al.*, 2004). There are several human databases, Mathivanan and colleagues examined eight of them. In all evaluations, the Human Protein Reference Database (HPRD) by Peri *et al.* (2003) comes out on top after benchmarking human protein interaction data (Mathivanan *et al.*, 2006).

Though the various PPI databases and detection methods have known problems with false positives and false negatives, they still represent a large dataset that is unbiased with respect to the evolutionary processes we are investigating.

2.1.2 Metabolic network data and assessment

Four publicly available datasets with metabolic network data, one for *E. coli*, one for baker's yeast *S. cerevisiae*, and two for human, were used as 'raw material' to construct the metabolic networks described in chapter 3. Below follows a brief description on how these 'raw material' datasets were generated, and what data they contained.

2.1.2.1 Data for the metabolic network of *Escherichia coli*

iJR904. The *E. coli* genome-scale metabolic (GSM) model iJR904 was constructed by Reed *et al.* (2003): 'i' stands for 'in silico model', JR are the initials of the first author scientist constructing the model, and 904 are the number of genes the model contains. Thus iJR904 contains 904 genes and 931 unique reactions. This model is an improvement to an earlier model (Edwards and Palsson, 2000a). It is a model rather than simply metabolism

data because the authors assessed the metabolic capacity of *E. coli* (see section 2.2) as well as gathered the unique reactions and the genes involved in them. Moreover, gene to protein to reaction associations (GPR) was gathered (Figure 2.1 is an example of GPR in the yeast model iND750 described below). For the research described in chapter 3 only the reactions in the iJR904 metabolic model with annotated genes were used, and hence the *E. coli* network data used in chapter 3 contains 904 genes and 873 unique reactions.

2.1.2.2 Data for the metabolic network of *Saccharomyces cerevisiae*

iND750. The yeast metabolic model was called iND750 by the creators in accordance to the name convention for computational models (see subsection 2.1.2.1 and (Reed *et al.*, 2003)), and thus 750 indicates that the model contained 750 genes and the associated transcripts, proteins and reactions (Figure 2.1) (Duarte *et al.*, 2004). The genome-scale network was a built-on the iFF708 model by Förster *et al.* (2003). The assembly was done manually so that genome annotations, gene expression data, enzymatic assays and physiological data were added to the genomic data. Compartmentalisation information was also added so that the model contained information for eight cellular locations: extracellular space, cytosol, mitochondrion, peroxisome, nucleus, endoplasmic reticulum, Golgi apparatus and vacuole. This information mainly came from the two yeast databases. However, when there was insufficient evidence for gene product localisation it was assumed to be in the cytosol. Moreover, unless there was additional evidence, the reactions catalysed by membrane proteins were presumed to take place in the cytosol. Proteins localised in the intercompartmental space of the mitochondria were classified as if they were localised in the cytosol as the two have equivalent metabolites and ion concentrations. According to the same reasoning proteins localised to the cell wall and periplasmic space were classified as if they were localised in the extracellular compartment (Duarte *et al.*, 2004). Finally, some genes have disappeared from the annotations since 2004 when the iND750 yeast dataset was built by Duarte *et al.* (2004). Therefore the dataset used for the yeast metabolic networks described in chapter 3 contained 743 unique genes and 810 unique reactions (Table 3.1).

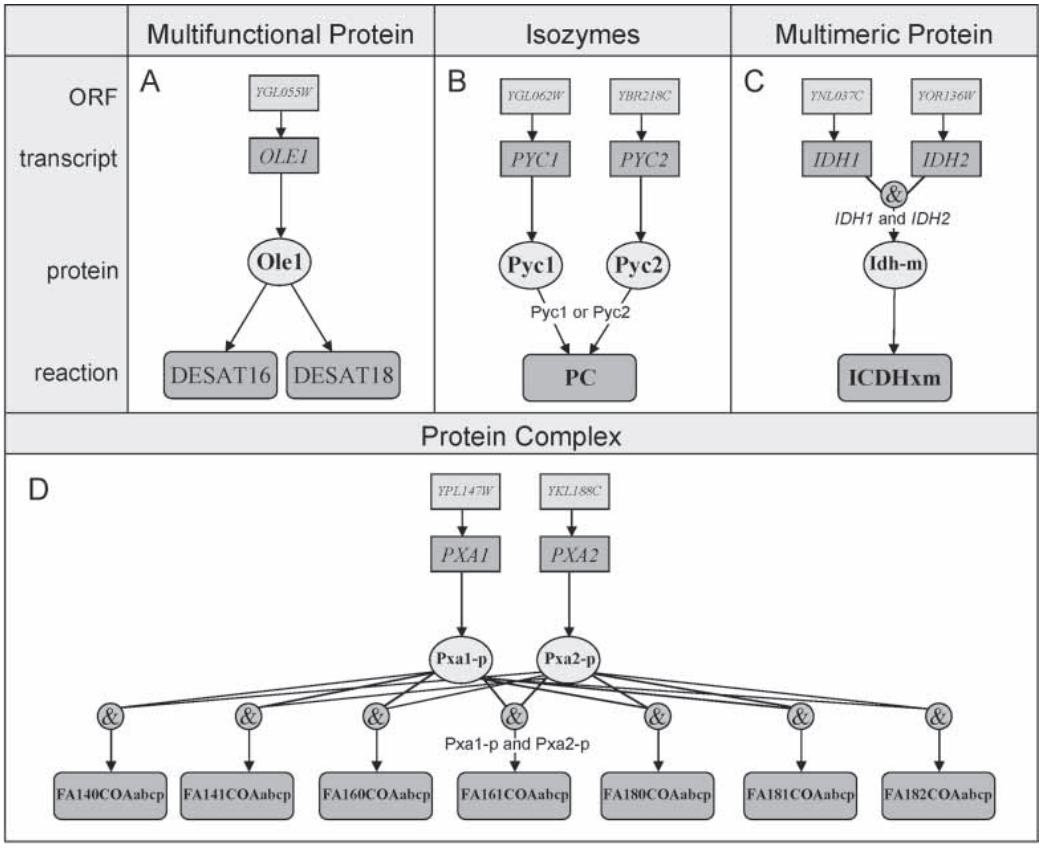


Figure 2.1: Different types of gene-protein-reaction associations in yeast metabolic network model iND750. Different examples of gene-protein-reaction associations in the metabolic network model of iND750. a) A gene encoding for a multifunctional protein, such as Ole1, which can catalyze more than one reaction. b) Isoenzymes Pyc1 protein and Pyc2 protein are two different proteins that can catalyze the same reaction independently from each other. c) Multimeric proteins, such as Idh, are formed from the association of two transcripts. d) Protein complexes, such as Pxa1-p and Pxa2-p in the figure, form a protein complex, and both proteins are required to catalyze the reactions. Figure from Duarte *et al.* (2004).

2.1.2.3 Data for the human metabolic networks

***Homo sapiens* Recon 1.** The first of the two human metabolic network datasets, called *Homo sapiens* Recon 1, was manually reconstructed, bottom-up, that is component-by-component from genomic as well as literature data, and again the proteins taking part in the reactions were localised to one of the eight cellular compartments; cytoplasm, mitochondrion, endoplasmic reticulum, Golgi apparatus, nucleus, lysosome, peroxisome and extracellular compartment. This network consists of 1,496 open reading frames (ORFs), 2,004 proteins, 2,766 metabolites and 3,311 metabolic and transport reactions (Duarte *et al.*, 2007). The data in the *Homo sapiens* Recon 1 dataset was used as a starting point to construct the metabolic network called Hs_D (as described in chapter 3), which consisted of 1475 unique genes (as only genes that could be matched to Ensembl were kept) and 2306 unique reactions respectively as empty reactions (i.e., reactions that were known to exist but which genes were unknown) were removed as they cannot be used for our PPI analyses.

EHMN. The second human metabolic network was also constructed manually from genome annotation, metabolic reaction information and literature (Ma *et al.*, 2007). The network was named Edinburgh Human Metabolic Network (EHMN). To create it Ma *et al.* (2007) generated their initial list of metabolic reactions from the Kyoto Encyclopedia of Genes and Genomes (KEGG) ligand database (Kanehisa Laboratories, 2011), and then all non-human reactions were manually removed. The team compared their network against the computationally generated HumanCyc network (SRI International, 2010), extracted all the enzyme identifier Enzyme Commission (EC) numbers from the latter, and compared them to the reactions in their own data. EC numbers present in the HumanCyc data that could also be verified to be part of human reactions through literature sources, that were not already present in their own data (six entries) were added to EHMN. Their final dataset consisted of 2,322 genes and 2,823 reactions. The EHMN metabolic dataset was used to construct the metabolic network called Hs_M as described in chapter 3. The Hs_M network consisted of 2,269 unique genes (after only genes matched to Ensembl were kept) and 887 unique reactions (again because empty reactions were removed).

Of the two human datasets, *Homo sapiens* Recon 1 (Duarte *et al.*, 2007) and EHMN (Ma *et al.*, 2007) only some (1,069 genes) overlap, and as

described above the techniques for obtaining and constructing the datasets differed. The *Homo sapiens* Recon 1 dataset is compartmentalised while the EHMN dataset is not. Other differences between the two datasets are that the *Homo sapiens* Recon 1 data contains many more reactions than the EHMN data, and also that all reactions in the EHMN dataset with the same enzyme identifier (EC number) were lumped together, whereas in the *Homo sapiens* Recon 1 data, each reaction had separate genes annotated to it. Therefore, for all these reasons, and because the two datasets have been assembled differently, pooling them was not a tempting idea. Moreover, as too few reactions would have remained if only the data overlapping between the two datasets had been used, both datasets were used separately as 'raw material' to create two separate metabolic networks (Table 3.1) for the metabolic research project described in chapter 3. Despite using two different datasets to create two separate human metabolic networks the results obtained from these are consistent with each other.

2.2 Flux Balance Analysis - Assessing robustness and fitness of metabolic networks

Flux balance analysis (FBA) is used to mathematically estimate the optimal flow of metabolites in metabolic networks, and this in turn indicates the fitness of a cell in a particular environment (Wagner, 2007). With FBA one can calculate an allowable solution space, and then calculate the optimal solution of this space, which is the maximum flux through a set of reactions since biomass production, i.e., cell growth under steady-state conditions, is associated with fitness (Orth *et al.*, 2010; Wagner, 2007). Since a metabolic network can be broken down to separate pathways (as described in 1.2.1), and each reaction through two metabolites has a metabolic flux we can get a list of differential equations, which are nontrivial to calculate computationally - especially for a big metabolic network. However, unless a cell is in a new environment, it is growing under steady-state, and therefore modelling can be done in steady-state. That is not to say that metabolites in the network and along a path do not change, although the average concentration of each metabolite is effectively in steady state. Therefore the complex differential equations from a substrate S through several metabolites $m_1, m_2, ..., m_n$ to product P can be expressed as a linear differential equation, which in turn can

be written down as a stoichiometric matrix where each column corresponds to the reactions from S to P and the rows are each metabolite m_1, m_2 to m_n (Wagner, 2007; Orth *et al.*, 2010). Assuming steady-state is the same as saying the average fluxes in the linear differential equation is zero, which in turn is the same as saying that the solution to the matrix is zero, which allows us to rewrite the matrix without the substrate S and product P listed in it. Now, that this is done we can consider it as a black box where we only know what comes out of the box. The differential equation solutions optimise a particular reaction. Usually this is the end product i.e. the biomass production and the cell's growth or "what comes out of the black box". However, "what goes into the box" or the start values of the equation can also be constraints set to assure that the end value obtained from the analysis is a realistic one. However, the optimal solution can be calculated without knowing the fluxes of the internal metabolites.

FBA can be understood metaphorically as illustrated in Figure 2.2, where we have a system with a source of energy at the top (food), and the running man at the bottom representing the biomass production. The first is converted into the second through a series of buckets (representing metabolites) and pipes (representing reactions). Flux in this analogy is given by the diameter of the pipes. We can solve for these diameters when maximizing biomass production given a set of input nutrients. It is also possible to exclude a reaction (conceptually blocking a pipe in Figure 2.2) and estimate the effect on biomass production (e.g. growth). Thus we can calculate the optimal, but not necessarily the actual biomass production of a cell (Orth *et al.*, 2010; Wagner, 2007). However, studies have shown that very often the prediction is close to the true experimental data results (Wagner, 2007; Edwards and Palsson, 2000b; Edwards *et al.*, 2001).

2.3 Permutation tests

To conclude whether the patterns we observe in our biological data are under a functional constraint or simply caused by chance we can perform permutation tests. For example if want to know if a certain type of interactions are more prevalent than we would expect by chance alone we can compare our real network against several (at least 1000, but sometimes 100,000 or even 1 million) randomly generated networks. We then count

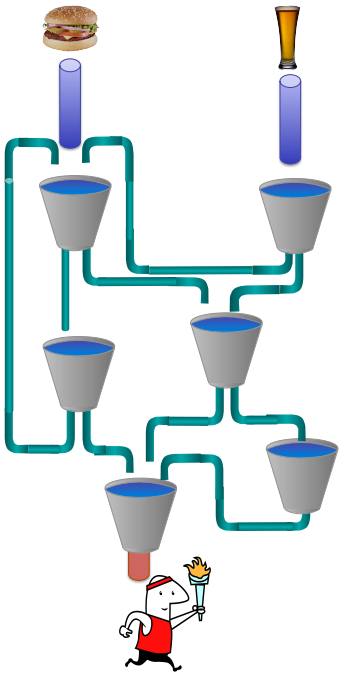


Figure 2.2: Cartoon representing a metabolic network in order to explain FBA. In the cartoon the hamburger and beer at the top represent the sources of energy or substrates S. The buckets are the different metabolites in the network while the pipes are the reactions, and the running man at the bottom represents the cell's or network's biomass production. Figure by Gavin C. Conant.

how many times we observe our interaction of interest in the real network, and how many times we observe these types of interactions in each of the rewired networks. Finally we count how many times for the rewired networks we observe at least as many or more interactions than in the real network. This gives us a p-value indicating if what we observed in the real data is statistically significant or is something that can be observed simply by chance. For example, if we have 1000 randomly rewired networks, and observe more of the interactions of interest in 4 of the rewired networks compared to the real network then the p-value would be $(4/1000) = 0.004$, which is highly significant since the thresholds for significance are usually set at 5% or 1 level respectively (i.e. we accept to fail to reject the null hypothesis at a 5% or 1% level respectively), which is why the p-value must be less than 0.05 or 0.01 respectively for the result to be statistically significant.

2.4 Maximum likelihood

Maximum likelihood estimation (MLE) is an approach for fitting experimental or empirical data to a predescribed statistical model. The model is framed in terms of a number of free (e.g., unknown) parameters, such as the topology of a phylogenetic tree, the probability of flipping a head from an unfair coin or the rate of nucleotide substitution in a DNA sequence. Given a dataset D, for any value of the parameters p one can calculate the likelihood L of observing D given those parameters:

$$L(D|p)$$

According to the maximum likelihood method we should prefer the parameter estimates that give us the highest likelihood of observing the data. While it is conceptually clear that the goal of MLE is to maximize equation 1, in practice the parameter space is very large and hence computationally challenging to scan, especially in cases such as phylogenetic ones where no analytic representation of equation 1 is possible. Thus, instead of analysing the entire parameter space, the computational algorithms used here (e.g. Phylogenetic Analysis by Maximum Likelihood (PAML)) move over the parameter surface only so long as the likelihood is increasing. The danger of this approach is that the algorithm may become trapped in a local, rather than the global, optimum. The best way to imagine this problem is to consider a marble

rolling over a surface. The marble rolls as long as the surface is slightly downhill. However, once it reaches a valley it will come to rest, even if that valley is not the lowest one on the landscape. In order to avoid such local optima, it is a good idea to start the maximum likelihood estimation from several distinct sets of starting parameters. If the same optimal solution is found by several searches starting from differing points in the parameter space, one can have some confidence that a global optima has been found.

MLE is usually not as simple as calculating the likelihood of a single coin toss, but rather that of a string of probabilities, and since each toss is independent the net likelihood is the product of the individual ones (Purcell, 2007; Siddall, 2007). When several small likelihoods are multiplied, the resulting total likelihood can become so small the computer can no longer distinguish it from zero. To avoid this computational limitation, most algorithms store the logarithm of the likelihood, rather than the likelihood itself. Of course, multiplication becomes addition in log space. Thus

$$a = bc$$

becomes

$$log(a) = log(b) + log(c)$$

Thus, when the likelihood approaches zero the log-likelihood will instead be a very large, negative value, which the computer will have no difficulty distinguishing from zero (Purcell, 2007).

As the model with the most parameters will always give the lowest negative log-likelihood compared to nested models with fewer parameters, in model fitting, the log-likelihood of a parameter-rich model is compared to that of a simpler null model using the ratio of the two likelihoods. However, in log space this ratio

$$a = b/c$$

becomes

$$log(a) = log(b) - log(c)$$

If the parameter-rich model does not improve on the null model, twice this difference in log-likelihood is distributed χ^2 with the number of degrees of freedom equal to the number of excess parameter in the alternative model (Purcell, 2007; Siddall, 2007; Yang, 2001), providing a useful test as to whether a proposed model fits a given dataset better than a simpler model.

2.4.1 *branch models* - a codon substitiution model for measuring evolutionary constraint

PAML is a software package implemented by Ziheng Yang (Yang, 1997, 2007) that containst several programs utilising maximum likelihood (ML) for the phylogenetic analyses of DNA and peptide sequences. One of the programs, CODEML was utilised to calculate the selective constraint K_a/K_s (or ω) for different species in an alignment. For this purpose the *branch models* were used as the different parameters of ω vary both along sites and species as the rates are being calculated (Yang and Nielsen, 1998).

Chapter 3

Patterns of indirect protein interactions suggest a spatial organisation to metabolism

The research described in this chapter has been published in *Molecular BioSystems* (Pérez-Bercoff *et al.*, 2011). Paper co-author Gavin C. Conant carried out the flux balance analysis (FBA), pathway analysis and also wrote the network rewiring software used in this study. (For a more detailed background regarding FBA see chapter 2, subsection 2.2, and for the rewiring algorithm see this chapter's subsection 3.2.8.) Analyses in *E. coli* were also carried out by paper co-author Gavin Conant.

3.1 Introduction

Although the analogy between enzymes and machines is a common one, it is rather more rare to note that an obvious extension, that of pathways and assembly lines, is also appropriate. Nonetheless, the idea that enzymes physically associate into complexes that reflect pathway structures is an old one (Yanofsky and Rachmeler, 1958; Mowbray and Moses, 1976; Rais *et al.*, 2000; Gaertner, 1978). One of the primary advantages of these associations is likely to be in facilitating metabolic channelling between sequential enzymes. Metabolic channelling is the collective name for a group of mechanisms that allow metabolites to move between enzymes without

being released into the bulk solvent (Verkman, 2002), reducing the degree to which pathway flux is diffusion-limited (Srere, 2000). One of the earliest examples of channelling was found in the tryptophan biosynthesis pathway, when Yanofsky and Rachmeler were unable to detect one of the pathway's necessary intermediates in cell extracts (Yanofsky and Rachmeler, 1958). It was only after the two catalytic domains in question were crystallized that an enclosed tunnel between the two active sites was identified, accounting for the intermediate's apparent absence (Hyde *et al.*, 1988).

More recently, it has become clear that macromolecular complexes able to facilitate channelling are common and can associate at specific cellular locations, such as the plasma membrane or mitochondrion (Seppet *et al.*, 2001; Brandina *et al.*, 2006; Campanella *et al.*, 2005; Graham *et al.*, 2007). Channelling is also widely phylogenetically distributed, with known examples from yeasts (Brandina *et al.*, 2006; Araiza-Olivera *et al.*, 2010), plants (Graham *et al.*, 2007; Winkel, 2004; Chuong *et al.*, 2004), mammals (Rais *et al.*, 2000; Seppet *et al.*, 2001) and bacteria (Mowbray and Moses, 1976; Hyde *et al.*, 1988). One of the important roles that these associations play is to partially isolate particular metabolites, such as ATP, so that functional units, such as myofibrils and ion pumps, respond to the local and not the global concentration of that metabolite (Saks *et al.*, 2008).

Metabolic channelling is at least in part an emergent property of two other cellular features: cells have very high protein concentrations (macromolecular crowding; (Ellis, 2001)), and and are highly spatially organised (Pielak, 2005; Ovádi and Srere, 1996). Indeed, an elegant experiment has recently been done showing that even when the plasma membrane has been disrupted much of the cell's protein machinery maintains its organisation and function rather than simply diffusing away (Hudder *et al.*, 2003). Cells use several mechanisms to spatially organise metabolism, including protein localization by signalling peptides (Dolezal *et al.*, 2006), and associations between enzymes and membranes (Stuart, 2008). Another potential mechanism for achieving subcellular organisation is protein-protein interactions, or PPIs (Ovádi and Saks, 2004). Here, we explore the role of PPIs in metabolism, dividing the interactions involved into two types: *direct* (protein interaction between proteins that participate in separate enzymes) and *indirect* (interaction between one protein not itself part of an enzyme that binds to proteins in at least two separate enzymes thus connecting these enzymes indirectly) (Chuong *et al.*, 2004). For each interaction type we ask

whether it is more common than expected among related reactions in the metabolic network. Our work follows that of Huthmacher and colleagues, who have argued for a global excess of direct protein-protein interactions (dPPIs) between pairs of enzymes with shared metabolites (Huthmacher *et al.*, 2007), as well as a correlation in enzyme-enzyme distance in the metabolic and protein interaction networks (Huthmacher *et al.*, 2008; Durek and Walther, 2008). Analyzing data from *Escherichia coli*, bakers yeast and humans, we revisit the role of dPPIs in structuring metabolism. We then introduce a role for indirect protein-protein interactions (iPPIs) that link pairs of enzymes through nonenzymatic intermediate proteins.

3.2 Methods

3.2.1 Protein interaction data

Statistics on the protein interaction networks used and their overlap with the metabolic networks are given in Table 3.1. Data on human protein-protein interactions (PPIs) was obtained from the Human Protein Reference Database (HPRD) release 7 (Keshava Prasad *et al.*, 2009; Peri *et al.*, 2003), and matched against Ensembl release 50 (Flicek *et al.*, 2008), (for details see chapter 4, subsection 4.2.1). Yeast and *E. coli* protein interaction data were obtained from the Database of Interacting Proteins (DIP) (Salwinski *et al.*, 2004).

3.2.2 Gene and protein identifiers used

Uniprot identifiers taken from the DIP database (Salwinski *et al.*, 2004) were mapped to gene identifiers from the *E. coli* genome and metabolic network (Reed *et al.*, 2003; Blattner *et al.*, 1997) using custom Perl scripts that queried the European Bioinformatics Institute (EBI) uniprot database (EMBL-EBI European Bioinformatics Institute, 2011). Yeast gene and protein identifiers were presented in standard format both by DIP (Salwinski *et al.*, 2004) and the metabolic network (Duarte *et al.*, 2004), such that no identifier mapping was required. Human protein identifiers from both the protein interaction (Keshava Prasad *et al.*, 2009; Peri *et al.*, 2003) and the metabolic networks (Ma *et al.*, 2007; Duarte *et al.*, 2007) were mapped to Ensembl identifiers as previously described in chapter 4 and (Bekaert and Conant, 2011).

Table 3.1: Statistics regarding the networks employed

Species	$\#Rxns^a$	$\#Genes^b$	Avg. $\#genes/rxn^{cd}$	Avg. $\#rxns/gene^{ed}$	Proteins[f]	Nonself interactions[g]	Prots in met. network[h]	Ints in met. network[i]
E. coli, Reed *et al.* (2003)	873	904	1.7	1.6	1862	7798	294	401
S. cerevisiae, Duarte *et al.* (2004)	810	743	1.7	1.8	4893	17169	269	329
H. sapiens, Ma *et al.* (2007)	887	2269	3.3	1.3	8876	32916	827	1735
H. sapiens, Duarte *et al.* (2007)	2306	1475	1.9	2.9	8876	32916	232	187

[a]Number of reactions in the metabolic network with annotated genes.
[b]Number of genes annotated in the metabolic network.
[c]The average number of genes annotated per reaction.
[d]The simple ratio genes/rxn and rxn/gene does not apply because it could be that every gene participates in, for example, only one reaction (rxns/gene = 1), while every reaction is catalysed by a group of, for example, ten genes (genes/rxn =10). It could be that every gene can do 10 reactions (rxns/gene=10), and each reaction is catalysed by a single gene (genes/rxn=10), or rxns/gene = 10 and genes/rxn = 10, or any combination between these extremes. Therefore these values are not inverses of each other.
[e]The average number of unique reactions a gene participates in.
[f]Number of proteins with annotated protein-protein interactions.
[g]Total number of protein-protein interactions in the network involving distinct proteins (i.e., nonself interactions).
[h]Number of proteins from the protein interaction network also present in the corresponding metabolic network. Used in only the network randomisations when analysing dPPIs.
[i]The number of unique protein-protein interactions that involve proteins that are both present in the metabolic network. Used in only the network randomisations when analysing dPPIs.

3.2.3 Mapping of metabolic and protein interaction networks

All three metabolic networks used are provided with gene identifiers for reactions with known enzymes. For the purposes of this study, any reaction without an annotated gene was omitted from all analyses except the flux balance analysis. For the remaining nodes, the proteins corresponding to said genes were mapped to the proteins identified in the protein interaction network using Perl. Table 3.1 gives the resulting statistics.

3.2.4 Robustness of iPPI enrichment to method of PPI detection

We repeated our analysis of the yeast metabolic network excluding all interactions determined by mass spectrometry and including only interactions obtained from yeast two-hybrid experiments. In neither case did we observe a value of m_S in the randomized networks as large as m in any of the five yeast networks ($p < 0.001$).

3.2.5 Pathway analysis

From the *Saccharomyces* Genome Database (SGD) (Cherry *et al.*, 1998), we downloaded pathway annotations for 531 yeast metabolic genes (a total of 181 pathways). We defined a network such that pairs of genes in the same pathway were joined by an edge. We then counted all instances where two genes in the same pathway shared a protein-protein interaction (i.e., a dPPI) or shared an interaction to a nonmetabolic protein (i.e., an iPPI). Cases where two genes were annotated as being involved in the same reaction by SGD were not counted toward the total of dPPIs and iPPIs. In other words, *intra*-reaction/*intra*-enzyme protein interactions were excluded (see Figure 2.1b and Bat1/Bat2 example in Results 3.3). Using the network randomisation approach described further in subsection 3.2.8 of this chapter, we asked whether the same number of dPPIs and iPPIs could be expected by chance. In neither case did the randomised networks show as many dPPIs or iPPIs as observed in the real network ($p < 0.001$). However, we note that, especially for the dPPIs, the pathways used are somewhat imperfect, since a gene may be annotated into a pathway without a corresponding reaction

annotation (meaning that we could again mistake an intra-enzyme protein interaction for a dPPI).
This section was performed by paper co-author Gavin Conant.

3.2.6 Gene Ontology analyses

Gene Ontology (GO) analyses for eukaryotic iPPIs were conducted as described by Boyle *et al.* (2004) using the Generic Gene Ontology (GO) Term Finder website (Sherlock and Weng, 2004).

3.2.7 Flux balance analysis

Flux balance analysis (FBA) is a common technique for bounding the space of potential metabolic fluxes using reaction stoichiometry and input metabolic constraints (Orth *et al.*, 2010). Briefly, the approach uses linear programming to bound the solution space of a system of homogenous linear equations such that a defined biomass objective function is maximised. Our custom software uses the GNU Linear programming toolkit (`http://www.gnu.org/s/glpk/`, (Makhorin, 2008)) to perform this computation: we note that our results are numerically identical to those produced by publically available FBA packages (e.g., the Systems Biology Research Tool (Wright and Wagner, 2008)). We thus inferred the combination of reaction fluxes that yielded maximal biomass production for the yeast and *E. coli* metabolic networks used here (Reed *et al.*, 2003; Duarte *et al.*, 2004). In yeast, we performed our analysis under six sets of input nutrient conditions, including aerobic growth with glucose, fructose, glycerol, ethanol and glutamine as the primary carbon source and anaerobic growth with glucose as the primary carbon source. In *E. coli* we used growth on glucose under anaerobic and aerobic conditions. For each condition we normalized all fluxes to the overall biomass flux under that condition. Finally, for each set of inputs, we also computed the change in biomass flux that resulted from individually constraining the flux through each reaction to half of its maximal value (simulating a computational knockdown since a knockout would give biomass zero, which is a non-trivial solution to the FBA problem). We estimated knockdown fitness f as the ratio of the constrained biomass flux to the unconstrained flux.
This section was performed by paper co-author Gavin Conant.

3.2.8 Comparison of metabolic and protein interaction networks by randomisation

In order to properly assess whether there is an excess of dPPIs or iPPIs, there are several potential confounding factors that any randomisation approach needs to avoid:

1. Different proteins can have very different numbers of protein interactions, and a randomisation scheme that fails to preserve this skew in the number of interaction partners might have unpredictable effects on the assessment of dPPI and iPPI excess.

2. One metabolic reaction might involve enzymes coded for by multiple and different genes, either as isoenzymes or multi-enzyme complexes.

3. One enzyme may catalyze more than one reaction.

4. Points 2 and 3 may be confounded, such that one multi-enzyme or isoenzyme complex can catalyze more than one reaction (*cf.*, the Bat1 and Bat2 enzymes in *S. cerevisiae*) (Voet and Voet, 2004).

5. It is desirable to preserve the interaction degree of the enzymes of the metabolic network, something that one obvious randomisation scheme, that of randomising the identities of the proteins in the protein interaction network relative to the metabolic network, fails to do.

We suggest that the randomisation approaches used by Huthmacher *et al.* (2007), do not preserve interaction degree in the protein interaction network, nor do they ensure that the isoenzyme/enzyme complex structure in the metabolic network is retained. We instead sought a randomisation approach that avoids all five of the issues listed in order to provide a clear assessment of dPPI and iPPI enrichment.

To do so, it is helpful first to more precisely define the question being asked. First, take the existing metabolic network M and an ensemble of protein interaction networks P_r where, across the entire ensemble of $P_r s$, each node has the same degree as does the real protein interaction network P. We define V^M as the edge set of M and V^P as the edge set of P such that x is the number of shared nodes between M and P. The expected number of dPPIs between M and a random network P_r is given by:

$$\sum_{V^M} P(V_{i,j}^{P_r} \in V^{P_r} | V_{i,j}^{M} \in V^{M}) \tag{3.1}$$

where $i < x$ and $j < x$. In other words, given each edge i, j in the metabolic network ($V_{i,j}^{M}$), we ask what the probability is that there is an equivalent edge in the ensemble of random protein interaction networks ($V_{i,j}^{P_r}$). When we take the sum over the entire metabolic network of these probabilities, we get the expected number of dPPIs. We can calculate the equivalent number for the real network P by replacing the probabilities in (equation 3.1) with an indicator variable that takes on value 1 if a given edge exists and 0 otherwise.

However, for the random networks it is clear that:

$$P(V_{i,j}^{P_r} \in V^{P_r} | V_{i,j}^{M} \in V^{M}) = P(V_{i,j}^{P_r} \in V^{P_r}) \tag{3.2}$$

since the P_r networks do not depend on M. Thus, when we compute (equation 3.1) for all P_r and compare that distribution to the real number of dPPIs, if we find that the real number of dPPIs falls within the range seen in the $P_r s$, we cannot reject the conclusion that (equation 3.2) is true of the real network as well. In other words, in that case our null hypothesis (that there are no more shared edges (dPPIs) than would be expected given the structures of M and P) cannot be rejected.

Conant and Wolfe (2006) have previously described a network rewiring procedure for creating the $P_r s$ that *exactly* maintains the interaction degrees of all nodes: only the interaction identities are changed. This algorithm starts with an existing protein interaction network with all self-interactions removed. Each edge in the network is then broken, leaving a "stub" for each of the two nodes it connects (thus there are $2V^P$ such stubs, 2 per edge in the original network). To create a randomised network, we use the following algorithm:

1. Randomly select two stubs, A and B, to join.
2. If stub A is from the same protein as is stub B, goto 5
3. If we have already created an interaction that joins stubs A and B (A/B) in a previous step, goto 5
4. Goto 8

5. If A and B are the first stubs chosen, goto 1
6. Else, break a random existing interaction C/D created at a previous step 8
7. For the pair A/C and for the pair B/D, goto 2
8. Create interaction A/B
9. Goto 1

The result of this algorithm is a randomised network where each node has exactly the same number of edges as it did in the original network: steps 2 and 3 of our procedure prevent the randomisation procedure from becoming trapped with a partial network where all remaining potential interactions are forbidden. In the worst case, the randomisation procedure will return the original network as the only possible network that allows the observed set of node degrees (e.g., in the case of a fully connected network). Thus, our approach is conservative in the sense that a failure to randomise the PPI network would result in our failure to reject the null hypothesis of no more dPPIs or iPPIs than expected by chance.

Using this protocol to analyse the ten yeast and *E. coli* networks shown in Table 3.2 leads us to infer an excess of dPPIs ($p < 0.05$), similar to the results of Huthmacher *et al.* (2007, 2008). However, this approach has a flaw. The protein interaction network contains many non-metabolic proteins. By randomising the entire network, we implicitly assume that the random networks will have just as many protein interactions between pairs of metabolic proteins as did the original network. If however the fact that two proteins are both enzymes increases the probability that they will share a protein interaction (violating equation 3.2 above), the randomised networks will have fewer interactions between pairs of metabolic proteins than did the original network, as is in fact observed for all three taxa ($p < 0.01$). Since dPPIs can only exist for proteins that are both members of the metabolic network, any tendency of the randomised networks to have fewer such interactions will be mistaken for evidence of more dPPIs in the real network. This problem is easily overcome by limiting the randomisations to interactions where both proteins are members of the metabolic network. Importantly, if there were no bias in interactions for metabolic proteins, this subset-based approach would be equivalent to randomising the entire

network.
This section was performed and written by paper co-author Gavin Conant.

3.3 Results

To study metabolic channelling in *E. coli*, baker's yeast and humans, we employed published metabolic and protein interaction networks (Table 3.1) (Zhu *et al.*, 2007; Reed *et al.*, 2003; Duarte *et al.*, 2004), including two different human metabolic networks, those of Ma *et al.* (2007), (Hs_M networks) and Duarte *et al.* (2007) (Hs_D networks). We studied reaction-centered metabolic networks, consisting of nodes that are metabolic reactions (Wagner and Fell, 2001). Two reactions are connected by an edge if they share a metabolite (Figure 3.1). Because there are a handful of ubiquitous metabolites such as hydrogen ions and water, we first counted the number of reactions each metabolite participates in (n). We then excluded the most highly connected metabolites from the network at five stringencies: $n \geqslant 10$, 15, 25, 50, and 100. We have named the networks based on this exclusion stringency. Thus the Hs_M_25 network is derived from the human network of Ma *et al.* (2007), with metabolites participating in 25 or more reactions excluded. Table 3.2 gives selected network statistics for the twenty networks examined. Except where noted, all of our conclusions hold across all five exclusion stringencies and all three taxa.

Any two proteins that catalyse distinct reactions, but share both a metabolite and a protein-protein interaction, are defined to possess a direct protein-protein interaction (dPPI). Huthmacher *et al.* (2007) detected an excess of dPPIs in yeast and *E. coli* compared to the number expected when randomizing either the metabolic network or both the metabolic and protein interaction networks. The authors randomised the metabolic network by exchanging reaction identities across the network and the protein interaction network by creating new network edges under an assumed probability distribution. Whether an excess of dPPIs was observed depended on the randomisation approach used (Huthmacher *et al.*, 2007).

While network randomisation appears to be the most appropriate way to assess whether there are more dPPIs than would be expected by chance, the exact method of constructing the required random networks involves controlling for a number of effects. We describe the issues associated with

Table 3.2: Metabolic network structure and dPPI prevalence. Selected statistics regarding the twenty metabolic networks used are given. dPPIs are defined as shared protein interactions between two enzymes that also share a metabolite.

Network	Edges[a]	Path length[b]	Clust. Coeff.[c]	l[d]	Average l_s[e]	Maximum l_s[f]	P[g]
Ec_10	2225	7.7	0.65	13	4.4	16	0.003
Ec_15	3168	5.3	0.63	15	6.0	16	0.002
Ec_25	5006	4.3	0.64	20	13.5	27	**0.052**
Ec_50	8355	3.6	0.66	25	21.4	36	**0.25**
Ec_100	15807	2.9	0.69	51	41.7	61	**0.08**
Sc_10	2004	7.1	0.66	6	4.2	11	**0.23**
Sc_15	2959	5.5	0.66	7	6.6	16	**0.49**
Sc_25	4375	5.0	0.67	7	7.6	18	**>0.5**
Sc_50	8749	3.9	0.70	12	14.1	30	**>0.5**
Sc_100	19196	2.8	0.73	33	29.3	47	**0.27**
Hs_D_10	4292	12.9	0.71	22	2.9	9	< 0.001
Hs_D_15	6555	9.3	0.69	27	3.2	10	< 0.001
Hs_D_25	11018	8.0	0.68	30	4.7	14	< 0.001
Hs_D_50	22844	5.4	0.74	34	5.9	20	< 0.001
Hs_D_100	44642	4.2	0.76	38	8.6	20	< 0.001
Hs_M_10	2726	4.5	0.71	174	83.0	104	< 0.001
Hs_M_15	3797	3.9	0.70	431	238.3	273	< 0.001
Hs_M_25	5139	3.5	0.69	435	241.5	292	< 0.001
Hs_M_50	7962	3.2	0.70	448	244.4	281	< 0.001
Hs_M_100	16213	2.7	0.73	459	250.8	286	< 0.001

[a]Number of edges in the undirected metabolic network considering only reactions with annotated genes.
[b]Average minimum path length: the average over all nodes of the average for each node of the minimum number of edges needing to be traversed to reach any other node, calculated with Dijkstras algorithm (Yoon *et al.*, 2006).
[c]Clustering coefficient of each network (Watts and Strogatz, 1998).
[d]Number of unique dPPIs, i.e., unique interactions between two enzymes, in the actual metabolic network.
[e]Average number of unique dPPIs seen in 1000 randomisations of the protein interaction network.
[f]Maximal number of unique dPPIs observed in any of the 1000 randomisations.
[g]P-value for the test of the hypothesis that there are no more dPPIs than would be expected given the respective structures of the two networks. Non-significant tests are shown in bold.

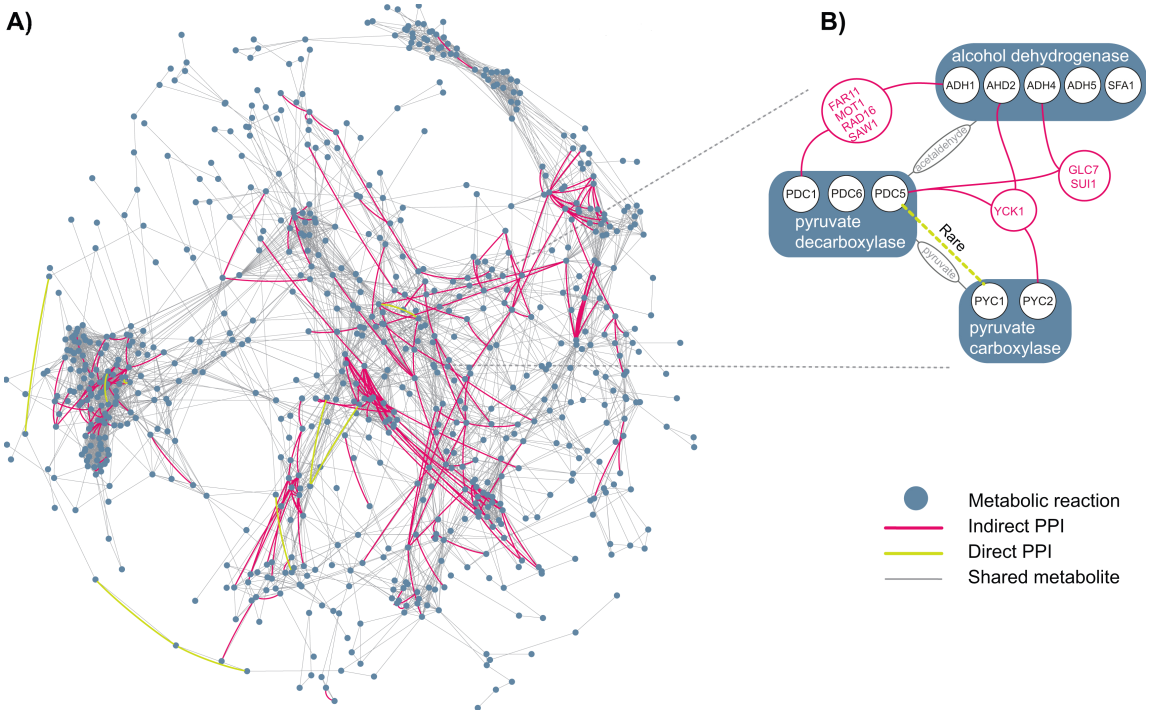


Figure 3.1: Indirect protein-protein interactions provide structure to the yeast metabolic network. a) The central component of the yeast metabolic network described by Duarte *et al.* (2004) is shown. Metabolites involved in more than 25 reactions are omitted (Sc_25). Nodes (blue) are reactions, joined by three types of edges. Thin grey lines indicate two reactions that share a common metabolite or metabolites. Light green lines indicate a direct PPI (dPPI) between at least two of the constituent enzyme proteins of a pair of reactions. Pink lines represent an indirect PPI (iPPI; meaning that there is a third, nonmetabolic, protein that both enzymes interact with). b) An enlarged view of a small part of the network, consisting of three enzymes involved in central carbon metabolism. Pyruvate decarboxylase and alcohol dehydrogenase are sequential steps in ethanol fermentation, and isozymes for these two reactions are joined by a total of seven indirect PPIs. Pyruvate decarboxylase and pyruvate carboxylase represent branch points in pyruvate metabolism. Note that the Yck1 protein actually joins all three reactions, although the link joining pyruvate carboxylase and alcohol dehydrogenase is not shown in A since these two reactions do not share a common metabolite. Despite the kinase (Yck1) and phosphatase (Glc7) example here, which would imply transient protein-protein interactions as phosphate groups are added to and removed from substrates respectively, it is quite possible these genes have multiple functions in accord with type I pleitropy (Wagner and Zhang, 2011). They may therefore also act as mediators to enable enzymes to be brought together into close proximity in metabolic pathways. Moreover, neither kinases nor phosphatases are over-represented amongst the iPPIs.

previous analyses and the precise approach used in subsection 3.2.8 of this chapter. Briefly, we kept the metabolic network unaltered and compared it to 1000 randomised protein-interaction networks in which the number of edges for each protein was kept constant but the identity of those edges was randomised. We then calculated the number of dPPIs found when comparing the original metabolic network to each randomised PPI network (subsection 3.2.8, chapter 3). One important refinement that we introduced to the randomisation was the exclusion of intra-reaction protein interactions (Duarte *et al.*, 2004; Gancedo and Flores, 2008). For example, in yeast, the enzymes Bat1 and Bat2 are encoded by genes produced by the yeast genome duplication (Byrne and Wolfe, 2005). These two proteins both catalyse the same two distinct reactions, namely the final steps of isoleucine and valine production (Voet and Voet, 2004). They also interact with each other according to the Database of Interacting Proteins (DIP) (Salwinski *et al.*, 2004). A naïve approach to identifying dPPIs would include this interaction, an inclusion that we would dispute, since the more parsimonious explanation is that the two proteins have descended from a single self-interacting protein through the genome duplication. To avoid these types of false signals, we therefore did not allow PPIs joining proteins from the same reaction to form part of a dPPI.

In *E. coli*, when we exclude metabolites involved in at least 10 or at least 15 reactions, we find an excess of dPPIs (Ec_10, Ec_15; $p \leqslant 0.003$; Table 3.2), but this enrichment is not observed for the three larger cutoffs (Ec_25, Ec_50, Ec_100; $p > 0.05$). Yeast does not have significantly more dPPIs than would be expected in any of the five networks shown in Table 3.2 ($p > 0.05$). Finally, we observe a significant overabundance of dPPIs for all ten human networks in Table 3.2 ($p < 0.001$). Our conclusion of no dPPI enrichment in yeast may differ from that of Huthmacher et al., because these authors do not appear to have excluded intra-enzyme interactions shared across multiple reactions (i.e., as in the Bat1/Bat2 example cited). We cannot currently assess why these three organisms show differing evidence of dPPIs. However, we do note that, of the three species, the set of protein interactions inferred for yeast may be the least biased by investigator choices, since many of these interactions were identified with high-throughput approaches. Since protein interactions from *E. coli* and humans are taken from literature studies, a tendency among researchers to look for protein interactions among members of the same pathways could potentially give

rise to a spurious indication of a general abundance of dPPIs. Given the ambiguous evidence for dPPIs and the relatively small number of them seen (Table 3.2), we have also explored the potential for more complex protein interactions that may facilitate channelling.

3.3.1 Indirect interactions between neighboring enzymes

Proteins that are not themselves part of the metabolic network might also act as mediators between enzymes. Durek and Walther (2008) have shown that non-metabolic proteins contribute to bringing reactions with shared metabolites into closer proximity in the protein interaction network. We extend this result by searching for an overabundance of indirect PPIs (iPPIs), where a pair of PPIs to a third non-metabolic protein join two enzymes sharing a metabolite (Figure 3.1b).

For each metabolic network, we determined m: the number of unique interactions between two enzymes and their mediator protein (the number of unique iPPIs). We then created 1000 randomised protein interaction networks as described above and calculated m_s: the number of unique iPPIs in each randomised network. We then compared the distribution of m_s to the value of m. For example, in the Sc_50 network, $m = 504$, while the largest value of m_s observed was 253. Indeed, for none of the fifteen eukaryotic networks did we find any values of m_s as large as the respective values of m ($p < 0.001$, Table 3.3). Moreover, because the yeast protein interaction data was derived from different types of high-throughput experiments, we can show that our results in yeast are robust to the method of protein interaction detection (Methods 3.2). In *E. coli* the situation is slightly more complex, as significant iPPI enrichment was observed only for the two most stringent thresholds of currency metabolite (for definition see Methods 1.2.1 and Figure 1.6) exclusion ($p \leqslant 0.001$; Ec_10 and Ec_15, Table 3.3). Given that an iPPI implies physical proximity between its three members, we conclude that iPPI enrichment will drive physical associations between related enzymes in a nonrandom fashion.

If the functional role of iPPIs is to physically associate metabolic enzymes, we would expect that their distribution in a network would be non-random. A consequence of our definition of iPPIs is that they represent a subset of the edges in their respective metabolic networks. We thus asked if the

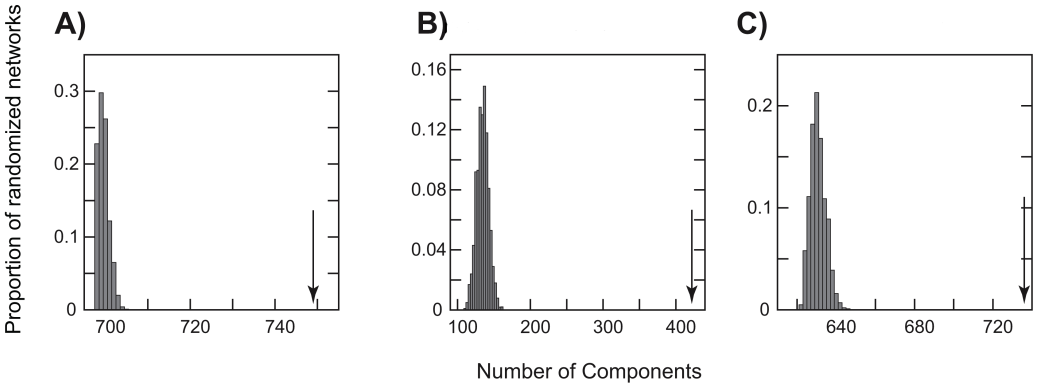


Figure 3.2: Indirect protein-protein interactions (iPPIs) produce more distinct metabolic subnetworks than would be expected by chance. On the x-axis is the number of components observed when a number of metabolic edges equal to the number of iPPIs is sampled at random from the metabolic network. The y-axis gives the proportion of simulations having that number of components (1000 simulations). The arrows in each panel give the number of network components produced in each metabolic network when only the metabolic edges corresponding to iPPIs are retained (749, 422 and 736, respectively). For each species, we selected the network with the most liberal threshold that still showed an excess of iPPIs. a) The *E. coli* Ec_15 network. b) The yeast Sc_100 network. c) The human Hs_M_100 network.

metabolic edges possessing iPPIs were unusual by examining the number of network components defined by iPPI-possessing edges. We compared this number of components to the number of components observed when the same number of metabolic edges were selected at random. (Note that, as with our previous analysis, this randomisation was performed only on the set of metabolic nodes with annotated genes, Methods 3.2). In all 17 real networks for which we identified iPPI enrichment the number of components seen was greater than that seen for any randomly selected sets of edges ($p < 0.001$, Figure 3.2). Further analyses counting the number of isolated nodes in the networks suggested that the primary driver of this phenomenon is that only a limited number of reactions possess iPPIs, meaning that the iPPI-induced metabolic networks have many more isolated nodes than do random ones ($p < 0.001$).

Table 3.3: Real metabolic networks have many more iPPIs than do randomised networks. iPPIs are defined as a shared protein interaction between two enzymes that share a metabolite and a third mediator protein. The number of such interactions in real metabolic networks is much larger than can be explained by chance ($m > m_s$).

Network	m^a	Average $m_s{}^b$	Maximum $m_s{}^c$	P^d
Ec_10	157	109.4	156	< 0.001
Ec_15	208	156.1	216	0.001
Ec_25	394	383.3	476	**0.35**
Ec_50	486	591.7	708	**>0.5**
Ec_100	870	1236.6	1464	**>0.5**
Sc_10	147	57.5	88	< 0.001
Sc_15	206	84.0	119	< 0.001
Sc_25	259	99.8	141	< 0.001
Sc_50	504	187.5	253	< 0.001
Sc_100	2192	466.1	605	< 0.001
Hs_D_10	352	62.9	96	< 0.001
Hs_D_15	369	69.5	113	< 0.001
Hs_D_25	386	73.2	109	< 0.001
Hs_D_50	399	78.6	120	< 0.001
Hs_D_100	444	90.5	126	< 0.001
Hs_M_10	2776	879.5	1050	< 0.001
Hs_M_15	6381	3277.8	3787	< 0.001
Hs_M_25	6411	3315.2	3706	< 0.001
Hs_M_50	6464	3356.4	3906	< 0.001
Hs_M_100	6564	3464.8	3980	< 0.001

[a]Number of unique iPPIs, i.e., unique interactions between two enzymes and a mediator protein, in the actual metabolic network.
[b]Average number of unique iPPIs seen in 1000 randomisations of the protein interaction network.
[c]Maximal number of unique iPPIs observed in any of the 1000 randomisations.
[d]P-value for the test of the hypothesis that there are no more iPPIs than would be expected given the respective structures of the two networks. Non-significant tests are shown in bold.

3.3.2 Analysis of yeast pathways for dPPIs and iPPIs

To explore whether the patterns we observed also existed at the level of metabolic pathways, we obtained a set of 183 metabolic pathways from the *Saccharomyces* Genome Database (Cherry *et al.*, 1998). Using a similar network approach, we asked whether there was an excess of dPPIs and iPPIs between members of the same metabolic pathway (Methods 3.2). For both dPPIs and iPPIs, we observe such an excess, with the number of dPPIs or iPPIs in the actual network being at least 2.5 times greater than the largest number seen in the randomised networks ($p < 0.001$). Although the annotation status of these pathways introduces some error into this analysis (Methods 3.2), it is encouraging that the general trends are similar to those seen when analysing metabolic networks.[a]

3.3.3 Functional annotation of mediator proteins

To understand the biological role of the mediator proteins (pink gene names in Figure 3.1b), we searched for Gene Ontology (GO) terms enriched among the mediator proteins (Boyle *et al.*, 2004). To do so, we used the largest eukaryotic networks because the other networks contain subsets of the mediator genes in these three (Sc_100, Hs_D_100 and Hs_M_100 respectively). In all three cases, the most significantly enriched molecular function term was "protein binding" (Bonferroni-corrected $p \leqslant 0.009$ across three hypothesis tests, and fold enrichment 1.95, 2.65 and 2.66 for the three networks respectively). In yeast, the most enriched biological process term was "localisation" (Bonferroni-corrected $p < 10^{-11}$ and fold enrichment 2.29), while in the Hs_D_100 and Hs_M_100 networks, the most enriched terms were "signalling" and "regulation of cellular process" respectively (Bonferroni-corrected $p < 10^{-34}$ and $p < 10^{-101}$ and fold enrichments 3.87 and 2.51 respectively). However, "localisation" was also significantly enriched in these datasets (Bonferroni-corrected $p < 10^{-17}$ with fold enrichment 3.44 for Hs_D_100 and 2.27 for Hs_M_100).

Next we asked whether the yeast mediator proteins were more likely to be

[a]This analysis was requested by the reviewers of the paper manuscript, and while the reaction data is similar to the metabolism reaction data in previous analyses, the connectivity information is defined differently (two reactions are connected if they are in the same pathway).

essential than the average gene. To do so, we asked what proportion of the mediator protein genes from the largest yeast network, Sc_100, were essential according to the Munich Information Center for Protein Sequences (MIPS) (Mewes *et al.*, 1999) Of these genes, 26% were deemed essential, compared to 18% of the genome at large and of genes with at least one protein interaction, both significant differences (χ^2 tests, $p < 0.01$).

3.3.4 Association of iPPI presence and high metabolic flux

We hypothesised that the reactions possessing iPPIs might be biased toward those carrying high flux, accounting for their non-random distribution in the network. Durek and Walther have shown that enzymes with high protein interaction degree show high flux (Durek and Walther, 2008): we asked if this result could be partly attributed to higher flux among enzymes with iPPIs. We thus computed the maximal flux through each *E. coli* and yeast reaction across several growth conditions using flux balance analysis (FBA; Methods 3.2.7) (Orth *et al.*, 2010). That is, for each of the two species all reactions with an iPPI were put into one bin, while reactions lacking iPPIs were put into a second bin. The fluxes were then compared between the bins so that for each specific reaction the reaction/flux was counted only once. This avoids the issue of having a few pathways enriched for iPPIs inflating the results. For both *E. coli* networks with iPPI enrichment (Ec_10 and Ec_15), the mean flux through reactions with an iPPI was at least 1.5-fold greater than that through reactions without one (Wilcoxon rank sum tests[b], $p < 0.04$). In yeast, no network showed a significant difference between reactions with and without iPPIs. However, many of the reactions in these networks showed no flux in our FBA analyses, either because the set of environmental conditions where that reaction occurs was not tested or because the biomass reaction used did not include that reaction's contribution (e.g., for micronutrients) (Förster *et al.*, 2003). The lack of these reactions reduced the sample size and hence power of our flux analysis. To partly compensate for this effect, we introduced a second analysis where we asked

[b]Reactions were broken down into two groups; those that had iPPIs, and those that lacked iPPIs. However, what was compared was the flux distributions using a Wilcoxon test i.e., the interactions are not compared, the fluxes are.

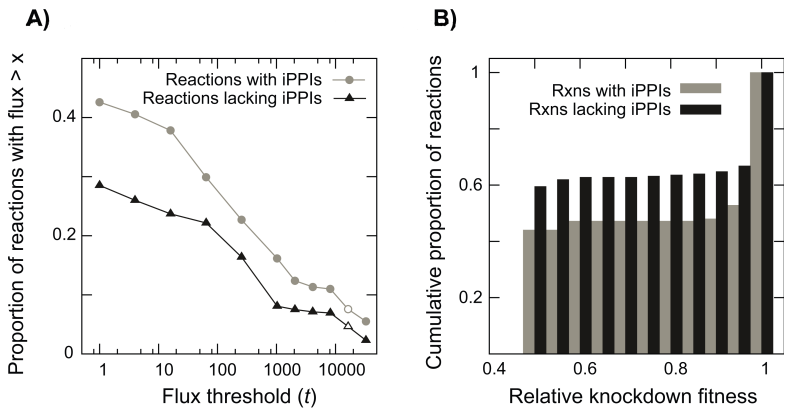


Figure 3.3: Reactions with iPPIs have higher flux and weaker knockdown effects. a) Reactions with iPPIs represent a greater proportion of high flux reactions than expected. On the x-axis is the threshold used to define high flux (t). On the y-axis is the proportion of reactions either having (grey) or lacking (black) iPPIs that have flux of at least t. Open symbols indicate comparisons where the difference in proportion between the reactions with and without iPPIs is not statistically significant ($p > 0.05$). b) Reactions with iPPIs have smaller knock-down effects than other reactions. On the x-axis is our measure of relative knockdown fitness: the ratio of biomass flux when a reaction is limited to one half its optimal flux to "wild-type" flux. On the y-axis is the cumulative proportion of reactions with relative knockdown fitness x for reactions that have (grey) or lack (black) iPPIs. Both panels show results from Sc_50, but other yeast networks are qualitatively similar.

whether the proportion of reactions having both an iPPI and high flux was greater than would be expected. We defined "high flux" according to a sliding threshold t, measured relative to the flux through the biomass reactions (Methods 3.2). Thus, when $t = 1$, a high flux reaction is a reaction with at least the flux of the biomass reaction. For a wide range of values of t ($0 \leqslant t \leqslant 128$), all five yeast networks show a greater proportion of reactions having both high flux and an iPPI than would be expected ($p < 0.05$; Figure 3.3a). As Figure 3.3a indicates, even for values of t where no significant excess of high-flux reactions with iPPIs is observed (due to the small number of total reactions with flux $> t$), the proportion of reactions having iPPIs that are high flux is always greater than that of reactions without iPPIs. Because of the clear difference in flux between reactions with and without iPPIs in *E. coli*, no equivalent test was carried out in that species.

We also computationally limited the flux through each reaction in the

two metabolic networks and calculated the resulting change in biomass flux. We defined the fitness effect of this knockdown as the ratio of the knockdown to original biomass flux (Methods 3.2). For Sc_15, Sc_25 and Sc_50, average knockdown fitness was *higher* for those reactions associated with an iPPI ($p < 0.05$; Figure 3.3b). No significant difference was observed in Sc_10, Sc_100, Ec_10 or Ec_15.

While it would be interesting to examine the metabolic roles of the iPPIs in humans, the goal of maximal biomass production used in yeast and *E. coli* is not an evolutionarily meaningful objective function in humans, since rapid cell division does not generally confer a fitness benefit in multicellular species. Shlomi and colleagues have instead provided some more computationally challenging methods for analysing the human metabolic network which may provide interesting insights into this problem in the future (Shlomi *et al.*, 2008; Jerby *et al.*, 2010).

3.4 Discussion

We offer several new observations supporting the contention that protein interactions play a global role in spatially structuring metabolism (Huthmacher *et al.*, 2007, 2008; Durek and Walther, 2008). First, were iPPIs simply a result of noise in the protein interaction data, it would be difficult to understand our observations that iPPIs involve a distinct subset of the reactions in the metabolic network and involve mediators that are more likely to be essential. Likewise, the association between high flux reactions and the presence of iPPIs suggests that the physical association of sequential enzymes may be reserved for reactions that require elevated flux. It is certainly the case that a number of existing examples of metabolic channelling involve such reactions (Mowbray and Moses, 1976; Rais *et al.*, 2000; Seppet *et al.*, 2001; Brandina *et al.*, 2006).

More generally, the advantages of spatially structuring cellular functions extend beyond simple increases in metabolic efficiency. For instance, micro-compartmentalisation may help to sequester metabolic intermediates with undesirable side effects (Brandina *et al.*, 2006; Ovádi *et al.*, 2004). Likewise, in signal transduction, spatial co-localisation increases efficiency and minimises cross-talk (Mochly-Rosen, 1995; Pawson and Scott, 1997). In fact, there are scaffolding proteins responsible for maintaining spatial co-

localisation (Pawson and Scott, 1997) and it is thus intriguing that a number of the mediator proteins here also have signalling annotations. We speculate that similar scaffolding structures may be at work in metabolism. Given these general Gene Ontology results, however, it is not yet clear how functionally specialised a given mediator protein will be.

One obvious question is why we occasionally obtain differing results depending on the stringency of currency metabolite removal. Thus, only Ec_10 and Ec_15 show significant iPPI enrichment in *E. coli*. We note that excluding more potential currency metabolites should *reduce* the number of spurious associations between reactions because only reactions with unique shared metabolites will be connected. Networks with more stringent currency metabolite removal (Ec_10, Sc_10 etc.) should therefore give better sensitivity in detecting dPPIs and iPPIs (at a cost of potentially excluding real iPPIs and a concomitant loss of statistical power). Importantly, unlike eukaryotic organisms such as yeast and humans, *E. coli* lacks mitochondria, nuclei and other membrane-bound compartments, facts reflected in the respective metabolic networks. We hypothesise that currency metabolite removal is therefore more important for this organism, since spurious associations between reactions that would be removed by compartmentalisation in eukaryotes (where, for instance, ATP in the cytoplasm is distinguished from that in the mitochondria) are not eliminated in the prokaryotic network. A similar tradeoff between statistical power and sensitivity probably explains why significant associations between iPPIs and knockdown effects are seen only for Sc_15, Sc_25, and Sc_50.

Our observations may also provide insights on other biological phenomena. For instance, the relative scaling between metabolic rate and body size is surprisingly similar across groups of related organisms and organ systems; a result that appears to be due to the fractal branching patterns of the transport networks involved (West *et al.*, 1999). It has recently been suggested that this scaling might also extend to the subcellular level (West *et al.*, 2002). If so, similar hierarchical structures within the cell, such as those implied above (Verkman, 2002; Pielak, 2005; Ovádi *et al.*, 2004) might be one source of that scaling. Likewise, the associations of iPPIs with high flux reactions also reaffirms the modular nature of the metabolic network (Ravasz *et al.*, 2002) with groups of enzymes working in functional isolation from each other. It is of course terribly tempting to employ the network structures of Figure 3.1 to infer this set of metabolic assembly lines.

However, such an effort is likely premature, given the incomplete nature of the protein interaction data (Gavin *et al.*, 2006; Ito *et al.*, 2001). Nonetheless, there is certainly more information embedded in the protein interaction and metabolic networks than has been analysed here. We suggest that the addition of protein abundance data might help infer the stoichiometry of the metabolic clusters relative to other parts of the cellular substructure, and we speculate that this organisation might show fractal patterns (Aon *et al.*, 2004).

3.5 Conclusions

Our analyses suggest that one role for the cell's protein interactions is to define the appropriate spatial configuration of metabolism. That this process involves indirect interactions suggests that such structures might have a modular nature (i.e., one localisation protein can spatially co-localise a number of enzymes with similar roles).

Chapter 4

Duplicability of genes encoding self-interacting proteins in the human genome

The research described in this chapter has previously been published in *BMC Evolutionary Biology* (Pérez-Bercoff *et al.*, 2010). The Hokamp paralogon dataset was specifically made by Karsten Hokamp for this study, and served as the starting point from which the Hokamp WGD genes were later generated. The Nakatani paralogon dataset was obtained by personal communication (between Yoichiro Nakatani and Takashi Makino in the laboratory), and the dataset served as the starting point to generate the Nakatani WGD genes for this study. Paper co-author Takashi Makino's Perl scripts were used to perform the GO analyses.

4.1 Introduction

Proteins have an impact on the cell through interactions with other components of the system. One type of interaction, PPI, has received much attention in the literature because of the possibilities of genome-wide surveys, such as two-hybrid screens, and the tractability of analysis. In particular, the evolution of PPIs and how this relates to other aspects of molecular evolution is interesting.

One special category of PPI is the interaction between identical copies of

a protein produced from the same gene (self-interaction) forming homomers. These comprise a significant fraction of the PIN due to genetic factors: the interacting partners are translated from the same mRNA, and so are *ipso facto* co-regulated and co-localised in the cell; as well as biochemical factors: identical proteins are expected to have high affinity for each other (Lukatsky *et al.*, 2007).

Gene duplication can act to shape the PIN because although an identical protein copy produced from a duplicate gene will perform the same interactions as the original, over time the protein sequences and the interactions they participate in will diverge (Levy and Pereira-Leal, 2008). In particular, the ancestral numbers of interactions may influence the dynamics of gain and loss of protein interactions after gene duplication (Zhang *et al.*, 2005). It has previously been shown that gene duplicability can be influenced by factors such as dosage-balance constraints (Papp *et al.*, 2003a), connectivity in interaction networks (Liang and Li, 2007) and function (Rambaldi *et al.*, 2008). The duplicability of genes also differs between SSD and WGD (Gout *et al.*, 2009; Makino *et al.*, 2009).

Several recent studies have examined the duplication of genes whose protein product can interact with a copy of itself (for simplicity we call these "self-interacting genes") (Levy and Pereira-Leal, 2008; Ispolatov *et al.*, 2005; Pereira-Leal *et al.*, 2007; Presser *et al.*, 2008). Pereira-Leal and colleagues investigated the evolutionary origins of protein complexes and concluded that they evolve through duplication of homodimers, that is through duplication of genes coding for self-interacting proteins (Pereira-Leal *et al.*, 2007). They showed that protein interactions amongst paralogous proteins occur more frequently than can be expected purely by chance in yeast, worm and fly. They also show that PPIs between homodimers and paralogous dimers (interactions between paralogous proteins) in these species are not independent, and conclude that the latter dimer type evolved from the former, something that had been suggested previously (Wagner, 2001). They argue that gene duplication and divergence are important forces driving the expansion of the eukaryotic proteome, and that multiple copies of identical subunits are an economical way of forming larger functional structures. Another study, which modelled the yeast protein interaction network before and after WGD found evidence for greater retention of self-interacting genes in duplicates (Presser *et al.*, 2008), and so is consistent with this hypothesis.

With a few exceptions (e.g., ref (Ispolatov *et al.*, 2005)) most previous

studies have used the yeast strain *S. cerevisiae* as a model organism. This is convenient, as thorough, full-scale proteomic interaction studies have been conducted in yeast using AP-MS to gain knowledge of protein complexes (Gavin *et al.*, 2006). Moreover there is complete genome sequence of several other yeast species for evolutionary comparisons, and extensive protein interaction data between protein pairs are available in the DIP (Xenarios *et al.*, 2000, 2001; Salwinski *et al.*, 2004).

In this study we investigate the duplicability of self-interacting genes in human. We also examine the impact of the extent of the duplication event by comparing SSD duplicate genes with WGD duplicate genes. Our results support the model of preferential retention of duplicated self-interacting genes. This result is robust against the method used to define duplicability. We also show a greater enrichment of self-interacting genes among WGD duplicates than SSD duplicates, and relate this to an overall higher connectivity of WGD genes in the PIN.

4.2 Methods

4.2.1 Filtering of human protein-protein interaction data

We obtained 37,107 PPIs involving 9303 genes from the Human Protein Reference Database (HPRD) release 7 (Keshava Prasad *et al.*, 2009; Peri *et al.*, 2003). We excluded interactions where either of the interacting partners could not be linked to an Ensembl release 50 identifier (Flicek *et al.*, 2008) as well as 13 genes that were too short or simple for the BLASTP sequence similarity search. Thus, the final dataset consisted of 34795 interactions, encoded by 8881 (1879 self-interacting) genes (Figure 4.1).

4.2.2 Definition of singleton and duplicate genes

An all-against-all protein BLAST (BLASTP) search (Altschul *et al.*, 1990) of all known and novel human peptides present in Ensembl Core release 50 (Flicek *et al.*, 2008) was performed to define singleton and duplicate genes in human. Singleton genes were defined as genes whose protein products lack any non-self hit with an E-value less than 0.1. A gene was considered to be a duplicated gene if its top, non-self hit had an E-value less than or equal to

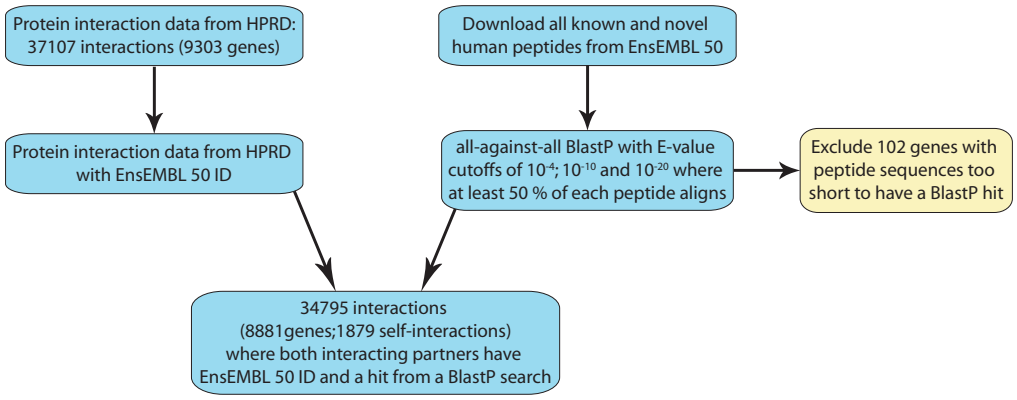


Figure 4.1: Data collection. Flow chart illustrating how the human interaction data were collected from HPRD release 7, and subsequently matched with the blastable Ensembl Core release 50 identifiers in order to extract the final 8881 genes involved in 34795 PPIs.

1 x 10^{-20}, and at least 50% of the two peptides aligned. Genes with BLAST hits at intermediate E-values (less than 0.1 but larger than 1 x 10^{-20}) were considered ambiguous genes, which could neither be classified as singleton or duplicate genes. The analyses were repeated with E-value thresholds of 1 x 10^{-4}, 1 x 10^{-10} and the results were consistent. Also, 102 genes lacked hits after the BLASTP search was performed (i.e., not even a self- hit). Absence of hits was caused by low complexity (simple sequence) masking or too-short peptide sequences and these were excluded from further analysis.

4.2.3 Definition of singleton and duplicate genes for the comparative study

Genes that have not recently duplicated in the human lineage (since the human-mouse split) were examined for mouse lineage-specific duplication using data from the Ensembl Compara release 50 (Flicek *et al.*, 2008) from human (*Homo sapiens*) and mouse (*Mus musculus*). A singleton gene was classified as a gene that had a one to one orthologous relationship between the two species, while a duplicate gene was a single human gene with at least two orthologous genes in mouse (1:many relationship; Figure 4.2).

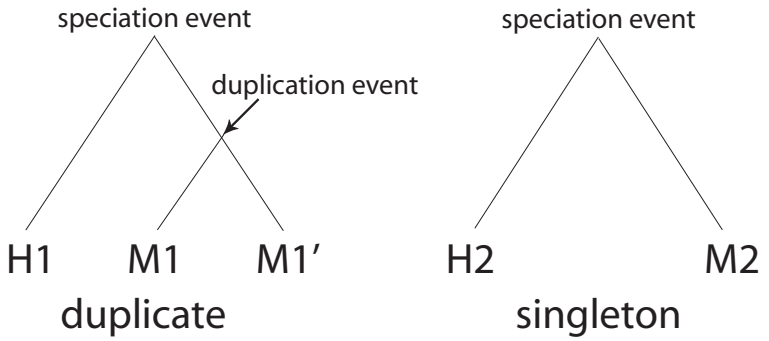


Figure 4.2: Definition of mouse-specific duplicated genes in human. Human genes (H) were classified as singletons if they had a one to one orthologous relationship with mouse (M), and as mouse-specific duplicate genes if the relationship with mouse was one to many *i.e.* if one human gene had at least two orthologous genes in the mouse lineage.

4.2.4 Comparison of WGD duplicate genes vs SSD duplicate genes

The frequency of protein self-interactions amongst human WGD-duplicate genes (also known as ohnologs (Wolfe, 2000)) were compared against SSD-duplicate genes. For this purpose two different paralogon datasets (by Nakatani *et al.* (2007) and Hokamp (2001) respectively) were used to extract WGD genes. 2877 WGD-duplicated genes with available PPI data were could be matched against the Nakatani dataset. 307 genes that were not classified by us as duplicated, based on our criteria of BLASTP sequence similarity and alignment length (above), were classified as WGD duplicate genes based on Nakatani's analysis which includes comparative gene synteny. Thus the final Nakatani based dataset (from here on the default WGD dataset when not otherwise noted) consisted of 2877 WGD genes and 2961 SSD-duplicated genes (Table 4.1). The second, Hokamp, dataset contained genes from Ensembl Core release 50 with available PPI data. *C. intestinalis* and Amphioxious (the latter from DOE Joint Genome Institute `http://www.jgi.doe.gov/`) served as outgroups, and the method used for detecting WGD genes was modified from (McLysaght *et al.*, 2002). Only paralogons consisting of at least 6 distinct duplicated genes were used. From the previous singleton-duplicate definitions (Table 4.2) some genes had to be re-assigned, and in the end the Hokamp dataset contained 2377 WGD genes

Table 4.1: Proportion of self-interacting duplicated genes generated by different mechanisms

Duplication mechanism	self-interacting	Total
WGD	717 (25 %)	2877
SSD	630 (21 %)	2961

Table 4.2: Frequency of self-interaction of singleton and duplicate (E-value threshold) genes under alternative parameters for paralog definition

Set	E-value	singletons	ambiguous	duplicates	genes/set	Tot genes
		self/nonself/tot		self/nonself/tot		
1	1 x 10^{-4}	433/2162/2595	104	1431/4751/6182	8777	8881
2	1 x 10^{-10}	433/2162/2595	375	1370/4541/5911	8506	8881
3	1 x 10^{-20}	433/2162/2595	755	1285/4246/5531	8126	8881

and 3492 SSD genes. A third and final WGD dataset was constructed from Nakatani and Hokamp WGD subset (Table 4.3).

We applied simple logistic regression (Crawley, 2005; McDonald, 2009) to measure the proportion of genes encoding self-interacting proteins over all genes having a particular number of interacting partners (degree k). In accordance with Ispolatov and colleagues (Ispolatov *et al.*, 2005) we found that self-interacting genes tend to have more interacting partners in general. We then measured the proportion of WGD genes over all duplicate genes (and SSD genes) with degree k.

4.2.5 Controlling for age and protein connectivity

From the previously conducted all-against-all BLASTP search we extracted the best non-self hit, with at least 50 % of the sequence aligning, as the closest paralog of each gene in order to create duplicate gene pairs. The genes

Table 4.3: Frequency of self-interactions amongst WGD and SSD duplicate genes under alternative WGD identification methods

Set	WGD dupl	SSD dupl	Tot genes
	self/nonself/tot	self/nonself/tot	
Nakatani	717/2160/2877	630/2331/2961	5838
Hokamp	583/1794/2377	769/2723/3492	5869
combined	402/1109/1511	481/1798/2279	3790

of each pair were classified according to previous gene type classifications (singleton, ambiguous, duplicate, Nakatani-WGD or SSD) with the intention of comparing the connectivity of WGD and SSD genes of similar age. Hence, the synonymous substitution rate (K_s) of gene pairs where both genes where either WGD or SSD genes. The synonymous substitution rate (K_s) was estimated using the Yang and Nielsen method (Yang and Nielsen, 2000), which is implemented in PAML 4 (Yang, 2007). Pairs with (K_s) values of -0 or 99 (saturated) were excluded. Two datasets, consisting of (2139) WGD pairs and (1519) SSD pairs respectively were created to be used as controls of the impact of duplication age for enrichment of self-interactions.

4.2.6 Gene Ontology analyses

Gene Ontology (GO) terms related to biological process and molecular function were examined for sets of self- and nonself-interacting genes, self-interacting WGD and SSD genes, and duplicated and singleton genes using GO slim (The Gene Ontology project website, 2010b) converted by human GO identifiers (The Gene Ontology project website, 2010a) (both available from `http://www.geneontology.org`). Expected values were estimated by simulation, and p-values were calculated as the difference between the expected and observed frequency of GO identifiers under the hypergeometric distribution. Finally the estimated p-values were adjusted by Bonferroni correction. The Perl scripts employed to perform the GO analyses were written by Takashi Makino.

4.3 Results and Discussion

4.3.1 Higher duplicability of self-interacting genes

We examined 34795 interactions between products of 8881 genes, 1879 of which are self-interacting (Figure 4.1).

The protein interactions follow a power-law distribution (Figure 1.3), a typical feature of biological networks (Albert, 2005; Barabási and Albert, 1999). Singletons were defined as human genes with no BLASTP hit in the human genome (other than self-hits) at an E-value threshold of 0.1. Duplicate genes had a non-self BLASTP hit in the human genome with an E-value less than or equal to 1 x 10^{-4}, 1 x 10^{-10} and 1 x 10^{-20} respectively (Table 4.2).

Table 4.4: Duplicate genes are enriched for self-interactions disregarding E-value cutoff used to define duplicate genes

Set	χ^2	df	p-value
1	45.24	1	1.74 x 10^{-11}
2	45.10	1	1.87 x 10^{-11}
3	45.02	1	1.96 x 10^{-11}

Genes with hits with intermediate E-values were excluded as ambiguous. Consistent with previous studies (Ispolatov *et al.*, 2005; Pereira-Leal *et al.*, 2007) we found that duplicate genes are enriched for self-interactions (for E-value 1 x 10^{-20}: $\chi^2 = 45.02$, p = 1.96 x 10^{-11}), and this result does not depend on the E-value used to define duplicate genes (Table 4.4).

Since homomers can give rise to heteromers through the duplication of self-interacting proteins, one possible outcome of the duplication of self-interacting genes is that the self-interaction is lost as the interaction between paralogs is favoured, perhaps because that permits greater evolutionary novelty. If this is the case it may reduce the apparent duplicability of self-interacting genes when duplications and interactions are measured in the same organism (Davis and Petrov, 2004; He and Zhang, 2006). Similarly, interactions may be gained after gene duplication. We measured duplicability in a sister lineage in order to exclude the possibility that the gene duplication interferes with the interaction status of the protein products.

We searched for mouse orthologs of the human genes using Ensembl Compara. We defined a singleton gene as a human gene that had a one to one ortholog in mouse, while a duplicate gene was a human gene with at least two co-orthologous genes in mouse (i.e., mouse lineage-specific duplication; Figure 4.2).

Thus only human genes that have not experienced a recent human lineage-specific duplication (within the last 90 - 100 myr) were considered, and duplicability was assessed by the status in the mouse genome (Table 4.5).

For our analysis we could only use genes with available interaction information, however the proportion of duplicated genes (mouse-specific duplication) in this reduced dataset (2.21 %) is comparable to the proportion for the whole human-mouse orthologue dataset (2.63 %) so we infer that no bias is introduced ($\chi^2 = 3.84$, df = 1, p > 0.05). We find that orthologs of human self-interacting genes have greater duplicability in the mouse lineage ($\chi^2 = 3.96$, df = 1, p = 0.047). The fact that the duplicated genes in

Table 4.5: Proportion of self-interacting singletons and duplicates defined either by BLASTP search (E-values 0.1 for singletons and 1 x 10^{-20} for duplicates as defined in section 4.2.2 and table 4.2) or by orthology status compared to the mouse genome.

Classification system		self-interacting	Total
Human BLASTP	Singletons	433 (16 %)	2595
	Duplicates	1285 (23 %)	5531
Mouse duplicability	1:1 orthologs	1682 (21 %)	7968
	1:many orthologs	51 (27 %)	186

this dataset have recently duplicated (since the mouse-human divergence) indicates that this is an ongoing phenomenon.

4.3.2 WGD genes are enriched for self-interactions by comparison with SSD genes

Previous studies have shown different properties for genes duplicated by SSD or WGD (Guan *et al.*, 2007; Hakes *et al.*, 2007a). In particular, dosage-balance constraints result in different duplication outcomes under WGD (biased retention) and SSD (lower duplicability). We compared genes duplicated by SSD or WGD to investigate whether the duplication mechanism influences the retention of self-interacting genes. It has been observed in yeast that genes that form heteromers (complexes of proteins encoded by different genes) have fewer paralogs than other genes, since the integrity of the complex depends on duplication of all, or none, of the genes in the complex (Papp *et al.*, 2003a). In the human data, 25 % of genes duplicated by WGD are self-interacting compared to only 21 % of SSD-duplicated genes (Table 4.1), a significant difference ($\chi^2 = 10.72$, df =1, p = 0.0011; Table 4.6). Because self-interaction does not depend on other genes, this observation is not immediately reconcilable with between-gene dosage-balance explanations for WGD gene retention. However, it was previously noted that self-interacting proteins tend to have more interacting partners (connectivity) than non-self-interacting genes (Ispolatov *et al.*, 2005), and we confirm this for our dataset (Logistic regression; Figure 4.3). Thus the higher fraction of self-interacting WGD genes may be due to an indirect effect of a greater number of interactions (and thus a greater tendency for dosage-balance constraints on these genes).

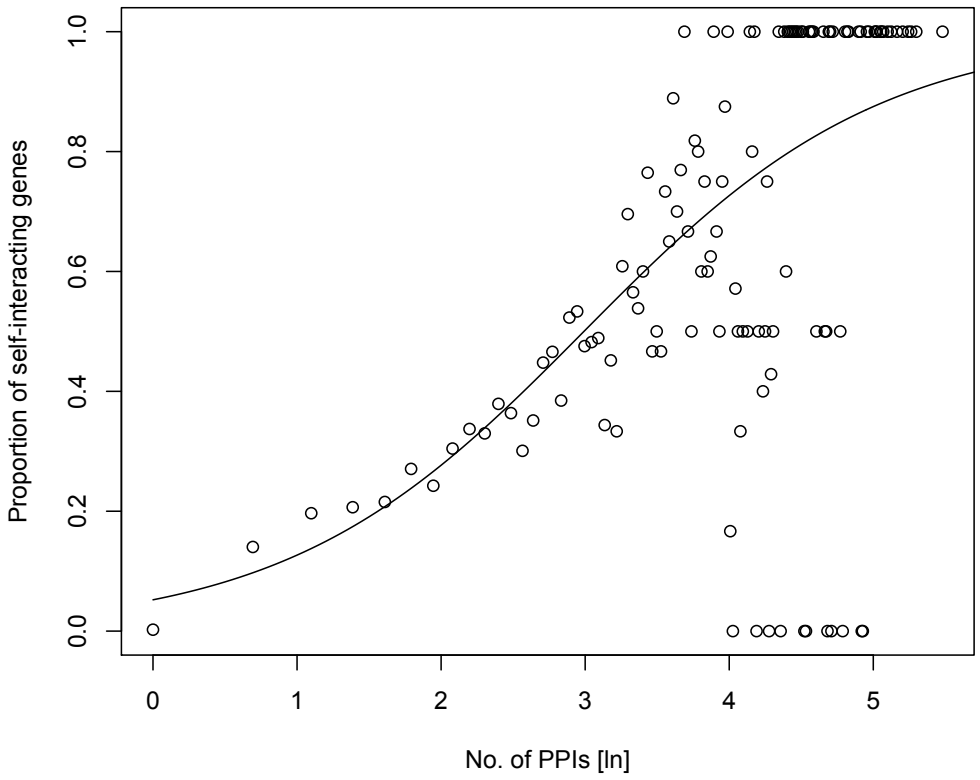


Figure 4.3: Increased proportion of self-interacting genes with increased protein connectivity.

The natural logarithm transformed protein connectivity (No. of PPIs; x-axis) plotted against the proportion of self-interacting genes (y-axis). The proportion of self-interacting genes over all genes increases with the number of PPIs of the genes i.e. self-interacting genes tend to have more PPIs in general compared to other, nonself-interacting genes. The `glm` function in `R` was used to fit the generalised linear model to the data.

Table 4.6: WGD genes are enriched for self-interactions disregarding of WGD detection method

Set	χ^2	df	p-value
Nakatani	10.72	1	0.0011
Hokamp	4.87	1	0.027
combined	15.07	1	1.04 x 10^{-4}

4.3.3 WGD genes have more interaction partners on average than SSD genes

To further investigate the relationship of protein interaction network connectivity and duplicability we examined the number of interacting partners of duplicated genes. We found that the genes with a higher number of interactions contain a greater proportion of genes duplicated by WGD, and that this is true irrespective of self-interaction ($p < 2 \times 10^{-16}$ and $p = 1.03 \times 10^{-7}$ respectively, logistic regression; Figure 4.4). However, it was previously noted in yeast that older genes tend to have more interaction partners (Kunin *et al.*, 2004; Prachumwat and Li, 2006), and when we control for age of duplication, we find no difference in connectivity between all WGD and SSD duplicates (Figure 4.5a) or between self-interacting WGD and SSD duplicates (Figure 4.5b). Furthermore, there is a bias of duplication timing for WGD genes i.e. WGD genes tend to be older than SSD genes (Figure 4.6). In particular, there are few WGD genes with K_s <1.5, which have low connectivity. These results indicate that K_s bias causes a difference in the total connectivity between WGD and SSD genes respectively.

We therefore considered whether the enrichment for self-interactions is truly a generality of duplicated genes, or is only a feature of WGD-duplicated genes which are older on average than SSD-duplicated genes and tend to have a higher number of interaction partners. Indeed, the enrichment for self-interaction among WGD genes compared to singletons is highly significant ($\chi^2 = 55.88$, df = 1, $p = 7.69 \times 10^{-14}$), but so too is the enrichment among SSD-duplicated genes ($\chi^2 = 19.92$, df = 1, $p = 8.08 \times 10^{-6}$). Thus we conclude that self-interaction is a general feature of duplicated genes that is influenced, but not determined by, the mode of duplication and the total number of interactions.

4.3.4 Self-interacting genes are enriched for developmental and essential biological processes, and WGD self-interacting genes are involved in metabolism

In order to understand the biological impact of duplication of self-interacting genes, we examined their functions using GO terms. We found that self-interacting genes are enriched for biological processes such as early

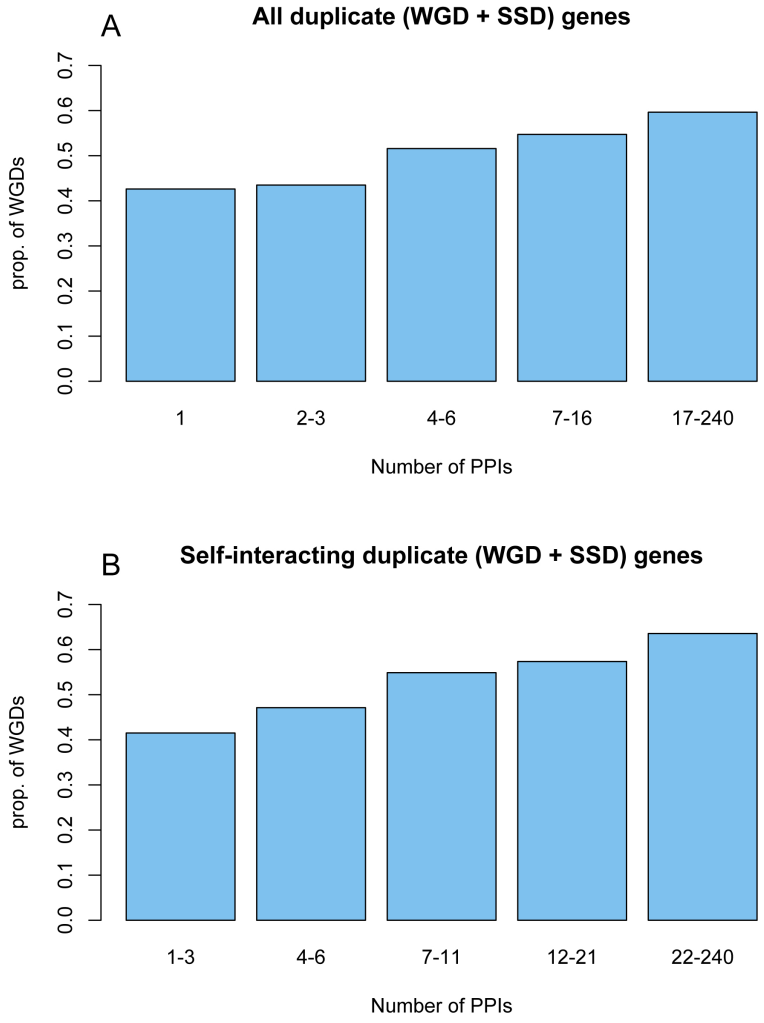


Figure 4.4: Relationship of duplication type and number of interactions. The proportion of WGD genes among all duplicate (WGD and SSD) genes increases with increased number of PPIs irrespective of self-interactions. a) Proportion of WGD genes among all duplicate genes with respect to the number of interactions. (Bins created to contain similar amounts of genes.) b) Proportion of self-interacting WGD genes among all self-interacting duplicate genes with respect to the number of interactions. (Bins created to contain similar amount of genes.)

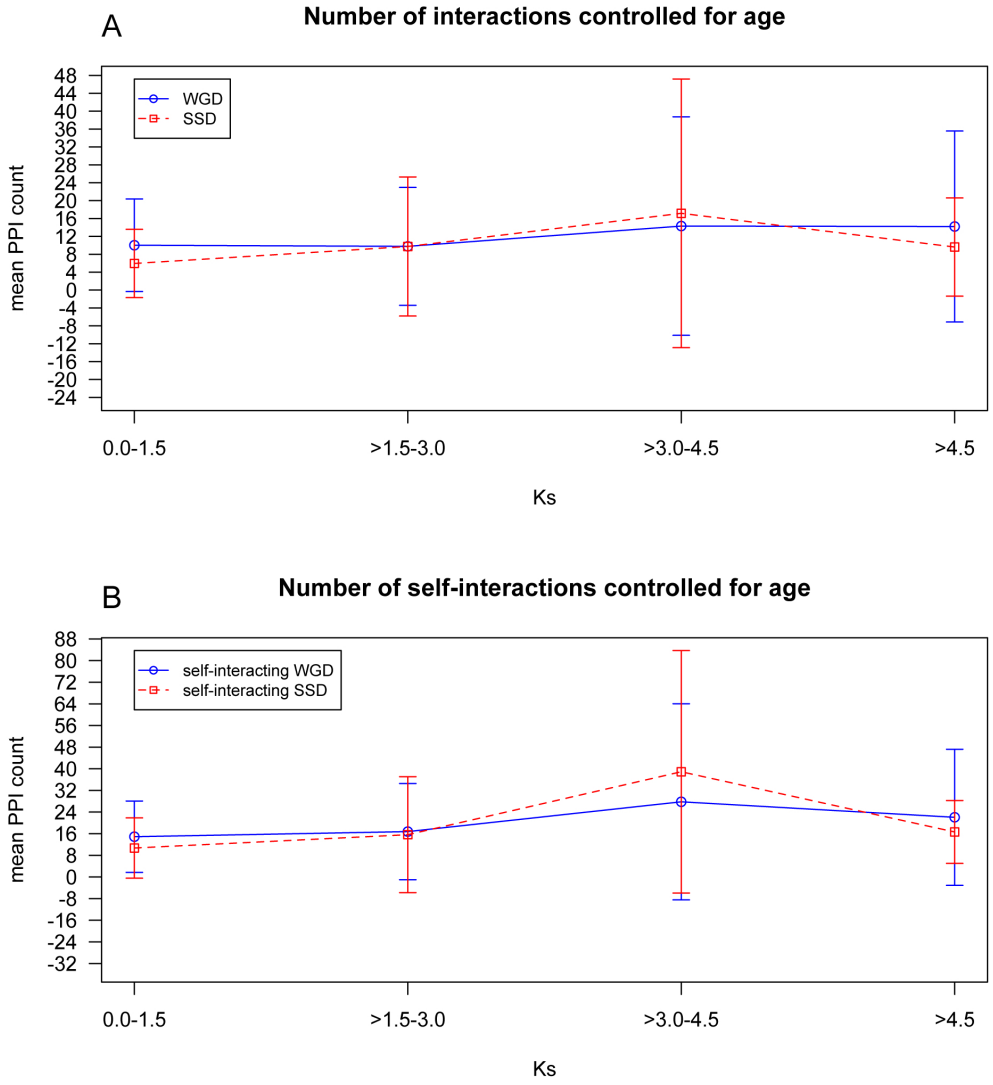


Figure 4.5: Number of duplications controlled for age. Relationship of synonymous divergence rate, and the number of PPI partners of each gene in the duplicate pair. The x-axis displays the synonymous substitution rate (K_s) between a) a duplicate pair or b) a self-interacting duplicate pair, while the y-axis is the mean value of the total number of PPIs of all genes in each K_s bin (category). The bars indicate the standard deviation (S.D.)

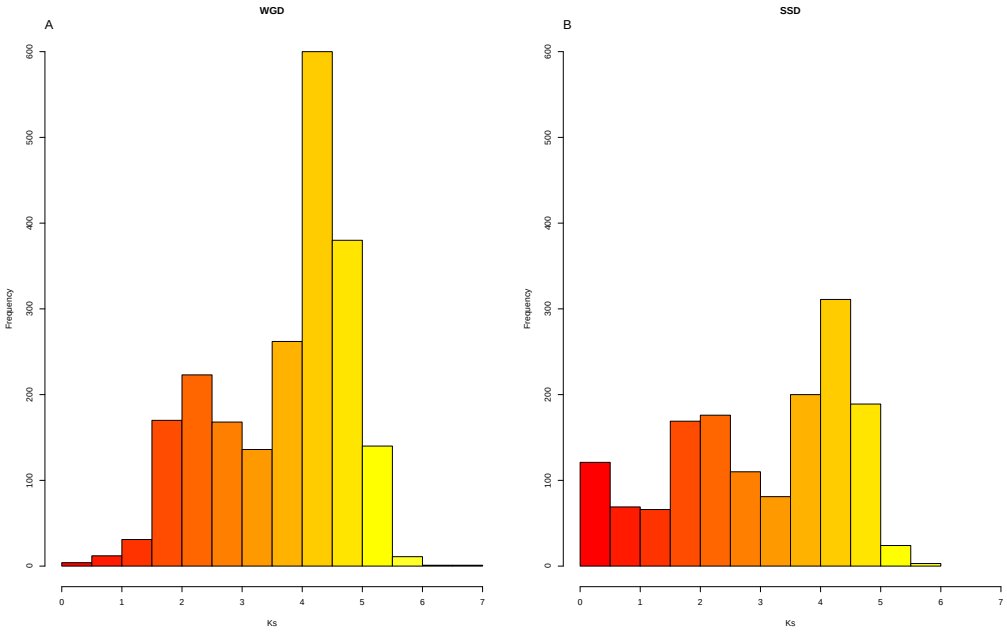


Figure 4.6: Distribution of synonymous substitution rate of WGD and SSD duplicate pairs.

The percentage of a) WGD and b) SSD duplicate genes with differing degrees of synonymous substitution rates K_s. The x-axis indicates the synonymous substitution rate (K_s) between a duplicate gene pair, and the y-axis the percentage of duplicates with that K_s.

development (GO:0030154 and GO:0007275), cell death, cell communication and response to stimulus, and molecular functions such as protein binding, kinase and transferase activity as well as receptor activity (Table 4.7). We then compared duplicated against singleton genes, and found that while the first set of genes are over-represented for cell communication and cell differentiation, and molecular functions including binding, receptor activity and channel activity; they are under-represented for metabolic and catabolic processes, nucleic acid binding and ligase activity (Table 4.8). Finally we compared self-interacting WGD and SSD genes against each other, and found that while the WGD genes are under-represented for "response to stimulus", "oxidoreductase", "antioxidant" and "hydolase activity" the same set of genes are over-represented for "regulation of biological processes", "metabolic processes", "kinase", "transferase" and "transcription regulation activity" (Table 4.9). Thus, genes involved in metabolic processes are under-represented among duplicate genes in general, but enriched in WGD-duplicated genes compared to SSD-duplicated genes. This is consistent with a recent report, which showed that after WGD metabolic genes are retained more often than non-metabolic genes due to selection for gene expression along the entire metabolic pathway (Gout *et al.*, 2009).

Table 4.7: Over-represented GO terms when self- and non-self-interacting genes are compared against each other

Biological Processes							
GO IDs	Term	Obs.	Exp.	S.D.	Z score	p-value[a]	Fold enrichment
GO:0006928	cell motion	143	90.359	8.0	6.5	1.25×10^{-8}	2.09
GO:0007154	cell communication	762	602.228	17.2	9.3	4.48×10^{-18}	1.49
GO:0007275	multicellular organismal development	484	403.84	15.2	5.3	6.05×10^{-6}	1.39
GO:0007610	behavior	110	68.806	7.4	5.6	7.57×10^{-7}	2.12
GO:0008219	cell death	290	170.469	11.0	10.8	9.11×10^{-24}	2.33
GO:0009987	cellular process	1141	1043.859	17.9	5.4	1.16×10^{-6}	1.22
GO:0030154	cell differentiation	283	221.16	12.1	5.1	1.35×10^{-5}	1.52
GO:0032501	multicellular organismal process	294	236.651	13.3	4.3	1.69×10^{-4}	1.45
GO:0043170	macromolecule metabolic process	967	891.187	18.0	4.2	5.97×10^{-4}	1.21
GO:0050789	regulation of biological process	1131	986.799	18.8	7.7	3.99×10^{-14}	1.30
GO:0050896	response to stimulus	480	357.361	15.4	8.0	1.55×10^{-14}	1.63
GO:0051704	multi-organism process	160	91.12	8.4	8.2	3.40×10^{-14}	2.47
Molecular Functions							
GO IDs	Term	Obs.	Mean	S.D.	Z score	p-value[a]	Fold enrichment
GO:0004871	signal transducer activity	134	85.963	7.8	6.2	3.02×10^{-7}	2.01
GO:0004872	receptor activity	248	201.966	11.6	4.0	3.56×10^{-3}	1.42
GO:0005488	binding	940	863.838	18.2	4.2	1.34×10^{-3}	1.20
GO:0005515	protein binding	1436	1232.226	17.3	11.8	3.12×10^{-31}	1.32
GO:0016301	kinase activity	237	134.222	9.6	10.7	2.39×10^{-21}	2.45
GO:0016740	transferase activity	269	189.553	11.5	6.9	7.45 x 10−10	1.74

[a]The estimated P values were adjusted by Bonferroni correction

Table 4.8: Over- and under-represented (italics) GO terms when duplicated genes are compared against singleton genes

Biological Processes							
GO IDs	Term	Obs.	Exp.	S.D.	Z score	p-value[a]	Fold enrichment
GO:0007154	cell communication	2009	1819.3	17.5	10.8	5.86 x 10^{-27}	1.80
GO:0006139	*nucleobase, nucleoside, nucleotide and nucleic acid metabolic process*	*1446*	*1630.6*	*16.6*	*-11.1*	*4.08 x* 10^{-25}	*0.76*
GO:0009058	*biosynthetic process*	*1522*	*1650.6*	*17.0*	*-7.6*	*2.21x* 10^{-12}	*0.86*
GO:0050789	regulation of biological process	3130	3001.9	18.3	7.0	5.91 x 10^{-11}	1.36
GO:0007275	multicellular organismal development	1346	1241.1	15.8	6.7	1.76 x 10^{-10}	1.63
GO:0032501	multicellular organismal process	787	712.5	12.4	6.0	2.34 x 10^{-8}	1.80
GO:0022904	*respiratory electron transport chain*	*3*	*14.9*	*1.9*	*-6.2*	*1.45 x* 10^{-6}	*0.07*
GO:0043170	*macromolecule metabolic process*	*2628*	*2726.8*	*18.2*	*-5.4*	*1.72 x* 10^{-6}	*1.00*
GO:0030154	cell differentiation	726	673.2	12.1	4.4	3.02 x 10^{-4}	1.58
GO:0007610	behavior	218	189.6	7.1	4.0	3.39 x 10^{-4}	2.28
GO:0006928	cell motion	300	267.4	8.3	3.9	6.67 x 10^{-4}	1.96
GO:0009056	*catabolic process*	*452*	*487.6*	*11.0*	*-3.2*	*1.70 x* 10^{-2}	*0.87*
GO:0005634	*nucleus*	*1*	*5.1*	*1.2*	*-3.4*	*4.85 x* 10^{-2}	*0.07*
GO:0016787	*hydrolase activity*	*1*	*5.1*	*1.2*	*-3.4*	*4.85 x* 10^{-2}	*0.66*
Molecular Functions							
GO IDs	Term	Obs.	Mean	S.D.	Z score	p-value[a]	Fold enrichment
GO:0005488	binding	2905	2686.899	18.3	11.9	3.85 x 10^{-29}	1.56
GO:0004872	receptor activity	714	608.911	12.3	8.6	2.06 x 10^{-19}	2.56
GO:0016301	kinase activity	496	420.462	10.0	7.5	4.99 x 10^{-14}	2.66
GO:0015267	channel activity	203	159.644	6.6	6.6	1.28 x 10^{-12}	5.76
GO:0015075	ion transmembrane transporter activity	300	247.312	8.1	6.5	2.48 x 10^{-11}	3.31
GO:0004871	signal transducer activity	315	265.132	8.2	6.1	2.76 x 10^{-9}	2.78
GO:0003676	*nucelic acid binding*	*1211*	*1290.829*	*16.5*	*-4.8*	*3.31 x* 10^{-5}	*0.89*
GO:0016787	hydrolase activity	820	813.748	13.8	4.1	7.61 x 10^{-4}	1.48
GO:0045182	*translation regulator activity*	*36*	*51.717*	*3.6*	*-4.3*	*2.28 x* 10^{-3}	*0.42*
GO:0003774	motor activity	68	55.557	3.7	3.4	1.09 x 10^{-2}	3.86
GO:0016874	*ligase activity*	*138*	*161.46*	*6.4*	*-3.7*	*1.21 x* 10^{-2}	*0.68*

[a]The estimated P values were adjusted by Bonferroni correction

Table 4.9: Over- and under-representation (italics) of GO terms in self-interacting WGD with respect to SSD genes

Biological Processes							
GO IDs	Term	Obs.	Exp.	S.D.	Z score	p-value[a]	Fold enrichment
GO:0050896	*response to stimulus*	*138*	*171.285*	*7.5*	*-4.4*	*2.47 x* 10^{-4}	*0.66*
GO:0050789	regulation of biological process	484	456.64	8.2	3.3	1.70 x 10^{-2}	1.15
GO:0043170	macromolecule metabolic process	399	372.058	9.1	2.9	3.91 x 10^{-2}	1.18
Molecular Functions							
GO IDs	Term	Obs.	Mean	S.D.	Z score	p-value[a]	Fold enrichment
GO:0016301	kinase activity	143	111.112	6.7	4.8	3.34 x 10^{-5}	1.93
GO:0016740	transferase activity	141	113.834	6.8	4.0	1.17 x 10^{-3}	1.72
GO:0030528	transcription regulator activity	134	107.982	6.7	3.9	1.67 x 10^{-3}	1.73
GO:0016491	*oxidoreductase activity*	*14*	*25.711*	*3.3*	*-3.6*	*1.93 x* 10^{-2}	*0.36*
GO:0016209	*antioxidant activity*	*0*	*4.858*	*1.5*	*-3.2*	*4.16 x* 10^{-2}	*0*
GO:0016787	*hydrolase activity*	*81*	*101.071*	*6.3*	*-3.2*	*4.37 x* 10^{-2}	*0.66*

[a]The estimated P values were adjusted by Bonferroni correction

4.4 Conclusions

We observed greater duplicability of human self-interacting genes and that WGD duplicate genes tend to be self-interacting more often than SSD duplicate genes. This latter observation probably relates to the higher overall connectivity of WGD genes in PINs. Highly connected genes are more likely to be subject to dosage balance constraints and so to be resistant to SSD, but preferentially retained after WGD (Papp *et al.*, 2003a; Makino *et al.*, 2009). This result is also consistent with studies in yeast, which showed preferential retention of interacting genes after WGD as a possible explanation of the PIN dynamics (Presser *et al.*, 2008). Moreover, consistent with previous observations in yeast (Kunin *et al.*, 2004; Prachumwat and Li, 2006), we found that protein connectivity is correlated with the time since gene duplication. Our results also support the hypothesis that duplication of self-interacting proteins should be selectively advantageous because it facilitates the evolution of complex protein structures (Levy and Pereira-Leal, 2008).

Chapter 5

A natural history of the human protein interaction network

The methods described in Methods subsections 5.2.5 and 5.2.6, and the insights gained from these experiments as described in Results 5.3.6 were performed and contributed by collaborator Gavin C. Conant. Furthermore, a slight modification to two Perl scripts written by Corey Hudson were used to perform the analysis described in Methods 5.2.2, and finally the Perl scripts written by Takashi Makino that were used in the study of chapter 4 to perform those GO analyses were also used to perform the GO analyses of this study.

5.1 Introduction

The elucidation of biological networks allows, in principle, researchers to understand the evolution of genetic structures larger than a single gene (Zhu *et al.*, 2007). However, such efforts are confounded by the fact that the network is usually only known for a single taxon, making evolutionary inferences challenging. A number of researchers have nonetheless explored network evolution: one of their main findings is the important role of gene duplication in that evolution (Teichmann and Babu, 2004). Early work on the yeast protein interaction network also found that while duplication creates redundancy in the set of protein interactions, that redundancy generally decays quickly after duplication (Wagner, 2001). This result led to the

natural question of the influence of the protein interaction structure on the patterns of gene duplication. It appears that proteins residing in the less dense parts of the protein interaction network are more likely to duplicate (Li *et al.*, 2006). In yeast in particular these duplicate effects also depend on the type of duplication, with genome-duplication produced duplicates tending to have more protein interaction partners than those produced by smaller duplications (Guan *et al.*, 2007; Hakes *et al.*, 2007a).

Of course, network structures have influences on other aspects of evolution. Two of the primary measurable features are its rate of evolution, or more strictly its constraint (i.e., the degree to which certain polymorphisms are filtered out of a population by purifying selection). Considerable work has gone into identifying predictors of these rates/constraint. Thus, in mammals, Vinogradov found that the second strongest predictor of a protein's rate of evolution (after its phylogenetic distribution) was its expression level (Pál *et al.*, 2006; Vinogradov, 2010). Moreover, at least for mammals, the evolutionary rate is more strongly influenced by the expression levels of co-expressed genes and genes in the same biological pathway than it is by the gene's own expression level (Vinogradov, 2010). More generally, studies in a number of taxa have shown an association between high expression and low rates of sequence evolution (Pál *et al.*, 2001, 2006; Drummond *et al.*, 2005; Vinogradov, 2010). Drummond and Wilke (2008) argued that this slow evolutionary rate of highly expressed proteins is due to the fitness costs of protein misfolding being greater for highly expressed genes, and this effect appears largely unaltered whether or not the gene in question has undergone a recent duplication (Drummond *et al.*, 2005). In mammals and in yeasts it has been found that extracellular proteins evolve faster than intracellular proteins (Julenius and Pedersen, 2006; Liao *et al.*, 2010). However, at least in yeast, this association may be a correlate of the generally rapid evolution seen among non-essential genes, since there are very few essential extracellular proteins (Julenius and Pedersen, 2006). In mammals essentiality and protein localization show independent predictive power for evolutionary rate (Liao *et al.*, 2010).

It is natural to ask how a protein's position in the interaction network influences its rate of evolution. Fraser *et al.* (2002, 2003) found that proteins with more protein interactions evolved more slowly than proteins with less interaction partners, and that this was not due to the proteins importance and fitness effect. However, this conclusion was disputed and at best the

association of evolutionary rate and interaction degree is quite weak (Jordan *et al.*, 2003; Bloom and Adami, 2003; Hahn *et al.*, 2004), certainly much weaker than the effect of expression (Drummond *et al.*, 2006). Indeed, Fernández and Lynch have recently proposed that in organisms with small population sizes, much of the structure of the interaction network might have evolved to maintain protein folds, raising the question as to whether an association of interaction degree and evolutionary rate is expected in these organisms (Fernández and Lynch, 2011). To the extent that evolutionary rate is constrained by interaction partners, the hypothesized mechanism is that proteins with many interaction partners have a larger proportion of the amino acid sequence is directly involved as binding sites, leading to selection against changes in the affected residues.

Of course, rather than inducing constrain, such interactions could also result in co-evolution between the two molecules (Fraser *et al.*, 2002). According to Fraser *et al.* (2004), co-evolution of expression of interacting proteins is more important factor for the prediction of protein-protein interactions (PPIs) than the co-evolution of the amino acid sequence of these proteins, at least in the case of the stable protein interactions that form complexes. Chen and Dokholyan (2006) define functional modules as spatially or chemically isolated group of biological components that have a common functional task, and they found that functional modularity governs the evolutionary rate of amino acid sequences and expression levels of proteins. (According to their definition 'functional modularity is the physical or functional interactions between proteins. The first refers to a physical protein interaction, while the latter can involve an interaction between a protein and a gene encoding another protein e.g., between a repressor protein and the gene being repressed.) These authors argue that functional modularity is more important for coordinated evolution between interacting proteins than are physical interactions, and that proteins within a functional module evolve at more similar molecular rates and have more similar expression levels than do two proteins in different modules. In keeping with the observation that protein interactions at best weakly constrain sequence evolution, it is also known that residues on a proteins surface show lower selective constrain than do internal ones, probably because the latter contribute more directly to proper protein-folding. In mammals it appears that many surface residues evolve essentially neutrally. Yet these proteins do not have fewer protein interactions than do proteins with constrained

surfaces (Conant, 2009). Despite these observations, it has been shown that surface residues involved in permanent protein interactions evolve slowly, while those involved in transient protein interactions can even have an increased substitution rate (Mintseris and Weng, 2005).

The degree to which interacting proteins undergo co-evolution between their binding sites (i.e., compensatory substitutions that maintain the PPI) is still debated, as is the contribution of such substitutions to the correlated evolution of proteins involved. Hakes *et al.* (2007b) argue that since interacting proteins experience the same environment and gene expression levels, the correlation in their evolutionary rates is sufficiently explained by these factors without needing to invoke correlated substitutions. However, the mirrortree method, that calculates correlation coefficient scores between interacting proteins, is still appropriate: indeed this approach has been used to predict PPIs in prokaryotes (Goh *et al.*, 2000; Pazos and Valencia, 2001) and eukaryotes (Hakes *et al.*, 2007b).

Here, we are interested in two primary questions. First, to what extent are the known human protein interactions of a recent (i.e., after the mammalian radiation) origin? Second, to what extent does selection act on network structure in the human protein interaction network (PIN)? To explore these questions, we used previously described human PPI data and sets of orthologous genes from 7 other mammals (Figure 5.1). We employ a parsimony approach to reconstruct part of the history of the human PIN, as well as looking for evidence of correlated evolution between interaction partners using a mirrortree approach (Figure 5.2). In addition to evidence for recently evolved PPIs, we show signals of weak but statistically significant co-evolution among the interacting proteins as well as confirming previous work that showed a tendency of interacting proteins to be under similar selective constraint (Vinogradov, 2010).

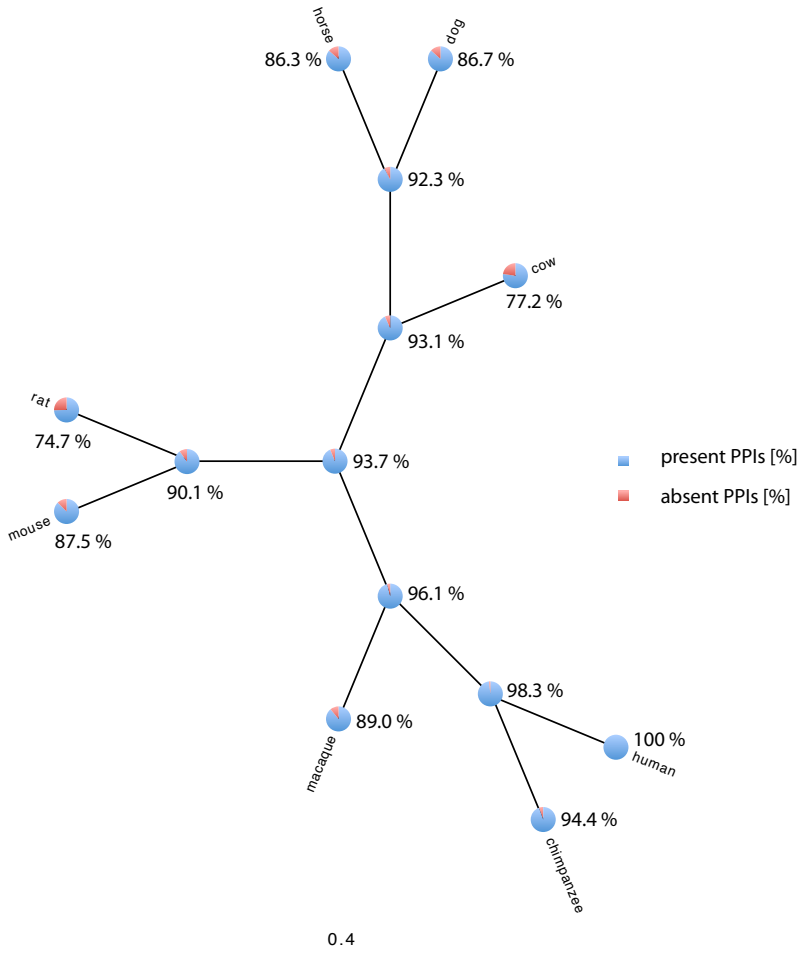


Figure 5.1: PPI presence and absence at the different nodes in the unrooted eutherian phylogenetic tree. At each node, we have shown the predicted percentage of the known human PPIs that were present at that node. Of course, the nature of the data used implies that this percentage will be 100% at the human tip node. The percentages at the other seven tip nodes were inferred by the presence or absence of the orthologues of the two human proteins making up the PPI (Methods 5.2.1). A parsimony approach was then used to infer the status of the interior nodes. The resulting phylogeny was visualized with FigTree (Rambaut, 2009).

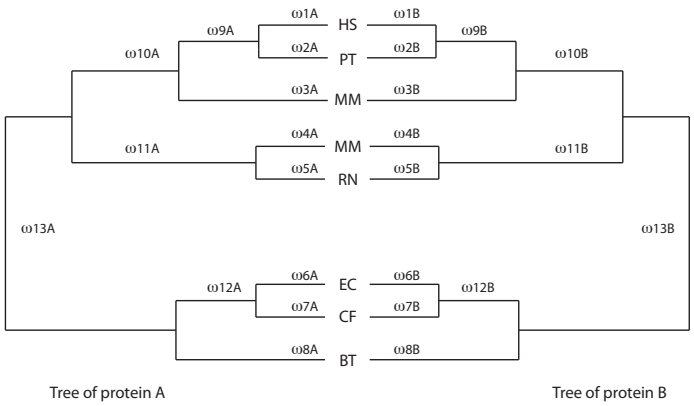


Figure 5.2: Mirrortree with protein-protein interactions. Interacting human proteins A and B are encoded by human genes A and B respectively. The mirrortree is constructed by finding the orthologues of human genes A and B in chimpanzee (PT branch), macaque (MM), mouse (MM), rat (RN), horse (EC), dog (CF) and cow (BT). The selective constraint K_a/K_s (or ω) for each branch was then calculated (Methods 5.2.2).

5.2 Methods

5.2.1 A parsimony-based reconstruction of the ancestral states of the human protein interaction network

We employed the set of human nonself PPIs from a previous study (see chapter 4). This set consists of 32,916 PPIs among 8876 genes. We next identified the orthologues of these 8876 genes from seven other mammals (*Pan troglodytes*, *Macaca mulatta*, *Mus musculus*, *Rattus norvegicus*, *Equus caballus*, *Canis familiaris* and *Bos taurus*) using a previously described tool (Conant, 2009)[a]. Using these data, we created a 'pseudo- alignment where each alignment column represented a human PPI. We then coded the status of that PPI in the other seven species based on the presence or absence of the orthologues of the two proteins involved in that genome. Our coding scheme

[a]Pairwise similarity searches between the genes from the different genomes were performed, and tandem duplicate results collapsed into single entries. Synteny is used to match orthologous genes to each other when multiple hits suggest several possible orthologues.

was as follows:

- If both orthologues were present for a species, we coded it as a 4.
- If an orthologue for the first (human) gene was absent, we coded a 1, likewise a 2 if the second orthologue was absent.
- If both orthologues were absent, we coded a 0.

We then estimated the most parsimonious reconstruction of the ancestral states of these PPIs for an unrooted eutherian phylogenetic tree using Mesquite (Maddison and Maddison, 2010). Using these reconstructions, we then estimated the number of new PPIs that arose on each branch of Figure 5.1.

5.2.2 Assessing co-evolution between interacting proteins using mirrortrees

We next attempted to assess the degree of co-evolution between interacting proteins by looking for shared changes in the selective constraint acting on their coding sequences. To do so, we used only nonself PPIs with 1:1 orthology across all 7 other species for both genes of the PPI pair. Using two Perl scripts (slight modifications of original scripts written by Corey Hudson) FASTA files were created from Ensembl release 50 (Flicek *et al.*, 2008) peptide sequences from all of the 8 species, and used as input in MUSCLE v3.6 (Edgar, 2004) in order to create peptide multiple sequence alignments (MSAs). These MSAs were then converted into nucleotide alignments, and a number of alignment quality checks were performed (Hudson and Conant, 2011). MSAs passing these filters were analyzed with codeml (model M1) in the PAML 4.4 software package (Yang, 2007). PAML produced maximum likelihood estimates of the ratio of nonsynonymous to synonymous substitutions per site (e.g., K_a/K_s), a method proven to be more adequate than pairwise distances for detection of correlated evolution between PPIs (Clark and Aquadro, 2010), for each branch of the phylogeny in Figure 5.1. We interpret K_a/K_s (hereafter ω) in terms of the average selective constraint acting on that branch. In order to avoid erroneous ω estimates resulting from codeml having become trapped in a local optimu, we started PAML from random initial conditions until the three analyses with highest likelihood differed by no more than 5% in their estimates of ω. In three

cases (ENSG00000110400, ENSG00000129038 and ENSG00000154767), the three runs of highest likelihood did not agree even after 100 total iterations of PAML from different conditions, and so these three genes (which participated in 11 PPIs in total) were omitted from further analysis. The result of this pipeline was a set of 12,472 PPIs, encoded by 5030 human genes for which we have estimates of selective constraint. Next the reduced PPIs dataset (of 12,472 PPIs) was used to construct mirrortrees (e.g., paired phylogenic trees, one from each protein with the same set of organisms present (Juan *et al.*, 2008)), one per PPI, and all based on the unrooted eutherian phylogenetic tree of the 8 mammals (Figure 5.2). We thus have thirteen paired branch ω estimates for each PPI pair.

5.2.3 Estimating co-evolution using the correlation of ω values in the phylogeny

We used the Spearman's rank sum correlation coefficient, ρ, in the paired ω values computed above as our estimator of co-evolution. As an aside, we note that because ω values have a highly non-normal distribution (a range from 0 to positive infinity but with a strong bias toward values near zero), we found that the Pearson's correlation coefficient, r, (which assumes normality in the distribution of ω) was inappropriate (Figure 5.3). The result is a subset of interaction pairs with a high ($r > 0.7$) Pearson correlation due to outlier effects (e.g., both genes share a single branch with a high ω value that drives up the Pearson correlation). Given that the Spearman correlation, ρ is not prone to over-stating correlations due to such single points, we chose to report all of our results in terms of this statistic.

To avoid the large sampling variances for cases where $\omega \geqslant 5$, branches with these large values were omitted from our analysis. We then required a minimum of six paired branches in order to compare the correlation statistic for a PPI gene pair. The result was correlation statistics for 7727 PPI pairs. To assess whether the correlations in ω values between the paired genes with a shared PPI were statistically significant, we sought to compare the values of ρ seen in this set of true PPIs and their associated ω values (dataset t-PPI) to the distribution of ρ values from a set of similar, non-interacting gene pairs. We created the datasets of non-interacting pairs using a randomization approach. We started by generating pairs of genes from the 5030 genes used above for which the two genes in each pair were not an interacting pair

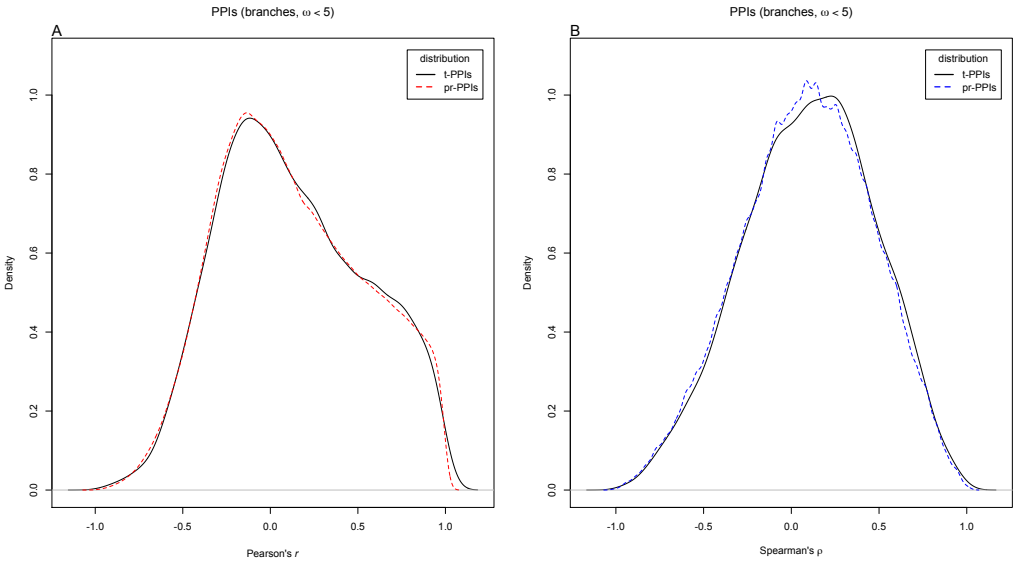


Figure 5.3: Distributions of Pearson's and Spearman's Rank Sums correlation coefficients of the true PPIs and pseudo random PPIs respectively. The true (filled lines) and pseudo random (dashed lines) PPIs. a) Distribution of the Pearson's correlation coefficient r contain a 'bump' at approximate $r = 0.7$ indicating that the data are non-normally distributed. This is caused by branches that are outliers compared to other branches in some of the mirrortrees, which then inflate Pearson's r to high values although there is no good correlation between the branches in those mirrortrees as can be observed for the same data in b) Spearman's Rank Sums correlation coefficients ρ for the same data are normally distributed and the corresponding mirrortrees do not get very high value at Spearman's $\rho = 0.7$ indicating there really is not a very high correlation between branches for those mirrortrees.

according to the PPI data. Once we obtained a set of 7727 unique gene pairs with no true protein interactions and a minimum of six branches in common (to mimic t-PPI), we defined a new pseudodata set (pseudo- random interactions or pr-PPI). We created 1000 such pr-PPI datasets and compared the mean value of Spearman's ρ from t-PPI to the distribution of 1000 mean values of ρ seen in the 1000 pr-PPI datasets.

In order to elucidate how each species contributes to the co-evolution signal, we repeated this analysis after removing the orthologous genes from that species. We then again compared the mean Spearman's rank correlation for the remaining branches for t-PPI to the distributions seen from the pr-

PPIs. We performed a similar analysis after removing five clades from the tree in Figure 5.1. These clades were: HS-PT clade (containing the human, chimpanzee and common HS-PT branches), HS-PT-MM clade (containing the human, chimpanzee, HS-PT, macaque and HS-PT-MM branches), MM-RN clade (containing the mouse, rat and common MM-RN branches), EC-CF clade (containing the horse, dog and EC-CF branches), and finally the EC-CF-BT clade (containing the horse, dog, EC-CF, cow and EC-CF-BT branches).

5.2.4 Shared signals of adaptive and co-evolution

We hypothesised that there might be cases of shared adaptive evolution among the PPI pairs. To explore this possibility, we looked for shared cases of $\omega > 1.0$ in paired branches using a χ^2 test. A similar analysis was performed with a threshold of $\omega > 0.5$, under the assumption that directional and purifying selection might have occurred on the same branch, limiting the divergence signal.

5.2.5 Association of the degree of selective constraint and protein interaction network position

Hudson and Conant (2011) have previously published an analysis of the selective constraints for 13,928 sets of mammalian orthologous genes. Briefly, these data consist of estimates of the ratio K_a/K_s (or ω), calculated with codeml (model M0) in the PAML 4.2 software package (Yang, 2007), for alignments (i.e., not per branch as in the earlier section 5.2.2) of one human gene for which at least six of the eight mammals considered had an identified orthologue (Conant, 2009), and for which no species had any tandem duplications for that gene (Hudson and Conant, 2011). We used these data to ask if proteins of similar selective constraint were more likely to interact with each other in the protein interaction network. To do so, we reduced the set of human PPIs to only those interactions where both proteins were found in the above orthology set. We then created a PPI network where each protein node was weighted by its value of ω denoted n_ω. Consider two nodes n and m that share a PPI. We define the edge weight, $e_{\Delta\omega}$, for that interaction as:

$$e_{\Delta\omega} = |n_\omega - m_\omega| \tag{5.1}$$

We then calculate the average edge weight, w, of the entire network with:

$$w = \frac{\sum_{e \in E} e_{\Delta\omega}}{n_E} \tag{5.2}$$

where E is the edge set of the network (i.e., all PPIs) and n_E is the number of edges.
This section was performed by collaborator Gavin Conant.

5.2.6 Statistical analysis of the network weights

To assess if the average weight was smaller than would be expected by chance, we randomly reassigned the set of node weights $n_{\Delta\omega}$ 10,000 times and recalculated the value of the weight w_R for each of those random networks. We then asked where w from the real network fell in the distribution of w_R.

One might think that the signal of similarity in selective constraint is an artifact of a few protein complexes that define many pairwise PPIs as well as having a similar selective constraint across the complex. Were this situation the case, we would expect that interacting nodes were similar not only in their constraint but also in their connection degree. To test this possibility, we applied a modification of the above approach, making the edge weight for a pair of nodes the difference in their respective degrees (number of PPIs). We then calculated the overall network weight as before. We compared this weight to the weight seen for a distribution of 1,000 randomly rewired PPI networks that have identical node degrees but where the interaction identities have been scrambled (Conant and Wolfe, 2006). Contrary to the above prediction, proteins in the real network are *less* likely to interact with nodes of a similar degree of constraint ($p = 0.013$).
This section was performed by collaborator Gavin Conant.

5.2.7 Gene Ontology analysis

Gene Ontology (GO) analysis was performed on three datasets. The first two datasets were created from the PPIs used to infer the ancestral states of the PPI network (Figure 5.1), while the third dataset was created from PPIs used for detecting signals of adaptive evolution. We define three sets

of genes (Figure 5.4):

1. Genes that participated in at least one primate-specific PPI, although not exclusively in primate-specific PPIs, against all other genes from the nonself-interacting PPIs (hereafter primate-present interactions; PrimPresI)

2. Genes involved *exclusively* in primate-specific PPIs (i.e., they cannot also be present in other type of PPIs) against all other genes from nonself-interacting PPIs (hereafter primate-unique interactions; PrimUniqI)

3. Genes from PPIs for which mirrortrees had both branches in any given species with molecular rates $\omega > 0.5$ against genes from remaining mirrortree PPIs (hereafter adaptive interactions; AdaptI)

Note: Primate specific PPIs are present along the primate clade, but no longer present at the most internal node i.e., where the primate, rodent and horse-dog-cow clades diverge.

In order to perform GO analyses for all three datasets genes were matched to GO slim (The Gene Ontology project website, 2010b) via conversion of Ensembl IDs to human GO identifiers (The Gene Ontology project website, 2010a), which were obtained from the Gene Ontology website (`http://www.geneontology.org`) (Ashburner *et al.*, 2000). Then using custom Perl scripts (a gift from Takashi Makino to PB, and same ones applied in chapter 4) p-values for GO terms were calculated under the hypergeometric distribution and adjusted for by Bonferroni correction to account for multiple testing.

5.2.8 Protein degree and position in the protein interaction network

We computed the connectivity (degree or k) of all proteins. Since the connectivity is the number of interactions for an individual protein we wanted to compare whether the genes from the PrimPresI dataset had significantly different average connectivity from the remaining genes. We also compared the average connectivity of the genes from the PrimUniqI against

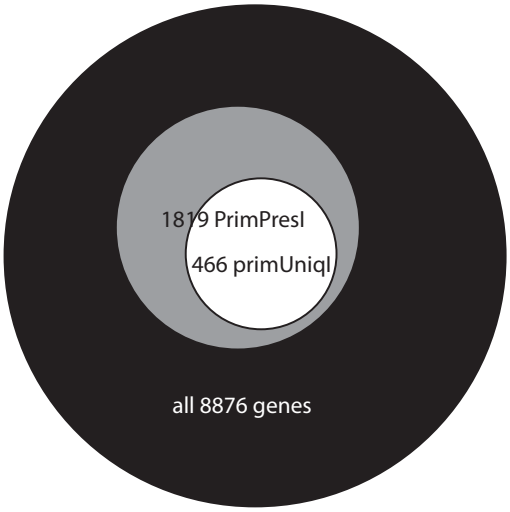


Figure 5.4: Gene sets used in the GO analyses of primate-specific protein interactions. The numbers in the Venn diagram are broken down as follows: 32,916 total PPIs are encoded by 8876 genes. Of those interactions, 2070 (encoded by 1819 genes) are found only along the primate-specific branch. Of the 1819 genes involved in those primate-specific interactions, 1,353 are also involved in other, nonprimate-specific interactions, and 466 are only involved in primate specific interactions.

all remaining genes. Next we calculated the pairwise connectivity difference between all PPI pairs since we wanted to compared the average connectivity difference of PPIs along the primate branch PPI (primate-specific PPIs) against the remaining PPIs i.e., those that are still present where the primate, rodent and horse-dog-cow clades diverge (nonprimate PPIs) in Figure 5.1.

5.3 Results

5.3.1 Reconstructing the ancestral states of the human protein interaction network

Given a set of previously described human PPIs (see chapter 4), we identified the orthologues of the genes involved in each PPI in seven other mammals (Methods 5.2.1). For analytical purposes, we assumed that if both orthologues of a pair of interacting proteins from human were present in another mammal, the PPI in question was also present (but see Discussion 5.4). Otherwise, it was assumed that the PPI was absent in that species. Using these data, we used parsimony to reconstruct the ancestral states of each human PPI in the phylogeny of Figure 5.1. Of course, given that our analysis is based on human, at the tip of the human branch 100% of the PPIs are present. However, at the internal node where the primate, rodent and dog-horse-cow clades diverge only 93.7% of the PPIs are inferred to have been present. We can thus identify and extract the 6.3% (or 2070) primate-specific PPIs and the 1819 genes that encode them.

5.3.2 Functional annotation of primate-specific PPI genes

To assess what these genes involved in primate-specific PPIs might be, we performed GO analyses on the genes encoding them. We first compared 1819 genes that were present in at least one primate-specific PPI (and potentially also in nonprimate PPIs; e.g., PrimPresI) compared to 7057 genes that were not involved in primate-specific PPIs (Figure 5.4). We found that the genes from PrimPresI were over-represented for biological process GO terms including "cell death", "response to stimulus" and "macromolecule metabolic processes". "Protein binding" is an over-represented molecular function, while under-represented functions included "channel activity" and "motor

activity" (Table 5.1). Using the same PPIs we also compared 466 genes encoding PPIs that are exclusively primate-specific (i.e., these genes are not part of any nonprimate PPI; PrimUniqI) against the remaining 8410 genes (Table 5.2). No biological process GO terms were over-represented among these genes, but five terms were under-represented: "cell communication", "regulation of biological process", "multicellular organismal development", "macromolecule metabolic process" and "cell differentiation". Similarly, one molecular function term "protein binding" was under-represented.

5.3.3 Protein degree of primate-specific PPI genes

We also asked if the proteins involved in primate-specific interactions ('primPPIs') differed from the remainder of the network in their network degree. Genes that participated in at least one primate-specific PPI ('primate genes' i.e., the PrimPresI dataset mentioned in Methods 5.2.7) are compared against those genes that never participate in primate-specific PPIs ('nonprimate genes'). Proteins with at least one primate-specific interaction had a much higher mean interaction degree than did other proteins $k_{PrimPresI} = 15.92523$ vs $k_{other1} = 5.489301$ (Table 5.3). This difference is statistically significant (W = 8923794, $p < 2.2 \text{ x } 10^{-16}$, Wilcoxon two-sample test), suggesting that the primate-specific PPIs are not wholly specific but rather that proteins with large numbers of interactions are more likely to also have a primate-specific interaction simply by chance. We further compared the connectivity of genes that exclusively participate in primate-specific PPIs ('exclusively primate genes' i.e., the PrimUniqI dataset) above against all remaining genes; $k_{PrimUniqI} = 4.034335$ vs $k_{other2} = 7.82711$ (Table 5.4). This difference is also statistically significant (W = 1328452, $p < 2.2 \text{ x } 10^{-16}$). We examined the most highly connected proteins in each set: Amongst the primate genes gene ENSG00000170027 has the highest degree ($k = 240$; 1 self-interaction + 239 nonself-interactions), and is a small-scale duplication (SSD) gene according to the classification from the previous study described in chapter 4. According to the Ensembl Genome Browser `http://www.ensembl.org` this protein is encoded by the YWHAG gene, a tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein. The protein of Ensembl gene ENSG00000173039 is also a SSD gene, and has the highest degree amongst the nonprimate proteins with $k_{max} = 99$ (no self-interaction). This protein is annotated as a RELA v-

Table 5.1: Over- and under-represented (italics) GO terms of genes present at least once in a primate-specific PPI. 1819 genes present in primate PPIs vs 7057 genes never observed in primate PPIs.

Biological Processes							
GO IDs	Term	Obs.	Exp.	S.D.	Z score	p-value[a]	Fold enrichment
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	542	490.776	16.4	3.1	3.26×10^{-2}	1.16
GO:0008219	cell death	233	156.533	10.5	7.3	1.11×10^{-10}	1.74
GO:0043170	macromolecule metabolic process	897	818.205	17.9	4.4	2.02×10^{-4}	1.14
GO:0050896	response to stimulus	432	328.055	14.6	7.1	4.50×10^{-11}	1.46
GO:0051704	multi-organism process	128	83.576	8.3	5.4	1.64×10^{-6}	1.82
Molecular Functions							
GO IDs	Term	Obs.	Exp.	S.D.	Z score	p-value[a]	Fold enrichment
GO:0003774	*motor activity*	*4*	*16.401*	*3.6*	*-3.5*	4.26×10^{-3}	*0.21*
GO:0005515	protein binding	1276	1138.038	16.7	8.3	3.58×10^{-15}	1.17
GO:0015075	*ion transmembrane transporter activity*	*45*	*72.377*	*7.4*	*-3.7*	2.86×10^{-3}	*0.57*
GO:0015267	*channel activity*	*19*	*45.917*	*5.8*	*-4.7*	2.38×10^{-5}	*0.36*

[a]The estimated P values were adjusted by Bonferroni correction

Table 5.2: Over- and under-represented (italics) GO terms of genes in exclusively present in primate-specific PPIs. 466 primate-specific genes vs 8410 nonprimate-specific genes.

Biological Processes							
GO IDs	Term	Obs.	Exp.	S.D.	Z score	p-value[a]	Fold enrichment
GO:0007154	*cell communication*	*82*	*131.136*	*9.3*	*-5.3*	*3.95 x* 10^{-7}	*0.57*
GO:0050789	*regulation of biological process*	*169*	*214.665*	*9.6*	*-4.7*	*3.03 x* 10^{-5}	*0.73*
GO:0007275	*multicellular organismal development*	*55*	*87.499*	*8.0*	*-4.1*	*3.84 x* 10^{-4}	*0.58*
GO:0043170	*macromolecule metabolic process*	*158*	*193.383*	*9.70*	*-3.6*	*3.00 x* 10^{-3}	*0.75*
GO:0030154	*cell differentiation*	*27*	*48.05*	*6.39*	*-3.3*	*7.60 x* 10^{-3}	*0.51*
Molecular Functions							
GO IDs	Term	Obs.	Mean	S.D.	Z score	p-value[a]	Fold enrichment
GO:0003674	unknown molecular function	32	15.454	3.7	4.4	3.76 x 10^{-3}	2.08
GO:0005515	*protein binding*	*232*	*276.514*	*9.2*	*-4.8*	*7.33 x* 10^{-5}	*0.80*

[a]The estimated P values were adjusted by Bonferroni correction

rel reticuloendotheliosis viral oncogene homolog A. Finally, amongst the genes that exclusively participate in primate-specific PPIs the gene with most connections is ENSG00000213416 (k = 72; 1 self-interaction + 71 nonself- interactions), a gene that exists in single-copy according to results from chapter 4. The protein is annotated as KRTAP4-12 keratin associated protein 4-12.

Table 5.3: Connectivity (node degree) statistics for genes involved in at least one primate-specific PPI, and genes never involved in primate-specific PPIs i.e. nonprimate PPIs.

Measure	primate genes	nonprimate genes
k_{min}	1.00	1.00
k_{max}	240.00	99.00
k_{mean}	15.93	5.49
Measure	*exclusively* primate genes	nonprimate genes
k_{min}	1.00	1.00
k_{max}	72	240
k_{mean}	4.03	7.83

Proteins tend to interact with other proteins of a different degree. We therefore investigated if this trend was consistent between the primate-specific PPIs ('primPPIs'), and the nonprimate PPIs ('nonprimPPIs'). For this purpose we calculated the absolute degree difference for all protein interactions in the two datasets:

$$\Delta k = |k_1 - k_2| \tag{5.3}$$

where k_1 and k_2 are the node degrees of the two proteins involved in a PPI. The average absolute node degree difference for primate-specific PPIs, $\Delta k_{primPPIs}$ is 29.1, while the average absolute node degree difference for nonprimate PPIs was larger, ($\Delta k_{nonprimPPIs}$ =39.4). This difference is statistically significant (Figure 5.5; W = 28577804, p = 1.233 x 10^{-15}, Wilcoxon two-sample test). Note however, that the maximum difference in degree for the two datasets differed little ($\Delta k_{maxprimPPIs}$ = 223 vs $\Delta k_{maxnonprimPPIs}$ = 239 ; Table 5.4), probably because the highly interacting gene YWHAG was a member of both PPI datasets.

Table 5.4: Absolute connectivity difference between proteins in primate-specific and nonprimate PPIs.

Measure	primate-specific PPIs	nonprimate PPIs
Δk_{min}	0.00	0.00
Δk_{max}	223.00	239.00
Δk_{mean}	29.12	39.37

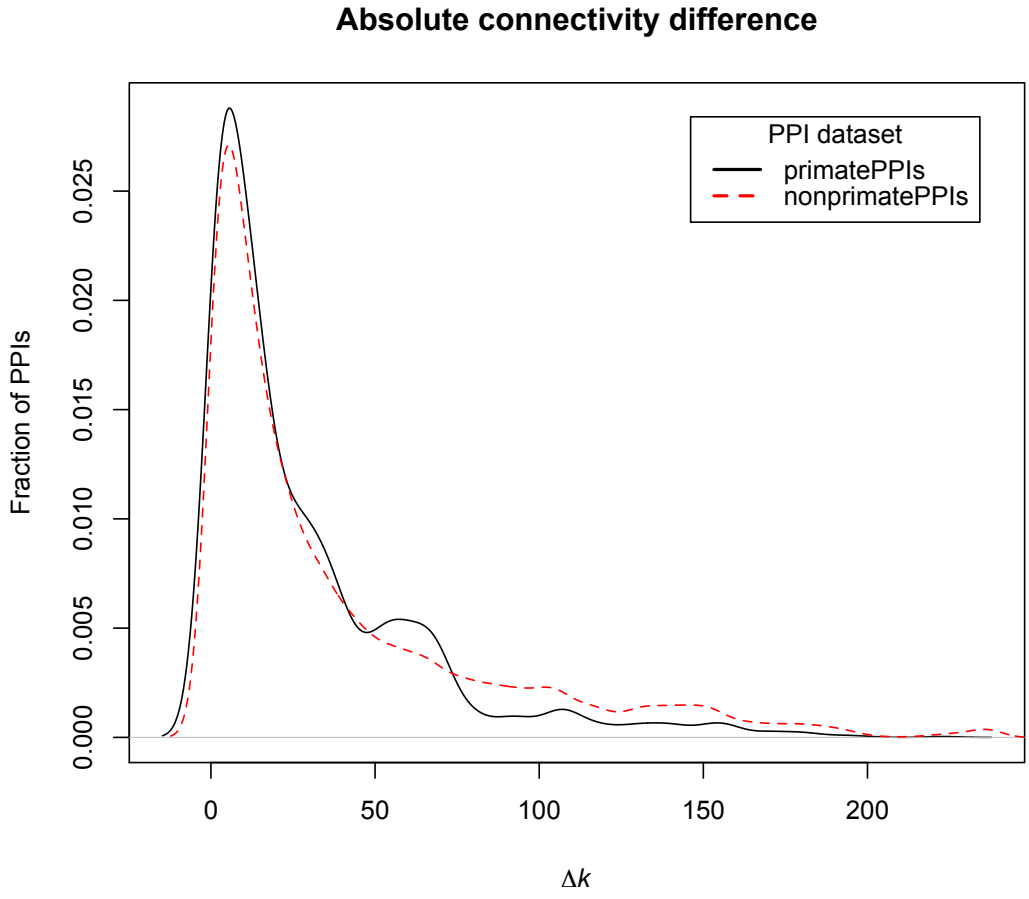


Figure 5.5: Comparison of distribution of absolute degree difference, Δk, between primate and nonprimate PPI pairs. The distribution of the difference in degree for a pair of interaction proteins (absolute degree difference, Δk; (x-axis) is shown. In black are the primate-specific interactions (primatePPIs), while red gives the remainder of the interactions.

5.3.4 Patterns of co-evolution support the ancient nature of most human PPIs

If a PPI is selectively important, one might expect that sequence changes on the surface of one member of the pair would need to be compensated for with changes in the other member. In other words, we might expect the protein pair to co-evolve. We therefore sought to assess the evidence for co-evolution among the PPI pairs by calculating the correlation coefficient of the selective constraint from the matched branches from each of the two genes. Next, in order to compare the distribution of the Spearman's ρ of the actual i.e., true PPIs (t-PPIs, see Methods 5.2.3) against a distribution of randomised, non-interacting gene pairs (pseudo-random or pr-PPIs; Methods 5.2.3). We thus created 1000 pr-PPI datasets with the same number of interactions as in t-PPI. We then compared the mean of the Spearman's ρ from t-PPI to the distribution of the means of Spearman's ρ from the 1000 pr-PPI datasets. None of the mean Spearman's ρs from the pr-PPI datasets were as large as that observed in t-PPI (Table 5.5). Omitting branches with $\omega \geqslant 5.0$ from the analysis did not change our conclusions (Table 5.5). We can therefore conclude that there is a significant association in selective constraint between pairs of interacting proteins.

The analysis presented in Figure 5.1 uses gene orthology (or its absence) as a proxy for the presence of a PPI, an approach with obvious flaws since finding the orthologues of the genes in a PPI does not necessarily mean the orthologous genes also interact in the sister lineage to form a PPI. To complement that work, we wanted to ask if the correlation in selective constraint between interacting genes tended to be weaker in parts of the tree most evolutionarily distant from humans, as one would predict if there were a significant number of evolutionarily novel PPIs in these data that do not exist in (to human distant) species such as horse and dog. Thus, we would predict that the signal of co-evolution might be stronger if only closer relatives of humans were considered. We thus sequentially removed each tip-branch from the phylogeny of Figure 5.1, and repeated our co-evolution analysis. In general, our results were in accord with expectations, i.e., the difference between the mean correlation from t-PPI and the mean of means in the pr-PPI datasets increased when more distant taxa were omitted from the analysis (Table 5.5). However, we observe two exceptions to this rule: human and macaque, a phenomenon that we do not currently have an explanation

for.

5.3.5 Weak evidence for shared instances of adaptive evolution between PPI partners

We sought to assess if there were pairs of genes involved in a PPI that both shared an instance of adaptive evolution (e.g., ω >1.0) along the same branch of the phylogeny. We thus computed the number of paired branches from our mirrortrees where both genes had $\omega > 1.0$. However, the number of such paired cases of $\omega > 1.0$ was not significantly greater than what would be expected given the overall number of cases of $\omega > 1.0$. However, when we lowered the threshold to $\omega > 0.5$, we found that, with exception for macaque, all branches we observed more such cases than we would expect by chance alone (Figure 5.6). Although this is not statistically significant for all cases. Finally we also performed a GO analysis on the 524 genes that had paired branches in the mirrortrees with $\omega > 0.5$. These 524 genes were over-represented for biological process GO terms "cell death" and "response to stimulus", and the molecular function GO terms "protein binding" and "receptor activity" (Table 5.6; 'AdaptI' dataset, see Methods 5.2.7).

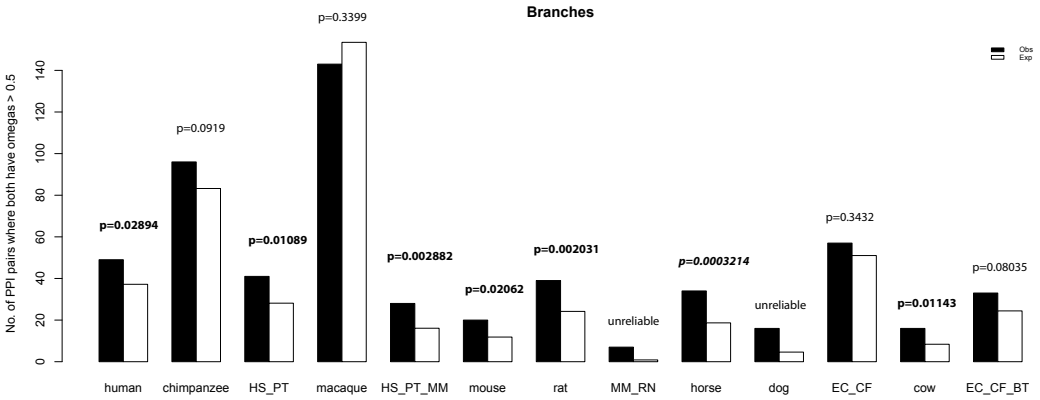


Figure 5.6: Branch-pairs where both branches have $\omega > 0.5$. We observe a trend, which indicates that if one branch in a mirrortree branch-pair has a weak selective constraint, ω, then its corresponding branch is likely to show a similar lack of constraint.

Table 5.5: Co-evolution signal detection using Spearman's Rank Sums correlation coefficient ρ comparison of real and pseudo datasets

Dataset[a]	$\overline{\rho_{pseudo}} \geqslant \overline{\rho_{real}}$	p-value[b]	$\overline{\rho_{REAL}}$[c]	$\overline{\rho_{FAKEs}}$[d]	Δ_{means}[e]
Full dataset (includes $\omega \geqslant 5$ branches), 7730 PPIs	0	< 0.001	0.1313473	0.122394398	0.008952902
Reduced dataset (only $\omega < 5$ branches), 7727 PPIs	0	< 0.001	0.1316623	0.122434393	0.009227907
HUMAN branch removed, (only $\omega < 5$), 7705 PPIs	0	< 0.001	0.1233236	0.110367582	0.012956018
CHIMP branch removed, (only $\omega < 5$), 7668 PPIs	0	< 0.001	0.1024526	0.099028617	0.003423983
MACAQUE branch removed, (only $\omega < 5$), 7303 PPIs	0	< 0.001	0.1261796	0.108342581	0.017837019
MOUSE branch removed, (only $\omega < 5$), 7173 PPIs	7	0.007	0.1322071	0.128107524	0.004099576
RAT branch removed, (only $\omega < 5$), 7132 PPIs	3	0.003	0.1369175	0.132354684	0.004562816
HORSE branch removed, (only $\omega < 5$), 6937 PPIs	1	0.001	0.1388224	0.130984564	0.007837836
DOG branch removed, (only $\omega < 5$), 6930 PPIs	5	0.005	0.1347147	0.128179691	0.006535009
COW branch removed, (only $\omega < 5$), 6785 PPIs	11	0.011	0.1331633	0.127107383	0.006055917
HS-PT clade removed, (only $\omega < 5$), 7563 PPIs	0	< 0.001	0.0951722	0.0837221	0.0114501
HS-PT-MM clade removed, (only $\omega < 5$), 6113 PPIs	0	< 0.001	0.06980665	0.053817904	0.015988746
MM-RN clade removed, (only $\omega < 5$), 5123 PPIs	91	**0.091**	0.1060473	0.110924196	-0.004876896
EC-CF clade removed, (only $\omega < 5$), 5893 PPIs	61	**0.061**	0.1407103	0.137595771	0.003114529
EC-CF-BT clade removed, (only $\omega < 5$), 3456 PPIs,	421	**0.421**	0.1652225	0.169144657	-0.003922157

[a]The number of PPIs (size n) of each dataset type varies because Spearman's ρ correlation coefficients could only be calculated for mirrortrees with at least 6 branches in common. For each real dataset type there are 1000 pseudo dataset (i.e. the randomisation is bootstrapped 1000 times), and each of the pseudo datasets are of the same size n as the equivalent real dataset.

[b]Number of times the mean of means of Spearman's ρ of the 1000 pseudo datasets are larger or equal to the mean value of the real dataset's Spearman's ρ. Non-significant tests are shown in bold.

[c]Spearman's ρ mean value for the real dataset

[d]Mean value of 1000 Spearman's ρ mean values

[e]Mean difference between the mean of the real dataset and the mean of means from the pseudo datasets

Table 5.6: Over- and under-represented (italics) GO terms of genes present in PPIs where proteins in the protein pair have $\omega > 0.5$. 524 genes with $\omega > 0.5$ for both branches vs remaining 4506 genes (of 5030 genes in total from 12472 PPIs for which mirrortrees could be constructed with reliable ML scores)

Biological Processes							
GO IDs	Term	Obs.	Exp.	S.D.	Z score	p-value[a]	Fold enrichment
GO:0008219	cell death	75	45.44	6.1	4.8	9.91 x 10^{-5}	1.81
GO:0050789	regulation of biological process	289	257.638	10.6	2.9	2.98 x 10^{-2}	1.15
GO:0050896	response to stimulus	142	86.132	8.1	6.9	5.33 x 10^{-10}	1.81
GO:0051704	multi-organism process	39	21.495	4.4	4.0	3.42 x 10^{-3}	2.03
Molecular Functions							
GO IDs	Term	Obs.	Mean	S.D.	Z score	p-value[a]	Fold enrichment
GO:0005515	protein binding	384	336.277	9.7	4.9	8.05 x 10^{-6}	1.18
GO:0004872	receptor activity	71	49.529	6.3	3.4	2.68 x 10^{-2}	1.55

[a]The estimated P values were adjusted by Bonferroni correction

5.3.6 Proteins interact with other proteins of similar constraint more often than expected

We developed a network-based test of whether selective constraints are distributed at random in the PPI network (see Methods 5.2.5), hypothesising that interacting proteins would tend to have similar selective constraints. Briefly, we compared the selective constraint, ω, between pairs of proteins that interact in the PPI network. The distribution of selective constraint in the network is highly non-random: the average difference in selective constraint between two interacting proteins is 0.10, as compared to an average of 0.12 seen when the ω values are distributed at random ($p < 0.0001$; Methods 5.2.6). This confirms the results of Vinogradov, obtained with a completely distinct approach (Vinogradov, 2010).

5.4 Discussion

Using human PPI data from the experimentally-validated and literature-curated Human Protein Reference Database (HPRD) (Peri *et al.*, 2003; Keshava Prasad *et al.*, 2009), we explored two questions: the age of these interactions, and the selective pressures acting on them. With regards to the first question, it appears likely that most modern human PPIs already existed at the time of the common ancestor of the eight mammals considered. Qian *et al.* (2011) have estimated the evolutionary rate of PPI apperance to be 2.6 x 10^{-10} new interactions per year in yeast. Based on this rate estimate, they inferred that a protein interaction known in humans had a 98% chance of being present in mouse, declining to 77% in fruit flies and to 66% in plants. Although their estimation of PPI presence of 98% is substantially higher than the 87.5% estimated here from our parsimony analysis, the two analyses (done in very different taxa) are in general agreement that most PPIs are evolutionarily ancient. Given this conclusion, it is not surprising that Pellegrini and colleagues were able to use phylogenetic profiles of gene presence and absence and PPI information in baker's yeast to predict protein interactions in other yeast species (Pellegrini *et al.*, 1999; Marcotte *et al.*, 1999). Although these methods cannot distinguish between co-evolution due to compensatory changes between interacting sites and simply co-variation due to similar expression levels, a high percentage of predicted interactions are true interactions (Goh *et al.*, 2000; Pazos and Valencia, 2001; Hakes *et al.*,

2007b). An important caveat, however, is that these approaches cannot distinguish between direct interactions between two proteins and proteins that are members of the same protein complex (Lovell and Robertson, 2010).

Of course, our approach is nevertheless potentially biased by the fact that we assess orthology and not interaction directly. Thus, we are in fact setting an upper bound on the number of interactions shared. Given this bias, we performed GO analyses on the set of unambiguously new interactions, the primate-specific interactions that comprise roughly 6% of the interactions surveyed. A number of core organismal processes (e.g., "multicellular organismal development", and "cell differentiation") were under-represented among these genes, presumably representing the conserved nature of these types of interactions across mammals. Similarly, the over-representation of terms such as "cell death" and "response to stimulus" might represent novel selective environments in primates, such as immune or aging-related selection for programmed cell death. We also observe differences in the connectivity of genes participating in primate-specific PPIs. While the higher degree of proteins participating in both primate (PrimPresI) and nonprimate interactions can be attributed to the greater likelihood of a high-degree protein having at least one interaction being primate specific, this rule cannot explain the higher than average degree of proteins participating exclusively in primate-specific interactions, and we currently have no hypothesis as to the origins of this trend.

Despite the biases in our parsimony analysis, our overall conclusion that most human protein interactions are ancient is supported by a second analysis, which shows that there is a signal of co-evolution among PPI pairs, even among mammals that are reasonably distant evolutionarily from humans (e.g., horse, dog and cow). Our analysis involved comparisons of the selective constraint (K_a/K_s i.e., ω) along matched pairs of branches in the eutherian phylogeny for interacting proteins (Figures 5.1 and 5.2). Of course, co-evolution has been previously used to identify sites with interacting residues. However, many of these methods are site-specific, and require a three-dimensional protein structures (Pollock *et al.*, 1999; Fares and Travers, 2006; Fares and McNally, 2006). Other methods detect co-evolution by calculating the pairwise distances between sequences of proteins known to interact, the values of which are then used as references for predicting other interactions (Goh *et al.*, 2000; Pazos and Valencia, 2001). Interestingly, although the 'binding neighbourhood' of interacting proteins

give the strongest co-evolutionary signal, co-evolution can also be detected at other sites in the two proteins (Kann *et al.*, 2009). Co-evolution can also extend beyond direct protein interactions: only after the network distance between two proteins exceeds 3 is the co-evolution signal lost (Liang *et al.*, 2010).

This analysis also shed light on our second question, that of the selective forces acting on the protein interactions. These data suggest significant and abiding selection that acts to maintain interactions. This conclusion is supported by our finding that the difference in average selective constraint between interacting proteins is smaller than expected, a result consistent with the findings of Vinogradov made with a completely distinct approach (Vinogradov, 2010). The obvious question raised by these observations is the exact nature of the selection at work. Fernández and Lynch have recently argued that many of the protein interactions in multicellular eukaryotes originated not through selection for novel functions but rather as a means for stabilizing existing functions in the face of the destabilizing forces of genetic drift (Fernández and Lynch, 2011). No results in the present work contradict this hypothesis as a general principle, but there are two points that suggest that it is probably not an exclusive process. First, co-evolution is not a strong prediction of a model where specific interactions are not under selection but instead there is generalised selection for enough interactions to maintain protein function. Secondly, we have (very weak) evidence for shared instances of directional selection in interacting pairs.

5.5 Conclusions

Using known protein interactions in human and their orthologous genes in 7 other species we constructed a phylogenetic profile of PPI presence/absence, and derived the protein network's internal nodes through parsimony, Then, using a modified mirrortree method (calculating selective constraints of each branch rather than calculating a correlation coefficient score for every amino acid pair in the MSAs) to detect co-evolution signal we can confirm that our estimation of PPI presence/absence through the ancestral nodes are significantly different from randomly generated pseudo PPIs. Thus validating our results, which seems to indicate that most human PPIs are relatively ancient, from the radiation near to placental mammals. We further found

that proteins of similar evolutionary constraint interact with each other more often than would be expected by chance alone. Still we need to bear in mind that the only experimentally validated PPIs come from human, while we inferred PPIs in the other seven species using gene orthology. This approach raises two issues. Firstly, depending on the species, genes that might be present could still not be detected due to poor genome coverage. Even though both orthologues of a human PPI are present in another species, this in itself does not automatically mean the protein products of these genes also interact in the sister lineage. However, given that we do find both orthologues to most of the human PPI pairs we have no reason to believe that these interactions first appeared in the human lineage. It seems much more intuitive then that the interactions are also old. Randomised gene pairings (such as those currently performed on the mirrortree subset of PPIs) of the orthology generated PPIs in the sister lineages would be a good indication as to whether these pairs are real, ancient and selectively meaningful. Furthermore, future analyses using mouse and even yeast PPI data could help elucidating which PPIs in human that are real and old. In combination with these experimental PPI datasets more genomes such as elephant, kangaroo, platypus and echidna could be added for more estimations of PPI presence/absence using gene orthology in order to shed light on how protein interactions have evolved along the mammalian lineage.

Chapter 6

Conclusions

Given that proteins interact with each other, DNA and RNA in order to perform the complex tasks that keep cells alive one needs to look at the intricate ways proteins interact with their surroundings in order to begin to understand protein function and as a consequence the constraints on their evolution. The three studies presented here attempt to shed some light to some of these questions. Therefore, in chapter 3 the role interactions have, between non-enzyme proteins (acting as mediators) and enzyme proteins, in the spatial organisation of metabolic networks, and their possible contribution to the network modularity, is investigated. Meanwhile, in chapter 4 the duplicability of self-interacting genes (that is, proteins encoded by the same gene interacting with each other i.e., forming protein 'homodimers') is investigated. It has previously been shown that after the duplication of self-interacting genes new interactions between the old gene's protein and the protein of the new duplicate gene (i.e., the now paralogous gene pair) will form as the two genes initially are identical. However, with time the two paralogues will diverge, and if the interaction between them remains a heterodimer has been formed, which is how protein complexes arise (Pereira-Leal *et al.*, 2007). If only one or a few genes encoding proteins in a protein complex or a pathway are duplicated while the remaining genes are not this will cause gene dosage imbalance in the cell, which can have detrimental effects. Homodimers are *de facto* balanced so this is not an issue. However, the self-interacting genes encoding the homodimers in general have more protein interaction partners, which could explain why the self-interacting genes more often seem to have arisen through whole genome

duplication (WGD) rather than small-scale duplication (SSD). Moreover, the high number of interaction partners can be explained by the age, as WGD genes in general are older than SSD genes. In fact, when age is controlled for the difference in connectivity between WGD and SSD genes disappears (see chapter 4). Finally from the research presented in chapter 5 we can conclude that most interactions in human are ancient interactions responsible for amongst other things old, non-human specific biological processes such as "multicellular organism development" and "cell development".

However, a caveat to all conclusions drawn from any work investigating protein interaction data is two-fold. Firstly, the protein interaction data are incomplete and contain false positives. Based on missing data and the false-positive rate it is therefore estimated that only 10% of the human protein interaction network (PIN) is completed Hart *et al.* (2006). Secondly, PIN datasets are many times biased (as will be elaborated later). Since proteins can have several interactions when calculating the false-positive rate it is appropriate to count interactions rather than proteins. When von Mering *et al.* (2002) benchmarked the quality of several yeast protein interaction datasets against a trusted, manually annotated dataset they found very little overlap between the different datasets. They also estimated that high-throughput methods contained at least 50% false positives. This was based on observations that the datasets contained interactions between functionally unrelated proteins, and from different cellular compartments. They further noticed that no dataset contained interaction information for more than 60% of all yeast proteins, and with the exception of the Y2H and synthetic lethality datasets, they were all biased to contain proteins that were abundantly present in the cell. In fact, according to the study by Bloom and Adami (2003) the correlation between slower evolutionary rate for highly connected proteins, as reported by (Fraser *et al.*, 2002, 2003), is simply an artifact caused by methods identifying more interactions for abundant proteins.

As mentioned above it is estimated that only approximately 10% of the protein interactions in human have been correctly identified (Hart *et al.*, 2006). Choosing a good quality dataset is therefore of utmost importance. When deciding on which dataset to use for the research studies described here (chapters 3- 4) the Human Protein Reference Database (HPRD) (release 7) was chosen because it is superior in numbers (identified interactions and unique number of proteins), and contains experimentally-validated and

literature-curated interactions (Peri *et al.*, 2003). Morever, in a benchmark it outperformed other human protein interaction databases in every aspect tested (Mathivanan *et al.*, 2006). As the HPRD data were downloaded for compilation in chapter 4 they contained 37,107 protein-protein interactions (PPIs), encoded by 9303 unique genes. After data curation, which amongst other things, involved only keeping HPRD entries that could be matched to Ensembl release 50 identifiers our PPI dataset consisted of 34,795 PPIs among 8881 unique genes (Figure 4.1). These numbers can be compared to other human PPI datasets e.g., Stelzl *et al.* (2005) dataset which contains 3186 PPIs among 1705 genes, or the CCSB-HI1 dataset containing ∼2800 interactions, which according to the authors' own estimation is ∼1% of the human PIN (Rual *et al.*, 2005). Another advantage with the HPRD dataset, over the human PPI dataset compiled by Ewing *et al.* (2007), which contains approximately as many interactions as the HPRD data, is that none of the methods for protein interaction identification used for the HPRD data are mass spectrometry (MS) methods, which is the case for the PPIs dataset of Ewing and colleagues. Although MS methods are advantageous in the sense that using a high-throughput methods allow for the processing of lots of data in one experiment, and specifically is capable of identifying protein complexes (although not the specific interactions between proteins within a complex) (von Mering *et al.*, 2002; Shoemaker and Panchenko, 2007) these methods also generate biased data counting more interactions for abundant proteins (Bloom and Adami, 2003). In an attempt to discard false positives several studies (including Ewing *et al.* (2007)) apply filters whereby considering only those interactions that have been identified by at least two different methods. For example, in the dataset by Stelzl *et al.* (2005) interactions are classified as low, medium or high confidence interactions. From that dataset, originally consisting of 3186 PPIs, only 911 PPIs remained as high-confidence interactions, thus reducing the dataset substantially. Although one could of course believe that it is better to have few interactions of good confidence than many of low confidence it is not as simple as that as the entire network topology alters substantially for each quality control that removes interactions, and as a consequence the results from these datasets change as well. An example of this effect is a study where negative correlation was found between the protein connectivity and the average connectivity of its neighbours (supporting a hub-centric network view). However, when the data was filtered to only keep interactions that have been identified by

two methods there was no correlation, and the correlation become positive if only interactions identified by three methods were kept in the dataset (Hakes *et al.*, 2008). Thus, not only are the conclusions drawn from the data not robust against the size of the dataset, but this is also problematic for another reason because even the original raw data is a small subset of all interactions (estimated to ∼10% in human and ∼50% in yeast). Therefore the filtered data is an even smaller subset of the original subset of all protein interactions, and since the results change with the dataset size it is obvious that no reliable conclusions can be made from these data subsets. Hakes *et al.* (2008).

In the publication together with the first release of HPRD the authors note that 44% of the proteins in their database have at least 3 interactions, and based on this they estimate the actual number of PPIs in the human PIN to be >200,000 (Peri *et al.*, 2003), while another group estimate there are between 154,000 to 369,000 human PPIs (Hart *et al.*, 2006). As mentioned above, the PPI dataset used in the studies of this thesis consists of 34,795 PPIs among 8881 genes matched to Ensembl release 50 identifiers. Estimations of how many protein interactions there may be in the PIN may vary since this estimation is not completely straightforward given that the degree distribution of protein networks follows a power-law (Figure 1.3), and thus although very many proteins will have few interactions there is no way of knowning how many 'few' interactions are. Moreover, a decreasing number of proteins will have very many interactions. However, we can calculate that since the Ensembl release 50 core data consists of 21,785 unique genes with peptide sequences, interactions have only been found for 8881 genes i.e., ∼41% of all genes, and even for those genes there could still be missing, yet unidentified interactions (as well as false positives). However, to do a direct comparison to the 44% of genes with at least 3 interactions in the original HPRD dataset there are 5392 genes or almost 61% of the 8881 genes in the HPRD PPI dataset used in the studies of this thesis that have at least 3 interactions. Meaning between the first and seventh release of HPRD the available data has, obviously, increased. Furthermore, of the 34,795 PPIs 32,916 were nonself-interactions encoded by 8876 genes (see chapters 3 and 5). Note that although 8876 genes had nonself-interactions 1879 of all genes also self-interact (see chapter 4). This means that of all genes in our dataset only 5 genes exclusively interact with protein copies of itself to exclusively form homodimers. However, since we only have interactions for ∼41% of all protein encoding genes, and probably only have ∼10% of all actual protein

interactions we cannot conclude whether it is truly unusual for proteins to only have self-interactions or not. As mentioned at the beginning of this chapter a previous study has shown that protein complexes evolve through the duplication of homodimers (i.e., self-interacting proteins) (Pereira-Leal *et al.*, 2007), which is why it would have been interesting to know with what frequency there are genes encoding proteins with only this type of interactions.

Another very important issue to remember when considering protein interaction networks is that they give a static picture of the interactions. As a concrete example gene ENSG00000170027 has the highest node degree (number of protein interactions) with 239 nonself-interactions and 1 self-interaction (see chapter 5). However, it is hard to imagine any protein having all these interactions at the same time. In fact, this protein is tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein. Thus, even from its name (containing the words "activation protein") it can be derived that this protein interacts with many other proteins at different times. If expression data was to be layered together with the protein interaction information we could further elucidate which interactions are static and which are transient. This would further increase our understanding of how proteins interact to run the cellular machinery.

Missing data and false positives, as elaborated above, cause problems, and scientists should be aware of what these issues are when they interpret their results, and seek to repeat their studies as improved protein interaction datasets become available. However, even though we currently have such an incomplete protein interaction dataset important research can still be performed. A very good example is the study by Rambaldi *et al.* (2008). They found that many genes involved in cancer often existed in single-copy, and encoded for proteins with many protein interactions. They further found it was these network features, rather than a common function or cellular location, that all the cancer genes investigated had in common. As illustrated in Figure 1.5, scale-free networks are more robust than random networks because the loss of a random node (as very many nodes have few interactions) will still keep the network structure intact. However, if a hub node (i.e., a highly connected node) is removed from the network, a substantial part of, or the entire network structure collapses. In the case of the cancer genes Rambaldi and colleagues identified this is what happens. A gene, encoding for a protein with many interactions, and existing only in single-copy becomes

defective, and thus the protein is either not encoded for or cannot bind to other proteins. Either way the network is dependent on these hub proteins, and when they cannot bind to other proteins the network structure falls apart, and the cell becomes cancerous. Thus, although we lack a full picture of the protein interaction network we do have good coverage of the human genome, and constantly other genomes too are being sequenced and their coverage improved. Thus we can still do our best possible research using these gene data together with the protein interaction data. For example, in chapter 3 protein and metabolic network data for *E. coli*, baker's yeast and human are used as raw material (see chapter 2) to study how proteins not themselves part of metabolic reactions bring metabolic reactions physically close, thus channelling reactions and contributing to the spatial organisation of metabolism. The results are consistent for all organisms, and since the protein interaction and metabolic network data are compiled by different researchers and for different species, it is unlikely that these datasets have the same consistent errors and biases we can be confident we have at least avoided some of the pitfalls inherent in using incomplete, erroneous and biased datasets. By using the available data with care, we can still make important biological discoveries when we apply a network perspective to our studies. That will help us make connections and draw conclusions that would not have been obvious if the genes and proteins had been studied as isolated entities instead of members of intricate networks. We always need to be cautious, and opened minded, prepared to re-evaluate our standpoints as new and better quality datasets are produced. Moreover, bioinformatics research can generate ideas for laboratory experiments conducted *in vitro* and/or *in vivo* (i.e., so called 'web labs'). Bioinformatics should not be overlooked simply because it consists of computational and statistical analyses. The data used for these analyses often ultimately come from wet lab experiments, and provide powerful opportunities to integrate the many individual lab-based observations to achieve higher level of understanding of biological functions. Looking only at genes, proteins or metabolic reactions in isolation in a test tube can prevent revealing what truly goes on in the cell. Knowing the minute details of a single protein complex is as important as understanding the general biological trends we can only observe and conclude from conducting genome-wide studies where the specifics of no single gene, protein or reaction is singled-out. (Although as pointed out earlier in this chapter unfortunately simply because data is large-scale does not exempt it from bias or guarantee

it to be of high quality, and as usual results can only be as good as the data used to generate it in the first place.) Together bioinformatics and 'wet lab' experiments contribute to our added biological understanding as long as we keep in mind that no experiment or method holds the absolute truth, but merely contributes to our added knowledge as we strive to understand the wonders of biological life.

This thesis contributes to the literature by showing how human genes that encode for proteins that form homodimers have a higher duplicability compared to other genes. This had previously been shown in simpler eukaryotes such as yeast, worm and fly, and in human (see chapter 4) using a dataset that was not publicly available for free (Ariadne Genomics, `http://www.ariadnegenomics.com`), which made it difficult to verify. This thesis shows that the same is true using the free and publicly available human dataset from Ensembl and HPRD. Furthermore this thesis presents indirect proof as to how certain protein interactions (namely *indirect* protein-protein interactions), that are not themselves part of metabolic reactions, help mediate such reactions by bringing enzymes in metabolic pathways into close proximity so that they can share metabolites and structure the organisation of metabolism. Finally this thesis identifies orthologues to the human genes encoding the protein-protein interactions giving strong indications that most protein-protein interactions present in human today are old at least to near the radiation of placental mammals.

Bibliography

Albert, R. (2005). Scale-free networks in cell biology. *J Cell Sci*, **118**(Pt 21), 4947–57.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–410.

Aon, M. A., O'Rourke, B., and Cortassa, S. (2004). The fractal architecture of cytoplasmic organization: scaling, kinetics and emergence in metabolic networks. *Mol Cell Biochem*, **256-257**(1-2), 169–84.

Araiza-Olivera, D., Sampedro, J. G., Mujica, A., Pena, A., and Uribe-Carvajal, S. (2010). The association of glycolytic enzymes from yeast confers resistance against inhibition by trehalose. *FEMS Yeast Res*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., SS, S. S. D., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., GM, G. M. R., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**(1), 25–29.

Barabási, A. L. (2009). Scale-free networks: a decade and beyond. *Science*, **24**(325), 412–413.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–12.

Barabási, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5**(2), 101–13.

Bekaert, M. and Conant, G. C. (2011). Copy number alterations among mammalian enzymes cluster in the metabolic network. *Mol Biol Evol*, **28**(2), 1111–1121.

Blattner, F. R., III, G. P., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997). The complete genome sequence of escherichia coli k-12. *Science*, **277**(5331), 1453–1462.

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

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004). Go::termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, **20**(18), 3710–3715.

Brandina, I., Graham, J., Lemaitre-Guillier, C., Entelis, N., Krasheninnikov, I., Sweetlove, L., Tarassov, I., and Martin, R. P. (2006). Enolase takes part in a macromolecular complex associated to mitochondria in yeast. *Biochim Biophys Acta*, **1757**(9-10), 1217–28.

Bromham, L. (2008). *Reading the story in DNA*. Oxford University Press.

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

Caetano-Anolles, G., Yafremava, L. S., Gee, H., Caetano-Anolles, D., Kim, H. S., and Mittenthal, J. E. (2009). The origin and evolution of modern metabolism. *Int J Biochem Cell Biol*, **41**(2), 285–97.

Campanella, M. E., Chu, H., and Low, P. S. (2005). Assembly and regulation of a glycolytic enzyme complex on the human erythrocyte membrane. *Proc Natl Acad Sci U S A*, **102**(7), 2402–7.

Chen, Y. and Dokholyan, N. V. (2006). The coordinated evolution of yeast proteins is constrained by functional modularity. *Trends Genet*, **22**(8), 416–419.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). Sgd: Saccharomyces genome database. *Nucleic Acids Res*, **26**(1), 73–39.

Chuong, S. D., Good, A. G., Taylor, G. J., Freeman, M. C., Moorhead, G. B., and Muench, D. G. (2004). Large-scale identification of tubulin-binding proteins provides insight on subcellular trafficking, metabolic channeling, and signaling in plant cells. *Mol Cell Proteomics*, **3**(10), 970–83.

Clark, N. L. and Aquadro, C. F. (2010). A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. *Mol Evol Biol*, **27**(5), 1152–1161.

Conant, G. C. (2009). Neutral evolution on mammalian protein surfaces. *Trends Genet.*, **25**(9), 377–81.

Conant, G. C. and Wolfe, K. H. (2006). Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biol*, **4**(4), e109.

Conant, G. C. and Wolfe, K. H. (2007). Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol.*, **3**(129), epub.

Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet*, **9**(12), 938–50.

Crawley, M. (2005). *Statistics An Introduction using R*. Wiley.

Davis, J. C. and Petrov, D. A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol*, **2**(3), E55.

Davis, J. C. and Petrov, D. A. (2005). Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet*, **21**(10), 548–51.

Dehal, P. and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, **3**(10), e314.

Dolezal, P., Likic, V., Tachezy, J., and Lithgow, T. (2006). Evolution of the molecular machines for protein import into mitochondria. *Science*, **40**(5), 314–318.

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

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*, **102**(40), 14338–14343.

Drummond, D. A., Raval, A., and Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*, **23**(2), 327–337.

Duarte, N. C., Herrgård, M. J., and Palsson, B. Ø. (2004). Reconstruction and validation of saccharomyces cerevisiae ind750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, **14**(7), 1298–309.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, **104**(6), 1777–82.

Durek, P. and Walther, D. (2008). The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles. *BMC Systems Biology*, **2**(100).

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**(5), 1792–1797.

Edwards, J. S. and Palsson, B. O. (2000a). The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A*, **97**(10), 5528–5533.

Edwards, J. S. and Palsson, B. O. (2000b). Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnol. Prog.*, **16**(6), 927–939.

Edwards, J. S., Ibarra, R. U., and Palsson, B. O. (2001). *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, **19**(2), 125–130.

Eisenberg, D. (1999-2010). Database of Interacting Proteins. `http://dip.doe-mbi.ucla.edu/`.

Ellis, R. J. (2001). Macromolecular crowding: obvious but underappreciated. *Trends Biochem Sci*, **26**(10), 597–604.

EMBL-EBI European Bioinformatics Institute (2011). European Bioinformatics Institute. `http://www.ebi.ac.uk/Tools/dbfetch/`.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J. P., Duewel, H. S., , I. S., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S. L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*, **3**(89), 89.

Fares, M. A. and McNally, D. (2006). Caps: coevolution analysis using protein sequences. *Bioinformatics*, **22**(22), 2821–2822.

Fares, M. A. and Travers, S. A. A. (2006). A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics*, **173**(1), 9–23.

Fernández, A. and Lynch, M. (2011). Non-adaptive origins of interactome complexity. *Nature*, **474**(7352), 502–505.

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., and Searle, S. (2008). Ensembl 2008. *Nucleic Acids Res*, **36**(Database issue), D707–14.

Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J. (2003). Genome-scale reconstruction of the saccharomyces cerevisiae metabolic network. *Genome Research*, **13**(2), 244–253.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, **296**(5568), 750–752.

Fraser, H. B., Wall, D. P., and Hirsh, A. E. (2003). A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol Biol*, **3**(11).

Fraser, H. B., Hirsh, A. E., Wall, D. P., and Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci U S A*, **101**(24), 9033–9038.

Freeling, M. and Thomas, B. C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.*, **16**(7), 805–814.

Gaertner, F. H. (1978). Unique catalytic properties of enzyme clusters. *Trends in Biochemical Sciences*, **3**(1), 63–65.

Gancedo, C. and Flores, C. L. (2008). Moonlighting proteins in yeast. *Microbiology and Molecular Biology Reviews*, **72**(1), 197–210.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**(6868), 141–7.

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**(7084), 631–636.

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., and Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J Mol Biol*, **299**(2), 283–293.

Gout, J. F., Duret, L., and Kahn, D. (2009). Differential retention of metabolic genes following whole-genome duplication. *Mol Biol Evol*, **13**, 13.

Graham, J. W., Williams, T. C., Morgan, M., Fernie, A. R., Ratcliffe, R. G., and Sweetlove, L. J. (2007). Glycolytic enzymes associate dynamically with mitochondria in response to respiratory demand and support substrate channeling. *Plant Cell*, **19**(11), 3723–38.

Guan, Y., Dunham, M. J., and Troyanskaya, O. G. (2007). Functional analysis of gene duplications in saccharomyces cerevisiae. *Genetics*, **175**(2), 933–43.

Hahn, M. W. and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol*, **22**(4), 803–806.

Hahn, M. W., Conant, G. C., and Wagner, A. (2004). Molecular evolution in large genetic networks: does connectivity equal constraint? *J Mol Evol*, **58**(2), 203–211.

Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G., and Robertson, D. L. (2007a). All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol*, **8**(10), R209.

Hakes, L., Lovell, S. C., Oliver, S. G., and Robertson, D. L. (2007b). Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc Natl Acad Sci U S A*, **104**(19), 7999–8004.

Hakes, L., Pinney, J. W., Robertson, D. L., and Lovell, S. C. (2008). Protein-protein interaction networks and biology–what's the connection? *Nat Biotechnol*, **26**(1), 69–72.

Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol*, **7**(11), 120.

Hartl, D. L. and Clark, A. G. (1997). *Principles of Population Genetics*. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, 3rd edition.

He, X. and Zhang, J. (2006). Higher duplicability of less important genes in yeast genomes. *Mol Biol Evol*, **23**(1), 144–51.

Hokamp, K. (2001). *A bioinformatics approach to (intra-)genome comparisons*. Ph.D. thesis, University of Dublin.

Horowitz, N. H. (1945). On the evolution of biochemical syntheses. *Proc Natl Acad Sci U S A.*, **31**(6), 153–157.

Hudder, A., Nathanson, L., and Deutscher, M. P. (2003). Organization of mammalian cytoplasm. *Mol Cell Biol*, **23**(24), 9318–26.

Hudson, C. M. and Conant, G. C. (2011). Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes. *BMC Evol Biol*, **11**(89).

Hughes, T., Ekman, D., Ardawatia, H., Elofsson, A., and Liberles, D. A. (2007). Evaluating dosage compensation as a cause of duplicate gene retention in paramecium tetraurelia. *Genome Biol*, **8**(5), 213.

Huss, M. and Holme, P. (2007). Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET Syst Biol*, **1**(5), 280–285.

Huthmacher, C., Gille, C., and Holzhutter, H. G. (2007). Computational analysis of protein-protein interactions in metabolic networks of escherichia coli and yeast. *Genome Informatics*, **18**, 162–172.

Huthmacher, C., Gille, C., and Holzhutter, H. G. (2008). A computational analysis of protein interactions in metabolic networks reveals novel enzyme pairs potentially involved in metabolic channeling. *Journal of Theoretical Biology*, **252**(3), 456–464.

Hyde, C. C., Ahmed, S. A., Padlan, E. A., Miles, E. W., and Davies, D. R. (1988). Three-dimensional structure of the tryptophan synthase alpha 2 beta 2 multienzyme complex from salmonella typhimurium. *J Biol Chem*, **263**(33), 17857–71.

Ispolatov, I., Yuryev, A., Mazo, I., and Maslov, S. (2005). Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res*, **33**(11), 3629–35.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, **98**(8), 4569–4574.

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol.*, **30**, 409–425.

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**(6833), 41–2.

Jerby, L., Shlomi, T., and Ruppin, E. (2010). Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol*, **6**(401).

Jonsson, P. F. and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, **22**(18), 2291–7.

Jordan, I. K., Wolf, Y. I., and Koonin, E. V. (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 Evol Biol*, **3**(1), doi:10.1186/1471–2148–3–1.

Juan, D., Pazos, F., and Valencia, A. (2008). Co-evolution and co-adaptation in protein networks. *FEBS Lett*, **582**(8), 1225–1230.

Julenius, K. and Pedersen, A. G. (2006). Protein evolution is faster outside the cell. *Mol Biol Evol*, **23**(11), 2039–2048.

Kanehisa Laboratories (1995-2011). KEGG LIGAND Database. `http://www.genome.jp/ligand/`.

Kann, M. G., Shoemaker, B. A., Panchenko, A. R., and Przytycka, T. M. (2009). Correlated evolution of interacting proteins: looking behind the mirrortree. *J Mol Biol*, **385**(1), 91–98.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A.,

Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database–2009 update. *Nucleic Acids Res*, **37**(Database issue), D767–72.

Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. (1993). Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc Natl Acad Sci U S A*, **90**(15), 7176–7180.

Kunin, V., Pereira-Leal, J. B., and Ouzounis, C. A. (2004). Functional evolution of the yeast protein interaction network. *Mol Biol Evol*, **21**(7), 1171–6.

Lenhinger, A. L., Nelson, D. L., and Cox, M. M. (1993). *Principles of Biochemistry*. Worth Publishers.

Levy, E. D. and Pereira-Leal, J. B. (2008). Evolution and dynamics of protein interactions and networks. *Curr Opin Struct Biol*, **18**(3), 349–57.

Levy, E. D., Pereira-Leal, J. B., Chothia, C., and Teichmann, S. A. (2006). 3d complex: a structural classification of protein complexes. *PLoS Comput Biol*, **2**(11), e155.

Levy, E. D., Erba, E. B., Robinson, C. V., and Teichmann, S. A. (2008). Assembly reflects evolution of protein complexes. *Nature*, **453**, 1262–1265.

Li, L., Huang, Y., Xia, X., and Sun, Z. (2006). Preferential duplication in the sparse part of yeast protein interaction network. *Mol Biol Evol*, **23**(12), 2467–73.

Liang, H. and Li, W. H. (2007). Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet*, **23**(8), 375–8.

Liang, H., Plazonic, K. R., Chen, J., Li, W. H., and Fernandez, A. (2008). Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet*, **4**(1), e11.

Liang, Z., Xu, M., Teng, M., Niu, L., and Wu, J. (2010). Coevolution is a short-distance force at the protein interaction level and correlates with the modular organization of protein networks. *FEBS Lett*, **584**(19), 4237–4240.

Liao, B.-Y., Weng, M.-P., and Zhang, J. (2010). Impact of extracellularity on the evolutionary rate of mammalian proteins. *Genome Biol Evol*, **2**, 39–43.

Light, S. (2006). *Investigations into the evolution of biological networks*. Ph.D. thesis, Stockholm University.

Lovell, S. C. and Robertson, D. L. (2010). An integrated view of molecular coevolution in protein-protein interactions. *Mol Biol Evol*, **27**(11), 2567–2575.

Lukatsky, D. B., Shakhnovich, B. E., Mintseris, J., and Shakhnovich, E. I. (2007). Structural similarity enhances interaction propensity of proteins. *J Mol Biol*, **365**(5), 1596–606.

Lynch, M., O'Hely, M., Walsh, B., and Force, A. (2001). The probability of preservation of a newly arisen gene duplicate. *Genetics*, **159**(4), 1789–804.

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. (2007). The edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, **3**, 135.

MacCarthy, T. and Bergman, A. (2007). The limits of subfunctionalization. *BMC Evol Biol*, **7**(213), 213.

Maddison, W. P. and Maddison, D. R. (2010). Mesquite: a modular system for evolutionary analysis. `http://mesquiteproject.org`.

Makhorin, A. (2000-2008). GLPK (GNU Linear Programming Kit). `http://www.gnu.org/s/glpk/`.

Makino, T. and Gojobori, T. (2006). The evolutionary rate of a protein is influenced by features of the interacting partners. *Mol Biol Evol*, **23**(4), 784–9.

Makino, T. and Gojobori, T. (2007). Evolution of protein-protein interaction network. *Genome Dyn*, **3**, 13–29.

Makino, T., Suzuki, Y., and Gojobori, T. (2006). Differential evolutionary rates of duplicated genes in protein interaction network. *Gene*, **385**, 57–63.

Makino, T., Hokamp, K., and McLysaght, A. (2009). The complex relationship of gene duplication and essentiality. *Trends Genet*, **12**, 12.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**(6757), 83–86.

Mathivanan, S., Periaswamy, B., Gandhi, T. K., Kandasamy, K., Suresh, S., Mohmood, R., Ramachandra, Y. L., and Pandey, A. (2006). An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, **7 Suppl 5**(7), S19.

McDonald, J. H. (2009). *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, Maryland, 2nd edition.

McLysaght, A., Hokamp, K., and Wolfe, K. H. (2002). Extensive genomic duplication during early chordate evolution. *Nat Genet*, **31**(2), 200–4.

Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. (1999). Mips: A database for genomes and protein sequences. *Nucleic Acids Res*, **27**, 44–48.

Meyer, A. and Van de Peer, Y. (2005). From 2r to 3r: evidence for a fish-specific genome duplication (fsgd). *Bioessays*, **27**(9), 937–945.

Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A*, **102**(31), 10930–10935.

Mochly-Rosen, D. (1995). Localization of protein kinases by anchoring proteins: a theme in signal transduction. *Science*, **268**(5208), 247–251.

Morowitz, H. J. (1999). A theory of biochemical organization, metabolic pathways, and evolution. *Complexity*, **4**(6), 39–53.

Mowbray, J. and Moses, V. (1976). The tentative identification in escherichia coli of a multienzyme complex with glycolytic activity. *Eur J Biochem*, **66**(1), 25–36.

Nakatani, Y., Takeda, H., Kohara, Y., and Morishita, S. (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res*, **17**(9), 1254–65.

Newman, M. E. J. (2008). *The New Palgrave Encyclopedia of Economics*, chapter Mathematics of networks. Palgrave Macmillan, Basingstoke, 2nd edition.

Ohno, S. (1970). *Evolution by Gene Duplication*. George Allen and Unwin.

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J Mol Evol*, **40**(1), 56–63.

Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nat Biotechnol*, **28**(3), 245–248.

Otto, S. P. and Whitton, J. (2000). Polyploid incidence and evolution. *Annu Rev Genet.*, **34**, 401–437.

Ovádi, J. and Saks, V. (2004). On the origin of intracellular compartmentation and organized metabolic systems. *Mol Cell Biochem*, **256-257**(1-2), 5–12.

Ovádi, J. and Srere, P. A. (1996). Metabolic consequences of enzyme interactions. *Cell Biochem Funct*, **14**(4), 249–58.

Ovádi, J., Orosz, F., and Hollán, S. (2004). Functional aspects of cellular microcompartmentation in the development of neurodegeneration: mutation induced aberrant protein-protein associations. *Mol Cell Biochem*, **256-257**(1-2), 83–93.

Pál, C., Papp, B., and Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, **158**(2), 927–931.

Pál, C., Papp, B., and Lercher, M. J. (2006). An integrated view of protein evolution. *Nat Rev Genet*, **7**(5), 337–348.

Papp, B., Pál, C., and Hurst, L. D. (2003a). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424**(6945), 194–7.

Papp, B., Pál, C., and Hurst, L. D. (2003b). Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet*, **19**(8), 417–22.

Park, J. and Barabási, A. L. (2007). Distribution of node characteristics in complex networks. *Proc Natl Acad Sci U S A*, **13**(104), 17916–17920.

Pawson, T. and Scott, J. D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. *Science*, **278**(5346), 2075–80.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, **14**(9), 609–614.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**(8), 4285–4288.

Pereira-Leal, J. B., Levy, E. D., Kamp, C., and Teichmann, S. A. (2007). Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol*, **8**(4), R51.

Pérez-Bercoff, Å., Makino, T., and McLysaght, A. (2010). Duplicability of self-interacting human genes. *BMC Evol Biol*, **10**(160).

Pérez-Bercoff, Å., McLysaght, A., and Conant, G. C. (2011). Patterns of indirect protein interactions suggest a spatial organization to metabolism. *Mol BioSyst*, page doi:10.1039/c1mb05168g.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V.,

Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, **13**(10), 2363–71.

Pielak, G. J. (2005). A model of intracellular organization. *Proc Natl Acad Sci USA*, **102**(17), 5901–5902.

Pollock, D. D., Taylor, W. R., and Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol*, **287**(1), 187–198.

Prachumwat, A. and Li, W. H. (2006). Protein function, connectivity, and duplicability in yeast. *Mol Biol Evol*, **23**(1), 30–9.

Presser, A., Elowitz, M. B., Kellis, M., and Kishony, R. (2008). The evolutionary dynamics of the saccharomyces cerevisiae protein interaction network after duplication. *Proc Natl Acad Sci U S A*, **105**(3), 950–4.

Prince, V. E. and Pickett, F. B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet*, **3**(11), 827–37.

Purcell, S. (2007). Maximum Likelihood Estimation. `http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_1.html`.

Qian, W. and Zhang, J. (2008). Gene dosage and gene duplicability. *Genetics*, **179**(4), 2319–24.

Qian, W., He, X., Chan, E., Xu, H., and Zhang, J. (2011). Measuring the evolutionary rate of protein-protein interaction. *Proc Natl Acad Sci U S A*, **108**(21), 8725–8730.

Rais, B., Ortega, F., Puigjaner, J., Comin, B., Orosz, F., Ovadi, J., and Cascante, M. (2000). Quantitative characterization of homo- and heteroassociations of

muscle phosphofructokinase with aldolase. *Biochim Biophys Acta*, **1479**(1-2), 303–14.

Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A., and Ciccarelli, F. D. (2008). Low duplicability and network fragility of cancer genes. *Trends Genet*, **24**(9), 427–30.

Rambaut, A. (2006-2009). Tree Figure Drawing Tool. `http://tree.bio.ed.ac.uk/`.

Rastogi, S. and Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol*, **5**(1), 28.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, **297**(5586), 1551–1555.

Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. Ø. (2003). An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biol*, **4**(9), R54.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**(7062), 1173–8.

Saks, V., Beraud, N., and Wallimann, T. (2008). Metabolic compartmentation - a system level property of muscle cells: real problems of diffusion in living cells. *Int J Mol Sci*, **9**(5), 751–67.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res*, **32**(Database issue), D449–451.

Sémon, M. and Wolfe, K. H. (2007). Consequences of genome duplication. *Curr Opin Genet Dev.*, **17**(6), 505–512.

Seppet, E. K., Kaambre, T., Sikk, P., Tiivel, T., Vija, H., Tonkonogi, M., Sahlin, K., Kay, L., Appaix, F., Braun, U., Eimre, M., and Saks, V. A. (2001). Functional complexes of mitochondria with ca,mgatpases of myofibrils and sarcoplasmic reticulum in muscle cells. *Biochim Biophys Acta*, **1504**(2-3), 379–95.

Sherlock, G. and Weng, S. (2004). Generic Gene Ontology (GO) Term Finder. `http://go.princeton.edu/cgi-bin/GOTermFinder`.

Shlomi, T., Cabili, M. N., MJ, M. J. H., Palsson, B. O., and Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol*, **26**(9), 1003–1010.

Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Comput Biol.*, **3**(3), 0337–0344.

Siddall, M. E. (2007). Maximum Likelihood. `http://research.amnh.org/~siddall/methods/ml.html`.

Srere, P. A. (2000). Macromolecular interactions: tracing the roots. *Trends Biochem Sci*, **25**(3), 150–3.

SRI International (2010). Humancyc: Encyclopedia of homo sapiens genes and metabolism. `http://humancyc.org/`.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege,

A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**(6), 957–68.

Stuart, R. A. (2008). Supercomplex organization of the oxidative phosphorylation enzymes in yeast mitochondria. *J Bioenerg Biomembr*, **40**(5), 411–417.

Teichmann, S. A. and Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nat Genet*, **36**(5), 492–496.

Templin, M. F., Stoll, D., Schrenk, M., Traub, P. C., Vöhringer, C. F., and Joos, T. O. (2002). Protein microarray technology. *Trends Biotechnol*, **20**(4), 160–6.

The Gene Ontology project website (1999-2010a). GO identifiers. `ftp://ftp.geneontology.org/pub/go/gene-associations/gene_association.goa_human.gz`.

The Gene Ontology project website (1999-2010b). GO slim. `ftp://ftp.geneontology.org/pub/go/GO_slimsandlink` and `goslim_goa.ob`.

Tillier, E. R. and Charlebois, R. L. (2009). The human protein coevolution network. *Genome Res*, **20**, 20.

Travers, S. A., Tully, D. C., McCormack, G. P., and Fares, M. A. (2007). A study of the coevolutionary patterns operating within the env gene of the hiv-1 group m subtypes. *Mol Biol Evol*, **24**(12), 2787–2801.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat Rev Genet*, **10**(10), 725–732.

Van de Peer, Y., Maere, S., and Meyer, A. (2010). 2r or not 2r is not the question anymore. *Nat Rev Genet*, **11**(2), 166.

Verkman, A. S. (2002). Solute and macromolecule diffusion in cellular aqueous compartments. *Trends Biochem*, **27**(1), 27–33.

Vinogradov, A. E. (2010). Systemic factors dominate mammal protein evolution. *Proc Biol Sci*, **277**(1686), 1403–1408.

Voet, D. and Voet, J. G. (2004). *Biochemistry*. John Wiley Sons, Inc., Hoboken, NJ.

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

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, **18**(7), 1283–92.

Wagner, A. (2007). *Robustness and Evolvability in Living Systems*, chapter 8-9. Princeton Studies in Complexity. Princeton University Press.

Wagner, A. and Fell, D. A. (2001). The small world inside large metabolic networks. *Proc Biol Sci*, **268**(1478), 1803–1810.

Wagner, G. P. and Zhang, J. (2011). The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat Rev Genet*, **12**(3), 204–213.

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

West, G. B., Brown, J. H., and Enquist, B. J. (1999). The fourth dimension of life: Fractal geometry and allometric scaling of organisms. *Science*, **284**(5420), 1677–1679.

West, G. B., Woodruff, W. H., and Brown, J. H. (2002). Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. *Proc Natl Acad Sci U S A*, **99 Suppl 1**, 2473–8.

Winkel, B. S. (2004). Metabolic channeling in plants. *Annu Rev Plant Biol*, **55**, 85–107.

Wolfe, K. H. (2000). Robustness - it's not where you think it is. *Nature Genetics*, **25**.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet*, **2**(5), 333–41.

Woolfit, M. and Bromham, L. (2003). Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol Biol Evol*, **20**(9), 1545–1555.

Woolfit, M. and Bromham, L. (2005). Population size and molecular evolution on islands. *Proc Biol Sci*, **272**(1578), 2277–2282.

Wright, J. and Wagner, A. (2008). The systems biology research tool: evolvable open-source software. *BMC Syst Biol*, **2**, 55.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). Dip: the database of interacting proteins. *Nucleic Acids Res*, **28**(1), 289–91.

Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J., Thompson, M. J., Marcotte, E. M., and Eisenberg, D. (2001). Dip: The database of interacting proteins: 2001 update. *Nucleic Acids Res*, **29**(1), 239–41.

Xiong, Z. and Pires, J. C. (2011). Karyotype and identification of all homoeologous chromosomes of allopolyploid brassica napus and its diploid progenitors. *Genetics*, **187**(1), 37–49.

Yamada, T. and Bork, P. (2009). Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nat Rev Mol Cell Biol*, **10**(11), 791–803.

Yang, Z. (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, **13**(5), 555–6.

Yang, Z. (2001). *Handbook of Statistical Genetics*, chapter 12, Adaptive Molecular Evolution, pages 327–350. John Wiley Sons, Ltd, 1st edition.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**(8), 1586–91.

Yang, Z. and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*, **46**(4), 409–418.

Yang, Z. and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, **17**(1), 32–43.

Yanofsky, C. and Rachmeler, M. (1958). The exclusion of free indole as an intermediate in the biosynthesis of tryptophan in neurospora crassa. *Biochim Biophys Acta*, **28**(3), 640–1.

Ycas, M. (1974). On earlier states of the biochemical system. *J Theor Biol.*, **44**(1), 145–160.

Yoon, J., Blumer, A., and Lee, K. (2006). An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics*, **22**(24), 3106–3108.

Zhang, Z., Luo, Z. W., Kishino, H., and Kearsey, M. J. (2005). Divergence pattern of duplicate genes in protein-protein interactions follows the power law. *Mol Biol Evol*, **22**(3), 501–5.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes Dev*, **21**(9), 1010–24.