Stony Brook University

Studies on Biological Evolution and Biological Networks: A Statistical Physics Approach

A Dissertation Presented

by

Koon-Kiu Yan

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Physics

Stony Brook University

December 2007

Stony Brook University

The Graduate School

Koon-Kiu Yan

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Sergei Maslov – Dissertation Advisor
Adjunct Professor, Department of Physics and Astronomy

Alexander Abanov – Chairperson of Defense
Associate Professor, Department of Physics and Astronomy

Adam Durst
Assistant Professor, Department of Physics and Astronomy

Jin Wang
Assistant Professor, Department of Chemistry

John Reinitz
Professor, Department of Applied Mathematics and Statistics

This dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Studies on Biological Evolution and Biological Networks: A Statistical Physics Approach

by

Koon-Kiu Yan

Doctor of Philosophy

in

Physics

Stony Brook University

2007

The availably of completely-sequenced genomes and various kinds of system-wide datasets have motivated a great deal of interests in the quantitative studies of biology. Owing to the large amount of data, statistical analysis is usually employed. In particular, tools or methods used in statistical physics can be useful in this kind of analysis. In this dissertation, I summarize my work on genome-wide or system-wide studies of biological evolution and biological networks.

Regarding biological evolution, we present a study on the genome-wide distributions of sequence identities of paralogous protein pairs

in various model organisms. We introduced a simple birth-and-death model based on gene duplication, gene deletion and point mutations to explain the common features in these distributions. Our mathematical framework revealed many important details including the relative rates of the evolutionary processes, previously unknown universality of intra-protein substitution rates and the consequences of whole genome duplications.

In the past decade, the idea of biological networks has emerged as a backbone to understand the complex interactions in biological systems. The studies presented in this dissertation cover three different aspects of biological networks: evolution, dynamics and algorithm. On the subject of network evolution, we quantified the topological divergence between paralogs in various protein networks and demonstrated that they provide certain functional redundancy. We also found that, at least in yeast, duplicated proteins lose their common regulators at a faster rate than common physical interaction partners. This can help explain how species with very similar gene contents can evolve novel properties in a relatively short timescale.

While the topology of biological networks serves as a starting point, it is important to study the underlying dynamical processes on networks. We present a study on the association and dissociation of proteins in a genome-wide protein interaction network. Like many biochemical reactions in a cell, physical interactions between

proteins are stochastic in nature. We studied how fluctuations in protein abundance lead to those in free protein concentrations and dimers concentrations. In addition to induced fluctuations, we studied the thermal noise of the system and found that it is affected by both the network topology and the heterogeneity in protein abundance. Our results suggest that undesirable cross-talk mediated by reversible protein interactions can be significantly suppressed.

From a practical point of view, very large networks appear in biology as a way to represent data from high-throughput experiments. In the final part of this dissertation, we present a network-based algorithm to predict and verify indirect regulatory interactions in a large-scale genetic regulatory network. This algorithm is tailored for large and heavily interconnected networks, which are of growing importance due to the rapid accrual of regulatory interactions. We applied the algorithm to the regulatory networks of several model organisms curated from literature, resulting in novel predictions along with calibrated reliability of existing ones.

To my beautiful wife Stella,

who shares with me tears and joys in these years.

Contents

List of Figures

List of Tables

Acknowledgements

This dissertation records the period of my life spent at Stony Brook. Life is not always easy. To a certain extent, I regard this period as a journey in a desert. However, life would not present difficulties beyond one can bear, and there is always a way out so that one can stand up under it. To me, I believe the people I meet here are by no chance coincidence, and they are the reasons why I can survive the journey.

First of all, I want to thank my thesis advisor, Sergei Maslov. I feel extremely privileged to have worked with such a smart and good-hearted scientist. It is Sergei who introduced me to the fascinating world of biology, taught me the ways of working with empirical data. I have to thank him for all advices in scientific writing, and patiently iterating my manuscripts. I have enjoyed our many discussions, and have benefitted greatly from his insight and guidance over the years. I certainly hope to continue working with him in the future.

I would like to thank Prof. John Reinitz, who has truly demonstrated a great devotion to science. I am greatly inspired and motivated by his lectures in AMS 691. It turns out some of the works in this dissertation can be directed back to the materials I learnt in his class. I have to thank him also for all his

career advices, which I believe, will always be useful down the road.

I thank my friends and classmates here, who help me to overcome the occasional loneliness associated with scientific work. I especially want to thank my dear officemates, Dmitri Volja, Sebastian Reyes, Dylan Walker, Huafeng Xie, with whom I have numerous funny moments and stimulating discussions. Of course, apart from the friendship, I want to thank Dylan and Huafeng for their fruitful collaborations. I learn a lot by discussing with them. A special thank to Anirban Chakraborti, who was very kind to me and offered me a lot of help.

A thank to Prof. Laszlo Mihaly and Pat Peiliker, who were there for me whenever an administrative issue came up.

I wish to thank my parents in Hong Kong for their unconditional love and support. I have missed many moments with them during these years of study, I wish I can repay them back in the future.

Most important of all, I want to thank my wife Stella for everything she has done. Even though I have dedicated this thesis to her, I know she deserves more than everything I can offer.

Chapter 1

Introduction

Over the past decade, the interface between statistical physics and biology has expanded rapidly. Apart from traditional connections, new research directions have emerged, stimulated by advancements in both disciplines.

In statistical physics, the theory of complex systems offers many new insights. Generally speaking, complex systems consist of many heterogeneous components that interact with each other. The overall behavior of such systems usually cannot be deduced from the understanding of a single individual. In fact, complex systems are distinguished by the emergence of collective behaviors. As many systems in our world are complex, the study of complex systems has connections with many branches of science: sociology, computer science, economics and certainly biology.

Stunning developments have been made in biology as well. New experimental techniques have yielded an explosion of biological data that are increasingly quantitative in nature. Examples are the availability of complete sequenced genomes, high-thoughput methods on protein interactions and various single-

cell experiments. With these data in hand, quantitative models of biological systems are now possible.

This thesis consists of a few studies lying on the interface between statistical physics and biology. In this introduction, I first highlight four important areas: **evolution**, **networks**, **noise** and **bioinformatics**, and proceed to outline our studies, which are motivated by the progress in these particular areas.

Evolution

Evolution is fundamental to life. In fact, it distinguishes life. Stochasticity is essential to evolution. It allows an organism to explore the configuration space of an infinitely complicated optimization problem, with the exploration subjected to natural selection. In such sense, modern life is the consequence of a stochastic process which has been ongoing for billions years.

There are many examples in statistical physics where random processes show predictable collective behavior. For instance, the energy of an equilibrium system may deviate from its mean value stochastically, but the deviations always follow a Gaussian distribution. It is natural for physicists to look for collective behavior behind evolutionary processes. Probably the earliest example was reported by Willis and Yule [1], who observed that the distribution of the number of species in a genus, family or other taxonomic group appears to follow a power law. Yule offered an explanation as follows. Species are added to genera by speciation. If we assume that this happens randomly at a constant rate, it follows that a genus with more species will gain a new species at a faster rate. Let us further assume that occasionally, the new species pro-

duced is so different from the others and becomes the founder of an entire new genus. Thus the number of genera increases steadily, so does the number of species within each genus. This model is usually called Yule process, and its solution is a power law distribution.

While the Yule process is based on macroscopic properties related to taxonomy, evolution can be studied at different levels. Perhaps the most fundamental level is that of molecules. The genome, the complete DNA sequence, contains all hereditary information. This includes both the genes (protein-coding sequences) and the non-coding sequences. An important subset of the genome called proteome, refers to the entire complement of proteins expressed by a genome. As information in these molecules is translated into 3D structures of proteins and then functional processes of an organism, the evolution of an organism can thus be studied in terms of the evolution of these molecules. The most important processes in molecular evolution include gene duplications, mutations and gene deletions.

Back in 1970, Ohno [2] proposed that gene duplication is an important source of raw material for molecular evolution. Whole gene duplications give rise to new protein-coding regions in the genome. An extra gene copy can be created by an unequal crossover, gene transposition and polyploidization etc. The two initially identical genes subsequently diverge from each other in their sequences and thus, as do the proteins encoded. The divergence in sequences between a pair of duplicated genes is caused by mutations. This occurs via a broad spectrum of processes including point substitutions, insertions, deletions (indels), and transfers of whole domains either from other genes in the same genome or even from genomes of other species. These changes in the DNA

sequence may or may not lead to changes in the amino-acid sequences of proteins encoded. This is because of the existence of a certain redundancy in the genetic code. Substitution leading to (not leading to) modification in the amino-acid sequence is called non-synonymous (synonymous) substitution. Gene deletions happen when genes are no longer required for the functioning of the organisms. They are either explicitly deleted from the genome or stop being transcribed and become pseudogenes whose homology to the existing functional genes is rapidly obliterated by mutations.

Evolution at a molecular level can be mathematically formulated as a stochastic process. It is interesting to look for collective behaviors or universal patterns behind the molecular information. Until recently, there were not enough data to address such issue in a quantitative way. At present, however, there are more than 3000 completely sequenced genomes. This enables system-wide study on a wide variety of genomes or proteomes. Interesting scaling behaviors are indeed discovered. Examples include the power law distribution of protein family sizes [3, 4], and the scaling between the number of transcriptional factors in a genome to the total number of genes in the genome [5, 6]. It is tempting to speculate that such universal statistical laws are results of simple evolutionary principles. In Chapter 2, we will propose a simple model of proteome evolution which explains certain universal features in the distribution of sequence identities among paralogous proteins.

Networks

In the past decade, a great deal of attention has been focused on networks (for a collection of important results, see for example Ref. [7]). In short, a network represents the interactions (links) between many heterogeneous individual components (nodes) of a system. From a statistical physics point of view, the network is now widely recognized as backbone of complex systems that determines the structures, functions and dynamics of systems. As a result of its generality, the idea of networks penetrate into a wide range of disciplines. Examples are the power grids and internets in engineering, various kinds of social networks between people, information networks such as citation networks and the WWW. Recent progress shows that network theory is gaining importance in biology. More and more examples shows that understanding individual genes or proteins is not enough to crack even a simple biological function. Biological processes typically involve coordinated activity of many components: genes, proteins, metabolites etc. Just as electrical systems are represented by circuit diagrams, biological networks offer a starting point to understand the complex interactions in biological systems. Together with the advances in high-thoughput methods, there are vast amount of data on biological networks, leading to the emergent of system biology [8].

There are many ways to classify biological networks. One is based on their specific functional role in an organism, like the network which regulates the cell cycle in fission yeast [9], and the SOS response network in E.coli [10] among other examples. These networks consists a variety of components including mRNA, proteins and various small molecules. The interactions between these

components may be different biological processes such as transcriptional regulation, protein-protein binding, phosphorylation and other post-translational regulations. Even though network sizes are relatively small ($\sim$ 10 nodes), the application of these networks toward quantitative prediction of biological systems is extremely challenging. This is certainly one of the most fruitful direction in the era of system biology.

Another way to look at biological networks is based on the type of individual components involved. A classic example is metabolic network, where all nodes are metabolites, and links are the corresponding chemical reactions. Networks of this classification are usually in genome-wide scales, as in this example, the network represents all the chemical reactions between all the metabolites in an organism. As a result of their complexity (hundreds or thousands nodes and links), one may not be able to speculate specific biological functions using these networks. However, large-scale networks are particularly useful in providing statistics. Studying structures and dynamics of these networks sheds light on the general design principles of biological processes, especially in the context of evolution.

In this thesis, we will focus on genome-wide protein networks. Our concern is the interactions between all the proteins expressed by an organism. In particular, we are interested in protein interaction networks and genetic regulatory networks. Protein interaction networks represent the physical binding between two proteins. The yeast-two-hybrid method and mass-spectroscopy are two major high-thoughput methods for studying these interactions (see Ref. [11] for a recent tutorial). System-wide protein interaction networks used in our studies include *H. pylori* [12], *S. cerevaise* [13, 14] and *D. melanogaster*

[15]. Genetic regulatory networks represent how the expression of a gene regulates the expression of some other genes. Since the expression of a gene results in the production of a protein that it codes, genetic networks are protein networks. (In case where there is no confusion, we may use the terms proteins and genes interchangeably.) Unlike interaction networks, the regulation of protein A by B does not imply the reverse, therefore regulatory networks are directed networks. Moreover, depending on whether the activity of a protein is activating or repressing the expression of another, edges in regulatory networks carry either a positive or a negative sign. One of the most well-known regulatory networks is the transcriptional regulatory network. It involves an important type of proteins: transcription factors. A transcription factor switches on/off the expression of a gene by binding directly to its upstream region. A transcriptional regulatory network thus represents how transcription factors regulate the activities of other proteins (and themselves, too). We will look at a system-wide data in *S. cerevaise*, obtained using the ChIP-chip technique [16].

Where do we start in the study of genome-wide protein networks? Like many other complex networks in statistical physics, one starts by looking at its topology. One of the most interesting observation is the existence of hubs (proteins with many neighbors) or more precisely, a broad degree distribution. The broad distribution, often taken as a power law distribution, can be explained by a simple statistical physics model based on gene duplication and divergence [17], which is a variation of the preferential attachment model recognized in many systems [18]. Other important statistical properties include degree correlation [19], and the frequent occurrence of small building units – network motifs [20]. It is important to bear in mind that the observed networks

are sharpened by evolutionary processes, and hence topological properties are therefore traces of evolution. Studying networks sheds light on the general design principles behind evolution. In addition, networks are the backbones of complex systems, allowing for further study of dynamical processes happening on these networks. In Chapter 3, we will look at protein networks from an evolutionary viewpoint. A dynamical process in a protein interaction network will be studied in Chapter 4.

Noise

Cellular events are performed by the constituent molecules. The relevant energy scale for these intermolecular interactions is comparable to the magnitude of k_BT (0.62kcal/mole or 0.03eV) , where k_B is the Boltzman constant and T is the room temperature. As a result, random thermal motion plays an essential role in biology. Indeed, the interplay between deterministic and thermal forces give rise to the complex behavior in biological processes. Described in a more concrete way, molecules come together by chance. Therefore all chemical reactions and physical interactions are stochastic in nature. In principle, stochasticity is significant only if the average number of molecules are low since individual reactions change the numbers of molecules by at most one or two. However this is indeed the case *in vivo*. Genes are usually present in one or two copies, many mRNAs are rare and often proteins are present in less than 100 molecules per cell.

There are two important questions concerning noise. First, how can we explain the robust physiology of a cell when the underlying molecular mecha-

nisms are random? Second, what kind of benefits can noise provide? The first question is related to the robustness of biological systems. Most biological processes involve a series of reactions; noise in the earlier step may propagate forward. It is therefore intriguing to study how nature can attenuate and tolerate the randomness. The robustness of biological systems is even more remarkable if one considers the precise regulation in the development of multi-cellular organisms. Regarding the advantages of noise, one important aspect is the generation of heterogeneous populations. A famous example is the phage-lambda infection process governed by the lysis-lysogeny decision circuit. Under favorable environmental conditions, most phages upon entering a bacterial cell choose to become lysogens. However, owing to stochastic fluctuations, it is possible that a small fraction lyse the bacterium. In general, genetically identical cells and organisms can exhibit great diversity in phenotypic effects even when they are exposed to the same environment. Such variations among population are believed to be helpful for the survival of the species.

Even though cellular randomness has long been predicted [21], quantitative measurements have been possible for only a few years. Owing to the advances in single cell experiments, astonishing details in cell-to-cell fluctuations are revealed. By tagging green fluorescent protein (GFP) as a reporter, one can measure the protein level of an arbitrary gene in a single cell by the fluorescent intensity. Repeated measurements across a population of cells give the cell-to-cell variation [22]. Further application of flow cytometry offers a high-thoughput strategy to measure protein concentration in large number of single cells [23, 24].

The recent studies of noise have been focused on the issue of gene expres-

sion and thus protein abundance. To quantify noise in protein abundance, autocorrelatons summarize both the magnitude and frequency of fluctuations. However, because of the limitation in obtaining temporal data so far, most studies employed the stationary averages and variances over an ensemble of cells. Noise level is usually quantified by the variance among a population of cells normalized by the mean or the square of the mean.

Recent studies of cellular noise have addressed the distinction between the so called intrinsic and extrinsic noise [22, 25]. Such classification always depends on the definition of system versus environment. Intrinsic noise refers to stochasticity inherent in the dynamics of the system, while extrinsic noise is originated from fluctuations in other cellular processes. If one regards a whole cell as a single system, intrinsic noise is due to stochastic fluctuations in production or degradation of individual proteins, and extrinsic noise corresponds to synchronous changes, for example, the variation in the cell size and changes in the outside environment.

Non-equilibrium statistical physics is an essential tool in studying fluctuations in biological systems. As a biological system can be viewed as a system of chemical reactions, deterministic kinetic rate equations are used to describe the dynamics of the ensemble averages. Moving from a deterministic approach to a stochastic process, we are concerned about the probability of the system being in a certain state and how this probability changes with time. This is done, in principle, by writing down the master equation. However, as biological systems of interest usually involve not just a few reactions, master equations are typically not tractable. Further techniques such as Fokker-Planck equation, Langevin equation and Monte-carlo simulation [26–28] are therefore important

to model biological systems with stochasticity. In Chapter 4, we will study the system-wide effects of fluctuations in protein interaction network using some of the mentioned methods.

Bioinformatics

Effective computational algorithms play central roles in biology. Indeed, sequence oriented algorithms such as sequence assembly and sequence alignment, have facilitated the complete sequencing of many genomes (including human) and the identification of coding genes. In the post-genomic era, a vast amount of sequences are already known and algorithms aimed at understanding the complex bimolecular interactions are drawing more and more research attention.

One of the most important technology in the post-genomic era is the DNA microarray (DNA chip). The uses of DNA microarrays enable one to measure the expression level of more than 10000 genes simultaneously. The expression level of a gene refers to its mRNA abundance. A high expression level means that the gene is switched on. The genome-wide collection of expression levels is usually called the expression profile. Expression profiles contain important information because they reflect the state of a cell. Suppose a genome has N genes, and each of them can only be "on" or "off". Then the expression profile would represent 1 out of the 2^N possible cellular states. The state of a cell depends on the environmental conditions, time, and the tissue it belongs to in multi-cellular organisms.

As genes are not independent of each other, expression profiles are results of

complex interactions. While the invention of DNA microarray techniques has lead to accumulation of a tremendous amount of expression profiles, these data do not immediately yield information about the genetic interaction networks. To extract a wealth of information from these huge and noisy data, effective algorithms are necessary. Correlation analysis is an important component in mining the data [29]. As genes expressed under similar conditions are likely to be functionally related, one can identify regulatory modules by looking at the similarity between genes in different conditions. However, correlation analysis cannot identify how the correlated genes are causally related. Perhaps the most intuitive way to study regulatory interactions between genes is to perturb the system [30]. For example, one can delete or overexpress a gene and observe the changes in expression profile. If the deletion of a gene g causes a reduction of expression level in a set of genes, g is positively regulating the set of genes. Based on this simple idea, different algorithms have been proposed including Boolean logic [31], Bayesian analysis [32], and topological analysis [33]. In Chapter 5, we will present a study along these lines.

As practical problems in bioinformatics or networks involve inference or combinatorial optimization, methods in statistical physics (more precisely spin glasses) turn out to be very useful. In fact, problems in bioinformatics motivate and enrich theoretical developments in statistical physics. An important example is the superparamagnetic clustering used in studying the microarray gene expression data [34]. The algorithm assigns a Potts spin S_i for each data point i, and the coupling constant between two spins is a decreasing function with respect to the distance between two corresponding data points. While the spins interact as an inhomogeneous ferromagnetic Potts model, clusters

appear naturally as regions of aligned spins and can be quantified by the spin-spin correlation. The temperature T controls the resolution of the clustering. When $T = 0$, all spins have the same value, the result is a single cluster. When $T \to \infty$, all spins are independent and therefore the procedure yields clusters with a single data point in each. Standard dendrograms can be generated by gradually tuning the temperature.

Outline

The study of life presents many interesting problems for physicists. The following chapters include several studies concerning biological evolution and biological networks. Chapter 2 is a study on proteome evolution [35]. We studied the genome-wide distribution of sequence identities of pairs of paralogous proteins in several model organisms. It is interesting, at least for statistical physicists, because the distributions in different organisms share a few common features. To explain the collective behaviors, we introduced a simple model based on basic evolutionary processes. Our mathematical framework revealed many important details including the relative rates of the evolutionary processes, universality of intra-protein substitution rates and traces of whole genome duplications.

Chapter 3, 4 and 5 are all related to networks. In Chapter 3, we studied the topology of protein networks from an evolutionary standpoint [36]. The emphasis is on the role of duplicated proteins on a protein network. Using system-wide data of various protein networks, we quantified their functional divergence with respect to their divergence in amino-acid sequences. We found

that in yeast, the rate of divergence in transcriptional regulatory network is faster than that in protein interaction network. This would help to explain how species with very similar gene contents can evolve novel properties in a relatively short timescale. The idea that a pair of duplicated proteins may act as backup for each other is further corroborated by analysis of data from gene-knockout experiments.

In Chapter 4, we extend the analysis from topology to dynamics and study the effects of noise in biological networks. In particular, we investigate the association and dissociation of proteins in a genome-wide protein interaction network [37]. In dynamical equilibrium, protein association and dissociation are in balance, resulting at a mixture of free proteins and dimers in a wide range of concentrations. We studied the effects of noise in the system, including thermal fluctuations and fluctuations as a result of noise in protein production and degradation. Tools such as the fluctuation-dissipation theorem from non-equilibrium statistical physics were used. We found that thermal fluctuations could be well suppressed as a result of the network topology and heterogeneity in protein abundance, and there are dramatic differences between intrinsic and extrinsic noise.

Chapter 5 is a practical bioinformatics study [38]. The regulatory interactions from high-throughput experiments (e.g. microarray) can be either direct or indirect in nature. Indirect regulations are important as they constitute the majority of experimental data. We developed a network-based algorithm to predict and verify indirect regulatory interactions in a large-scale genetic regulatory network. This algorithm is tailored for large and heavily interconnected networks, which are of growing importance due to the accrual of data

from high-throughput experiments. We applied the algorithm to the regulatory networks of several model organisms curated from literature, resulting in novel predictions and calibrated the reliability of existing ones.

Chapter 2

A Stochastic Model of Proteome Evolution

2.1 Background

Proteins lie at the heart of all cellular processes, it is therefore important to understand how evolution shaping the defined proteome of an organism. At the molecular level, the evolution of a proteome can be understood in terms of its full repertoire: the set of all protein coding genes in the corresponding genome. As we have remarked in Chapter 1, the most significant processes in molecular evolution are gene duplications, gene deletions and changes in amino-acid sequences.

The recent availability of complete genomic sequences of a diverse group of living organisms allows one to study these basic mechanisms on an unprecedented scale. Using sequence alignment algorithms, one could define the similarity between any two proteins. In molecular evolution, two proteins simi-

lar in their amino-acid sequences are referred as homologs. Homologs originate from a common ancestry and are further classified into orthologs and paralogs. Orthologs are separated by a speciation event, i.e. when a specie diverges into two species, the two divergent copies of a single gene are orthologs. On the other hand, paralogs are separated by a gene duplication event, i.e. if a gene in a genome is duplicated, the two copies (still located in the same genome) are called paralogs. In this study, we focus on the set of all paralogs within an individual genome.

The sequence similarity of a pair of paralogs is quantified by a percentage called percent identity (PID). For the set of all paralogous pairs in a genome, the set of their PID is a dynamic entity that changes due to duplications, deletions, and local changes in amino-acid sequences of its constituent proteins. For example, duplication events constantly create new pairs of paralogous proteins with PID=100%, while subsequent substitutions, insertions and deletions result in their PID drifting down towards lower values. The distribution of PID (PID histogram) thus contains indirect information about past duplications, deletions, and sequence divergence events that took place in the ancestral genome.

In this chapter, we start by presenting empirical observations of the PID histograms of six model organisms. They are: prokaryotic bacteria *H. pylori, E. coli*, a single-celled eukaryote *S. cerevisiae* (baker's yeast), and multi-cellular eukaryotes *C. elegans* (worm), *D. melanogaster* (fly), and *H. sapiens* (human). Even though the organisms are diverse in nature, their histograms share a number of common features. We propose a simple stochastic model involving only gene duplications, deletions and amino-acid substitutions to explain the

features. Using our underlying mathematical framework, we then extract the average rates and some other intrinsic parameters of these basic evolutionary processes. The implications of our analysis will be discussed.

2.2 Distribution of Sequence Identities of Paralogous Proteins

As an operational definition, two proteins are referred as paralogs if they have a statistically significant sequence similarity. Therefore the set of all paralogous pairs in a genome can be identified by performing pairwise alignment to all possible protein pairs in a genome. In this study, we use the BLAST algorithm (Blastp) [39] to perform the sequence alignment.

Not every pair of BLAST hits are regarded as paralogs. Generally speaking, a BLAST output is taken as a pair of paralogs if (1) it is statistically significant (E$< 10^{-10}$), (2) the length of the aligned region constitutes at least 80% of the length of the longer protein. The second filter enables us to avoid pairs of multi-domain proteins homologous over only one of their domains. Table 2.1 shows the sizes of the genomes and some general statistics of the output. The details of the genomes and the exact procedures are described in Appendix A.

With all the paralogous pairs, one could easily obtain the distribution of sequence identities (PID histogram). Fig. 2.1 shows the histogram $N_a(p)$ of amino-acid sequence identities p of *all pairs* of paralogous proteins in the six genomes. The p-dependence of these histograms has three distinct regions I,II,III.

Organism	Proteome size N_{genes}	% of proteins with paralogs	BLASTP hits	Number of pairs in $N_a(p)$	Number of pairs in $N_d(p)$
H.pylori	1590	14	3228	260	148
E.coli	4288	33	16768	2614	1013
S.cerevisiae	5885	29	43915	2297	1025
C.elegans	19099	36	204398	46463	5545
D.melanogaster	14015	30	557047	17621	3238
H.sapiens	25319	37	1330721	31078	6595

Table 2.1: Statistics of genomes used in this study. The first column is the name of the organism, the second column – the number of protein-coding genes in its genome, N_{genes}, the third column is the percentage of proteins with at least one paralog, the fourth column - the total number of distinct BLAST hits generated before we applied subsequent filtering, the fifth column - the number of paralogous pairs included in $N_a(p)$, and the sixth column - in $N_d(p)$.

- *Region I*: There is a sharp and significant upturn in the PID histogram above roughly 90-95% compared to what one expects from extrapolating $N_a(p)$ from lower values of p. Apparently the constants (or possibly even mechanisms) of the dynamical process shaping $N_a(p)$ are different in this region.

- *Region II*: This region covers the widest interval of PIDs $30\% < p < 90\%$. $N_a(p)$ in this region can be approximated by a power-law form of $p^{-\gamma}$ with $\gamma \approx 4$ (shown as a dashed line in Fig. 2.1. The best fits to the power-law form in the Region II are listed in Table 2.2 and (with the exception of yeast and human) they fall in the $3-5$ range.

- *Region III*: In this region $p < 25-30\%$ the histogram $N_a(p)$ starts to deviate down from the $p^{-\gamma}$ powerlaw behavior. This decline is an

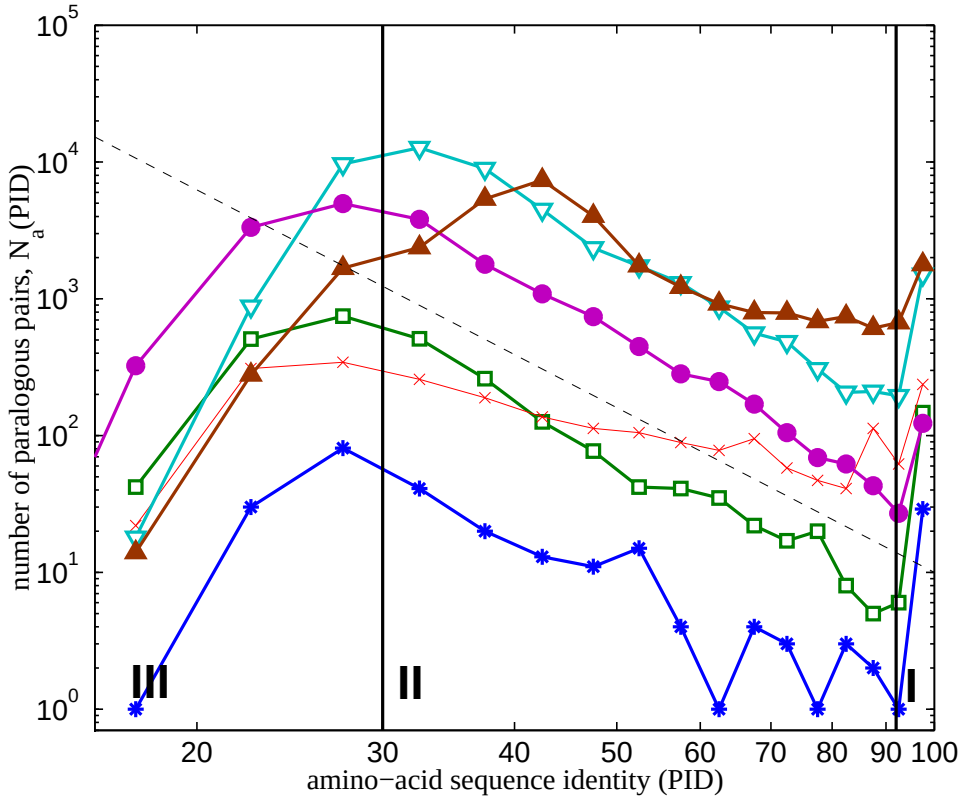

Figure 2.1: Histograms of all amino-acid sequence identities for all pairs of paralogous proteins in complete genomes of *H. pylori* (blue stars), *E. coli* (green open squares), *S. cerevisiae* (red crosses), *C. elegans* (cyan open triangles), *D. melanogaster* (magenta filled circles) and *H. sapiens* (brown filled triangles). The dashed line is a power-law p^{-4}. Note the logarithmic scale of both axes. Vertical lines separate regions I, II and III described in the text.

artifact of the inability of sequence-based algorithms such as BLAST to detect some valid paralogous pairs with low sequence identity. This explanation is corroborated by the observation that the exact position of the downturn of $N_a(p)$ in the region III is determined by the E-value cutoff.

As the proteomes have different sizes $N_{\rm gene}$ (from 1,600 in *H.pylori* to 25,000

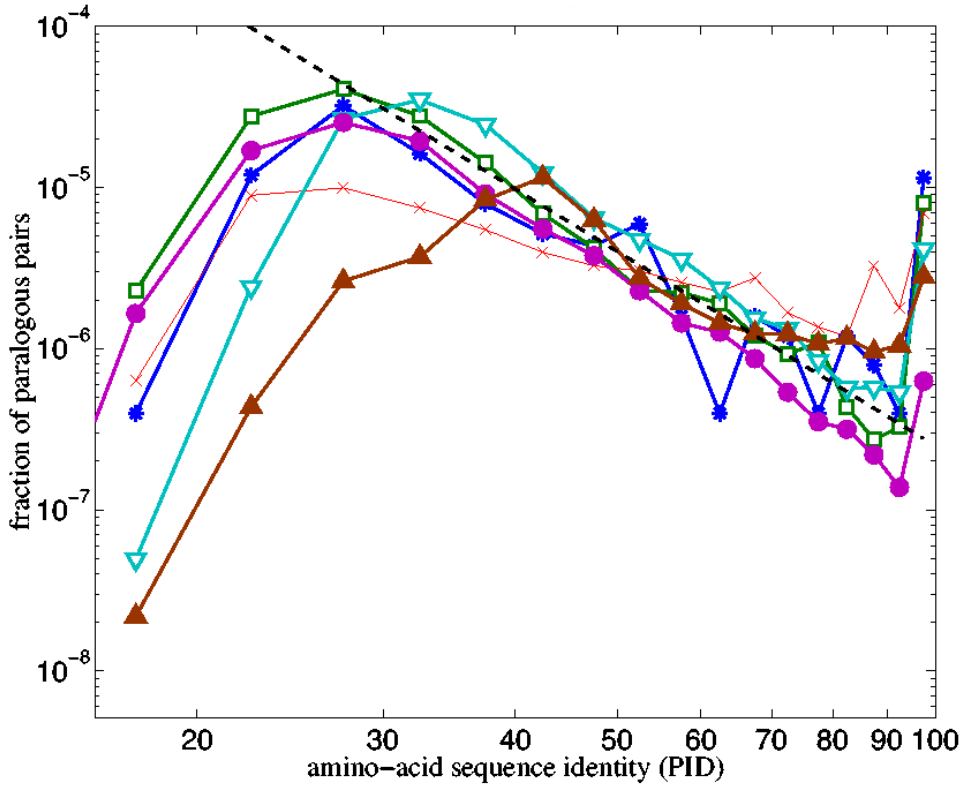

Figure 2.2: Normalized histograms for all amino-acid sequence identities. Each histogram in Fig. 2.1 is normalized by the number of gene pairs, i.e. $n_a(p) = 2N_a(p)/N_{\rm gene}^2$. The histograms of the six genomes are approximately collapsed to an near-universal shape.

in *H. sapiens*, see Table 2.1), it is meaningful to normalize the histogram by the number of gene pairs, approximately $N_{\rm gene}^2/2$. The near-universality in the shapes of the PID histograms is perhaps best illustrated by the normalized histograms as shown in Fig. 2.2.

2.3 A Model of Proteome Evolution

Birth-and-death model

The near-universality in shapes of PID histograms suggests that different proteomes are driven by similar mechanisms. To explain the features of the PID histogram, we introduce a simple stochastic model defined as follows.

- Gene Duplication. Random gene duplication happens at rate α_{dup}.
- Gene Deletion. Random gene deletion occurs at rate α_{del}.
- Mutation. This includes amino-acid substitutions, insertions and deletions. The processes cause the sequence identity of any given pair of paralogous proteins to decay with time. The decay is described by $\frac{dp}{dt} = -v(p)$. For our immediate purposes we will leave it unspecified.

Random gene duplication and deletion events refer to the birth and death of new protein coding genes. As shown in Fig. 2.3, when a gene A is duplicated to A' a new pair of paralogs with PID=100% is created (dotted line) and thus added to the rightmost bin of the PID histogram. The bin $N_a(p = 1)$ increases in rate $\alpha_{\text{dup}} N_{\text{genes}}$, where N_{genes} is the total number of protein-coding genes in the genome. Furthermore the freshly created gene A' inherits both paralogous partners (B and C) of the gene A. The PIDs of these two newly created paralogous pairs A'-B and A'-C (dashed lines) are also added to the respective bins in the histogram. Thus a duplication of any of one the two paralogous genes with PID=p among other things results in the creation of a new pair of paralogs with the same PID. This process increases $N_a(p)$ at a rate

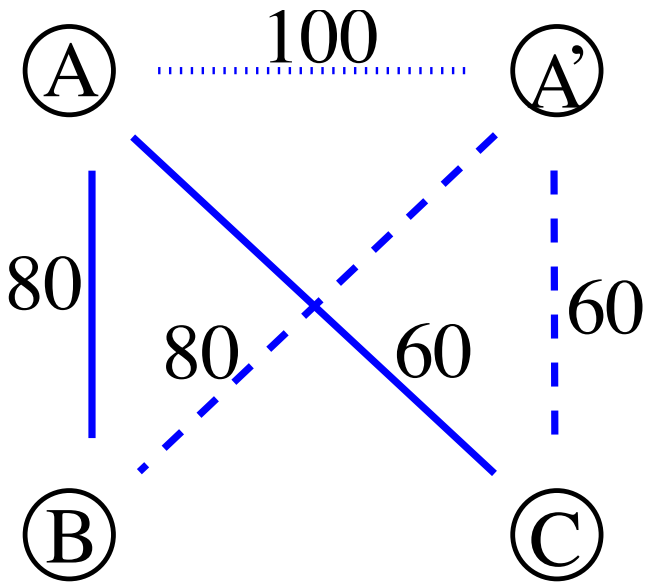

Figure 2.3: Illustration of gene duplication on PID histogram. A single gene duplication event $A \to A'$ gives rise to three new paralogous pairs: $A' - A$, $A' - B$ and $A' - C$. Immediately after the duplication the pair $A - A'$ has the PID=100% , while PIDs of $A' - B$ and $A' - C$ are equal to those of $A - B$ and $A - C$. Thus the PID of every previously existing paralogous pair involving A gets duplicated along with the duplication $A \to A'$.

$2\alpha_{\rm dup} N_a(p)$. The factor two comes from the fact that duplications can happen in any of the two paralogs. Similarly the deletion of any of the two genes in this paralogous pair decreases $N_a(p)$ at the rate $2\alpha_{\rm del} N_a(p)$.

To understand the effect of mutation, we summarize the effects of actual amino-acid substitutions, insertions and deletions via an effective "substitution rate" μ. Consider two paralogous proteins with PID=$p\times$100%, in the simplest possible case, changes in their sequences happen uniformly at all amino acid positions at a constant rate μ. The PID of this paralogous pair changes according to the equation $dp/dt = -2\mu p$. The factor p comes from the

observation that only changes in still identical parts of two sequences lead to a further decrease of the PID, while the factor two is because substitutions can occur in any one of the two proteins. This equation results in an exponentially decaying PID: $p(t) \sim \exp(-2\mu t)$, where paralogous pairs are drifting to the left hand side of the PID histogram. More generally the drift of PID could be described by the equation $dp/dt = -v(p)$. When the substitution rate varies for different amino acids within the same protein, the relation between p and the drift velocity $v(p)$ is no longer linear. The negative drift of PIDs generates a p-dependent flux of paralogous pairs down the PID axis given by $v(p)N_a(p)$. As shown in Fig. 2.4, the net flux into the PID bin of the width Δp centered around p is given by $N_a(p+\Delta p)v(p+\Delta p) - N_a(p)v(p)$. Adding up contributions of all three processes, one gets

$$\begin{aligned} \frac{\partial N_a(p,t)}{\partial t} &= \frac{\partial}{\partial p}[v(p)N_a(p,t)] \\ &+ 2\alpha_{\rm dup}N_a(p,t) - 2\alpha_{\rm del}N_a(p,t) + \alpha_{\rm dup}N_{\rm genes}\delta(p-1), \quad (2.1) \end{aligned}$$

where $\delta(p-1)$ is 1 if $p = 1$ and 0 otherwise.

In our model the total number of genes $N_{\rm genes}$ in the genome exponentially grows (or decays) according to $dN_{\rm genes}/dt = (\alpha_{\rm dup} - \alpha_{\rm del})N_{\rm genes}$. When the genome size of an organism remains approximately constant ($\alpha_{\rm dup} = \alpha_{\rm del}$), one can find the stationary asymptotic solution of Eq. 2.1. In general, the stationary solution for $N_a(p,t)$ does not exist for an exponentially growing or shrinking genome. However, one could define a normalized histogram $n_a(p,t) = 2N_a(p,t)/N_{\rm gene}^2$. Eq. 2.1 is then written as

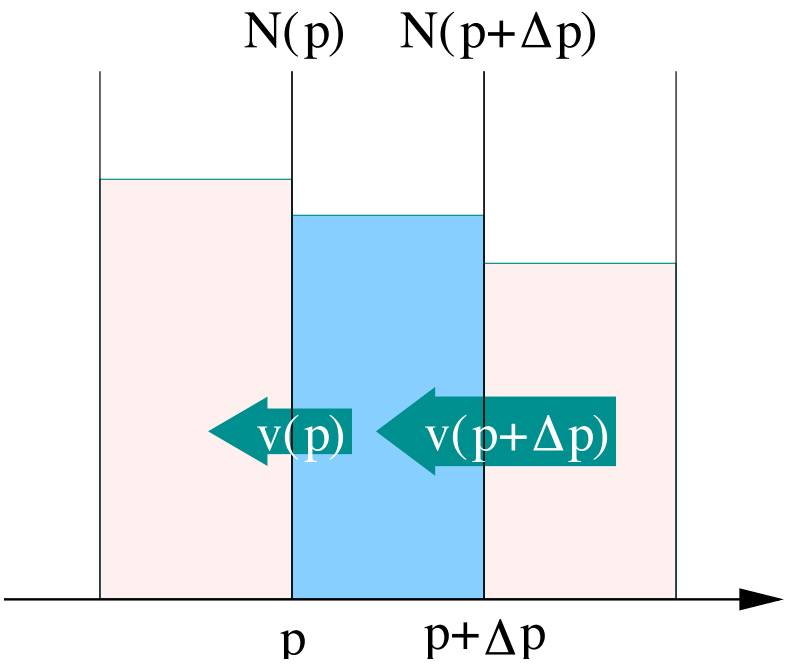

Figure 2.4: Illustration of amino-acid substitution on PID histogram. The PID decay generates a negative flux $v(p)N_a(p)$ down the PID-axis. The net flux into a given bin Δp is given by $v(p+\Delta p)N_a(p+\Delta p) - v(p)N_a(p)$.

$$\frac{\partial n_a(p,t)}{\partial t} = \frac{\partial}{\partial p}[v(p)n_a(p,t)] + \frac{2\alpha_{\rm dup}}{N_{\rm genes}}\delta(p-1). \tag{2.2}$$

The new equation describing the dynamics in the normalized histogram is much simpler, and there is a steady state solution given by

$$n_a(p) \sim N_a(p) \sim 1/v(p). \tag{2.3}$$

The conjecture that the normalized PID histogram $n_a(p,t) = 2N_a(p,t)/N_{\rm gene}^2$ indeed is nearly stationary during the course of evolution is corroborated by the fact that all six $n_a(p)$ curves in various genomes used in our study approx-

imately lie on top of each other in Fig. 2.2.

With the equations in hand, we are in the position to explain the near-universal power law behavior as shown in region II of the histogram. Comparing the empirical form of $N_a(p) \sim 1/p^4$ with Eq. 2.3, one concludes that the drift velocity in real genomes must obey $v(p) \sim p^4$. Such a non-linear dependence of $v(p)$ could be explained by the variability of the effective substitution rate within proteins (intra-protein variability). Assuming the intra-protein variability of substitution rates μ is described by a probability distribution $\rho(\mu)$, one gets the following expression for $p(t)$ and $v(t)$:

$$p(t) = \int_0^\infty \rho(\mu) e^{-2\mu t} d\mu \tag{2.4}$$

$$v(t) = -\frac{dp(t)}{dt} = \int_0^\infty 2\mu\rho(\mu) e^{-2\mu t} d\mu. \tag{2.5}$$

Eq. 2.4 could be looked as a generalization of the previously discussed exponential decay of $p(t)$ derived for a constant substitution rate μ. It simply weighs these exponentials by their likelihood of occurrence $\rho(\mu)$. For any given $\rho(\mu)$ one could eliminate time from Eqs. 2.4 and 2.5 and express v as a function of p. In the absence of an analytical expression relating $v(p)$ and $\rho(\mu)$ one is limited to use a trial-and-error method. We start with the Gamma distribution

$$\rho(\mu) \sim \mu^{\theta-1} \exp(-\mu/\mu_o). \tag{2.6}$$

which has been predominantly used in the literature [40, 41]. Inserting the defined $\rho(\mu)$ into Eqs. 2.4 and 2.5 one gets $p(t) = (1 + 2t\mu_o)^{-\theta}$ and $v(t) =$

$2\theta\mu_o(1+2t\mu_o)^{-(\theta+1)}$ which leads to

$$v(p) \sim p^{(\theta+1)/\theta} \tag{2.7}$$

$$N_a(p) \sim p^{-(\theta+1)/\theta}. \tag{2.8}$$

Intrinsic Parameter θ in Intra-Protein Substitution

Based on Eq. 2.7, the empirically detected power law $N_a(p) \sim 1/p^\gamma$ can be generated by our model if the intra-protein substitution rate distribution $\rho(\mu)$ follows a gamma-distribution with $\theta = 1/(\gamma - 1)$. The apparent p^{-4} dependence suggests that $\theta \approx 0.33$.

The Gamma-distribution $\sim \mu^{\theta-1}\exp(-\mu/\mu_o)$ is traditionally used to model and fit the distribution of substitution rates in individual families of proteins (this tradition goes back to [42]). Our approach extends this to a proteome-wide scale and demonstrates that beyond its role as a *ad hoc* fitting function, the Gamma distribution indeed provides an excellent quantitative description of variability of intra-protein substitution rates. The genome-wide value of the exponent $\theta \approx 0.33$ obtained in our analysis is consistent with its previous estimates in large protein families. For example, the fitting performed by Refs. [40, 41] resulted in the exponent θ in the $0.2 - 0.4$ range.

It is important to emphasize that our result - the power-law form of $N_a(p)$ depends only on the *intra-protein* variability of substitution rates at different amino-acid sites within the same protein. Such variability should not be confused with a much larger protein-to-protein variability of average substitution rates. Indeed, different proteins encoded in the same genome are known

to have vastly different average rates of amino-acid substitutions. Some sequences, such as those of ribosomal proteins, remain virtually unchanged over billions of years of evolution, while others evolve at a much faster pace. In fact, the very importance of a protein is sometimes quantified by its average rate of evolution as more essential proteins involved in core cellular processes tend to evolve at slower than average rates.

2.4 Extracting the Rates of Gene Duplication and Deletion

So far we have focused on the power law behavior in region II of PID histograms. In this section, we will explain features in region I and use our mathematical formalism to extract the relative rates of evolutionary processes in different genomes. The results are summarized in Table 2.2. In particular, the power law exponent we discussed are presented in the third column. Let us start by introducing the idea of true duplicated pairs in contrast to the set of all paralogous proteins.

Distribution of sequence identities of true duplicated pairs

Even though paralogs are originated from gene duplication, it is important to note that a pair of paralogs may not be a direct result of a single gene duplication event. Consider a simple example. The family of four evolutionary related proteins A,B,C,D contributes six paralogous pairs to $N_a(p)$. This family was actually created by three subsequent duplication events: first A

Organism	Proteome size	γ	$\alpha^*_{\rm dup}/\bar{\mu}$	$\alpha_{\rm dup}/\bar{\mu}$	$\alpha_{\rm del}/\bar{\mu}$	$\alpha^*_{\rm del}/\bar{\mu}$
H.pylori	1590	3.1	0.73	0.032	0.16	67
E.coli	4288	4.4	1.37	0.038	0.10	64
S.cerevisiae	5885	1.8	1.61	0.24	0.24	27
C.elegans	19099	4.2	3.16	0.27	0.37	41
D.melanogaster	14015	4.4	0.35	0.084	0.22	30
H.sapiens	25319	2.4	2.82	0.85	0.16	19

Table 2.2: Parameters of proteome evolution. The first column contains the name of the organism, the second column – $N_{\rm genes}$, the number of genes in its genome, the third column is the value of the exponent γ in the best fit with $p^{-\gamma}$ to $N_a(p)$ in the region II. The fourth, fifth, sixth and seventh columns are correspondingly the ratios $\alpha^*_{\rm dup}/\bar{\mu}$, $\alpha_{\rm dup}/\bar{\mu}$, $\alpha_{\rm del}/\bar{\mu}$, and $\alpha^*_{\rm del}/\bar{\mu}$ defined and measured as described in the text.

duplicated to give rise to B, then B duplicated to C and finally C duplicated to D. Thus only three out of total six paralogous pairs are directly produced in gene duplication events. The actual number of duplicated pairs could be even smaller if some intermediate genes were deleted in the course of the evolution. In general a family consisting of F proteins contributes at or around $F(F-1)/2$ paralogous pairs to $N_a(p)$, but only $F-1$ duplicated pairs, which we quantify by $N_d(p)$.

Nothing in the BLAST output for a given paralogous pair contains any information if it should or should not be included in the $N_d(p)$. However, using the set of all sequence identities of proteins for a given family, one could tentatively reconstruct the course of duplication events that led to the appearance of this family. Generally speaking, this is a complicated task involving building the most parsimonious phylogenetic tree for every family in a genome. In this study, we use a much simpler alternatives based on the Minimum Spanning Tree (MST) algorithm (see Appendix A). For a family consisting of F

proteins, this algorithm generates $F-1$ duplication events in its past history, contributing to the set $N_d(p)$. Numbers of pairs included in $N_a(p)$ and $N_d(p)$ distributions in different organisms are listed in the Table 2.1.

The dynamics of the distribution of duplicated pairs $N_d(p,t)$ is described by simply excluding the duplication term $2\alpha_{\rm dup}N(p,t)$ from the equation for $N_a(p,t)$. Indeed, this term is caused by PIDs of non-duplicated paralogs A'-B and A'-C (dashed lines in Fig. 2.3) generated when a gene A was duplicated. However, only the actual duplicated pair A-A' with initial PID of 100% (dotted line in Fig. 2.3) is included in the distribution of duplicated pairs $N_d(p)$. Thus the dynamics of N_d is described by

$$\frac{\partial N_d(p,t)}{\partial t} = \frac{\partial}{\partial p}[v(p)N_d(p,t)] - 2\alpha_{\rm del}N_d(p,t) + \alpha_{\rm dup}N_{\rm genes}\delta(p-1). \qquad (2.9)$$

Once again the stationary solution exists for the normalized distribution. However, in this case the correct normalization factor is given by $N_{\rm genes}$ and not $N_{\rm genes}^2/2$ as for $N_a(p)$. Thus the normalized PID histogram of duplicated pairs $n_d(p,t) = N_d(p,t)/N_{\rm genes}$ evolves according to

$$\frac{\partial n_d(p,t)}{\partial t} = \frac{\partial}{\partial p}[v(p)n_d(p,t)] - (\alpha_{\rm dup}+\alpha_{\rm del})n_d(p,t) + \alpha_{\rm dup}\delta(p-1). \qquad (2.10)$$

According to our empirical findings the average rate of sequence divergence of paralogous proteins in most organisms is described $v(p) = 2\bar{\mu}p^{\gamma}$, where $\bar{\mu}$ is the substitution rate averaged over all amino-acid positions in all proteins, and $\gamma \approx 4$ is the exponent related to the intra-protein variability of μ. By

solving the steady state solution of Eq. 2.10, one arrives at

$$N_d(p) \sim n_d(p) \sim \frac{1}{p^\gamma} \exp\left(-\frac{\alpha_{\text{dup}} + \alpha_{\text{del}}}{2\bar{\mu}(\gamma-1)p^{\gamma-1}}\right). \tag{2.11}$$

We will make use of this equation for extracting information on α_{dup} and α_{del} later on.

Deletion rate of recent duplicates

A very pronounced and reproducible feature in all organism-wide histograms is an abrupt drop as is lowered from 100% down to about 90-95% (region I in Fig. 2.1). The drop is as large as 30-fold in prokaryotes and is around 3-to-10 fold in eukaryotes. One of the most plausible explanation for this initial drop in the region I is that freshly duplicated genes are characterized by a much higher deletion rate, i.e. $\alpha^*_{\text{del}} \gg \alpha_{\text{del}}$ [43]. Functional roles of such genes have not had enough time to diverge from each other making each of them more disposable than an average gene in the genome. Indeed, for *S. cerevisiae* it was empirically demonstrated [44] that the deletion or inactivation of genes with a highly similar paralogous partner in the genome is up to 4 times more likely to have no consequences for the survival of the organism than the deletion/inactivation of genes lacking such a partner. A similar analysis along this line will be discussed in Chapter 3.

The N_a dynamics in the region I is described by

$$\frac{\partial N_a(p,t)}{\partial t} = \frac{\partial}{\partial p}[2\bar{\mu}N_a(p,t)] + (2\alpha_{\text{dup}} - 2\alpha^*_{\text{del}})N_a(p,t) + \alpha_{\text{dup}}N_{\text{genes}}\delta(p-1) \quad . \tag{2.12}$$

while the normalized distribution n_a obeys

$$\frac{\partial n_a(p,t)}{\partial t} = \frac{\partial}{\partial p}[2\bar{\mu} n_a(p,t)] + (2\alpha_{\rm dup} - 2\alpha^*_{\rm del}) n_a(p,t) + \frac{2\alpha_{\rm dup}}{N_{\rm genes}}\delta(p-1) \quad . \quad (2.13)$$

Here $2\bar{\mu} = v(100\%)$ is the average substitution rate in freshly duplicated pairs and $\alpha^*_{\rm del}$ is the deletion rate inside region I. For $\alpha^*_{\rm del} \gg \alpha_{\rm del}$, the equation has an exponentially decaying stationary solution given by $n_a(p) \sim \exp(\alpha^*_{\rm del} p/\bar{\mu})$. This functional form is consistent with the empirical data for p just below 100% and the best fits to $\alpha^*_{\rm del}/\bar{\mu}$ are listed in the seventh column of the Table 2.2.

Ref. [43] analyzed the distribution of silent substitution numbers per silent site K_s between pairs of recently duplicated genes. Under the same "drift and deletion" hypothesis used to derive the Eq. 2.12, the distribution of all duplicated pairs N_d in terms of K_s should also have an exponential decaying form $N_d(K_s) \sim \exp(-\alpha^*_{\rm del} K_s/\bar{\mu_s})$, where $\bar{\mu_s}$ is the average drift velocity of K_s immediately following the duplication event. Fits to this exponential functional form performed in Ref. [43] resulted in $\alpha^*_{\rm del}/\bar{\mu_s} \sim 7-24$. Our estimates $\alpha^*_{\rm del}/\bar{\mu} \sim 20-70$ are consistent with those of [43] provided that the $\bar{\mu}/\bar{\mu_s}$ ratio is in $0.1-1$ interval.

Long- and short-term duplication rates.

The number of paralogous pairs with PID$\simeq$100% also contains information about the raw duplication rate $\alpha^*_{\rm dup}$ in the genome. This rate is subsequently trimmed down to its long-term stationary value $\alpha_{\rm dup}$ by the removal of a large fraction of freshly created pairs as described in the previous subsection. New

pairs with PID=100% are created at a rate $\alpha^*_{\rm dup} N_{\rm genes}$, while they leave the bin containing PID=100% at a rate $2\bar{\mu} N_a(100\%)/\Delta p$. Here Δp is the width of the bin and $N_a(100\%)$ is the number of pairs in this last bin. The width of the bin is assumed to be small enough so that the removal of genes from the bin due to deletion is negligible in comparison to that due to the drift in their sequences. Thus $\alpha^*_{\rm dup}/\bar{\mu} = 2N_a(100\%)/(N_{\rm genes}\Delta p)$. The average duplication rates calculated this way are presented in the fourth column of Table 2.2 . They are compatible with $\alpha^*_{\rm dup}/\bar{\mu}_s$ calculated in [45], where the same idea was applied to $N_d(K_s)$.

The rate $\alpha^*_{\rm dup}$ includes the creation of some extra duplicated pairs which are then quickly (on an evolutionary timescale) eliminated from the genome during a "trial period" for PID>90%. We have already demonstrated that such a deletion happens at a very high rate $\alpha^*_{\rm del}$ and thus has to be treated separately from the background deletion rate $\alpha_{\rm del}$. The duplication rapidly followed by a deletion does not change the overall distribution of paralogous pairs. Therefore, the long-term average duplication rate $\alpha_{\rm dup}$ used in Eqs. 2.3 and 2.11 is in fact considerably lower than the raw duplication rate $\alpha^*_{\rm dup}$. An approximate way to calculate it is to use power-law fits to $N_a(p)$ in the region II to extrapolate it up to 100%. Such extrapolated value $N^{\rm ext}_a(100\%)$ could then be used to calculate the long-term average duplication rate as $\alpha_{\rm dup}/\bar{\mu} = 2N^{\rm ext}_a(100\%)/(N_{\rm genes}\Delta p)$ (see the fifth column of Table 2.2).

Long-term stationary deletion rate

Finally let us estimate the average of deletion rates. We performed a two-parameter fit to the $N_d(p)/N_a(p)$ ratio with $A\exp(-B/(\gamma-1)p^{\gamma-1})$ (see Eqs. 2.3 and 2.11) in the $30\% < p < 90\%$ interval (region II in Fig. 2.1). Here A and $B = (\alpha_{\rm dup} + \alpha_{\rm del})/(2\bar{\mu})$ are the two free fitting parameters. The exponent γ used in the fitting formula itself was obtained from the best fit to $N_a(p)$ in the same region with the power-law form $p^{-\gamma}$ (see column 3 in Table 2.2). The ratio $\alpha_{\rm del}/\bar{\mu}$ was extracted from the best-fit value of B. Using the previously estimated values of $\alpha_{\rm dup}/\bar{\mu}$ as shown in the fifth column, the ratio $\alpha_{\rm del}/\bar{\mu}$ are calculated (the sixth column of the Table 2.2).

Genome size dependence of average duplication and deletion rates

Our data indicate that the long-term stationary duplication rate $\alpha_{\rm dup}$ is of the same order of magnitude as the stationary deletion rate $\alpha_{\rm del}$ (compare columns 5 and 6 in the Table 2.2). This is to be expected since any large discrepancy in these rates would generate much larger differences in genome sizes than actually observed in these model organisms. However, as was proposed by [43] both of these rates are considerably smaller than their raw counterparts $\alpha^*_{\rm dup}$ and $\alpha^*_{\rm del}$ that include only recently duplicated pairs. Our rates for the fruit fly *D. melanogaster* are consistent with an earlier observation [45] of an abnormally low average duplication rate in this organism. According to our data $\alpha^*_{\rm dup}/\bar{\mu}$ is about 9 times lower than that in the genome of *C. elegans*. The long-term stationary duplication rate $\alpha_{\rm dup}/\bar{\mu}$ in the fly is also the lowest in all

eukaryotic genomes used in this study but is only 3 times lower than that in the worm.

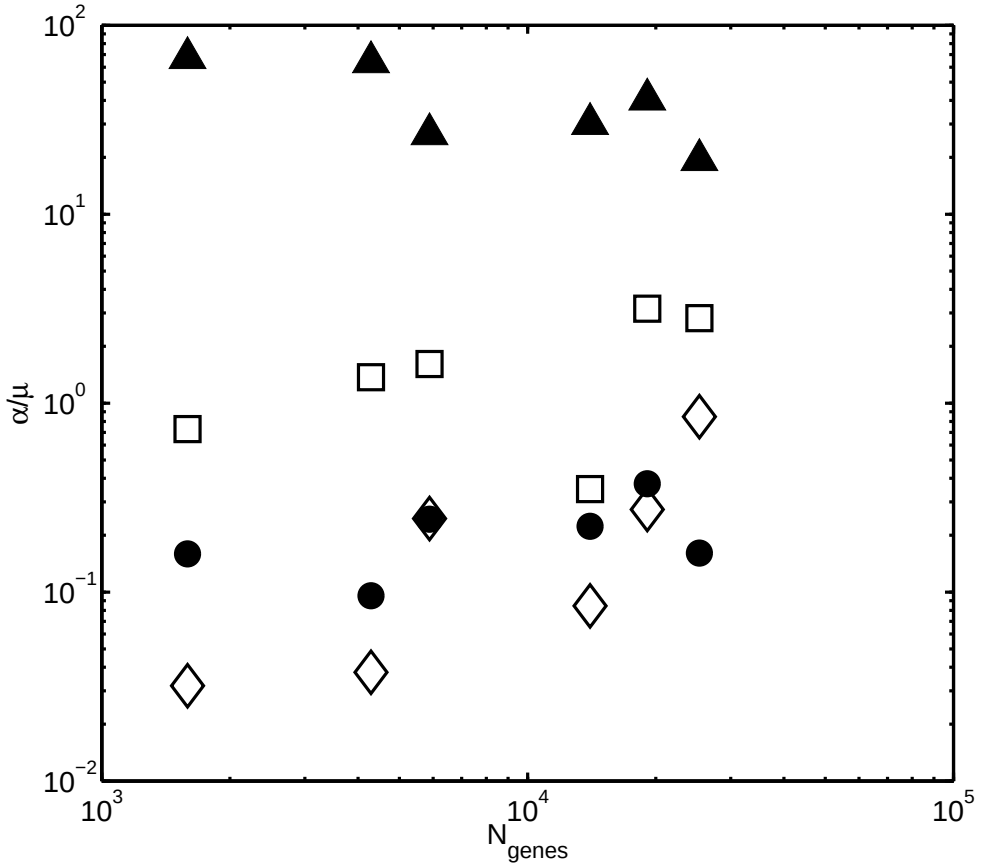

Figure 2.5: Correlation between the number of genes in an organism and its duplication/deletion rates. Evolutionary parameters $\alpha_{\rm dup}/\bar{\mu}$ (open diamonds), $\alpha_{\rm del}/\bar{\mu}$ (filled circles), $\alpha^*_{\rm dup}/\bar{\mu}$ (open squares), and $\alpha^*_{\rm del}/\bar{\mu}$ (filled triangles) plotted versus the total number of genes $N_{\rm genes}$ in an organism. Organisms in the order of increasing number of genes are *H. pylori, E. coli*, *S. cerevisiae*, *D. melanogaster, C. elegans*, and *H. sapiens*. As explained in the text, more complex organisms (those with larger $N_{\rm genes}$) tend to be characterized by higher values of the first three ratios but lower values of the last ratio.

Intriguingly $\alpha_{\rm del}/\bar{\mu}$, $\alpha_{\rm dup}/\bar{\mu}$, and $\alpha^*_{\rm dup}/\bar{\mu}$ ratios are all positively correlated

with the complexity of the organism quantified by the total number of genes in its genome (see correspondingly filled circles, open diamonds, and open squares in Fig. 2.5). This means that either the per-gene duplication rate in more complex organisms is consistently higher than in their simpler counterparts or that their average amino-acid substitution rate is lower. It is likely that both trends operate simultaneously. A plausible explanation for the latter trend is that more sophisticated mechanisms of DNA copying and repair of higher organisms lead to lower average amino-acid substitution rates.

On the other hand, consistent with findings of Ref. [46], we find that the deletion rate of recent duplicates, $\alpha^*_{\rm del}/\bar{\mu}$, (filled triangles in Fig. 2.5) exhibits a negative correlation with the number of genes in the genome. A likely explanation of this correlation proposed in Ref. [46] is via the decrease in the effective population size N_e in more complex organisms.

2.5 Effects of Whole Genome Duplications on the Histogram of Sequence Identities

Two of the organisms used in our study (*S. cerevisiae* and *H. sapiens*) are characterized by a dramatically lower value of the power-law exponent γ (1.8 for yeast and 2.4 for human) and the overall poor quality of the power law fit to $N_a(p)$. One plausible explanation is in terms of Whole Genome Duplications (WGD) in lineages leading to these genomes. It is well established [47] that baker's yeast underwent a WGD event, which most likely occurred about 100 millions years ago. While the subject remains controversial, it is now com-

monly believed that vertebrate lineage leading to human also underwent one or several large-scale duplication events [48, 49]. In the immediate aftermath of a whole genome duplication event the PID distribution change as follows: $N_a(p) \to 4N_a(p)$ for p<100%, while $N_a(100\%) \to 4N_a(100\%) + N_{\rm genes}$. Indeed, every ancestral paralogous pair A-B would give rise to 3 new pairs with the same PID: A-B', A'-B, and A'-B'. At the same time the bin containing the PID=100% would in addition get $N_{\rm genes}$ (or fewer for a large segmental duplication) of freshly created duplicated pairs of the type A-A' and B-B'. The subsequent spread of this enormous peak at PID=100% towards lower values of PID accompanied by a rapid deletion of redundant copies would result in an effective flattening of the $N_a(p)$ histogram in its upper range and thus lower effective value of the exponent γ.

To further test this hypothesis we analyzed the recent sequenced genome [50] of a ciliate *Paramecium tetraurelia*. This organism underwent as many as four separately identifiable WGD events [50]. We used our standard methods to construct the PID histogram $N_a(p)$ from the all-to-all alignment of its nearly 40,000 genes (see Appendix A). Solid diamonds in Fig. 2.6 correspond to its full PID histogram consisting of 103,828 paralogous pairs. Ref [50] identified the lists of putative pairs of duplicated genes generated in each of the four WGD events in the lineage leading to this genome. By dropping one randomly-selected gene from these WGD pairs, we generated the set of four progressively more narrow PID histograms. These histograms are also shown in Fig. 2.6: 41,890 pairs excluding the genes generated in the latest WGD event (solid squares), 25,342 pairs excluding those generated in the latest two WGD events (solid circles). 22.287 pairs excluding the genes generated in the last three

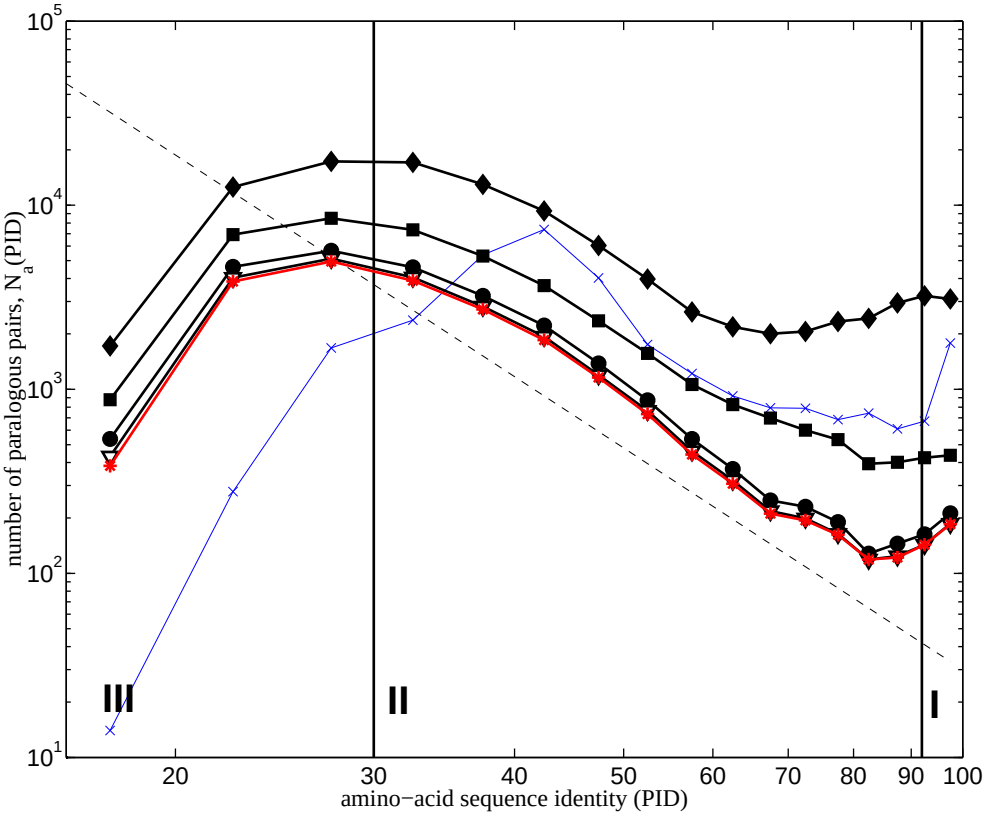

Figure 2.6: Histograms of sequence identities in *P. tetraurelia* and the relationship with WGD. The histogram of sequence identities of 103,828 paralogous pairs among 39642 proteins in the genome of *Paramecium tetraurelia* (solid diamonds) detected by an all-to-all blastp alignment. Other histograms shown in this plot correspond to a progressive removal of new proteins created in the four WGD events [50]: 41,890 pairs among 27,616 proteins excluding those generated in the latest WGD event (solid squares), 25,342 pairs among 23,618 proteins excluding those generated in the latest two WGD events (solid circles). 22,287 pairs among 22,853 excluding those generated in the last three WGD events (open triangles), and 21,417 among 22,635 proteins excluding those generated in all four WGD events (red stars). One can see that by progressive elimination of pairs generated in WGD events the functional form of the $N_a(p)$ histogram in *P. tetraurelia* approaches the universal scaling form: $N_a(p) \sim p^{-4}$ (dashed line). For comparison we copy from Fig. the histogram of 31,078 pairs among 25,319 *H. Sapiens* proteins (blue x's).

WGD events (open triangles), and 21,417 excluding those generated in all four WGD events (red stars). One can see that by progressive elimination of pairs generated in WGD events the functional form of the $N_a(p)$ histogram in *P. tetraurelia* approaches the universal scaling form: $N_a(p) \sim p^{-4}$ (dashed line). The analysis of *P. tetraurelia* provides an additional strong support to our conjecture that the unusually flat PID histograms in human and yeast are caused by WGD events in lineages leading to these two organisms (for comparison, Fig. 2.6 also reproduces the histogram of *H. sapiens*).

2.6 Conclusion and Outlook

We have introduced a stochastic birth and death model of proteome evolution. Several versions of such models were previously [3, 4, 51], used in the context of power law distribution of protein family sizes. Our model extends these attempts by concentrating on the evolution of sequence identities as opposed to just the number of proteins in families.

The idea of quantifying evolutionary parameters using the histogram of some measure of sequence similarity of duplicated genes in itself is not new. It was discussed by Gillespie (see [52] and references therein) and later applied by Lynch *et.al.* [43] to measure the deletion rate of recent duplicates. There are two important differences between our methods and those of the Ref. [43]: 1) We use the histogram of amino-acid sequence identities as opposed to that of silent substitutions used in Ref. [43]. This allows to dramatically extend the range of evolutionary times amenable to this type of analysis. 2) We combine the study of a highly redundant dataset of all paralogous pairs with that of

protein pairs that were actually created by a duplication event. The shape of the histogram of sequence identities in the former set is to a first approximation independent of duplication and deletion rates and thus allows us to study fine properties of amino-acid substitution rates.

Probably the best demonstration of universality by our birth-and-death model is value of θ. Our model is based on a simplified picture of genome evolution. In particular, we implicitly assumed the neutrality of gene duplication and deletion events and thus the homogeneity of duplication and deletion rates for different proteins. Such an assumption is, strictly speaking, not true. Families containing essential genes were recently shown to be characterized by higher average duplication and deletion rates [53]. However, we would like to emphasize that the validity of the exponent goes beyond the validity of the approximation. The advantage of using the histogram of sequence identities generated by the all-to-all alignment lies in its remarkable universality and robustness. When the formalism is applied to individual families one can see that family-to-family variation of (and correlations between) the duplication rate $\alpha_{\rm dup}$, the deletion rate $\alpha_{\rm del}$, and the average substitution rate μ_0 affect only the pre-factor in the powerlaw form of $N_a(p)$. Thus the exponent $\gamma = 1 + 1/\theta$ describing this power law is very robust and depends only to the exponent θ quantifying the *intra-protein* variability of amino-acid substitution rates.

The exact mechanisms behind are not entirely clear. Chances are that it is dictated more by the protein physics rather than by organism-specific evolutionary mechansims. A possible path towards an explanation of the exponent θ from purely biophysical principles starts with the results of Ref. [54], which models the effects of (correlated) multiple amino-acid substitutions on stability

of the native state of a protein.

Chapter 3

Evolution of Molecular Networks

3.1 Background

Evolution modifies an organism on multiple levels, ranging from sequences of individual molecules, their coordinated activity in the cell (molecular networks), all the way up to the phenotype of the organism itself. In the previous chapter, we looked at evolution at the level of protein sequences. In this chapter, we move to a higher level and look at evolutionary changes at the level of protein networks, including protein interaction networks and transcriptional regulatory networks.

Evolution in protein networks is the result of changes in individual proteiins. Immediately after a duplication event the pair of freshly duplicated proteins is thought to be identical in both sequences and functional roles in the cell. With the presence of two copies, the duplicates are allowed to accu-

mulate various mutations as the selection pressure is relaxed. The functions of the ancestral gene might then be shared by two duplicates, and each of them might independently develop new functions. This process causes functional divergence for the duplicates. Even though the picture is intuitive, the term "function" is subtle and difficult to quantify. Networks thus offer a more concrete approach in quantifying functional divergence.

The fate of duplicated genes can be studied by looking at the position of paralogs in a protein network. Right after a gene duplication event, the two copies share the same set of neighbors. With subsequent mutations, the number of common neighbors reduces and each of them has independently gained new neighbors. We use a concept of overlap to quantify this divergence. For a pair of paralogs, the overlap Ω is defined as the number of common neighbors they have in the network. To take into account the original number of neighbors, normalized overlap could be used (see Fig. 3.1 for illustration).

In this chapter, we quantify the functional divergence of paralogs using the idea of overlap for a number of system-wide datasets. We start by looking at baker's yeast *S. cerevisiae*. To this end we measure: 1) The similarity of positions of paralogs in the transcription regulatory network given by overlap in transcription regulators, 2) The similarity of the set of binding partners. These measures reflect, correspondingly, the upstream and downstream properties of molecular networks around duplicated genes. We then repeat the analysis using species-wide data on protein interaction networks in a bacterium *H. pylori* and a fruit fly *D. melanogaster*. We end this chapter by looking at the functional roles of duplicated proteins, more precisely the ability to substitute for each other. This is demonstrated using data from a single gene knockout

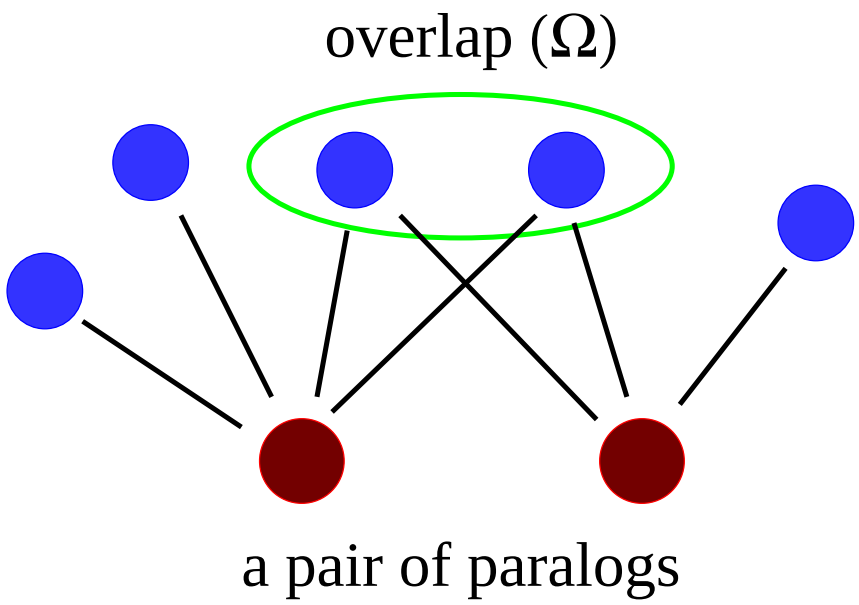

Figure 3.1: Illustration of the concept of overlap in a molecular network. For a pair of paralogs the overlap Ω is defined as the number of common neighbors they have in the network. In the case of transcription network the regulatory overlap Ω_{reg} counts transcription factors regulating both paralogs, while for the physical interaction network the interaction overlap Ω_{int} counts their common binding partners. The pair of paralogs used in this illustration has the overlap $\Omega = 2$ out of the total of 5 distinct neighbors of the pair. That corresponds to a normalized overlap of $2/5 = 0.40$.

experiment in *S. cerevisiae* and a RNAi experiment in a nematode worm *C. elegans*.

3.2 Divergence of Duplicated Genes in Networks of *S. cerevisiae*

The first measure of the divergence of duplicated genes compares sets of their transcriptional regulators. Such a set contains information about different

conditions under which a given gene is expressed, and thus reflects the spectrum of its functional roles in the cell. In this context, the overlap of a pair of paralogs is given by the number of transcription factors that bind to upstream regulatory regions of *both* these genes. The system-wide data for the transcription regulatory network in yeast was taken from the ChIP-on-chip experiment by Lee *et al.* [16] which investigated *in vivo* binding patterns between 106 yeast transcription factors and upstream regulatory regions of all 6270 yeast genes. Using the set of paralogous proteins in yeast (see Appendix A for details), we study for each pair of paralogs, the relation between the overlap and their sequence similarity. As shown in Fig. 3.2A, one can see that the regulatory overlap has a tendency to decrease as a function of PID. While multiple overlaps dominate the distribution for PID $\geq$ 80%, they gradually disappear at lower PIDs.

Fig. 3.2B shows the average value of the regulatory overlap as a function of PID. The regulatory overlap in this plot is normalized by a proxy to the ancestral connectivity of a gene, estimated as the total number of distinct transcription factors that are involved in regulation of at least one of the pair of proteins (see Fig. 3.1). The correlation between the normalized regulatory overlap Ω_{reg} and the PID is highly statistically significant: the Pearson correlation is 0.34 (P-value around 10^{-70} for 2275 data points). Even for the lowest value of PID=20% the average Ω_{reg} significantly exceeds its value in non-paralogous proteins. One interesting feature of the graph in Fig. 3.2B is that even pairs of proteins whose amino acid sequences are 100% identical to each other on average have only about 30% overlap in their upstream regulation. Such low regulatory overlap of recently duplicated genes can be partially

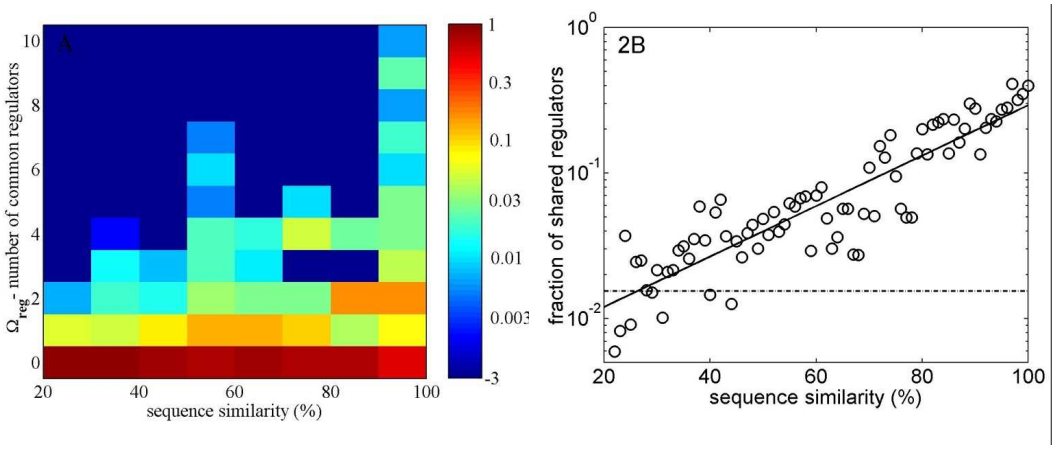

Figure 3.2: Divergence of the upstream transcriptional regulation of duplicated genes in yeast. A) The distribution of the regulatory overlap Ω_{reg} of paralogous proteins. The Y-axis – Ω_{reg} – is the number of transcription factors that cis-regulate *both* genes encoding a pair of paralogous proteins. The X-axis is the percent identity (PID) of amino acid sequences of these two proteins. The colorbar shows the likelihood of finding a given Ω_{reg} in a given PID bin (note the logarithmic scale). B) The average regulatory overlap Ω_{reg} normalized by the total number of regulators of either one or the other paralog plotted as a function of the PID. Error bars are estimated by the square root of the total number of shared regulators in a PID bin. The solid line is the best fit to the exponential form: $\Omega_{reg} \sim \exp(\gamma PID)$ with $\gamma = 0.03$. The dashed horizontal line at 0.015 is a null-model expectation of the normalized overlap of two randomly selected proteins (not necessarily paralogs).

attributed to false positives and false negatives present in the dataset. It might also be sometimes caused by an incomplete duplication of the upstream regulatory region of a gene, or by a burst of very rapid evolution of the regulatory region immediately following the duplication event. The second feature of the Fig. 3.2B is a gradual decline of the average regulatory overlap over the whole range of sequence similarities. The data in Fig. 3.2B can be fitted with an exponential decay with a rate corresponding to an average 3% loss of common regulators of a paralogous pair for every 1% decrease in their amino

acid sequence identity. Thus already at PID=80% about half of the common regulations present at PID=100% are lost.

There are studies which are similar but nicely complement our findings. One is by Gu *et.al.* which reported the decline of similarity between microarray profiles of paralogs [55]. In fact, due to a more direct information about transcriptional regulation contained in the ChIP-on-chip dataset of [16] compared to microarray experiments, our analysis extends the gradual decline to much lower PID than was detected in Ref. [55]. Another study by Papp *et.al.* [56] has reported a rapid decline in the number of shared regulatory motifs of duplicated genes (a short piece of DNA in which transcription factors bind to). This study is carried out as a function of a much faster silent substitution rate K_s compared to our PID. Indeed, in their analysis Papp *et al.* logarithmically binned the K_s into four broad bins: below 0.01, 0.01-0.1, 0.1-1, and above 1. Since the reliability of the measured silent substitution rate dramatically decreases at high values of K_s, the whole long-time behavior (i.e. that for PID $< 75\%$ which in yeast roughly corresponds to $K_s > 1$) of the regulatory overlap remained inaccessible to the analysis of Ref. [56].

The rate of divergence between sets of *upstream* transcriptional regulators of paralogous proteins has an obvious *downstream* counterpart: it is the rate at which paralogous transcription factors loose their downstream targets. Unfortunately, an attempt to quantify this rate using the same dataset that we used above for the rate of upstream divergence would be limited to only 4 paralogous pairs formed by 106 transcriptional regulators studied in [16].

We now consider the second measure of the divergence, systematically comparing functional roles of duplicated (paralogous) proteins in the physical in-

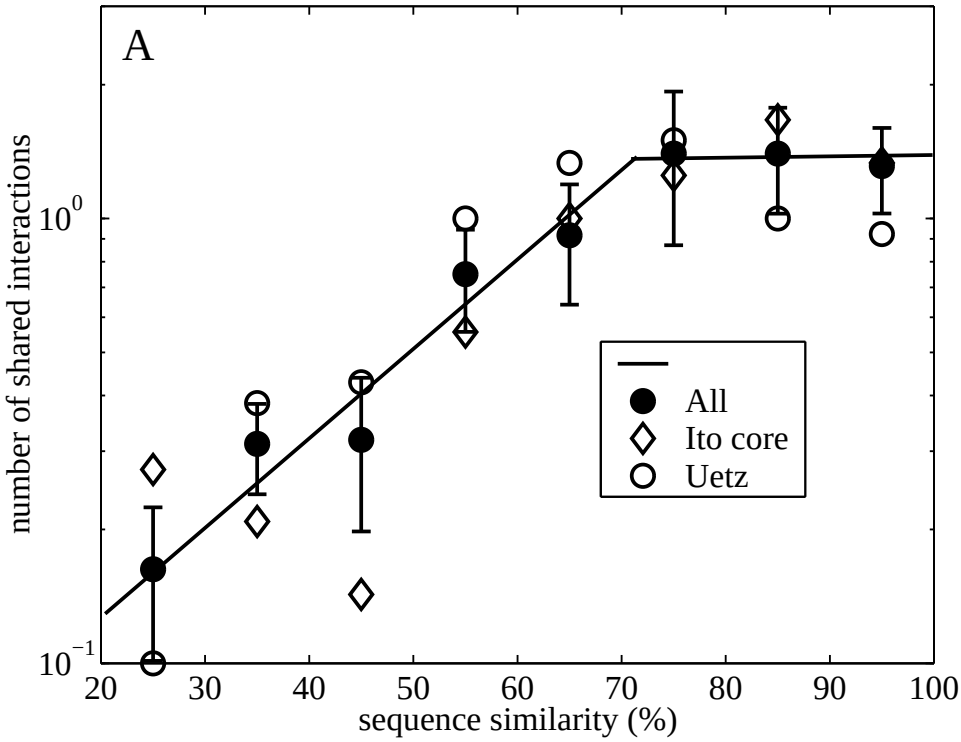

Figure 3.3: Divergence of the downstream function of duplicated genes. The average value of the interaction overlap Ω_{int} – the number of physical interaction partners shared by a pair of paralogous proteins – as a function of the similarity of their amino acid sequences. The physical interaction data are taken from the set of Uetz *et al.* [13] (crosses), core dataset of Ito *et al.* [14] (diamonds), and the non-redundant combination of the two (filled circles). Note the apparent plateau for PID's between 60% and 100% in both datasets. Solid lines are guides for the eye. The average interaction overlap in the combined dataset for a random (usually non-paralogous) pair of proteins is equal to 8×10^{-3} (off-limits in the figure).

teraction network. The functional similarity of a pair of proteins is in part reflected in the "interaction overlap" Ω_{int} given by the number of other proteins that physically interact with both of them (See Fig. 3.1). In our study we use the system-wide information about protein-protein physical interactions obtained by combining two high-throughput two-hybrid experiments: Uetz [13] and Ito [14]. Fig. 3.3 shows the average value of the interaction

overlap Ω_{int} between pairs of paralogous proteins as a function of PID – their amino-acid similarity. Again Ω_{int} typically decreases with decreasing PID, reflecting the gradual loss/change of binding partners of proteins in the course of evolution. A similar analysis, but as a function of the silent substitution rate (K_s) was previously reported by Wagner [57]. In agreement with that study, we find that paralogous proteins are more likely to share interaction partners than one expects by pure chance alone (see the caption to the Fig. 3.3). Our set of yeast paralogs contains 189 paralogous pairs such that both paralogs physically interact with at least one other protein in the combined dataset of Refs. [13, 14]. Out of these pairs 60 (30%) share at least one interaction partner. The correlation between the Ω_{int} and the PID in the combined two-hybrid dataset is highly statistically significant: the Pearson correlation is 0.36 (P-value around 5×10^{-6} for 189 data points). The most interesting observation from Fig. 3.3 is that the divergence in the set of binding partners becomes systematic only for PID$<$ 70%, while above 70%, it remains roughly constant in both Uetz, Ito and combined datasets .

Having presented different measures of upstream and downstream divergence of duplicated genes in yeast *S. cerevisiae* we are now in a position to discuss them in a wider context. Comparing Fig. 3.2 to Fig. 3.3 one concludes in yeast the upstream regulation of genes evolves more rapidly than downstream functions of their protein products. Indeed, the overlap in the set of binding partners remain virtually constant down to PID of 70%, at which point their average regulatory overlap has dropped to about 40% of its maximum. This is in accordance with a view which puts regulatory changes as one of the main driving forces of evolution.

3.3 Divergence of Physical Interactions of Paralogous Genes in *H. pylori* and *D. melanogaster*

The analysis of evolution of molecular networks advocated in this paper requires a *large* (preferably genome-wide) and *unbiased* (i.e. no anthropogenic selection present in databases) dataset describing a molecular network in a given species. Apart from yeast, which is arguably the best studied model organism, system-wide two-hybrid physical interaction assays were published for a simple bacterium *H. pylori* [12], and a fly *D. melanogaster* [15]. In Fig. 3.4 we used these two datasets to quantify the decay of the average interaction overlap as a function of amino-acid sequence similarly. The correlation between Ω_{int} and PID is highly statistically significant in both cases: the Pearson correlation of 0.43 (P-value around 3×10^{-4} for 65 data points) for *H. pylori*, and 0.19 (P-value around 10^{-26} for 2843 data points) in *D. melanogaster*.

Our basic conclusions agree for all quite diverse organisms used in this study: paralogous proteins are much more likely to share binding partners than expected by pure chance alone. Furthermore, the number of common interaction partners goes down as PID of their amino acid sequences decreases. In yeast and *H. pylori* we see the evidence of an initial plateau at which the average overlap appears to be independent of PID. On the other hand in *D. melanogaster* there is no evidence of such plateau, which makes the average rate of loss of common binding partners (about 4.5% for every 1% of change in PID) quite high in this organism. However, in the absence of system-wide data on transcription factors' binding in *D. melanogaster* and *H. pylori* we could not quantify rates of upstream changes in these two organisms, and consequently

cannot compare them to the corresponding downstream rates.

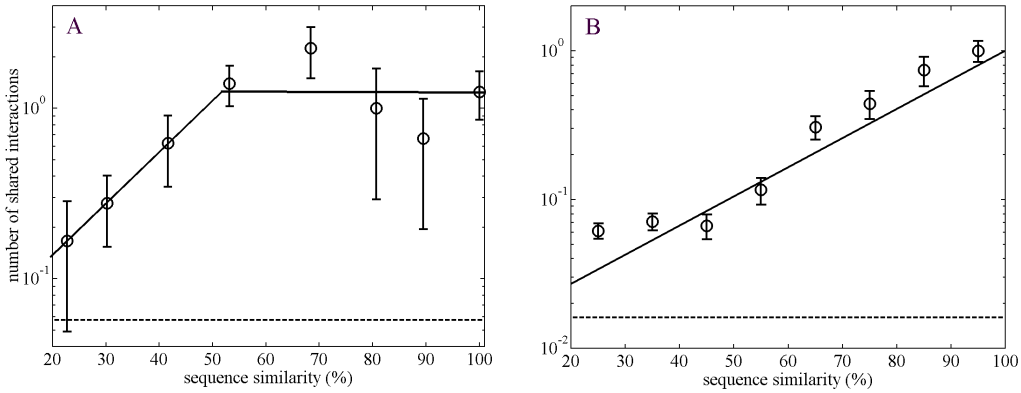

Figure 3.4: Divergence of the physical interaction neighborhoods of duplicated genes in *H. pylori* and *D. melanogaster*. The average value of the interaction overlap Ω_{int} of paralogous proteins in *H. pylori* (A) and *D. melanogaster* (B) as a function of the amino acid sequence similarity. The physical interaction data are taken from Ref. [12] for *H. pylori* (A) and from Ref. [15] for *D. melanogaster*. Note the apparent plateau for PID's between 50% and 100% in panel A and its absence in panel B. Dashed horizontal lines show the average interaction overlap for a random (usually non-paralogous) pair of proteins. The solid line is the best fit to the exponential form: $\Omega_{int} \sim \exp(\gamma PID)$ with $\gamma = 0.045$.

3.4 Functional Divergence: Robustness against Knockout

Apart from counting neighbors in networks, to quantify the extent of divergence/redundancy of paralogs, one could examine phenotypes of null-mutants lacking one of them. This is an alternative but more direct appraoch. Gu *et.al* used a systematic gene-deletion study in yeast to compare the fraction of essential genes (so that their null-mutants have lethal phenotype) between genes

with and without paralogs in the genome [44]. It was found that the fraction of essential genes is approximately 4 times higher among singleton genes than among ones protected by a highly similar paralog. It was also demonstrated that such protection by a paralog persists down to rather low levels of its amino-acid sequence similarity (PID) with the deleted protein. Using a more recent and larger systematic study [58] of viability of null-mutants in yeast (see Appendix A), we confirm these findings. As shown in Fig. 3.5, proteins having paralogs similar to themselves are less likely to be essential. Notice that the fraction of essential proteins shows a dramatic increase as the PID to their closest paralog falls below 50%. Thus paralogous proteins with sequence similarity above 50% can typically substitute for each other. Interestingly, the magnitude of this protective effect is the strongest in the nucleus, where the largest fraction of essential proteins resides.

One might expect the protective effect of paralogs to be unique to single-celled organisms such as yeast. Indeed, in multicellular organisms duplicated proteins are often expressed only in specific tissues and therefore unable to substitute for each other. However, using a systematic study of RNAi (RNA Interference) phenotypes in a nematode worm *C. elegans* [59] (see Appendix A for details of the data) we found such protection to be equally strong in this multicellular organism (See Fig. 3.6). As in Fig. 3.5, the x-axis in Fig. 3.6 is PID – the similarity of amino acid sequences between a given protein and its closest related paralog (all singleton proteins without paralogs are clumped into the 0% PID bin). The y-axis is the fraction of tested proteins whose elimination by the RNAi technique was found [59] to give rise to a nonviable phenotype (embryonic or larval lethality or sterility). In worm the protection

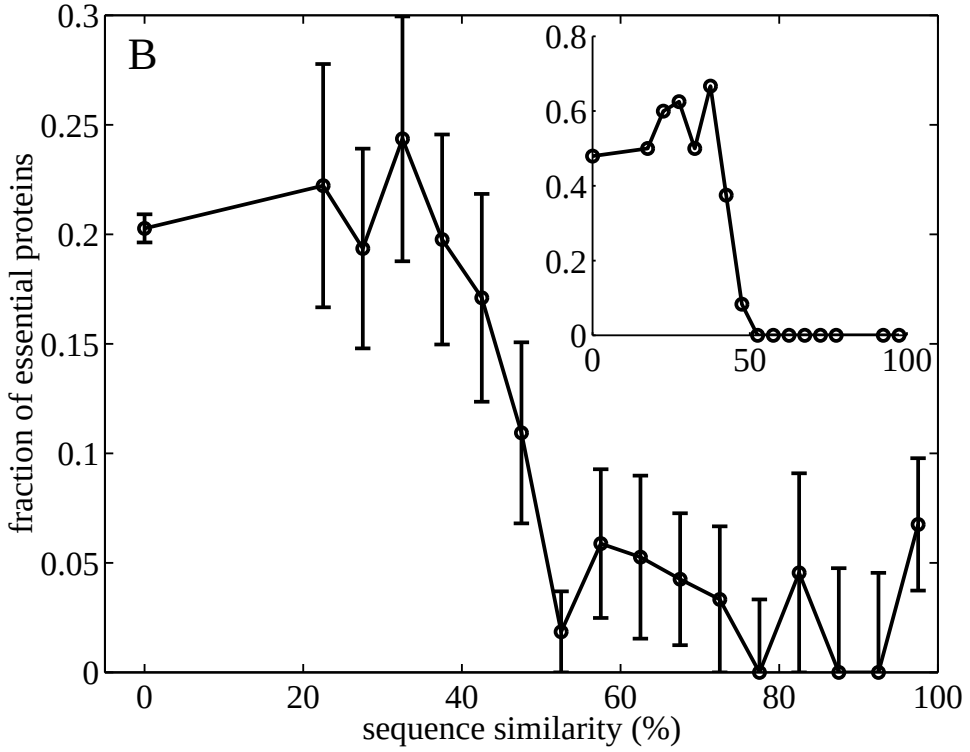

Figure 3.5: Protective effect of paralogs in a *S.cerevisae*. The fraction of essential (lethal null-mutant) proteins among all proteins tested in Ref. [58] as a function of PID to their most similar paralog in the yeast genome. Proteins with no paralogs (singletons) are binned at 0% PID. The inset (note the change of scale on the y-axis) shows the fraction of essential proteins in the subset of all proteins known to be localized in the yeast nucleus. Here the effect becomes even more pronounced so that all 18 nuclear proteins protected by a paralog with at least 50% similarity were found to be non-essential.

of having a paralog starts to gradually weaken for PID $< 70\%$. In both worm and yeast there seems to be a four-fold drop in the fraction of essential proteins between PID=0% and 100%.

In the inset to Fig. 4 we kept all successfully cloned genepairs, while in the main panel we dropped those genepairs whose product was predicted [59] to target mRNA product of more than one gene in the genome. It is

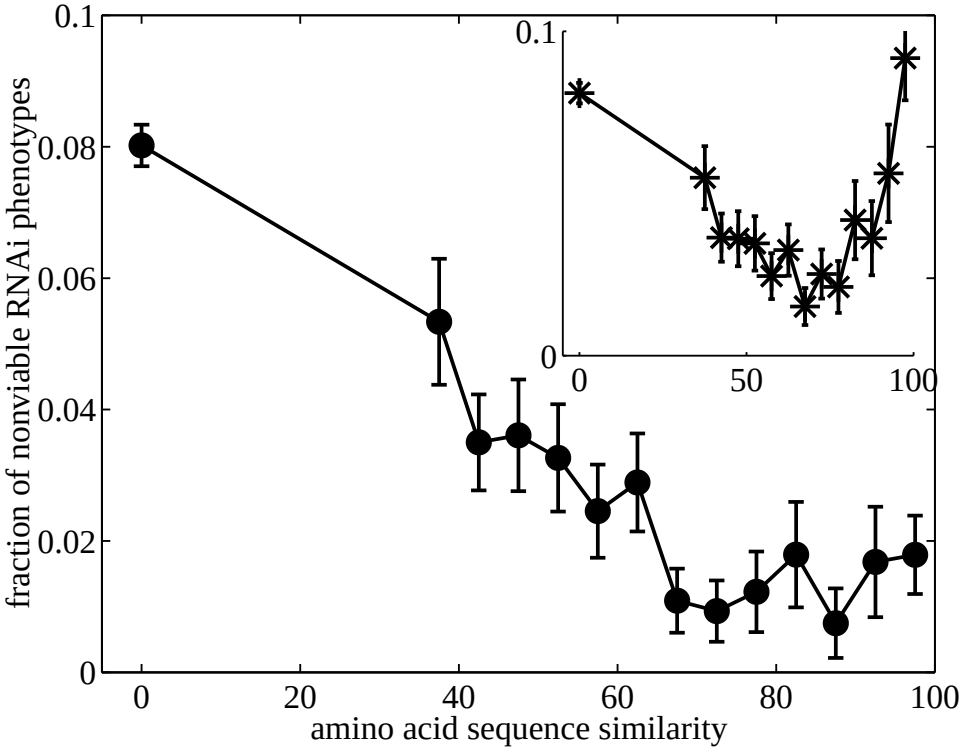

Figure 3.6: Protective effect of paralogs in a nematode worm *C. elegans*. The fraction of essential (non-viable RNAi phenotype [59]) proteins among all tested worm proteins as a function of PID to their most similar paralog in the worm genome. The plot in the inset uses all RNAi phenotypes reported in Ref. [59], while the main panel drops RNAis that are predicted to target mRNA products of more than one gene. Note that while the graph in the main panel is qualitatively similar to that in Fig. 3B, in the inset the fraction of essential proteins at PID=100% rises to its level for singleton proteins. Thus when mRNAs of highly similar paralogs are eliminated along with the target mRNA, the protective effect totally disappears.

instructive that the fraction of essential genes as a function of PID shown in the inset to Fig. 4 has a well pronounced minimum around PID=70% and then subsequently starts to rise for higher values of PID. The tentative explanation for this behavior is that unlike single-gene deletion technique used in yeast, the RNAi technique is based on RNA complementarity and can eliminate several

different mRNAs with similar sequences. Therefore, paralogous genes with nearly identical DNA sequences prove to be useless from the point of view of protection against RNAi since their mRNA products would be eliminated at nearly the same rate as the intended targets. This neatly explains why in the inset to Fig. 4 the fraction of nonviable phenotypes for genes with a 100% identical paralog in the genome approaches that of unprotected genes without paralogous partners (keep in mind that in this plot we use amino acid sequence identity of proteins and not of their mRNA precursors.) This observation also reinforces the point of view that the decline in the fraction of essential genes vs PID shown in Figs 3.5 and 3.6 is indeed caused by protective effects of paralogs and cannot be explained by a possible tendency of nonessential genes to duplicate more frequently.

3.5 Conclusion and Outlook

Our results showed evidences of duplication and divergence in molecular networks. For all molecular networks studied in this work, we found that even the most distantly related paralogous proteins on average have more similar positions within a network than a randomly selected pair of proteins. That means that paralogous proteins are likely to at least partially retain their functional redundancy for extremely long time after the duplication event. This is further supported by the analysis of yeast knockout experiment and the worm RNAi experiment.

There are numerous examples that redundancy resulting from gene duplication is important in achieving robustness in biological systems [60]. However,

it is not clear to what extent biological robustness is maintained by genetic redundancy. It is likely that redundancy is not the most important factor. For example, using the same RNAi dataset as we did, Ref. [61] found that out of the 16000 genes, more than 7500 singleton genes show no detectable phenotypic effect. Apart from redundancy, network organization is widely recognized as a way to achieve robustness. Even though there are models which capture the robust behaviors of biological systems, a general quantification of biological robustness is still not available.

Our results also indicate that the genetic regulation of paralogous proteins changes faster than both their amino acid sequences and the set of their protein interactions partners. It is tempting to extend this observation to pairs of homologous proteins in different species (orthologs) that diverged from each other as a result of a speciation (as opposed to a gene duplication) event. This would help to explain how species with very similar gene contents can evolve novel properties on a relatively short timescale. However, such an inter-species comparison of molecular networks has to wait for the appearance of whole-genome data on molecular networks in closely related model organisms.

Chapter 4

Dynamics and Noise in Protein Binding Network

4.1 Background

Recent high-throughput experiments have revealed networks of protein protein interactions (PPI) that are interconnected on a genome wide scale. In the last chapter, we have studied PPI networks from an evolutionary perspective. In this chapter, we go further away from the basic topology and study the dynamical processes on these networks.

An important element associated with every protein in a PPI network is its abundance. Experiments in yeast [24, 62] showed that protein abundance is highly heterogeneous. Average protein concentrations range between 50 to 10^6 molecules per cell with a median value around 3000. As binding is a reversible process, a cell consists of a mixture of free proteins and dimers in dynamical equilibrium. Equilibrium is attained when the forward reaction

(binding) rate balances the backward reaction (unbinding) rate. The reaction rates are proportional to the product of concentrations of reactants. This is commonly known as the law of mass action (LMA).

What happens if the equilibrium is perturbed? Suppose the total concentration of a certain protein is increased by a significant amount, say two-fold, the free concentrations (and at the same time dimers concentrations) of its neighbors will be perturbed, and changes may propagate to its next nearest neighbors and so on. Recent studies [63, 64] show that on average, the magnitude of cascading changes in equilibrium free concentrations exponentially decays with the distance from the source of perturbation. Therefore in general undesirable cross-talks presented by such a highly connected network are suppressed. Of further interest is that a significant number of pathways are found, along which perturbations can survive over a substantial length (up to 4 steps). These pathways are suggested as possible candidates for functional signalling.

While Refs. [63, 64] have focused on large perturbations (real signals) to the equilibrium state, protein concentrations are subject to stochastic fluctuations in the course of their production and degradation. This affects the concentrations of free proteins and dimers. Apart from this, thermal fluctuations exist in the binding and unbinding processes. All these result in dynamical fluctuations in the LMA equilibrium state. In this work, we analytically and numerically studied the individual effect of thermal noise and noise originated from protein production and degradation in a PPI network, we found that the network affects the two kinds of noise quite differently.

First, we will present the general formalism, the PPI network of a model

organism *S. cerevisiae* (baker's yeast) will be used for detailed calculation. Readers are asked to refer to Appendix B for details of the datasets. While its systemwide interaction network and protein abundance are well documented, the binding energies between proteins are not extensively measured. In the absence of genome-wide information regarding the value of dissociation constants, we assume them to be the same $K_{ij} = K_d$. In the following study, we use $K_d = 10\text{nM} \sim 340$ molecules/cell in yeast, which is comparable to the average of the dissociation constants found in the PINT database [65]. With the use of a homogeneous dissociation constant, however, one cannot infer the biology of individual proteins. Nonetheless, we can focus specifically on the role of network topology and protein concentrations. It should be stressed that the general formalism can further be applied for cases involving heterogeneous dissociation constants.

4.2 General Formalism of Temporal Variation

We represent a system of N distinct types of interacting proteins by an $N \times N$ adjacency matrix A ($A_{ij} = 1$ if protein i, and j could interact to form heterodimer ij, and 0 otherwise). We are interested at the concentration of free protein i (F_i) and dimers ij (D_{ij}). At any time, they are related to the concentration, C_i, by the mass conservation equation

$$C_i = F_i + \sum_j A_{ij} D_{ij}. \tag{4.1}$$

Followed from the law of mass action, the concentrations satisfy the chem-

ical kinetic equation

$$\frac{dD_{ij}}{dt} = r_{ij}^{(\text{on})} F_i F_j - r_{ij}^{(\text{off})} D_{ij}, \tag{4.2}$$

where $r^{(\text{off})}$ and $r^{(\text{on})}$ are the dissociation and association rate constants respectively. Eq. 4.2 describes the dynamics through which the system attains equilibrium. The steady state solution for each heterodimer ij is

$$\bar{F}_i \bar{F}_j = K_{ij} \bar{D}_{ij}, \tag{4.3}$$

where $K_{ij} = r_{ij}^{(\text{off})} / r_{ij}^{(\text{on})}$ is referred as the dissociation constant.

In this chapter, we are interested in cases where the system is close to equilibrium. Assume the total concentrations are constant, a linearization of Eq. 4.2 and the mass conservation implies

$$-\frac{1}{r_{ij}^{(\text{off})}} \frac{d}{dt} \delta D_{ij} = \frac{1}{K_{ij}} (\bar{F}_i \sum_{k_1} A_{jk_1} \delta D_{jk_1} + \bar{F}_j \sum_{k_2} A_{ik_2} \delta D_{ik_2}) + \delta D_{ij}. \tag{4.4}$$

By labeling the indices using edges instead of vertices, Eq. 4.4 can be represented by a $E \times E$ matrix Γ where E is the number of edges in the network.

$$-\frac{1}{r_{\mu}^{(\text{off})}} \frac{d}{dt} \delta D_{\mu} = \sum_{\nu} \Gamma_{\mu\nu} \delta D_{\nu}, \tag{4.5}$$

where $\Gamma\mu\nu$ is defined as

$$\Gamma_{\mu\nu} = \begin{cases} 1 + \frac{D_\mu}{F_j} + \frac{D_\mu}{F_i} & \text{if } \mu = \nu, \\ \frac{D_\mu}{F_j} & \text{if } \mu \neq \nu \quad \text{but connected via j,} \\ \frac{D_\mu}{F_i} & \text{if } \mu \neq \nu \quad \text{but connected via i,} \\ 0 & \text{if } \mu \neq \nu \quad \text{and not connected.} \end{cases} \tag{4.6}$$

Here I assume the dimer μ is formed by proteins i and j. From now on, we use Greek indices to indicate dimers (edges), and Latin to label proteins (nodes).

Eq. 4.2 is the dynamical equation for dimer concentrations. As a result of mass conservation, it is equivalent to describe the system by free concentrations and rewrite Eq. 4.2 as

$$\frac{dF_i}{dt} = \sum_j r_{ij}^{(\text{off})} A_{ij} D_{ij} - \sum_j r_{ij}^{(\text{on})} A_{ij} F_i F_j, \tag{4.7}$$

and linearize around the equilibrium. For mathematical simplicity, we further impose that all K_{ij}s are identical. Indeed if that is the case, one can write

$$-\frac{1}{r^{(\text{off})}} \frac{d}{dt} \delta F_i = \sum_j \Lambda_{ij} \delta F_j, \tag{4.8}$$

where the $N \times N$ matrix Λ is given by

$$\Lambda_{ij} = \frac{\bar{F}_i}{K_d} A_{ij} + \delta_{ij} \frac{C_i}{\bar{F}_i}. \tag{4.9}$$

The i^{th} diagonal element of Λ is the sequestration level of protein i. The perturbation of a highly sequestered protein will relax rapidly. However, it will

not contribute much to the relaxation of its neighbors since its cross section $\bar{F}_i/K_d$ is rather low. As we only consider scenarios near equilibrium, we will, from now on, drop the bar on equilibrium concentrations.

The relaxation of the system from any fluctuation $\delta\mathbf{D}$ (or $\delta\mathbf{F}$) towards the steady state is determined by the spectral properties of the matrix Γ (or Λ). These matrices can be symmetrized by the following similarity transforms,

$$S_\Gamma = D^{-1/2}\Gamma D^{1/2} \tag{4.10}$$

$$S_\Lambda = Q^{-1/2}\Lambda Q^{1/2}, \tag{4.11}$$

where D and Q are diagonal matrices with $D_{\mu\mu}$ are the concentration of dimer μ and Q_{ii} are the concentration of free protein i. Due to the symmetrization, the eigenvalues of Γ and Λ are therefore real. In fact, the eigenvalues are all positive, and the system will therefore always return to equilibrium. The relaxation of any perturbation can be decomposed into eigenmodes. Each eigenmode has its own characteristic decay time $\tau^{(\alpha)}$ given by $(r^{(\mathrm{off})}\lambda^{(\alpha)})^{-1}$, where $\lambda^{(\alpha)}$ is the corresponding eigenvalue. The eigenmodes are useful in understanding the system. An important property is that they are actually quite localized to a few vertices. This could be verified by the calculating the so called participation ratio for each eigenmode. The participation ratio (PR) is defined as $(\sum x_i^4)^{-1}$ where x_i are the components of a normalized eigenvector. The value of PR $= 1$ if the eigenvector is concentrated on a single component, but takes the value n if the eigenvector spreads uniformly over all the n components. For the Λ matrix we defined in Eq. 4.9, there are 1439 nodes, but the mean PR is only about 2.7. We will explain the implication of

localized eigenmodes in the later part of this chapter.

4.3 Thermal Noise: Two Proteins Case

Association and dissociation between proteins are kinetic proceses. As a result, the concentrations of free proteins and dimers always fluctuate near the equilibrium values given by Eq. 4.3. We refer to these fluctuations as thermal noise. To estimate the thermal fluctuations in dimer/free concentrations for a given cell, we assume the total concentrations of proteins are constant. This is because the total concentration of a protein in a cell fluctuate rather slowly, compared to the characteristic time scale of the relaxation of the system, as a result of binding and unbinding.

Before going to the general network case, we start by looking at how two proteins (say A and B) bind to form dimers. The deterministic equation is a simplified version of Eq. 4.4. To include the effect of stochasticity, a fictitious random force is added. Following the argument by Bialek *et.al.* [66], the random force can be considered to be caused by the variation of the association and dissociation rate constants. By linearizing Eq. 4.2 with additional variations of $r^{(\mathrm{on})}$ and $r^{(\mathrm{off})}$, one arrives at

$$\frac{1}{r^{(\mathrm{off})}}\frac{d}{dt}\delta D_{AB} = -(\frac{F_A}{K_d} + \frac{F_B}{K_d} + 1)\delta D_{AB} + (\frac{\delta r^{(\mathrm{on})}}{r^{(\mathrm{on})}} - \frac{\delta r^{(\mathrm{off})}}{r^{(\mathrm{off})}})D_{AB}. \qquad (4.12)$$

This equation is similar to Eq. 4.4, with an extra term describing the stochasticity. Multiplying the whole equation by $r^{(\mathrm{off})}$, we have $\delta r^{(\mathrm{on})} F_A F_B$ and $\delta r^{(\mathrm{off})} D_{AB}$ respectively. As $r^{(\mathrm{on})} F_A F_B$ and $r^{(\mathrm{off})} D_{AB}$ are the increment and

decrement to D_{AB} via forward and backward reactions, the two extra terms can be regarded as the variance of these contributions.

The rate constants are related to the binding energy g of the dimer AB via detailed balance given by

$$\frac{r^{(\text{off})}}{r^{(\text{on})}} = \exp(\frac{1}{kT}g). \tag{4.13}$$

By substituting the log-derviative of Eq. 4.13 to Eq. 4.12, we have

$$-\frac{1}{r^{(\text{off})}}\frac{d}{dt}\delta D_{AB} = (\frac{F_A}{K_d} + \frac{F_B}{K_d} + 1)\delta D_{AB} + \frac{\delta g}{kT}D_{AB}. \tag{4.14}$$

With δg as the thermodynamic conjugate of δD (assume the system is in a unit volume), we define the dynamic susceptibility of the system by transforming Eq. 4.14 to the frequency space, arriving at

$$\alpha(\omega) = \frac{\delta D_{AB}}{\delta g} = \frac{D_{AB}}{kT}\frac{1}{F_A/K_d + F_B/K_d + 1 - i\omega/r^{(\text{off})}}. \tag{4.15}$$

Using the Fluctuation-Dissipation theorem which relates the dissipative part of the dynamic susceptibility to the power spectrum of the spontaneous fluctuations, the power spectrum of δD is given by

$$|\delta D_{AB}(\omega)|^2 = \frac{2kT}{\omega}Im(\alpha(\omega)), \tag{4.16}$$

where $Im(\alpha)$ denotes its imaginary part. The steady state fluctuation is given

by the area under the spectrum.

$$\begin{aligned}\langle \delta D_{AB}^2 \rangle &= \int_{-\infty}^{\infty} \frac{2kT}{\omega} Im(\alpha(\omega)) \frac{d\omega}{2\pi} \\ &= kT\alpha(0) \\ &= D_{AB} \frac{1}{F_A/K_d + F_B/K_d + 1}. \end{aligned} \tag{4.17}$$

An intuitive way to understand Eq. 4.17 is to view the average fluctuations as an balance between "kick out" and "drag back". D_{AB} is a source of randomness which kicks the system out of equilibrium, while at the same time, the term $(F_A/K_d + F_B/K_d + 1)^{-1}$ brings the system back.

To quantify the noise for the stochastic variable D_{AB}, it is instructive to use the so called Fano factor η, defined as the ratio between the variance and the mean. In this case, we have

$$\eta = \frac{1}{F_A/K_d + F_B/K_d + 1}. \tag{4.18}$$

For Poisson process where the birth/death events are independent, the Fano factor is exactly one. Note that in Poisson process, when the number is more than average, there will be more candidates for death and thus the system will be brought back to normal. Eq. 4.18 suggests that the Fano factor in thermal fluctuations is narrower than a Poisson process. This is because once a dimer breaks down, not only there are more candidates for binding, the probability of binding also increases by the law of mass action. This additional effect further enhances the tendency to return to normal and therefore give a narrower distribution. This also explains why large F_A and F_B give a lower

Fano factor and $\eta \to 1$ if both A and B are highly sequestered.

To explore further this example, let us assume $C_A < C_B$, label the proteins of type A and define for each of them a time series $x_i(t)$ as

$$x_i(t) = \begin{cases} 1 & \text{if i forms a dimer with protein B} \\ 0 & \text{otherwise} \end{cases} \tag{4.19}$$

Define $X = \sum_i x_i$, we have $\langle X \rangle = D_{AB}$ and $\sigma_X^2 = D_{AB}\eta$. Let us assume x_is are i.i.d. (which is not true and we will return immediately). σ_X^2 will then be $\sum_i \sigma_{X_i}^2$. From ergodicity the ensemble average $\sigma_{X_i}^2$ is equal to the time average $\langle x_i^2 \rangle_t - \langle x_i \rangle_t^2$. As x_i can only be 0 or 1, the expression is the same as $\langle x_i \rangle_t (1 - \langle x_i \rangle_t)$, which is equal to $\frac{D_{AB}}{C_A}(1 - \frac{D_{AB}}{C_A})$. The last step follows again from ergodicity. Therefore if x_is are i.i.d., σ_X^2 is equal to $D_{AB}(1 - \frac{D_{AB}}{C_A})$.

The variables x_is are not independent. The reason is there are only finite amount of B, the binding of a B molecule to a particular A protein will decrease the chance of another A protein being bound. As a result, for two A proteins i and j, $x_i(t)$ and $x_j(t)$ are correlated in the sense $\langle x_i x_j \rangle_t < \langle x_i \rangle_t \langle x_j \rangle_t$. The actual Fano factor σ_X^2 / D_{AB} is bounded by $(1 - \frac{D_{AB}}{C_A})$, which is a tighter bound compared to 1.

Mathematically, one can arrive at the same result using Eq. 4.18. As $C_A < C_B$ and thus $F_A < F_B$, we have

$$\begin{aligned} \eta &\leq \frac{1}{1 + F_B/K_d} \\ &= 1 - \frac{D_{AB}}{C_A} \\ &= 1 - \frac{D_{AB}}{D_m}. \end{aligned} \tag{4.20}$$

Here D_m is the maximum number of possible dimers formed by proteins A and B, which is $min(C_A, C_B)$. Such an upper bound is reached if $C_B \gg C_A$, which is exactly a scenario where the A molecules can be regarded as independent due to the abundance of B.

Relation with chemical Langevin equation

Eq. 4.12 is reminiscent of the chemical Langevin equation introduced by Gillespie [67]. Indeed, one could write down the chemical Langevin equation in this case as

$$\frac{d}{dt}D_{AB} = F_A F_B r^{(\text{on})} - D_{AB} r^{(\text{off})} + \sqrt{F_A F_B r^{(\text{on})}}\zeta_1 - \sqrt{r^{(\text{off})} D_{AB}}\zeta_2, \quad (4.21)$$

where ζs are two independent white noise defined as $\lim_{dt\to 0} N(0, 1/dt)$. Here $N(a, b)$ stands for a Gaussian variable with mean a and variance b. A linearization gives us

$$\frac{d}{dt}\delta D_{AB} = -(\frac{F_A}{K_d} + \frac{F_B}{K_d} + 1)\delta D_{AB} r^{(\text{off})} + \sqrt{F_A F_B r^{(\text{on})}}\zeta_1 - \sqrt{r^{(\text{off})} D_{AB}}\zeta_2. \quad (4.22)$$

Compare Eq. 4.22 with Eq. 4.12, we can identify the binding term $\frac{\delta r^{(\text{on})}}{r^{(\text{on})}} D_{AB} r^{(\text{off})} dt$ as $N(0, F_A F_B r^{(\text{on})} dt)$ and the other $\frac{\delta r^{(\text{off})}}{r^{(\text{off})}} D_{AB} r^{(\text{off})} dt$ as $N(0, D_{AB} r^{(\text{off})} dt)$.

Chemical Langevin equation is an approximation to the chemical master equation. In this example, it assumes that the increase in dimer number due to association and the decrease in dimer number due to dissociation follow two independent Gaussian distribution. The stochastic process (the number of dimer with respect to time) is completely the result of the molecular collisions.

This is the reason why temperature does not enter the average magnitude of fluctuations as shown in Eq. 4.17. Of course, since dissociation constants are functions of temperature, the Γ matrix and thus σ depend implicitly on temperature.

The essence of our approach is to translate the Gaussians as a result of a fictitious variation in binding energy and apply the Fluctuation-Dissipation theorem. It relies on only the macroscopic kinetics and bypasses the microscopic description. Alternatively, one can find the steady state fluctuation using the linear noise approximation [26, 27], which is introduced in biological context by Paulsson [68, 69]. The final result by Paulsson is an equation which relates the covariance matrix σ, the magnitudes of fluctuation sources (ζs) and the dissipation matrix Γ, which is also commonly referred as the Fluctuation-Dissipation theorem in literature.

4.4 Thermal Noise: General Network Case

One can generalize from an isolated dimer to interconnected dimers. In this case, Eq. 4.14 is replaced by

$$\frac{1}{r_{ij}^{(\text{off})}}\frac{d}{dt}\delta D_{ij} = -\frac{1}{K_{ij}}(F_i \sum_{k_1} A_{jk_1}\delta D_{jk_1} + F_j \sum_{k_2} A_{ik_2}\delta D_{ik_2}) - \delta D_{ij} + D_{ij}\frac{\delta g}{kT}, \tag{4.23}$$

or in terms of the Γ matrix,

$$\frac{d}{dt}R^{-1}\delta\mathbf{D} = -\Gamma\delta\mathbf{D} + \mathcal{D}\delta\mathbf{g}/kT, \tag{4.24}$$

where R is a diagonal matrix whose entries are the $r^{(\text{off})}$s of dimers and $\mathcal{D}$ is another diagonal matrix formed by the equilibrium dimers concentrations. The components of the column vector $\delta\mathbf{g}$ can again be thought as the variation of binding energy of the dimers.

Following the previous procedures, the dynamic susceptibility is defined by Fourier transforming Eq. 4.24:

$$\alpha(\omega) = (\Gamma - i\omega R^{-1})^{-1}\mathcal{D}\frac{1}{kT}. \tag{4.25}$$

Using the Fluctuation-Dissipation theorem, which is multi-dimensional in this case, the power spectrum of $\delta\mathbf{D}$ is therefore

$$(\delta D_\mu \delta D_\nu)_\omega = \frac{ikT}{\omega}(\alpha^*_{\nu\mu} - \alpha_{\mu\nu}). \tag{4.26}$$

where $\alpha^*_{\nu\mu}$ is the complex conjugate of $\alpha_{\nu\mu}$.

A direct application of Eq. 4.26 gives us the correlation function

$$\langle \delta D_\mu(t)\delta D_\nu(t')\rangle = \int (\delta D_\mu \delta D_\nu)_\omega \exp(-i\omega(t-t'))\frac{d\omega}{2\pi}. \tag{4.27}$$

More importantly, we have the steady state covariance matrix σ defined as

$$\sigma_{\mu\nu} = \langle \delta D_\mu \delta D_\nu\rangle = \int (\delta D_\mu \delta D_\nu)_\omega \frac{d\omega}{2\pi}. \tag{4.28}$$

The integral can be evaluated using the dispersion relations and the fact that

the imaginary part of $\alpha(\omega)$ is an odd function. The final result becomes

$$\sigma_{\mu\nu} = kT\alpha_{\mu\nu}(0) = (\Gamma^{-1}\mathcal{D})_{\mu\nu}. \tag{4.29}$$

As we are dealing with the steady state fluctuations, the decaying time scales $r^{(\text{off})}$s do not enter the final expression.

$\sigma_{\mu\nu}$ appears to be asymmetric as Γ is not a symmetric matrix. However, recall Eq. 4.10, $\sigma_{\mu\nu}$ is in fact symmetric with respect to interchange in μ and ν. Using Eq. 4.29, the steady state fluctuation $\langle \delta D_\mu^2 \rangle$ of a dimer μ is given by $\Gamma^{-1}_{\mu\mu} D_\mu$. Thus the relative fluctuations $\sqrt{\langle \delta D_\mu^2 \rangle}/D_\mu$ scales with $1/\sqrt{D_\mu}$. Or in the other words, the Fano factor η of dimer μ is $\Gamma^{-1}_{\mu\mu}$.

Properties of thermal fluctuations are implicitly stored in the matrix Γ. Of particular interest is the inequality $\Gamma^{-1}_{\mu\mu} < \Gamma^{-1}_{\mu\mu}$, resulting a lower bound for η (see Appendix C for a detailed mathematical proof)

$$\eta_\mu \geq (\frac{F_i}{K_d} + \frac{F_j}{K_d} + 1)^{-1}. \tag{4.30}$$

Here we label the dimer ij by the index μ. The bound is interesting because for proteins i and j in the PPI network with free concentrations F_i, F_j and dimer concentration D_{ij}, the right hand side of the inequality resembles the Fano factor of thermal fluctuations in a isolated system of two physically interacting proteins with the same set of concentrations (see also Eq. 4.18). In the other words, the network as a whole amplifies the thermal fluctuations for a pair of interacting proteins compared to a case in which the two proteins are isolated.

Can amplification occur without limit? The answer is obviously no by

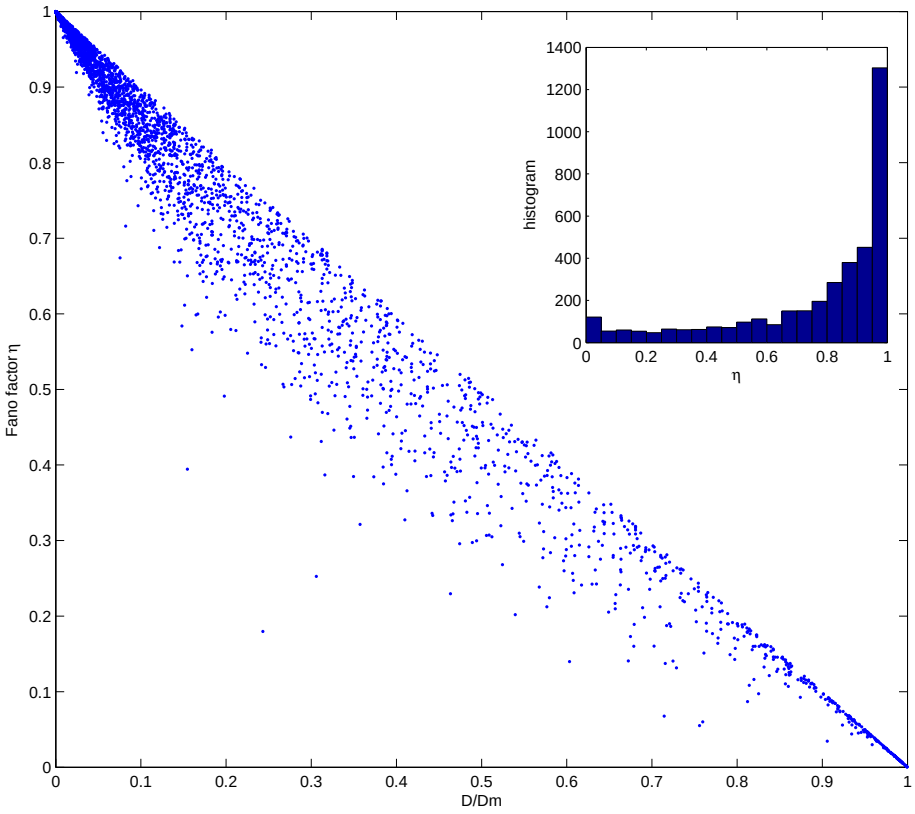

Figure 4.1: Thermal fluctuations in protein interaction network. Fano factor η of a dimer is defined as the ratio between the variance and the mean of its concentration. The calculation of η is described in the main text. The maximal η is 1, which is the case of a Poisson process. The inset is the histogram of η. About one third of the dimers lie at the bin closest to 1. For each dimer ij, D_m is defined as $min(C_i, C_j)$. The ratio D/D_m is a measure of occupation. Dimers with low occupations have lower Fano factors. It is clear from the plot that all points lie below the diagonal defined by the equation $y = 1 - x$.

numerically calculating the Fano factors. The inset of Fig. 4.1 shows histogram of the Fano factors for all dimers in our curated protein interaction network. As we have explained in the case of two proteins, the Fano factors of thermal fluctuations in reversible binding processes have values less than one. Out

of the 3880 dimers in our network, one third of them lie at the bin with $\eta > 0.95$. However, if one repeats the analysis using a network in which the total concentrations are reshuffled, over one half of the dimers will fall at the last bin. The same is true if one randomizes the network by swapping the edges while preserving the degree distribution [19]. Real-life PPI networks may, in general, have narrower thermal fluctuations compared to their random counterparts.

From the values of Fano factors, thermal fluctuations are narrower compared to Poissonian fluctuations. It is interesting to compare thermal fluctuations with the fluctuations in the production or degradation of proteins. In single gene expression model [70], the Fano factor is given by $1 + b$ where b is the average number of proteins translated per mRNA molecule. In other words, such fluctuations are always larger than thermal fluctuations. In the later part of this chapter, we will discuss the effects of fluctuations due to protein production and degradation.

Motivated by the Eq. 4.20, for each dimer, we plot in Fig. 4.1 η against the ratio between its equilibrium concentration and its maximal possible concentration $D_m = min(C_i, C_j)$. We call this ratio the occupation factor. It is important to note that even though D_m is the maximum possible concentration for a particular dimer, it is not possible for a set of dimers to reach their corresponding D_m simultaneously. Therefore, the actual maximum concentration should be effectively smaller. As we can observe in Fig. 4.1, it is clear that for every dimer μ, $\eta_\mu \leq 1 - D/D_m$.

A hint to explain this is to use the hypothetical i.i.d. scenario discussed in the previous section for a case of two isolated interacting proteins. With the

presence of other proteins, say C, D, E etc, one can still define $x_i(t)$ such that $x_i(t) = 1$ if protein i of type A is bound to a B molecule and 0 otherwise. If x_is are i.i.d., the variance of X (which is defined as $\sum_i x_i(t)$), σ_X^2, is equal to $D_{AB}(1 - D_{AB}/C_A)$, thus the corresponding η is $1 - D_{AB}/C_A$, i.e. the observed upper bound. Like the case for two proteins only, there is a discrepancy between the real η and the bound since the A proteins, say i and j, compete with each other for Bs, resulting at $\langle x_i x_j \rangle_t < \langle x_i \rangle_t \langle x_j \rangle_t$. On the top of that, in the presence of C, D, E etc, the real" η deviates further from the bound as the extra proteins compete with B in binding with protein j. To sum up, for any dimer μ which is formed by $\mu 1$ and $\mu 2$, we have

$$\frac{1}{1 + F_{\mu 1}/K_d + F_{\mu 2}/K_d} \leq \eta_\mu \leq 1 - \frac{D_\mu}{D_m}, \tag{4.31}$$

where $D_m = min(C_{\mu 1}, C_{\mu 2})$.

Eq. 4.31 suggests a coordinate transform

$$\eta_\mu = (1 - \zeta)\eta_{min} + \zeta \eta_{max}, \tag{4.32}$$

where η_{min} and η_{max} are the two bounds respectively. Fig. 4.2 is a histogram of the parameter ζ. The parameter quantifies the thermal fluctuations of a dimer relative to the two extremal values. When $\zeta = 0$, $\eta = \eta_{min}$. When $\zeta = 1$, $\eta = \eta_{max}$. The pileup against the upper limit suggests that, despite the i.i.d. assumption, the Fano factor of a dimer is very close to the upper limit. Besides, there exists dimers whose Fano factors are close to the lower limit, i.e. they are rather isolated from the network. Indeed most of them are

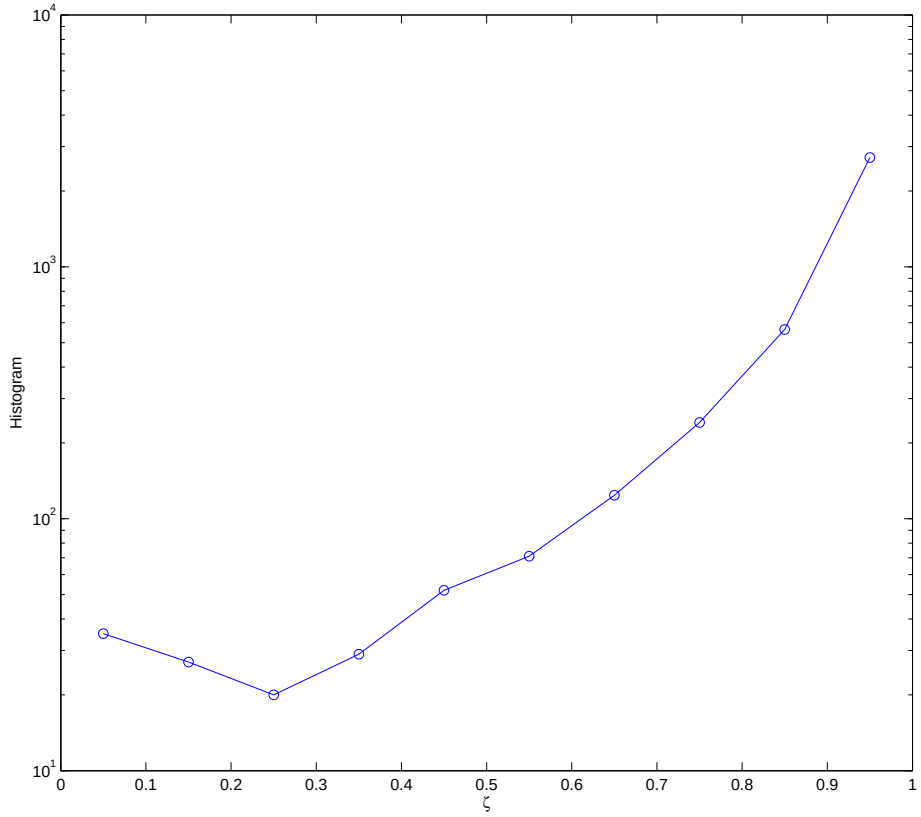

Figure 4.2: Histogram of ζ, a parameter to quantify thermal fluctuations between two extremal values. When $\zeta = 0$, $\eta = \eta_{min}$. When $\zeta = 1$, $\eta = \eta_{max}$.

hanging at periphery of the network. It is important to note that the isolated is not only in the sense of topology. It is incorporated by the concentrations. For those dimers with low ζs, their thermal fluctuations are likely due to their own association and dissociation, rather than the influence of other dimers in the network.

Power spectrum of thermal fluctuations

As a result of the damping effects induced by the matrix Γ, it is interesting to look at the power spectrum of our stochastic variable, δD_μ. Being consistent with the chemical Langevin equation, the stochastic effects of binding and unbinding are assumed to be white noise, i.e. the power spectrum is flat. Using our formalism, the noise power spectra of dimers can be easily found numerically using Eq. 4.25 and 4.26. A more general approach could be found in [71].

Fig. 4.3 shows the noise spectra for a few dimers. Most dimers have power spectra like the red and blue curves. The main observation is the existence of a cut-off frequency. Below the cut-off, the spectral density is apparently flat. As the frequency increases above the cut-off, the noise power decreases with $1/\omega^2$. This is usually called the low-pass-filtering property. One can fit the power spectra using the function

$$S(\omega) = \frac{\theta}{(\omega/r^{(\text{off})})^2 + \omega_o^2}. \tag{4.33}$$

Here, ω_o is interpreted as an effective time scale of the relaxation process. When ω_o is large, fluctuations are rapidly damped, the noise level is therefore low. At the same time, a system with a large ω_o has a fast response fast and thus the system has enough time to catch up provided that the noise frequency is not extremely high. The flat region of the spectrum is therefore wide. Note that the area under the power spectrum is the average fluctuation. From these observations, it can be concluded that a large ω_o leads to smaller fluctuations.

Green data points in Fig. 4.3 refer to a dimer with two relaxation time

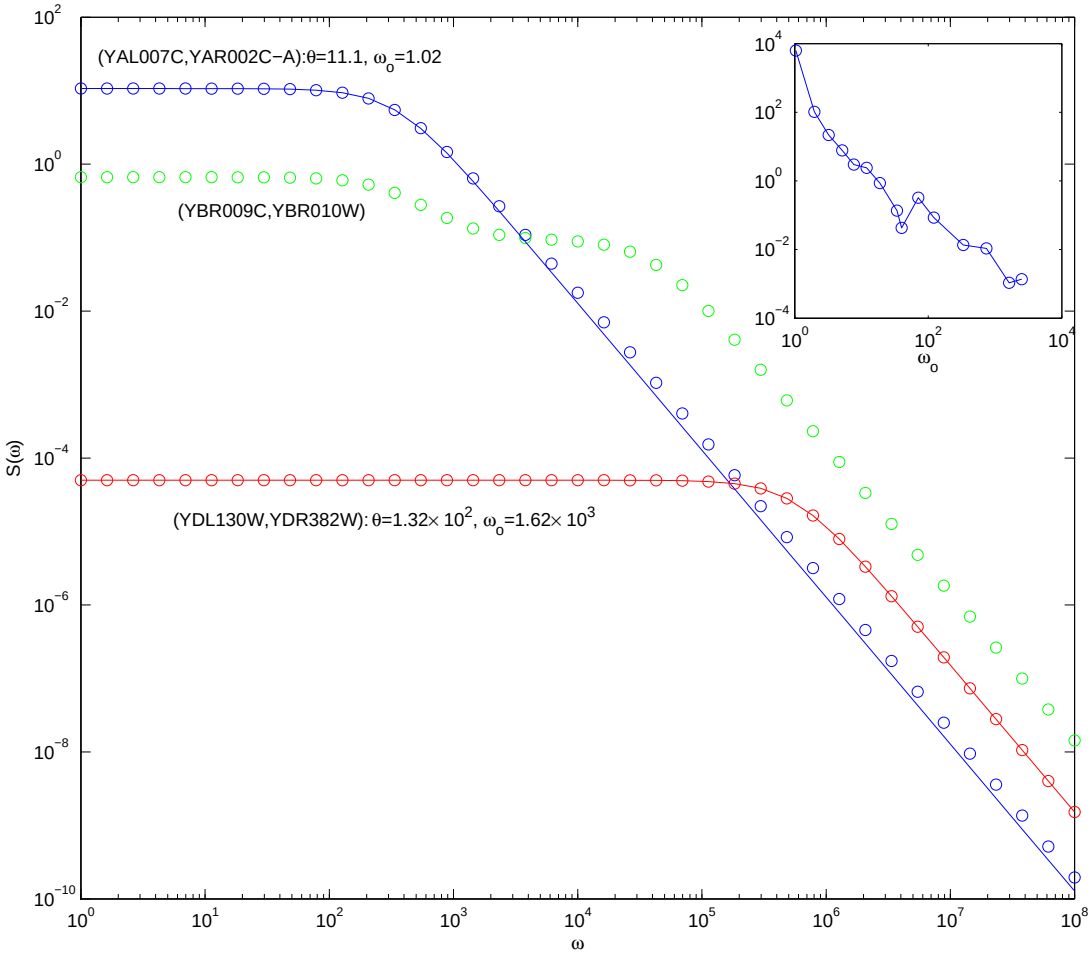

Figure 4.3: Examples of power spectra for three dimers. Data points are obtained numerically using Eq. 4.25 and 4.26. The red and blue data points are typical cases. The lines are results of the best fit to the function $S(\omega) = \dfrac{\theta}{(\omega/r^{(\text{off})})^2 + \omega_o^2}$. The values of the fitting are shown in the legend. The red curve has a shorter effective relaxation time compared to red. The green data refer to a dimer with two relaxation time scales as described in the main text. The inset is the distribution of effective relaxation rates ω_o for all dimers.

scales. Indeed, one can project the fluctuation of a dimer on the eigenvectors of Γ. As the eigenvectors are in general localized, the fluctuation is likely to lie along a particular eigenvector, and its relaxation time-scale will be likely reflected by the eigenvalue of that particular eigenvector. However, even though it is rare, there are dimers (e.g. dimer as shown in Green data points) whose

fluctuations involve significantly more than one eigenvector in a significant way.

In general, one can obtain the effective relaxation rates of all dimers by fitting all the power spectra using Eq. 4.33. The effective relaxation rates follow a broad distribution which resembles the distribution of eigenvalues of the matrix Γ (see inset of Fig. 4.3).

4.5 Effects of Noise in Total Concentrations

Apart from thermal noise, total concentrations of proteins fluctuate due to stochasticity in production and degradation. Or in an ensemble point of view, the total concentration of a protein varies from cell to cell, which is recently measured experimentally by [24]. Since the function of a free protein molecule can be very different from a dimer, it is significant to decompose the noise into more biologically meaningful components: concentrations of individual free proteins and concentrations of individual heterodimers.

Consider a case where there is an abrupt but small change in total concentrations, i.e. δC is a step function. To incorporate heterogeneous dissociation constants, one can use Eq. 4.5 and study the effects on dimer concentrations. But for mathematical simplicity, let us assume the dissociation constants are identical and look from the view of free proteins. By taking into account the change in total concentration δC in Eq. 4.1, Eq. 4.8 can be rewritten as

$$-\frac{1}{r^{(\mathrm{off})}}\frac{d}{dt}\delta F_i = \sum_j \Lambda_{ij}\delta F_j - \delta C_i. \tag{4.34}$$

Instead of looking at the relaxation of δF_i, it is more relevant to study the relative perturbation, $\delta F_i/F_i$. Note that F_i is the equilibrium concentration, and therefore a constant. Using the definition of Λ and dividing Eq. 4.34 by F_i, we have

$$-\frac{1}{r^{(\text{off})}}\frac{d}{dt}\frac{\delta F_i}{F_i} = \sum_j (\frac{F_j}{K_d}A_{ij} + \delta_{ij}\frac{C_i}{F_i})\frac{\delta F_j}{F_j} - \frac{\delta C_i}{C_i}\frac{C_i}{F_i}. \tag{4.35}$$

It is interesting to point out that the relaxation of $\delta F_i/F_i$ is governed by the transpose to the Λ, which determines the relaxation of δF.

As fluctuations in C are rather slow compared to the time required to attain equilibrium, we are interested in the long time limit of the solution of Eq. 4.35 which is given by

$$\frac{\delta F}{F} = (\Lambda^T)^{-1}\mathcal{Z}\frac{\delta C}{C}, \tag{4.36}$$

where $\mathcal{Z}$ is a diagonal matrix with elements $\mathcal{Z}_{ii} = C_i/F_i$. The role of $(\Lambda^T)^{-1}\mathcal{Z}$ is a transfer function, which maps the relative change in total concentrations to relative changes in free concentrations.

Eq. 4.36 implies that there could be very different consequences if two proteins are increased by the same percentage in their total concentrations. Because of the $\mathcal{Z}$ matrix, if the total concentration of a highly sequestered protein is increased, the effect could then be magnified. In addition, by spectrally decomposing Λ into its left and right eigenvector $|l_k\rangle$ and $|r_k\rangle$, we have

$$\frac{\delta F}{F} = \sum_k \frac{1}{\lambda_k}|l_k\rangle\left\langle r_k|\mathcal{Z}\frac{\delta C}{C}\right\rangle, \tag{4.37}$$

where λ_ks are the eigenvalues. Therefore, if a fluctuation $\delta C_i/C_i$ is dominated by an right eigenvector of Λ with a small eigenvalue, the resultant response $\delta F/F$ is more significant. Moreover, the response could spread further apart spatially on the network if the corresponding left eigenvector has a large participation ratio. Eq. 4.37 allows one to extract the proteins whose concentrations may change dramatically by a fluctuation in a particular protein. This is a more mathematical approach of finding concentration-coupled proteins introduced in Ref. [64].

As it has been already emphasized in Ref. [63], real-ife PPI networks would be more prone to propagating perturbations than their randomized counterparts. While this allows effective propagation of signals, it suffers from the effects of noise. It is interesting to compare the thermal noise case, where real-life networks have narrower thermal fluctuations compared to randomized versions.

Several studies [22, 72, 73] have distinguished the so-called intrinsic and extrinsic noise. The intrinsic noise is due to the independent stochastic fluctuations in production and degradation, therefore lacks correlation between proteins. On the other hand, the extrinsic noise corresponds to correlated shifts in abundance of several proteins, which could be attributed to, e.g. the ribosome-mediatd noise or the variation in cell size. Despite experimental measures of protein abundance [24], it is hard to separate the effect of these two factors.

To study the effects of extrinsic and intrinsic noise on binding network, we consider two cases where the protein concentrations are either all coherent (extrinsic) or independent (intrinsic). Suppose the concentrations of all pro-

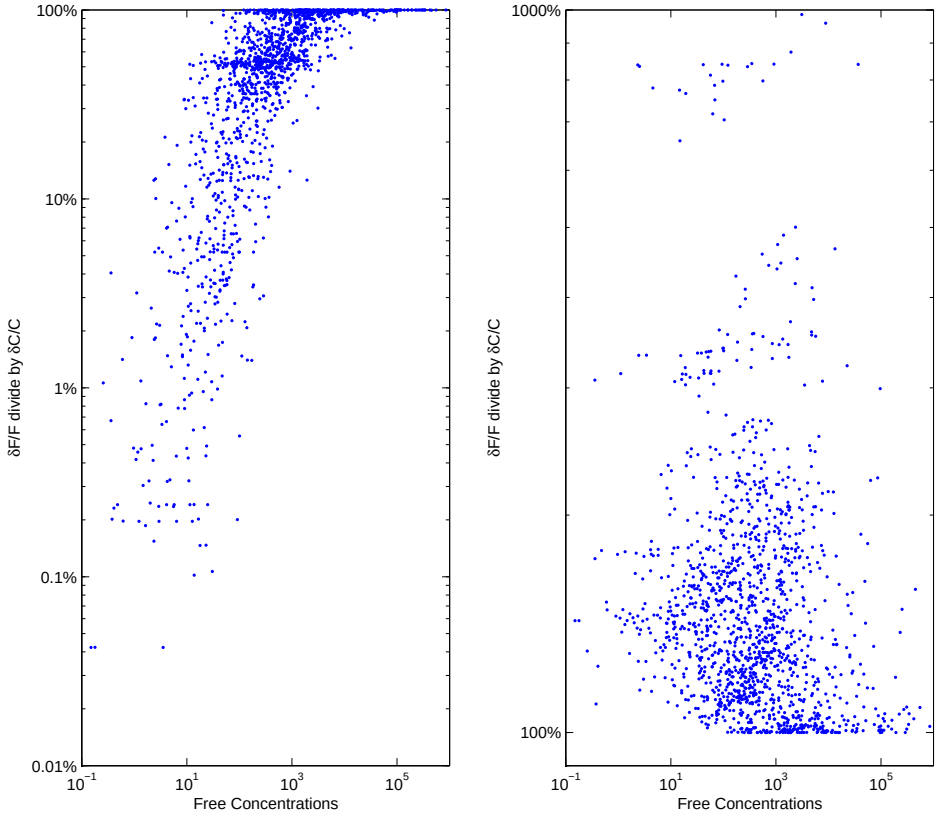

Figure 4.4: Relative response in free concentrations as a result of fluctuations in total concentrations. The response is calculated by Eq. 4.36 where the fluctuation in total concentration is assumed to be 20% for each protein. The relative changes in the steady state free concentrations $\delta F_i/F_i$, normalized by 0.2 in this case, are plotted against the original free concentrations F_i. Response to intrinsic fluctuations (right panel) are in general orders of magnitude higher than the response to extrinsic fluctuations (left panel).

teins change by 20% (around the average magnitude from experiment [24]), the relative perturbation in free concentrations can be found using Eq. 4.36. As shown in Fig. 4.4, in case of correlated fluctuations (left panel), the resultant changes in free concentrations are tiny. It means that the fluctuations

in total concentrations tend to cancel each other. In contrast, intrinsic fluctuations (right panel) contributed from different proteins can sometimes add up and cause considerable changes in the steady state free concentrations, which could be orders of magnitude higher than the corresponding changes for extrinsic fluctuations.

It is also interesting to look at the case for the random counterpart. We found that if the concentrations are reshuffled, the intrinsic noise can be not added up to the level as in case of the real-life network. This is consistent with the conclusion that real-life network is more likely to propagate noise in total concentrations.

So far, we have focused on the steady state response to total concentrations. More generally, one could study the average dynamical fluctuations by Fourier transforming Eq. 4.34 to the frequency space

$$\delta F(\omega) = (\Lambda - i\frac{\omega}{r^{(\text{off})}}I)^{-1}\delta C(\omega). \tag{4.38}$$

The power spectrum is thus

$$\begin{aligned} S(\omega) &= \langle \delta F(\omega)\delta F(\omega)^{\dagger}\rangle & (4.39)\\ &= (\Lambda - i\frac{\omega}{r^{(\text{off})}}I)^{-1}\langle \delta C(\omega)\delta C(\omega)^{\dagger}\rangle(\Lambda^{T} + i\frac{\omega}{r^{(\text{off})}}I)^{-1}. & (4.40) \end{aligned}$$

Given the fluctuations in total concentration $\langle \delta C(\omega)\delta C(\omega)^{\dagger}\rangle$, the noise in free concentrations could be evaluated by the area under the spectrum. While the fluctuations in single gene expression are well studied, analytical models on expression of many interrelated genes are intractable. We will not go into

details in this study.

4.6 Conclusion and Outlook

We have presented the mathematical formalism for studying dynamical processes in protein binding networks governed by the law of mass action. In particular, we use the formalism to study the thermal noise and the noise due to fluctuations in total concentrations. It turns out that real-life PPI networks respond quite differently to these two kinds of noise. While real-life PPI networks suffer the problem of noise due to fluctuations in total concentrations, it usually has smaller thermal fluctuations. It is in general true that thermal noise is rather mild, however, as the number of proteins in a cell could be rather small and biological systems can sometimes be so sensitive, its effect is still not negligible.

In this chapter, our analysis is mostly statistical. In fact, apart from system-wide analysis, it is fruitful to look at how dynamical fluctuations (change in concentrations) are transmitted in a pathway through successive binding and unbinding. More interestingly, one can aim at how noise propagates or attenuates along the pathway. This allows one to generalize the studies in [63, 64] to include fluctuations in all frequencies.

As a simple example, one could consider binding as a kind of post-translational regulation. Suppose B is a binding partner of A. Without B, the noise of the free concentration of A is just the noise of its total concentration, which is the result of stochasticity in transcription, translation and degradation. On the other hand, if A is regulated by its binding partner B, this kind of noise is re-

duced. A larger reduction is even possible if A is highly sequestered. However, the use of B may induce extra noise sources, including the fluctuations in the total concentration of B and thermal effect. In general, the result changes in noise amplitude are frequency dependent.

The mathematics we presented give the power spectrum in concentrations for different components in a pathway. As noise in different frequencies may have different effects in the system, for example, high frequency noise is usually less important as the system is not fast enough to respond, the area under the spectrum may not be the best measure of noise. It is instructive to start from simple topologies (linear cascades, stars) and quantify the noise using the spectra. Methods along in this line have been used in simple genetic networks [74, 75].

Chapter 5

Large-scale prediction and verification of indirect regulatory interactions in model organisms

5.1 Background and Introduction

The development of high-throughput experimental techniques has lead to accumulation of tremendous amount of data describing regulatory interactions in model organisms. In order to understand biological processes at a system-wide level, effective computational algorithms are necessary [76, 77].

The regulatory interactions from high-throughput experiments can be either direct or indirect in nature. We call the regulation from a regulator to a target protein *direct* if it is mediated by a direct molecular mechanism such as

transcriptional regulation of a target protein's level by a transcription factor or phosphorylation by a kinase. Conversely, regulations involving any number of intermediate proteins are referred to as *indirect*. In fact, indirect regulations are vastly more common than the direct ones and thus easier to detect experimentally.

There are many approaches to study regulation interactions. Methods like Boolean function [31], Bayesian analysis [32] are proved to be very useful. However, these methods are conceptually more complicated, and they are applicable only if the available data contains specific details like carefully chosen perturbation conditions and temporal information. In practice, one may have a large and noisy dataset of regulations (including both direct and indirect) specifying only the signs. In this case, simpler method based on network topology can be more useful. The idea is to represent the data in terms of a directed network in which edges carry signs. A directed edge tells which gene regulates which other gene, and the sign represents whether the regulation is activation (positive) or inhibition (negative). In spite of its simplicity, topological analysis has been proved to be a powerful tool in extracting a wealth of information from noisy regulatory data [33]. In this work, we develop a network-based algorithm which allows one to verify existing indirect regulations and to predict missing ones. The algorithm is applicable to large and heavily connected networks combining direct and indirect regulatory interactions.

Consider a protein i regulating (either directly or indirectly) a protein k which in its turn is known to regulate (again directly or indirectly) a protein j, then it is rather likely to also have an indirect regulatory interaction between i and j. This simple observation could be further extended in two ways.

Firstly, indirect regulations could propagate along longer protein cascades, thus a series of regulations $i \rightarrow k_1 \rightarrow k_2 \rightarrow j$ increases the likelihood of an indirect regulation $i \rightarrow j$. Secondly, having multiple parallel pathways reinforce the predictability. Therefore, if protein i regulates proteins k_1, k_2 and each of them regulates protein j, it is even more likely to find an indirect regulation from i to j.

Naively, to predict or verify an indirect regulation between protein i and protein j, one could simply count the number of paths connecting i and j. However, this counting scheme has to take into account two important observations. First of all, paths should be weighted differently according to their lengths. Inferences based on cascades is less reliable, and thus such should contribute less to the likelihood. Secondly, the inferred sign of the indirect regulation from different paths should agree with each other. In general, if a protein i and a protein j are connected by a multi-step path, the sign of the resultant indirect regulation between i and j is given by the product of signs of all intermediate edges. It is natural to assume that the effect of a positive path (whose edges give a positive product) and the effect of a negative path (whose edges give a negative product) contradict and to some extent cancel each other.

In the next section, we will show that this central idea of predicting likely indirect regulations could be easily incorporated using a matrix formalism. Obviously, the likelihood can serve as a quantitative measure of the reliability of any regulation in a dataset. Thus one could also verify already known regulations based on this calculated likelihood. A regulation with a high likelihood is deemed reliable. On the other hand, indirect regulations missing from the

dataset could be reliably predicted. As always, there is a tradeoff between the number of predictions and their quality. Due to the usage of multi-steps indirect paths from one protein to another, our method is tailored for heavily interconnected networks in which such multiple regulations are common.

We applied our algorithm to the set of genetic regulations extracted from contents of the entire PubMed database (14,000,000 abstracts) and 47 full text journals. The automatic extraction of interactions was made possible by the Medscan algorithm based on Natural Language Processing (NLP) techniques [78]. Both direct and indirect regulatory interactions were collected for four model organisms: *Homo sapiens*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Drosophila melanogaster* (see Table 5.1 for details). As reflected in their inter-connectedness index $IC = \langle k^{in}k^{out}\rangle/\langle k^{in}\rangle$, all these networks are globally connected ($IC > 1$). In particular, since the network of human proteins is the largest and the most heavily connected ($IC \simeq 60$) among the networks used in this study, we will illustrate our algorithm using mostly this network.

Large-scale network analysis of indirect regulatory interactions in yeast was recently studied in [33, 79, 80]. These works focused on the classification of regulations as either direct or indirect and subsequently pruning of indirect regulations. Pruning of indirect regulations is a useful procedure from the point of network simplification. However, being developed for relatively sparse networks, these algorithms assume all links are equally reliable and neither of these algorithms performs well for heavily interconnected networks considered in this study. Owing to the accrual of data from high-throughput experiments, heavily entangled networks are evidently unavoidable. Indeed, to effectively

Table 5.1: Regulatory networks of the four model organisms. The IC (interconnectedness) index, defined as $\langle k^{in}k^{out}\rangle/\langle k^{in}\rangle$, quantifies whether a single perturbation typically spreads over a network ($IC > 1$) or dies out ($IC < 1$). Here k^{in} and k^{out} stands for the in and out-degree of the nodes respectively. Golden sets consist of frequently reported regulations which are used as highly-trustable references in this study.

Organisms	Number of Proteins	IC	Number of links		Size of Golden set	
			positive	negative	positive	negative
Homo sapiens	7853	61.9	36426	16436	3442	1671
Saccharomyces cerevisiae	1218	3.42	1208	813	125	85
Arabidopsis thaliana	490	2.84	426	252	42	25
Drosophila melanogaster	569	1.39	410	203	46	25

study large and heavily connected networks, one is forced to weigh links by their reliability. In principle, our algorithm for links verification could be efficiently used for pruning a network. However, we choose to focus on prediction and verification of novel indirect regulations.

5.2 Algorithm for Prediction and Verification

Matrix formalism

In this work, we represent the dataset of all known direct and indirect regulatory interactions in a given organism as a directed network. In matrix notation, it is fully defined by an adjacency matrix A taking the values

$$A_{ij} = \begin{cases} +1 & \text{if } i \text{ positively regulates } j, \\ -1 & \text{if } i \text{ negatively regulates } j, \\ 0 & \text{if } i \text{ is not known to regulate } j. \end{cases} \tag{5.1}$$

To predict new indirect regulations and to quantify the reliability of the existing ones, we use another matrix X given by

$$\begin{aligned} X &= A^2 + \lambda A^3 + \lambda^2 A^4 + \lambda^3 A^5 \cdots, \\ &= \frac{A^2}{I - \lambda A} \end{aligned} \tag{5.2}$$

where λ is a parameter to be discussed later. X_{ij} includes the contribution of all paths from i to j. $(A^n)_{ij}$ is the net number of paths (number of positive paths minus the number of negative paths) of length n from node i to node j, the sign of X_{ij} is based on whether positive paths or negative paths dominate. If positive (negative) paths dominate, X_{ij} is positive (negative), and it is likely that i is indirectly activating (repressing) j.

The parameter λ in Eq. 5.2 is basically a free parameter which could be optimized later to provide the best performance for the algorithm. Generally speaking, λ determines the weights of different paths. If λ is chosen to be less than one, the contribution from long paths is exponentially suppressed. In this work, we have chosen different λ's for different networks in order to optimize the performance of our algorithm. We will first present our results using the *optimal* value of λ. The definition of the optimal λ and its determination will be addressed later on.

Calibration of reliability

We have argued that the absolute magnitude of matrix elements of X is a measure of reliability of indirect regulations. Following the matrix formalism, we calculate X for four different regulatory networks: *H. sapiens*, *S. cerevisiae*,

A. thaliana and *D. melanogaster*.

In our algorithm, every non-zero element of X possesses certain predictive power. We collect all possible predictions by picking out all non-zero X_{ij}'s. The validity of our algorithm is evident if pairs i and j with large value of $|X_{ij}|$ are likely to correspond to more reliable regulations. To show this is indeed the case, one needs to use "golden set" containing completely trustable regulations, which however is not readily available. For this purpose, we define the golden set to be regulations which are frequently reported in the literature (for details of the cutoff, see Appendix D). These regulations form the most reliable part within the original network. In fact, the values of the median value of $|X|$ for all the non-zero matrix elements and those within the golden set are 3.9×10^{-3} and 3.5 respectively.

Fig. 5.1 shows a more detailed calibration of the matrix elements. We define a predictive set of size n using the n predictions with the largest values of $|X_{ij}|$. If all the possible predictions are used, the size of the set is huge (up to 10^7). The number of predictions covered in the golden set is counted and normalized by the corresponding number obtained by a set of n random predictions. As shown in Fig. 5.1, the overlap between the golden set and the best 100 of our predictions is $10,000$ (sic!) times better than what is expected by pure chance alone. The advantage decreases when predictions with smaller values of $|X_{ij}|$ are included. In case all possible predictions are used, the predictive set is only sightly better (2-fold) than a random set. This is expected since predictions with smaller values of $|X_{ij}|$ are much less likely to be reliable.

Large $|X_{ij}|$ is a result of confirmation by multi-step paths from i to j,

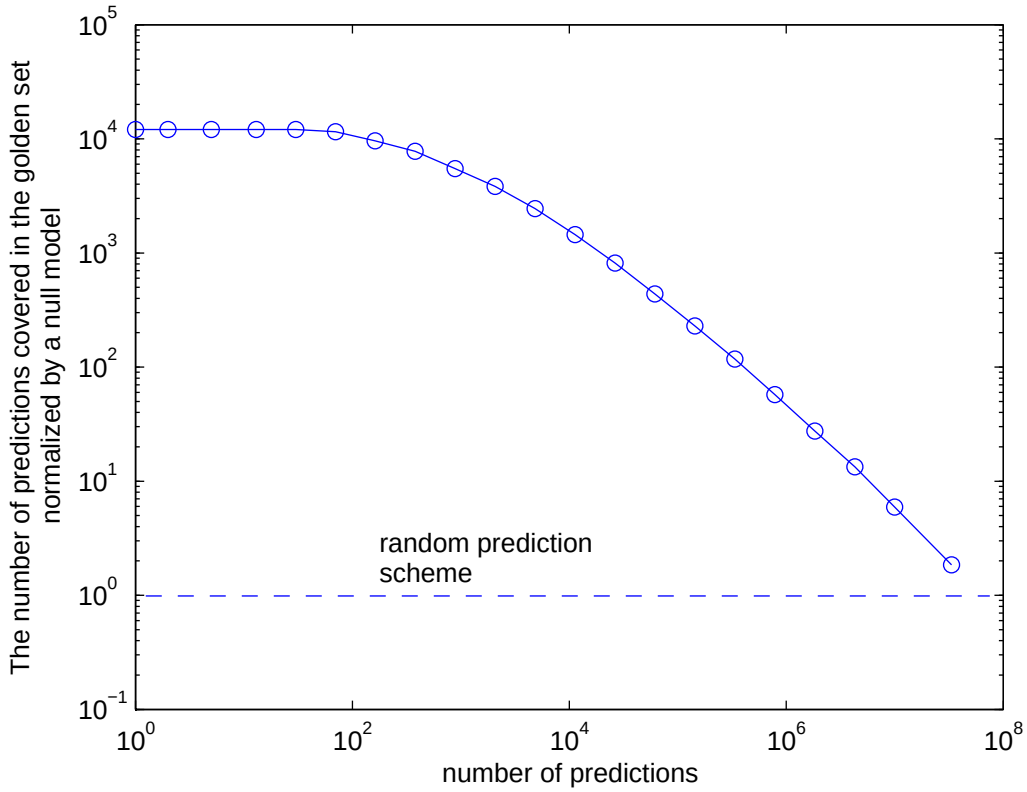

Figure 5.1: Advantage of our prediction scheme over random predictions. The x-axis shows the number of predictions ranked by the values of $|X_{ij}|$. The vertical axis shows the the number of predictions in the golden set normalized by the null-model expectation. The ratio decreases as the size of the predictive set increases.

therefore such predictions are likely to be indirect in nature. To prove that it is indeed the case, we separate the golden set into direct and indirect subsets based on the information whether a regulation is transcriptional regulation or not. Such information information is again obtained from literature using the Medscan algorithm (see Appendix D). In agreement with our expectation, the predictions are biased toward the indirect subset as shown in Fig. 5.2 (see caption for details).

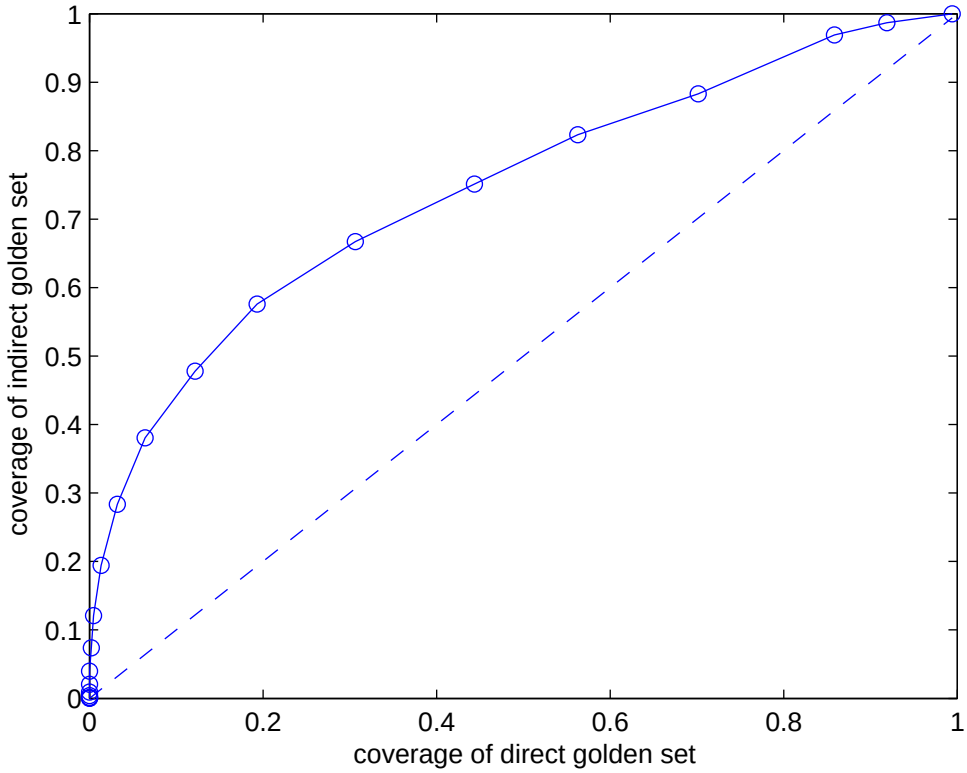

Figure 5.2: The coverage of the direct (indirect) subset for a given set of predictions is defined as the number of verified predictions normalized by the size of the direct (indirect) golden set. Data points closer to the origin refer to predictions with larger average value of $|X_{ij}|$, and therefore presumably more reliable. As reflected by the convexity of the curve, those regulations are more likely to be indirect rather than direct.

An important use of the matrix elements is to determine whether the regulations are positive or negative. Under our formalism, regulations corresponding to large *positive* matrix elements are likely to represent positive regulations. In order to calibrate the reliability for a set of predictions, we define the average quality by counting the fraction of prediction whose inferred sign agrees with that reported in the golden set. Fig. 5.3 shows the tradeoff between the

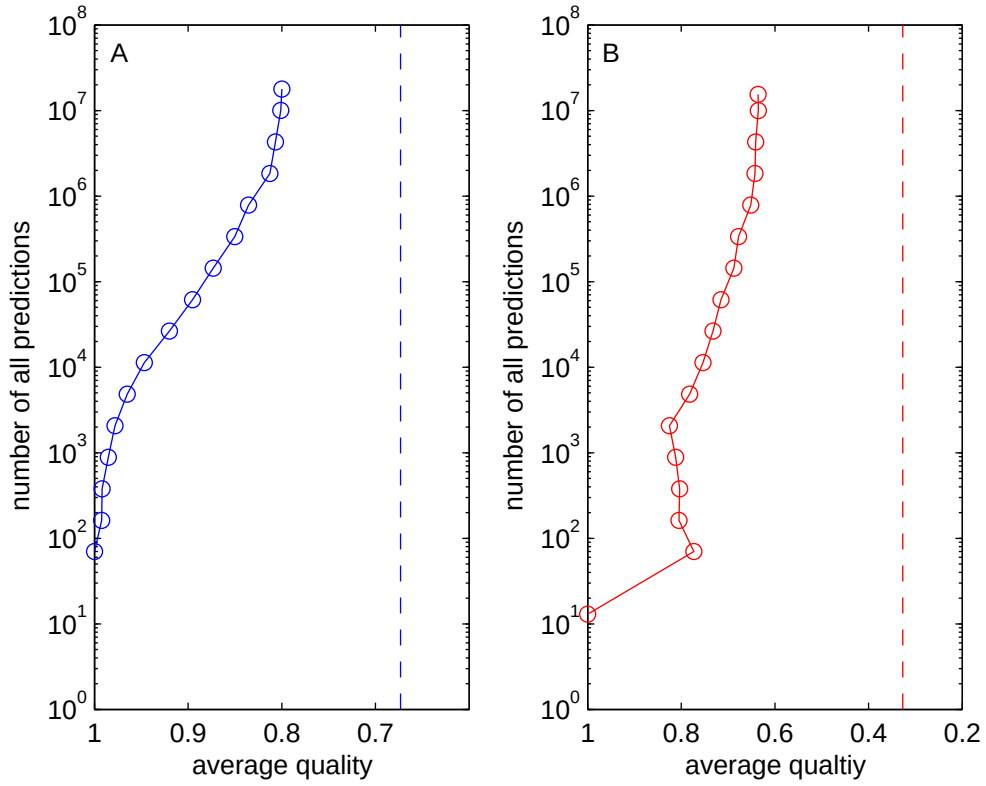

Figure 5.3: The tradeoff between the number of predictions and their average quality (panel A for positive predictions and B for negative predictions). For a set of predictions, the average quality is defined as the fraction of predictions whose sign agrees with that in the golden set. The dotted line is the quality expected for a null model as described in the main text.

number of predictions and the average quality. As shown in Fig. 5.3A, a set of predictions with average quality 100% offers about 100 predictions of positive regulation. However, if one is willing to downgrade the quality to 95%, the number of predictions is up to 5000. By including all the positive entries in X, we are offered a huge number of predictions, but with a relatively low quality. However, even in that case, the average quality is still much better than

a null model, which is defined as the fraction of positive regulations among all the regulations in the golden set. Thus the quality of our null model for positive (negative) regulations in human is $3442/(3442+1671) = 0.67$ ($1671/(3442+1671)$=0.33). They are shown as dashed lines in Fig. 5.3. Using negative matrix elements, one could also predict negative regulations. Large negative elements of X are indeed more likely to have negative signs in our golden set (see Fig. 5.3B).

To understand better the quality of our sign predictions, we study the Receiver Operating Characteristic (ROC) curves. Fig. 5.4A is the ROC curve for positive-sign predictions. It shows the sensitivity against specificity in different predictive sets as described by varying the $|X_{ij}|$ threshold. For positive-sign prediction, sensitivity is defined as the fraction of regulations in the positive golden set which are predicted to be positive by our algorithm. Specificity, on the other hand, is defined as the fraction in the golden *negative* set that are predicted to be positive by our algorithm. Data points close to the origin consist of predictions with large X_{ij}. The most important observation is the convexity of the curve, which means that the sign of interaction predicted by our method is more likely to be correct than expected by pure chance. In fact for a totally random predicted set, the ROC curve would be a straight line $y = x$. The area under a ROC curve is commonly used to quantify the performance of an algorithm. Using the negative X_{ij} to predict negative regulations, one could similarly define sensitivity and specificity resulting another ROC curve as shown in Fig. 5.4B.

Making use of the ROC curves, we could address the primary assumption behind our definition of the golden set: the larger is the number of papers

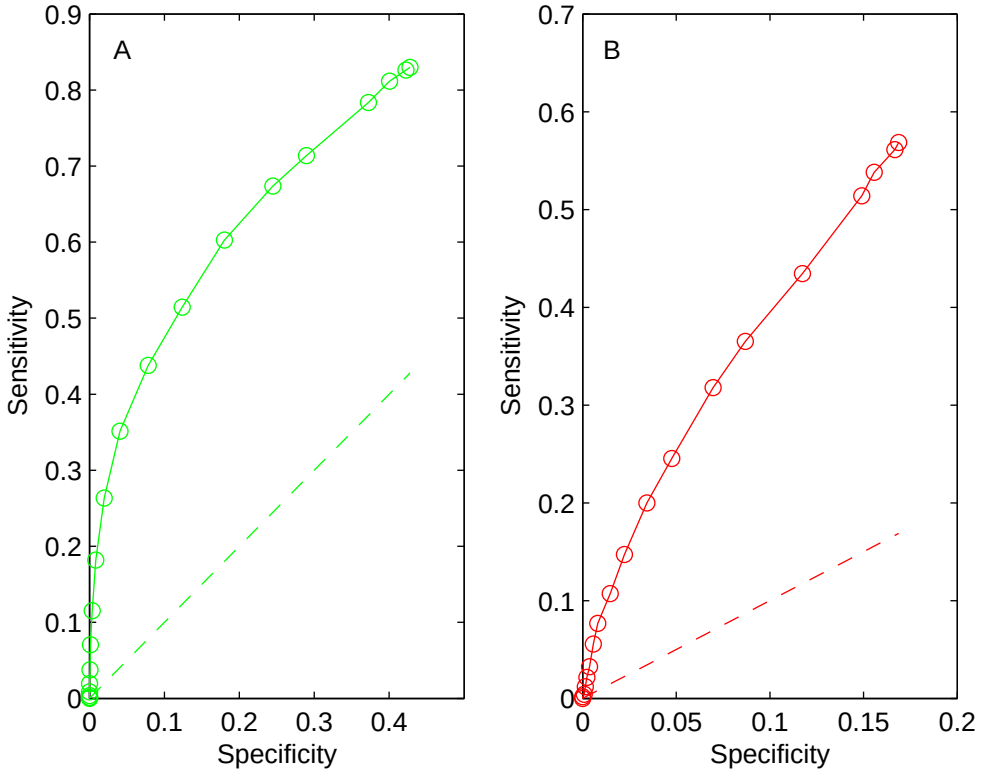

Figure 5.4: ROC curves for sign predictions using positive X_{ij} (panel A) and negative X_{ij} (panel B). Each data point corresponds to a predictive set defined by a particular threshold of X_{ij}. The dotted lines are $y = x$ which is the null-model expectation. The area under the ROC curve to the left of the solid line measures the performance of our algorithm.

reporting a given interaction, the more reliable it is. We define different golden sets by varying the publication cutoff. Golden sets arising from a high cutoff consists of regulations with the largest number of papers reporting it. The size of the set is thus smaller but it is supposed to be more reliable. By comparing the area of the ROC curves obtained from different golden sets, we find that indeed the ROC curve from a high-cutoff golden set encloses a larger area (see

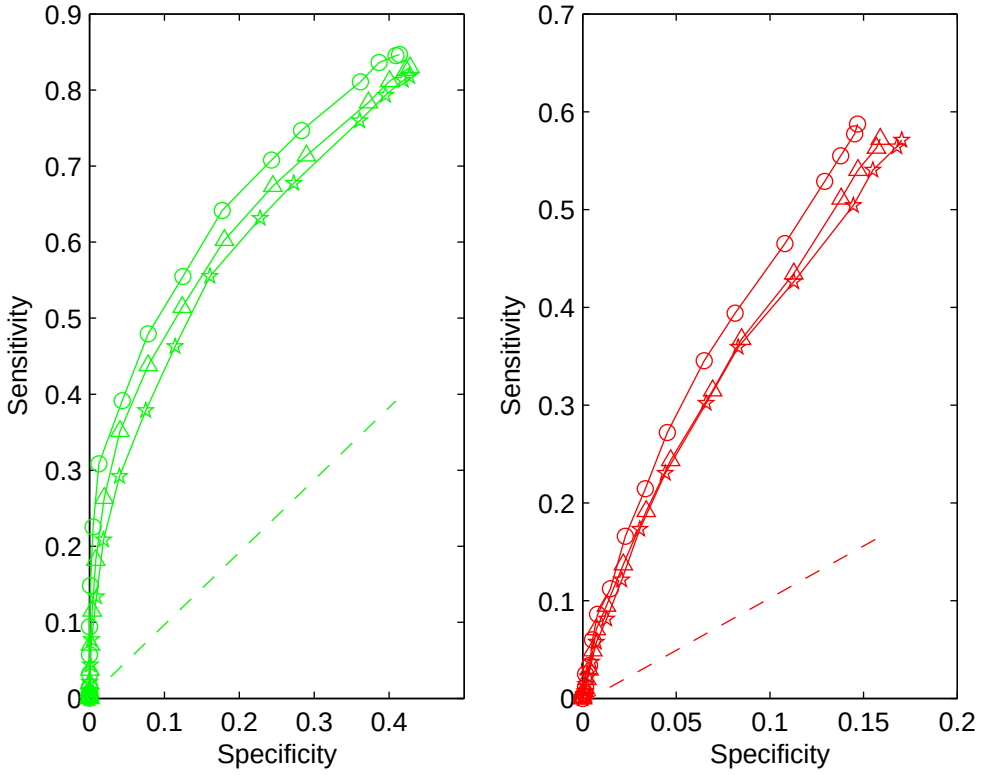

Figure 5.5: ROC curves of the human regulatory network using golden sets of different cutoffs. A golden set is defined by regulations which are highly reported in literature. An interaction belonging to the golden set with cutoff 5% is among the top 5% of the dataset in terms of the number of papers in reporting. Data points labeled by $\circ$, $\triangle$ and $\star$ are the results of golden sets whose sizes are 5%, 10% and 20% of the original network. The ROC curves (positive and negative) corresponding to a high-cutoff golden set enclose larger areas.

Fig. 5.5).

The optimal value of λ

With ROC curves in hand, we are in a position to choose an appropriate λ for Eq. 5.2. As a common practice, the quality of a ROC curve is quantified

by the area under the curve (see Appendix D for the estimation of the area). The optimal λ is thus the one whose ROC curve encloses the largest area. However, the direct comparison of different areas may be ambiguous. For example, compare the ROC curves from Fig. 5.4, the one on the left panel encloses a larger area while at the same time, the length covered in the x-axis is longer. To overcome the problem, we introduce a cutoff in the x-axis, and integrate area from 0 up to the cutoff. In this study, the cutoff is chosen to be 0.1. As the beginning region of the ROC curve refers to the highly reliable predictions, the introduction of the cutoff restricts ourselves in comparing the most reliable predictions. Thereafter, we define a quantity θ to measure the overall performance of the algorithm, which is the ratio between the area under the ROC curve from 0 to the cutoff and the corresponding area under the straight line $y = x$. The ratio could be understood as the advantage of our algorithm over random predictions.

The performance of a particular λ in Eq. 5.2 could be quantified by the resultant θ. In Fig. 5.6, we plot θ against different λ's for positive and negative ROC curves in the human dataset. In short, the optimal λ is the one which gives the largest θ. From Fig. 5.6, the optimal λ for positive and negative predictions are 0.025 and 0.030 respectively.

5.3 Validation of New Predictions

The practical application of our algorithm is to generate novel predictions of indirect regulations. Every non-zero matrix element of X stands for a prediction. However, predictions could fall into two categories: those covered

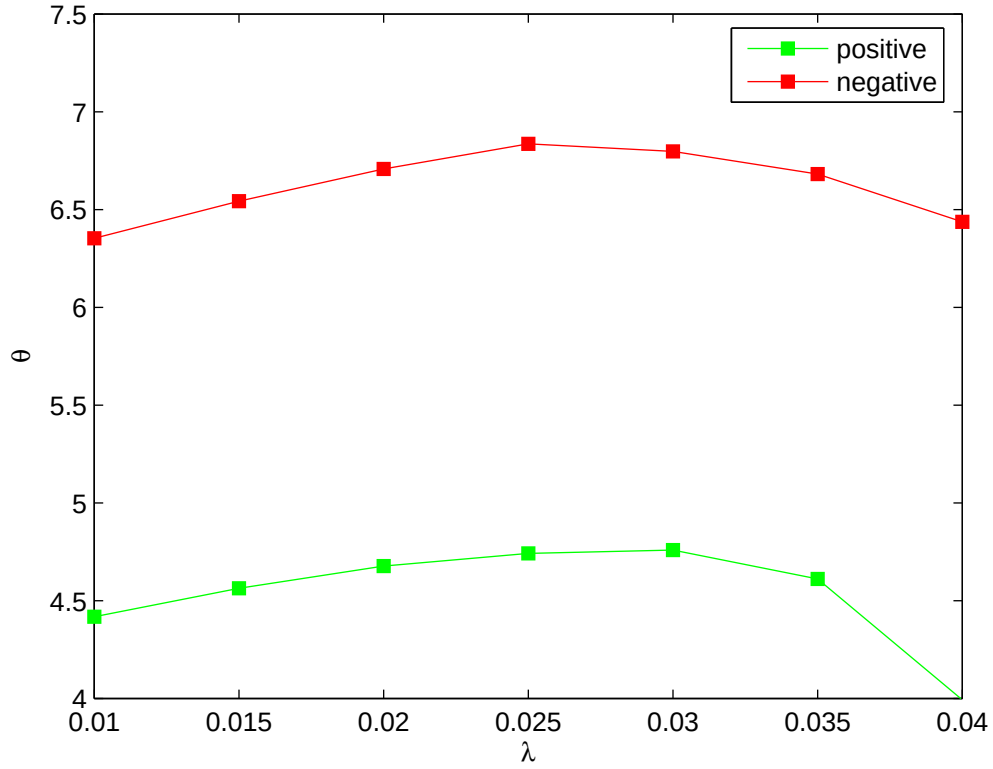

Figure 5.6: Determination of the optimal value of λ. The optimal λ maximizes θ defined by the ratio between the area under the ROC curve from 0 to 0.1 and the corresponding area under the straight line $y = x$. For human network, the optimal λ for positive and negative predictions are 0.025 and 0.030 respectively.

in the golden set and those not. Using the predictions covered in the golden set, we have calibrated the reliability. Next, we are going to focus on the predictions missing from the golden set. First of all, we do not consider these regulations as defects. In fact, being in the same predictive set, they possess the same quality as those covered in the golden set. Therefore, we could use them as real predictions of missing regulations and expand the original dataset with these predictions.

Table 5.2 shows the number of the these new predictions offered by our algorithm for the four model organisms. Two different quality cutoffs 95% and 75% are used. The number of predictions offered varies among the datasets, this is because the datasets have different number of nodes, links and topologies. However, in all cases, one could gain more predictions by lowering the quality cutoff. We would like to stress that the term quality is only calibrated within a dataset, therefore it is not meaningful to compare the new predictions in human and yeast even though the apparent qualities are the same. In fact, predictions from human dataset are the most reliable, because our algorithm is benefited from the heavily connected nature of the human dataset.

Without experimental verification, it is hard to validate our new predictions. To demonstrate our new predictions indeed make biological sense, we compare our new predictions to a dataset of human regulatory interactions. The dataset is also obtained from literature using the Medscan algorithm but all the regulations are not included in Table 5.1 and the matrix A (see Appendix D for details). We find that a significant fraction of our new predictions coincide with this dataset. As shown in Table 5.2, we have generated 2500 new predictions with an average quality of 95% for the human network. Among them 750 are indeed verified in the extra dataset. The corresponding P-value with respect to a random model is less than 10^{-100}. It is important to point out again that our algorithm predicts the signs of these 750 regulations, however, these have to be waited for future validation.

Table 5.2: Number of new predictions offered by our algorithm in regulatory networks of different organisms.

Organisms	95% sign quality	75% sign quality
Homo sapiens	2500	1.8×10^7
Saccharomyces cerevisiae	190	7100
Arabidopsis thaliana	85	13000
Drosophila melanogaster	650	1400

5.4 Conclusion

To sum up, we have developed a novel algorithm which allows one to verify already known indirect regulations, infer their signs (if it is not known), and to predict the new ones, which have not yet been experimentally detected. As an input it uses a network consisting of all presently known regulatory interactions (both direct and indirect). Our algorithm also allows one to make an educated guess about which of the interactions in the original network are direct and which are indirect in cases when this information is not readily available (as e.g. in microarray experiments following a perturbation localized on one or several genes). Thus it contributes to a popular topic of reconstructing direct regulatory network from microarray data [32, 81].

Bibliography

[1] Yule GU: **A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis**. *Philos. Trans. R. Soc. London B* 1925, **213**:21–87.

[2] Ohno S: *Evolution by gene duplication*. Berlin: Springer-Verlag 1970.

[3] Huynen MA, van Nimwegen E: **The frequency distribution of gene family sizes in complete genomes.** *Mol Biol Evol* 1998, **15**(5):583–589.

[4] Qian J, Luscombe NM, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model.** *J Mol Biol* 2001, **313**(4):673–681.

[5] Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E, Westbrock-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GK, Wu Z, Paulsen IT, Reizer J, Saier MH, Hancock RE, Lory S, Olson MV: **Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen.** *Nature* 2000, **406**(6799):959–964.

[6] van Nimwegen E: **Scaling laws in the functional content of genomes.** *Trends Genet* 2003, **19**(9):479–484.

[7] Newman M, L BA, Watts DJ: *The structure and dynamics of networks*. Princeton 2006.

[8] Alon U: *An introduction to system biology*. Chapman and Hall/CRC 2007.

[9] Tyson JJ, Csikasz-Nagy A, Novak B: **The dynamics of cell cycle regulation.** *Bioessays* 2002, **24**(12):1095–1109.

[10] Krishna S, Maslov S, Sneppen K: **UV-induced mutagenesis in Escherichia coli SOS response: a quantitative model.** *PLoS Comput Biol* 2007, **3**(3):e41.

[11] Shoemaker BA, Panchenko AR: **Deciphering protein-protein interactions. Part I. Experimental techniques and databases.** *PLoS Comput Biol* 2007, **3**(3):e42.

[12] Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, Chemama Y, Labigne A, Legrain P: **The protein-protein interaction map of Helicobacter pylori.** *Nature* 2001, **409**(6817):211–215.

[13] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403**(6770):623–627.

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

[15] Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RLJ, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM: **A protein interaction map of Drosophila melanogaster.** *Science* 2003, **302**(5651):1727–1736.

[16] Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in Saccharomyces cerevisiae.** *Science* 2002, **298**(5594):799–804.

[17] Ispolatov I, Krapivsky PL, Yuryev A: **Duplication-divergence model of protein interaction network.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2005, **71**(6 Pt 1):061911.

[18] Barabasi A, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**(5439):509–512.

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

[20] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks.** *Science* 2002, **298**(5594):824–827.

[21] Schroedinger E: *What is Life?* Cambidege Univ. Press 1944.

[22] Elowitz MB, Levine AJ, Siggia ED, Swain PS: **Stochastic gene expression in a single cell.** *Science* 2002, **297**(5584):1183–1186.

[23] Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N: **Noise in protein expression scales with natural protein abundance.** *Nat Genet* 2006, **38**(6):636–643.

[24] Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS: **Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise.** *Nature* 2006, **441**(7095):840–846.

[25] Swain PS, Elowitz MB, Siggia ED: **Intrinsic and extrinsic contributions to stochasticity in gene expression.** *Proc Natl Acad Sci U S A* 2002, **99**(20):12795–12800.

[26] van Kampen NG: *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library 2007.

[27] Keizer J: *Statistical Thermodynamcis of Nonequilibrium Processes*. Springer-Verlag 1987.

[28] Gillespie DT: **Exact stochastic simulation of coupled chemical reactions**. *J Phys Chem.* 1977, **81**:2340–2361.

[29] Bergmann S, Ihmels J, Barkai N: **Similarities and differences in genome-wide expression data of six organisms.** *PLoS Biol* 2004, **2**:E9.

[30] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102**:109–126.

[31] Ideker TE, Thorsson V, Karp RM: **Discovery of regulatory interactions through perturbation: inference and experimental design.** *Pac Symp Biocomput* 2000, :305–316.

[32] Pe'er D, Regev A, Elidan G, Friedman N: **Inferring subnetworks from perturbed expression profiles.** *Bioinformatics* 2001, **17 Suppl 1**:S215–24.

[33] Wagner A: **How to reconstruct a large genetic network from n gene perturbations in fewer than n(2) easy steps.** *Bioinformatics* 2001, **17**(12):1183–1197.

[34] Blatt M, Wiseman S, Domany E: **Superparamagnetic clustering of data.** *Phys Rev Lett* 1996, **76**(18):3251–3254.

[35] Bock Axelsen J, Yan KK, Maslov S: **Parameters of proteome evolution from the histogram of amino-acid sequence identities of paralogous proteins**. *Biol Direct* 2007, **2**:32.

[36] Maslov S, Sneppen K, Eriksen KA, Yan KK: **Upstream plasticity and downstream robustness in evolution of molecular networks.** *BMC Evol Biol* 2004, **4**:9.

[37] Yan KK, Walker D, Maslov S: **Noise and fluctuations in PPI networks governed by the law of mass action**. *to be submitted* 2007.

[38] Yan KK, Maslov S, Mazo I, Yuryev A: **Prediction and verification of indirect regulatory interactions in densely interconnected regulatory networks**. *q-bio.QM* 2007, :arXiv:0710.0892v2.

[39] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403–410.

[40] Yang Z: **Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites.** *Mol Biol Evol* 1993, **10**(6):1396–1401.

[41] Grishin NV, Wolf YI, Koonin EV: **From complete genomes to measures of substitution rate variability within and between proteins.** *Genome Res* 2000, **10**(7):991–1000.

[42] Uzzell T, Corbin KW: **Fitting discrete probability distributions to evolutionary events.** *Science* 1971, **172**(988):1089–1096.

[43] Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290**(5494):1151–1155.

[44] Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH: **Role of duplicate genes in genetic robustness against null mutations.** *Nature* 2003, **421**(6918):63–66.

[45] Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH: **Extent of gene duplication in the genomes of Drosophila, nematode, and yeast.** *Mol Biol Evol* 2002, **19**(3):256–262.

[46] Lynch M, Conery JS: **The origins of genome complexity.** *Science* 2003, **302**(5649):1401–1404.

[47] Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**(6634):708–713.

[48] Lundin LG: **Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse.** *Genomics* 1993, **16**:1–19.

[49] McLysaght A, Hokamp K, Wolfe KH: **Extensive genomic duplication during early chordate evolution.** *Nat Genet* 2002, **31**(2):200–204.

[50] Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Camara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Le Mouel A, Lepere G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Betermier M, Weissenbach J, Scarpelli C, Schachter V, Sperling L, Meyer E, Cohen J, Wincker P: **Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia.** *Nature* 2006, **444**(7116):171–178.

[51] Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV: **Birth and death of protein domains: a simple model of evolution explains power law behavior.** *BMC Evol Biol* 2002, **2**:18.

[52] Gillespie DJH: *The Causes of Molecular Evolution*. Oxford University Press 1994.

[53] Shakhnovich BE, Koonin EV: **Origins and impact of constraints in evolution of gene families.** *Genome Res* 2006, **16**(12):1529–1536.

[54] Roland CB, Shakhnovich EI: **Divergent evolution of a structural proteome: phenomenological models.** *Biophys J* 2007, **92**(3):701–716.

[55] Gu Z, Nicolae D, Lu H, Li W: **Rapid divergence in expression between duplicate genes inferred from microarray data**. *Trends in Genetics* 2002, **18**:609–613.

[56] Papp B, Pál C, Hurst L: **Evolution of cis-regulatory elements in duplicated genes of yeast**. *Trends in Genetics* 2003, **19**:417–422.

[57] Wagner A: **The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes.** *Mol Biol Evol* 2001, **18**(7):1283–1292.

[58] Giaever G, Chu A, Ni L, Connelly C, etal: **Functional profiling of the Saccharomyces cerevisiae genome**. *Nature* 2002, **418**:387–391.

[59] Kamath R, Fraser A, Dong Y, Poulin G, Durbin R, Gotta M, Kanapink A, Le-Bot N, Moreno S, et al:: **Systematic functional analysis of the Caenorhabditis elegans genome using RNAi**. *Nature* 2003, :231–237.

[60] Wagner A: *Robustness and Evolvability in Living Systems*. Princeton University Press 2005.

[61] Conant GC, Wagner A: **Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans.** *Proc Biol Sci* 2004, **271**(1534):89–96.

[62] Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast.** *Nature* 2003, **425**(6959):737–741.

[63] Maslov S, Sneppen K, Ispolatov I: **Spreading out of perturbations in reversible reaction networks**. *New J. Phys.* 2007, **9**(8):273.

[64] Maslov S, Ispolatov I: **Propagation of large concentration changes in reversible protein-binding networks.** *Proc Natl Acad Sci U S A* 2007, **104**(34):13655–13660.

[65] Kumar MDS, Gromiha MM: **PINT: Protein-protein Interactions Thermodynamic Database.** *Nucleic Acids Res* 2006, **34**(Database issue):D195–8.

[66] Bialek W, Setayeshgar S: **Physical limits to biochemical signaling.** *Proc Natl Acad Sci U S A* 2005, **102**(29):10040–10045.

[67] Gillespie DT: **The chemical Langevin equation**. *J Chem Phys.* 2000, **113**:297–306.

[68] Paulsson J: **Summing up the noise in gene networks.** *Nature* 2004, **427**(6973):415–418.

[69] Paulsson J: **Models of stochastic gene expression**. *Physics of Life Reviews* 2005, **2**:157–175.

[70] Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A: **Regulation of noise in the expression of a single gene.** *Nat Genet* 2002, **31**:69–73.

[71] Warren PB, Tanase-Nicola S, ten Wolde PR: **Exact results for noise power spectra in linear biochemical reaction networks.** *J Chem Phys* 2006, **125**(14):144904.

[72] Pedraza JM, van Oudenaarden A: **Noise propagation in gene networks.** *Science* 2005, **307**(5717):1965–1969.

[73] Raser JM, O'Shea EK: **Noise in gene expression: origins, consequences, and control.** *Science* 2005, **309**(5743):2010–2013.

[74] Simpson ML, Cox CD, Sayler GS: **Frequency domain analysis of noise in autoregulated gene circuits.** *Proc Natl Acad Sci U S A* 2003, **100**(8):4551–4556.

[75] Austin DW, Allen MS, McCollum JM, Dar RD, Wilgus JR, Sayler GS, Samatova NF, Cox CD, Simpson ML: **Gene network shaping of inherent noise spectra.** *Nature* 2006, **439**(7076):608–611.

[76] Li H, Wang W: **Dissecting the transcription networks of a cell using computational genomics.** *Curr Opin Genet Dev.* 2003, **13**(6):611–616.

[77] Herrgard M, Covert M, Palsson B: **Reconstruction of microbial transcriptional regulatory networks**. *Curr. Opin. Biotechnol.* 2004, **15**:70–77.

[78] Novichkova S, Egorov S, Daraselia N: **MedScan, a natural language processing engine for MEDLINE abstracts.** *Bioinformatics* 2003, **19**(13):1699–1706.

[79] Tringe SG, Wagner A, Ruby SW: **Enriching for direct regulatory targets in perturbed gene-expression profiles.** *Genome Biol* 2004, **5**(4):R29.

[80] Kyoda K, Baba K, Onami S, Kitano H: **DBRF-MEGN method: an algorithm for deducing minimum equivalent gene networks from large-scale gene expression profiles of gene deletion mutants.** *Bioinformatics* 2004, **20**(16):2662–2675.

[81] Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7**(3-4):601–620.

[82] Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195–197.

[83] Kruskal J: **On the shortest spanning tree of a graph and the traveling salesman problem.** *Proc. Am. Math. Soc.* 1956, **7**:48–50.

[84] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34**(Database issue):D535–9.

Appendix A

Details of Proteomes and Generation of Paralogous Proteins

Generating lists of paralogous proteins

The proteomes of *H. pylori* strain 26695 and *E.coli* strain K12-MG1655 were downloaded from the Comprehensive Microbial Resource (CMR, http://cmr.tigr.org) version 1.0. Sequences of *S. cerevisiae* proteins are from the Saccharomyces Genome Database (SGD, http://www.yeastgenome.org) version number 20031001. The *D. melanogaster*'s sequences are from the Berkeley Drosophila Genome Project (http://www.fruitfly.org), release 3.1. *C. elegans* and *H. sapiens* were from Wormbase (http://www.wormbase.org), release WS127 and the NCBI database (ftp.ncbi.nlm.nih.gov/genomes/H_sapiens), build 34.1 respectively.

The initial set of paralogous pairs for each of the organisms was identified by an all-to-all alignment of sequences of its proteins to each other using the BLASTP program [39]. For *H. pylori, E. coli, S. cerevisiae*, and *D.melanogaster* genomes, the E-value threshold of 10^{-10} was employed. This corresponds to p-values of the order of 10^{-12} (for *H. pylori*) and lower. Due to larger genome sizes of *C. elegans* and *H. sapiens* an even more conservative E-value of 10^{-30} was used to reduce the number of hits generated by the algorithm.

The "raw" datasets for worm, fly and human often contain multiple overlapping protein sequences predicted by different gene models of the same gene (including but not limited to different splicing variants). To avoid spurious hits we first mapped entries in raw datasets to unique gene IDs. This was easy to accomplish in the fly and worm datasets, where names of different gene models differ from each other by the last letter. In human genome,

this was done by mapping the gi numbers of sequences in the raw dataset to unique GeneID (LocusLink) identifiers from the Entrez Gene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene). Subsequently, if multiple BLAST hits were connecting the same pair of gene IDs we kept the one with the longest aligned region. This way we were guaranteed that one and only one pair of splicing (or gene model) variants per pair of gene IDs would contribute to the PID histogram.

In all genomes, only pairs in which the aligned region constituted at least 80% of the length of the longer protein were kept [45]. This excludes contribution from pairs of multi-domain proteins paralogous over only one of their domains.

Initially, the PID histogram in *S. cerevisiae* had two very sharp peaks at 51% and 70%. A close inspection revealed that these peaks are produced by evolutionary related subfamilies of nearly identical transposable elements. To correct for this obvious artifact in *S. cerevisiae* we removed 108 proteins encoded by known transposable elements listed in the Saccharomyces Genome Database and their homologs.

The overall shape of the PID histogram in regions I and II is not sensitive to the E-value cutoff chosen. Similarly, the results are nearly independent on the type of the BLOSUM substitution matrix used (in the end we opted for the BLOSUM45.) Finally, we verified that our results are independent of the alignment algorithm utilized to calculate PIDs. Indeed, in the fly dataset we have recalculated PIDs for all paralogous pairs detected by BLAST using much more sophisticated Smith-Waterman algorithm [82]. The resulting histogram is virtually indistinguishable from that based on the blastp output [35].

Study of P. tetraurelia

The proteome of *P. tetraurelia* were downloaded from the ParameciumDB (http://paramecium.cgm.cnrs-gif.fr/db/index). The lists of putative pairs of duplicated genes generated in each of the four WGD were downloaded from the supplementary materials of Ref. [50]. E-value of 10^{-30} was used due to the large size of the proteome. The procedures used to generate the paralogous pairs are described as above.

Identification of true duplicated pairs by the minimum spanning tree algorithm

We are naturally not in possession of the set of protein pairs that actually underwent duplication in the course of evolution of a given genome. The iden-

tification of the most likely candidates for these "true" duplicates is in general a rather complicated task which involves building the actual phylogenetic tree for every family in a genome. However we could make a much simpler educated guess by connecting paralogous proteins with the Minimum Spanning Tree (MST). The MST is a tree maximizing the sum of PIDs along its edges (or, to agree with its name, minimizing its opposite sign value). For a family consisting of F proteins such a tree has exactly $F - 1$ edges representing our best guess about the actual duplication events. One can prove the truth of this by induction: when a freshly duplicated pair is created with PID=100% it extends the previously existing Minimum Spanning Tree of a family by one edge. Assuming a constant rate of divergence for all paralogous pairs in a given family, the set of duplicated pairs would continue to form the Minimum Spanning Tree at all times. We used the Kruskal algorithm [83] to approximately detect the MST.

Appendix B

Details of Various Datasets Used

The transcriptional regulatory network of yeast used in Chapter 3

The network was taken from Ref. [16], which reported the so-called "ChIP-on-chip" study of *in vivo* binding of 106 transcription factors to upstream regulatory regions of genes encoding all 6270 of yeast proteins. Since the number of transcriptional regulators in this dataset is quite large, the probability that by pure chance the same transcription factor would be incorrectly detected among upstream regulators of *both* duplicated genes is small (of order of 1%). Thus the contribution of false positives of the dataset of Ref. [16] to the regulatory overlap Ω_{reg} is quite insignificant. We therefore use a P-value cutoff equal to 10^{-2} (12854 regulations) less conservative than the 10^{-3} cutoff (4418 regulations) of Lee *et al.* [16].

Protein interaction network of yeast used in Chapter 3

As a source of information about binding partners of yeast proteins we combined the data from two independent high-throughput two-hybrid experiments: the core dataset of Ito *et. al.* [14] (806 interactions among 797 proteins) and the extended Uetz *et. al.* dataset [13], downloaded from the website of this group (1446 interactions among 1340 proteins). The resulting network consists of 1734 proteins joined by 2111 non-redundant interactions. Using this combined dataset we found that even 100% identical proteins share on average only 30% of their binding partners. However, unlike for upstream regulation, the set of interaction partners of a protein is fully determined by its amino acid sequence. Therefore, an imperfect overlap in the set of binding

partners of identical proteins has to be attributed to false positives/negatives inevitably present in high-throughput two-hybrid experiments. The relatively high rate of false negatives in genome-wide two-hybrid experiments is further corroborated by the fact that datasets used in our study coming from two independent experiments [13, 14] have only 141 interactions in common.

Protein interaction network of *H. pylori* in Chapter 3

The two-hybrid assay of protein-protein inetractions in *H. pylori* used in Fig. 3.4 was obtained from the supplementary materials of Ref. [12]. It contains 1465 interactions between 732 proteins.

Protein interaction network of *D. melanogaster* used in Chapter 3

Our analysis of the interaction overlap between paralogous proteins in *D. melanogaster* is based on the full dataset of the high-throughput two-hybrid experiment [15]. It consists of 20671 protein-protein physical interactions involving 7002 of fly proteins.

Dataset of the yeast knockout experiment

The system-wide data on viability of *S. cerevisiae* null-mutants used in our study was obtained from Ref. [58] in which 1103 essential (non-viable null-mutants) and 4678 non-essential (viable null-mutants) yeast proteins were reported. The lists of viable and non-viable null-mutants as discovered in Ref. [58] were downloaded from the Saccharomyces Genome Database (http://www.yeastgenome.org/).

Dataset of the RNAi phenotypes in *C.elegans*

Our analysis of protective effects of paralogs in *C. elegans* is based on the set of 15587 viable and 1170 non-viable (embryonic or larval lethality or sterility) RNAi phenotypes reported in [59]. The information about worm paralogs is obtained from the EuGenes database (http://iubio.bio.indiana.edu:8089/) and consists of 30036 paralogous pairs involving 10071 worm proteins (blastp with 10^{-30} cutoff and no requirements on the length of aligned region). In Fig. 3.6 we used 13884 RNAi phenotypes for which we were able to uniquely map the genepair name to the worm protein name used in EuGenes.

Datasets used in the study of Chapter 4

The curated PPI network data used in our study is based on the 2.020 release of the BIOGRID database [84]. We kept only pairs of physically interacting proteins that were reported in at least two publications. That left us with 5798 non-redundant interacting pairs. Further restrictions for both proteins to have experimentally measured total abundance [62] narrowed it down to 4185 interactions among 1740 proteins. We further took the strongly connected component, resulting at a network with 1439 nodes and 3880 edges.

Appendix C

Proof of Eq. 4.30

Here we give a proof of Eq. 4.30,

$$\eta_\mu \geq (\frac{F_i}{K_d} + \frac{F_j}{K_d} + 1)^{-1}, \tag{C.1}$$

where μ is the dimer formed by proteins i and j.

Using Eq. 4.29, η_μ is given by the matrix element $(\Gamma^{-1})_{\mu\mu}$. Recall the definition of Γ from Eqs. 4.4 and 4.5, the RHS is the reciprocal of $\Gamma_{\mu\mu}$. We are going to prove that $(\Gamma^{-1})_{\mu\mu} > (\Gamma_{\mu\mu})^{-1}$.

First of all, Γ can be symmetrized by Eq. 4.10, i.e.

$$S = D^{-1/2}\Gamma D^{1/2}, \tag{C.2}$$

where D is a diagonal matrix as defined in Eq. 4.10. As S is symmetric, we can further diagonalize it by an unitary matrix U.

$$S = U\mathbf{D}U^{-1}, \tag{C.3}$$

where $\mathbf{D}$ is another diagonal matrix formed by the eigenvalues of Γ, and we know they are all positive.

Denote $Q = D^{1/2}U$, we can write down the matrix element $\Gamma_{\mu\mu}$ as

$$\Gamma_{\mu\mu} = \sum_k Q_{\mu k}\mathbf{D}_{kk}(Q^{-1})_{k\mu}. \tag{C.4}$$

Similarly,

$$(\Gamma^{-1})_{\mu\mu} = \sum_k Q_{\mu k}(\mathbf{D^{-1}})_{kk}(Q^{-1})_{k\mu}. \tag{C.5}$$

Using $Q = D^{1/2}U$ one can simplify Eqs. C.4 and C.5, arrive at

$$\Gamma_{\mu\mu} = \sum_k (U_{\mu k})^2 \mathbf{D}_{kk} \tag{C.6}$$

$$(\Gamma^{-1})_{\mu\mu} = \sum_l (U_{\mu k})^2 (\mathbf{D}_{kk})^{-1}. \tag{C.7}$$

Here, we have used the unitarity of U.

As $\mathbf{D}_{kk}$ are positive, for a convex function f, we have

$$f(\sum_k (U_{\mu k})^2 \mathbf{D}_{kk}) < \sum_k (U_{\mu k})^2 f(\mathbf{D}_{kk}). \tag{C.8}$$

The inequality $(\Gamma^{-1})_{\mu\mu} > (\Gamma_{\mu\mu})^{-1}$ follows as the function $f(x) = 1/x$ is indeed convex.

Appendix D

Methods and Datasets Used in Chapter 5

Collections of regulatory networks

The regulatory networks for different model organisms are obtained by the Medscan algorithm based on Natural Language Processing (NLP) [78]. The term "regulation" refers to the general influence of the activity of one protein by another. Therefore, apart from transcriptional regulations (which are direct regulations), regulations might be results of any post-transcriptional or post-translational interactions between proteins.

Regulations are extracted from over 14 million PUBMED abstracts and 47 full text journals. Properties of regulations including the sign (positive or negative) and its nature (direct and indirect) are parsed whenever the information could be extracted. The number of times a regulation is reported in literature is kept for the definition of golden sets. Details of each network is shown Table 5.1.

Apart from the data as shown in Table 5.1, we have extracted an additional set (35672) of human regulations. The regulations are not included with the datasets in Table 5.1 because their signs could not be parsed. In this study, we use them as independent validation for the new predictions generated by our algorithm.

Definition of golden sets

For each organism, the corresponding positive (negative) golden set is defined by the top 10% most frequently reported positive (negative) regulations. The size of each golden set could be found in Table 5.1. Comparing the whole human dataset and its corresponding golden set, the average number of times

that an interaction is reported are 3.35 and 22.6 respectively. The ratios between the two numbers are roughly the same for the other organisms.

Estimation of the area under a ROC curve

For each ROC curve, we fit the data point by the function $y = Ax^B$ using the MATLAB function `fminsearch`, which is based on the Nelder-Mead method in non-linear optimization. The area under the fitted curve is numerically evaluated in MATLAB by the function `quadl` using the adaptive Lobatto quadrature.

To exclude the data points far from the origin, which are results of less reliable predictions, we introduce a cutoff in the x-axis. Area is integrated from 0 up to the cutoff. In this study, a cutoff of value 0.1 is used.