Abstract

Computational analysis of biological networks: Measuring evolutionary rewiring and predicting regulatory relationship

Chong Shou

2011

Biological networks represent various types of molecular organizations in a cell. In the previous decade, large amount of network data have accumulated that facilitates our knowledge of the composition, topological structure, and functional significance of biological systems. Recently, great scientific achievement has been made to unravel inter- and intra- species variations at both molecular and system levels. Understanding how biological networks evolve could eventually help explain the general mechanism of cellular system. To this end, this thesis investigates the evolution of biological networks in terms of network rewiring. It compares rewiring rate differences among the common types of biological networks utilizing experimental data across species. Then it applies the rewiring rate formulism to show that regulatory networks generally evolve faster than non-regulatory collaborative networks, which is consistent among all species compared. It goes on to address network data quality issue and to computationally model the process of network rewiring with a simulation algorithm. Currently, building high quality biological networks is still the main goal in the system biology community. The final part of this thesis introduces a novel approach to predict transcription factor (TF) target genes in yeast, with significantly better prediction power than previously reported methods. It identifies histone sensitive and insensitive TFs to be distinct and biologically meaningful clusters.

Computational analysis of biological networks:

Measuring evolutionary rewiring and predicting regulatory relationship

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Chong Shou

Dissertation Directors: Mark B. Gerstein and Michael Snyder

May 2011

Acknowledgements

It is a pleasure to thank all those who made this thesis possible. First, I owe my deepest gratitude to Prof. Mark B. Gerstein for his professional guidance and supervision. He provided me with opportunities to work on many different projects, supported my research with insightful suggestions, and helped me finish my doctorate as a foreign student.

I am also grateful to my co-advisor Prof. Michael Snyder for his academic and financial support through all stages towards my doctorate, proposed ideas, and intriguing discussions.

I would like to thank my thesis committee members, Asst. Prof. James Noonan and Prof. Hongyu Zhao, for their advice and support on my work.

I am indebted to many of my colleagues in the Gerstein Lab and Snyder Lab, especially Philip M. Kim, Chao Cheng, Nitin Bhardwaj, Hugo Y.K. Lam, Ugrappa Nagalakshmi, Zhi Lu, Koon-Kiu Yan, Kevin Yip, Tara Gianoulis, Mihali Felipe, Steve Hartman and Miyoung Hong.

This thesis would have never been possible without the support and love from my parents and my wife Shanshan. You were always there when I had doubts and difficulties, encouraging me toward the completion of this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Biological systems are intricate, robust and regulated networks, which describe the collection of myriad fundamental molecular events. Cellular proliferation, differentiation, and environmental interactions each requires the production, assembly, operation, and regulation of many thousands of components, and they do so with remarkable fidelity in the face of many environmental cues and challenges [1]. Understanding the topology and underlying mechanism of these biological networks has become a major topic in functional –omics, which evidently helps summarizing, explaining and predicting experimental observations.

Modern molecular biology is established on the discovery, analysis, and manipulation of macro-molecules: DNAs, RNAs and proteins, which are the key components of all biological networks. Analyzing these molecules generally includes three major stages: sequencing, alignment, and evolutionary insights. Sequencing sets the stage of molecular biology, which unravels the chemical composition and linear structure of proteins in 1950s [2, 3], RNAs in 1960s [4], and finally DNAs in 1970s [5]. Recent advances in sequencing technology now enable rapid,

efficient, and cost-saving decoding of large genomic DNAs [6] and transcriptomic RNAs [7, 8]. With conserved sequences became available for two or more species, scientists started to develop computer algorithms that align the character sequences to identify regions of similarity [9, 10]. Upon the alignments of sequences from multiple species or individuals from within a species, evolutionary and population genetic analyses study the substitution rates, evolutionary constraints, and genetic variations.

Research of molecular sequences not only provides ample information and tools of DNAs, RNAs and proteins, it also serves as an example of future research direction of biological networks. In the past ten years, the advent of high-throughput techniques has facilitated the discovery and identification of many different types of biological networks which describe different aspects of the cellular system [1]. Similar to sequence alignments, researchers then compare and align networks from different species to uncover conserved signaling pathways and functional groups of molecules [11]. It is thus expected that evolutionary and population genetic analyses of biological networks will follow, much resembles the third stage of sequence research.

The types of biological networks currently include, but are not limited to, protein interaction, genetic interaction, transcription factor-target regulatory, miRNA-target regulatory, kinase-substrate phosphorylation, and metabolic pathway. Each of these networks represents a specific type of relationship among molecules, and has its distinct large-scale construction approach.

Protein interaction networks represent physical globular binding among protein molecules. Protein complexes that comprised individual binding proteins are the functional units of

biological processes. The first high-throughput physical interaction maps were generated using Yeast Two-Hybrid (Y2H) systems, which identifies binary interactions, in yeast *S. cerevisiae* and then other organisms including *D. melanogaster*, *C. elegans*, and *H. sapiens* [12-16]. Tandem Affinity Purification (TAP) technique followed by mass spectrometry is later used to identify large protein complexes *in vivo* [17, 18]. Protein interaction network is currently the most data abundant biological network, mostly due to that the above approaches are homogeneity and efficiency in detecting all potential interactions.

Genetic interaction networks denote epistatic relationship between genes in that a gene's function is affected by one or many other genes. The most common type is synthetic lethality in which mutations do not cause loss of viability individually, but are lethal when combined. Many approaches have been developed to detect genetic interactions, such as Synthetic Genetic Array (SGA), dipoid-based Synthetic Lethality Analysis with Microarrays (dSLAM) and more recently synthetic dosage-suppression analysis [19].

Transcription factor-target regulatory networks represent physical binding of transcription factors to the upstream DNA motifs of their target genes, regulating the level of transcription. Chromatin immunoprecipitation (ChIP) based techniques followed by microarray (ChIP-chip) or more recently direct sequencing (ChIP-seq) have been widely used to map genomic locations of transcription factor binding sites (TFBSs) [1]. Although the techniques are high-throughput and applicable to almost all transcription factors (TFs) and organisms, uncovering regulatory relationship networks exhaustively under all conditions and cell-types seems forbidding. Under the assumption that TFs recognize and bind to their specific motifs, many computational methods

tried to predict TFBSs using motifs discovered from previous ChIP experiments.

miRNA-target regulatory networks are relatively new type of biological network, which emerge from the discovery of microRNAs and their widespread activity in all multi-cellular organisms. miRNAs are small 23 nucleotides, on average, RNAs that regulate gene expression by binding to complementary sequence regions of their target mRNAs, which leads to the degradation of mRNAs [20]. High-throughput experimental approaches unraveling miRNA-target regulatory networks are not yet existed. However, multiple computational methods predicting miRNA targets using complementary sequence matching are widely accessible to build regulatory networks with reasonable accuracy. The comprehensive mapping of transcription regulation network including TFs, miRNAs, and target genes may render interesting regulatory motifs and hierarchical structures [21, 22].

Kinase-substrate phosphorylation networks represent phosphorylation events in proteome which play key roles in signaling pathways. It is estimated that ~30% of yeast and human cellular proteins are phosphorylated *in vivo* [1, 23-25]. Proteome chip technology has detected 1,325 substrates out of 4,400 yeast proteins being phosphorylated by 82 kinases *in vitro* [26]. Mass spectrometry is another widely used method allowing the identification of substrate phosphorylated residues [24]. Understanding the regulatory roles of kinases and phosphatases in a global phosphor-regulatory network reveals novel functional co-operations and a core interaction backbone [27].

Metabolic pathway networks are the collection of biochemical reactions which metabolize dietary compounds into the final nutritional products and energy, catalyzed by enzymes. Common

metabolic pathways, such as glycolysis and citric acid cycle, are well-characterized in the history of biochemistry, and they are highly conserved even comparing vastly distant species, indicating fundamental similarity of all living organisms [28].

Biological networks are composed of nodes (DNAs, RNAs, proteins, and metabolites) and edges (particular relationships between a pair of nodes). Studying the topology of networks gives us mathematically and biologically interesting findings. The number of connected neighbor nodes to a particular node is called its degree. Almost all biological networks studied are scale-free networks whose degree distribution follows a power law, compared to Poisson in random networks [29]. The major characteristic of a scale-free network is that only a few nodes, called hubs, have large degree, and most other nodes are connected to the network through the hubs. This characteristic contributes to the efficient connectivity and robustness of the network [29].

Computational biologists have found that hubs in yeast protein interaction and TF-target regulatory networks tend to be essential genes and under higher evolutionary constraints, removal of which results the cell unviable [30, 31]. Another important type of nodes is called bottlenecks, which act like “bridges” connecting two network clusters. Bottlenecks are also shown to be enriched of essential genes [32]. These findings are consistent with the topological importance of hubs and bottlenecks in maintaining the connectivity of scale-free networks.

Regulatory networks are usually associated with a special hierarchical structure that upper layer “regulators” manage their lower layer “targets”, but not vice versa [33]. The directional relationships form a top-down hierarchical regulatory structure much like those we have in governments and corporations [33]. Regulators in the middle tend to have more targets, and more

likely to collaborate with each other, indicating their heavy functioning duties [33, 34]. Establishing the regulatory hierarchical structure and relating it to biological features will improve our understanding of the complex cellular controlling and response system.

Relating three-dimensional protein structure to protein interaction network generates two classes of interaction types: transient and permanent [35]. Transient interactions are those using interacting sites in protein structure that also been used by other interactions. It is possible that transient interactions are only active and functional under certain conditions. On the contrary, permanent interactions are ones which may contribute to the formation of stable protein complexes. Linking the classification to other biological properties may uncover interesting implications in biology.

1.0.1 Evolutionary rewiring of biological networks

In chapter two [36], we present a unified formulism to measure network rewiring rate for all types of biological networks. In the past decade, we have accumulated a large amount of biological network data and expect even more to come. In the near future, we anticipate being able to compare many different biological networks as we commonly do for molecular sequences. It has long been believed that many of these networks change, or "rewire", at different rates. We have developed such a formalism based on analogy to simple models of sequence evolution, and used it to conduct a systematic study of network rewiring on all the currently available biological networks. We found that, similar to sequences, biological networks show a decreased rate of

change at large time divergences, because of saturation in potential substitutions. However, different types of biological networks consistently rewire at different rates. Using comparative genomics and proteomics data, we found a consistent ordering of the rewiring rates: transcription regulatory, phosphorylation regulatory, genetic interaction, miRNA regulatory, protein interaction, and metabolic pathway network, from fast to slow. This ordering was found in all comparisons we did of matched networks between organisms. We also investigated how readily our formalism could be mapped to other network contexts; in particular, we showed how it could be applied to analyze changes in a range of "commonplace" networks such as family trees, co-authorships and linux-kernel function dependencies.

1.0.2 Computational simulation of network rewiring

In chapter three [36], we describe a computational method to simulate the course of network rewiring, which is then used to support the analysis and conclusions in chapter two. It is generally agreed in the biological network community that current network data from large-scale experiments are suffered from a significant extent of false positives and false negatives. To evaluate the extent to which our rewiring rate measure is affected by unreal connections and incomplete data, we simulate data noise by rewiring the networks. Rewiring rate calculation is applied to these simulated networks, and perturbation analyzed to estimate the affect of data noise. Our network rewiring model includes four sources of rewiring, add edge, remove edge, add node, and remove node. Sensitivity analysis is performed to assess the relative importance of four

parameters, which is helpful in understanding the possible underlying mechanism of network rewiring.

1.0.3 Predicting TF regulatory network

In chapter four [37], we introduce a novel method to predict TF regulatory network incorporating chromatin modification information, and achieve better prediction power than previously reported methods. Transcription factors are key regulators of gene expression. Although experimental efforts are on the way to unravel binding profiles of more TFs in multiple conditions, cell lines, and species, it is almost impossible that we will exhaust all the combinations by experiments. Therefore, understanding TF binding mechanism from currently available data sets and making high quality predictions are the major issues in TF regulatory network research. A number of experimental and computational methods have been developed to identify target genes of transcription factors in yeast. Chromatin modifications affect transcription by changing the accessibility of transcription factors to chromatin and recruiting transcription factors. We propose a machine learning method that integrates transcription factor binding motif and chromatin modification profiles, which captures both condition-specificity and transcription factor-specificity of chromatin modifications, and substantially improves the prediction of transcription factor target genes in yeast. We found that transcription factors could be clustered into histone-sensitive and insensitive ones. The target genes of the histone-sensitive transcription factors have stronger signals of histone modification, while those of insensitive ones have weaker

ones. The two clusters also differ in degree of connectivity in protein-protein interaction network, position in the transcriptional regulation hierarchy as well as other features, indicating possible differences in their transcriptional regulation mechanisms. The model also shows potential application in distinguishing between direct and indirect transcription factor-DNA interactions.

In chapter five, we conclude the thesis with possible future directions.

Chapter 2

Analyzing evolutionary rewiring of biological networks

2.1 Introduction

With the advent of large-scale genomic and proteomic technologies in discovering interacting and regulatory relationships in cells, many types of biological networks, though incomplete, have been constructed in various eukaryotic species [12-14, 16, 21, 22, 26, 38-49]. The kinds of networks currently include, but are not limited to, protein interaction, genetic interaction, transcription factor-target regulatory, miRNA-target regulatory, kinase-substrate phosphorylation, and metabolic pathway. Biological networks have been used to explain differences between closely related species that share high sequence similarities [38, 39, 41]. For example, human and chimpanzee genomic sequences are found to have only 1.23% differences in SNPs and 3% in indels [50]. However, the subtle sequence divergence is hardly sufficient to explain phenotypical,

behavioral and social differences between the two species. As a result, biological networks (organizations of molecules) are proposed to play a central role in speciation complementary to individual molecules [38, 39, 41]. However, it is still largely unknown how fast biological networks evolve.

Biological network research has followed the path of sequence research to some degree. In the past three decades, biological sequence research has experienced three stages: initial sequencing data generation, pairwise alignment and evolutionary rate analysis. Simple models such as the Jukes-Cantor model [51] describe evolutionary sequence divergence in terms of time. In fact, various biological sequences evolve at different rates depending upon their functional importance [52, 53]. Genomic sequence analyses in various species have helped us to learn levels of conservation among genomic regions and genes [54-56]. Similarly, proteomic sequence and structure analyses show that protein regions have varied evolutionary constraints [57, 58]. Analogous to sequence analysis, the development of biological network research has three similar stages: network construction by large-scale experiments and computational predictions [12-14, 16, 21, 22, 26, 38-49], pairwise network comparison to find conserved edges as interologs or regulogs [11, 59] and building general network alignment tools [60, 61], and finally investigating levels of conservation and evolutionary change on biological networks.

One of the advantages of network study is that we can make analogies to draw intuition. For example, in commonplace social contexts, we readily observe that some “network” relationships change faster than others. Personal acquaintance networks may change in days, friendship networks and co-worker networks in months or years, while family networks change over

decades. This intuition of network stability differences could be quantified and compared by the rewiring rate that reflects the nature of network relationships. Similarly, in cellular systems biological networks may rewire at various rates during evolution.

Increasingly we have seen many approaches to compare biological networks across organisms, uncovering interesting relationships of network evolution and the functional implications [41, 62-68]. Due to current limitations of network construction technologies and the large evolutionary distance between the species compared, the overlap between current network datasets is small. Nevertheless, the estimation of the rewiring rate in protein interaction networks is possible [62]. Various methods were used in different studies inconsistent for direct comparison, with each focused on one of the biological network types. Also, most of the studies were species specific that did not compare species with large evolutionary divergence.

Given that previous studies have set the stage, now is an opportune time to quantify network rewiring in all these comparisons in a unified way. In the past three years, more data has become available for a greater number of species covering many types of biological networks [21, 22, 38, 39, 41]. The comprehensive set of network data allows systematic comparison of rewiring rates of biological networks and drawing more robust conclusions by using a set of species pairs.

We show here the rewiring rates of several types of biological networks in eukaryotes. The approach used is consistent across network types and robust to network data quality. We observed that the rewiring rate is characteristic of the type of edge (relationship between node entities) in both biological and commonplace networks. This analysis gives an initial picture of biological network rewiring and provides intuition and useful tools for the future when more network data

becomes available.

2.2 Results

2.2.1 Quantifying network rewiring rate

To calculate the rewiring rate of biological networks, we first established node orthology between two species, and then defined edge orthology as a conserved relationship between orthologous entities across different species, which is a generalization of "interologs" in protein interaction network and "regulogs" in TF regulatory network [11, 59]. One species network is considered reference, and three sets of nodes are identified. Common nodes (CNs) are nodes present in both networks, loss nodes (LNs) only in reference network and gain nodes (GNs) only in the other compared network. Four types of rewired edges are then identified and counted including gain or loss edges between CNs, loss edges involving LNs, and gain edges involving GNs (see Figure 2.1). The rewiring rate was measured by the total number of rewired edges (R) between two networks normalized by the combined network size, the total number of possible edges if two networks were both "complete" (C), and divergence time (T). Total number of rewired edges (R) counts all non-conserved edges (interologs, regulogs or other type of "logs") in two networks. The total number of possible edges (C) has five components: total possible edges of complete networks consisting of only common nodes (CNs), nodes that are only present in one of the two networks (GNs or LNs), and total possible edges between the two (between CNs and LNs, or CNs and GNs) (see Figure 2.2). The measure is in essence percentage edge change of

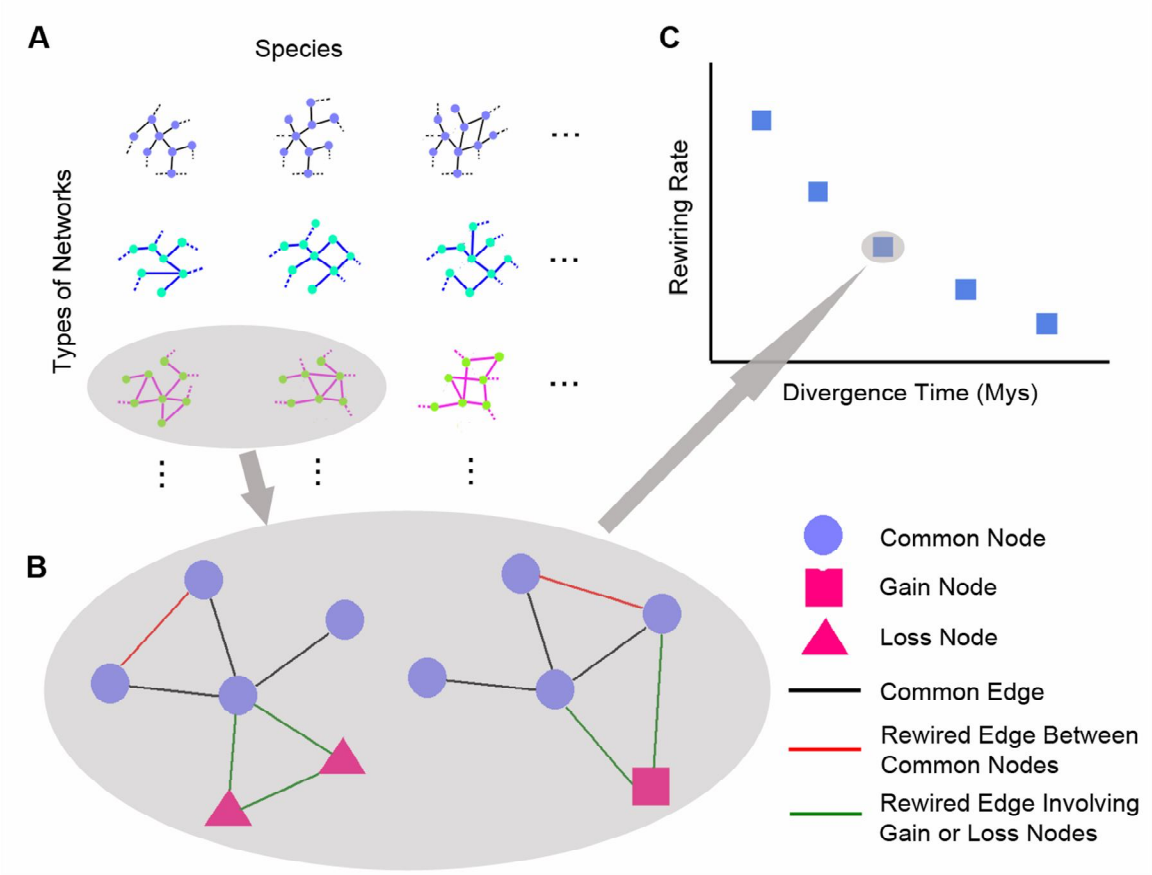

Figure 2.1: Measuring network rewiring by comparing networks of species pairs.
(A) Types of biological networks with currently available data for different species are collected. Selected types of commonplace networks with multiple time-point data are also collected. (B) For each network type, we perform edge rewiring analysis for pairs of species. Three types of nodes are first identified as CNs, GNs and LNs. Four types of rewired edges are then identified and counted including gain/loss edges between CNs (red) and those involving GNs or LNs (green). Rewiring rate from comparing the networks is calculated (see Materials and Methods). (C) Rewiring rate calculated from schematic (B) corresponds to a typical result point.

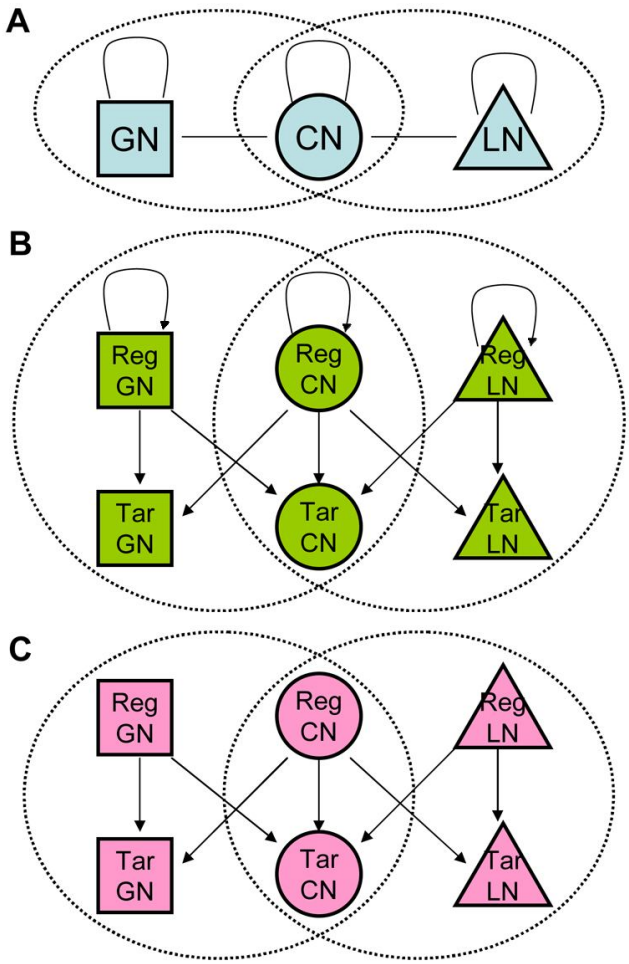

Figure 2.2: Schematic of total number of possible edges calculation in rewiring rate.
(A) For collaborative networks including protein interaction network, genetic interaction network and metabolic networks. Solid circles represent sets of nodes, as common nodes (CN), gain nodes (GN) and loss nodes (LN); dashed circles conceptually represent individual networks. Lines represent complete number of undirected edges between node sets, with each corresponding to a term in total number of possible edges summation. (B) For TF target regulatory network and kinase-substrate phosphorylation network. TFs or kinases are shown as regulators (Reg), and TF target genes or substrates as targets (Tar). Arrows represent complete number of directed edges between node sets. (C) For miRNA target regulatory network. miRNAs are shown as regulators (Reg) and their target genes as targets (Tar).

network in a given time period. We have collected data for each type of network for different species (see Table 2.1), and calculated rates for different time divergence species pairs (see Figure 2.1).

2.2.2 Log-Log linear relationship between divergence time and rewiring rate

For all types of biological networks, we observed faster rewiring rates for smaller divergence species pairs and slower rewiring rates for larger divergence species pairs, with a strong negative linear relationship between rewiring rate (per edge per Mys) and divergence time (Mys) in Log-Log scale (see Figure 2.3, Table 2.2). It was thus inappropriate to use the rewiring rate calculated from a specific species pair as a general measure for a network type. Using species pairs with different divergence times could result in large differences. However, different species pairs with similar divergence times tended to have close rewiring rates. This indicated that our rewiring rate measure was dependent upon divergence time but not on species.

We then asked the question whether the observed negative linear relationship in Log-Log scale between rate and divergence time in networks is parallel to what is seen in nucleotide sequence evolution. For sequence evolution, we use the equation $P = \frac{3}{4} - \frac{3}{4}e^{-8\alpha T}$ from the Jukes-Cantor model, where P is the percentage of sequence change and T is divergence time [51]. Though it is a simple model with only one parameter (α), Jukes-Cantor model captures the core relationship between P and T, and is sufficient in this case for comparing sequences with networks. P/T is the approximation of the instantaneous sequence evolutionary rate (dP/dT) and

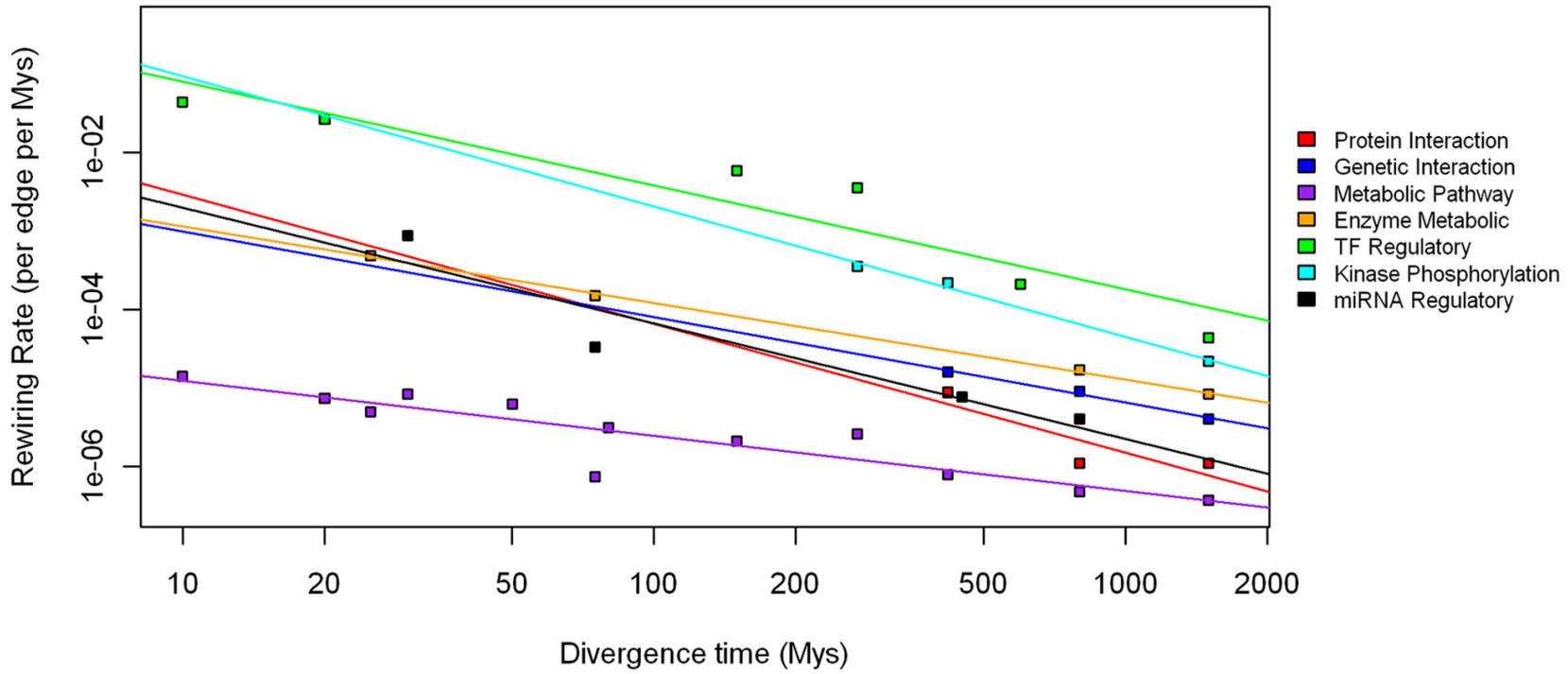

Figure 2.3: Ordering of extent of biological network rewiring.
Rewiring rates calculated for seven types of real biological networks (each with a different color) are shown as points on the Log-Log scale plot. Each rewiring rate corresponds to a divergence time of its two species comparison.

Species A	Species B	Estimated Divergence Time (Mys)	Networks Used for Rewiring Analysis
S. cerevisiae	*S. mikatae*	10	Metabolic pathway, TF,
S. cerevisiae	*S. paradoxus*	10	Metabolic pathway,
S. cerevisiae	*S. bayanus*	20	Metabolic pathway, TF,
H. sapiens	*M. mulatta*	25	Metabolic pathway, Metabolic enzyme
C. elegans	*C. briggsae*	30	Metabolic pathway, miRNA
D. melanogaster	*D. pseudoobscura*	50	Metabolic pathway,
H. sapiens	*M. musculus*	75	Metabolic pathway, Metabolic enzyme, miRNA
S. cerevisiae	*C. glabrata*	80	Metabolic pathway,
S. cerevisiae	*K. lactis*	150	Metabolic pathway, TF,
S. cerevisiae	*D. hansenii*	270	Metabolic pathway,
S. cerevisiae	*C. albicans*	270	Metabolic pathway, TF, Phosphorylation,
S. cerevisiae	*S. pombe*	420	PPI, Genetic, Metabolic pathway, Phosphorylation,
H. sapiens	*D. rerio*	450	miRNA
C. elegans	*D. melanogaster*	600	TF
H. sapiens	*D. melanogaster*	800	PPI, Genetic, Metabolic pathway, Metabolic enzyme, miRNA
H. sapiens	*C. elegans*	800	PPI, Genetic, Metabolic pathway, Metabolic enzyme, miRNA
S. cerevisiae	*D. melanogaster*	1500	TF
S. cerevisiae	*H. sapiens*	1500	PPI, Genetic, Metabolic pathway, Metabolic enzyme, Phosphorylation,

Table 2.1: Estimated divergence times between species pairs
All species pairs used in this study for calculating rewiring rates comparing species networks are listed with estimated divergence time in evolution. The types of networks used for each of these species pairs are also listed.

Biological Network	Linear Regression Model	Correlation Coefficient	P-val
Transcription factor regulatory network	$\log(r) = (-1.32 \pm 0.63) \times \log(t) + (0.22 \pm 1.43)$	-0.95	0.004**
Kinase phosphorylation network	$\log(r) = (-1.66 \pm 1.97) \times \log(t) + (0.62 \pm 5.45)$	-0.99	0.06*
miRNA regulatory network	$\log(r) = (-1.47 \pm 1.61) \times \log(t) + (-1.24 \pm 3.68)$	-0.94	0.06*
Protein interaction network	$\log(r) = (-1.64 \pm 11.83) \times \log(t) + (-0.90 \pm 34.43)$	-0.87	0.3
Genetic interaction network	$\log(r) = (-1.09 \pm 1.59) \times \log(t) + (-1.92 \pm 4.61)$	-0.99	0.07*
Metabolic enzyme network	$\log(r) = (-0.97 \pm 0.09) \times \log(t) + (-1.98 \pm 0.22)$	-0.99	0.0005***
Metabolic pathway network	$\log(r) = (-0.70 \pm 0.24) \times \log(t) + (-4.21 \pm 0.5)$	-0.90	0.00006****

95% confidence intervals for the fitted parameters are computed for linear models.

Table 2.2: Linear regression models of biological network rewiring rate and divergence time
For each type of biological network, rewiring rates (r) from different species pairs are regressed with divergence time (t), both in Log scale. Pearson correlation coefficient is also calculated.

can be used for direct comparison with rewiring rate of networks. A negative linear relationship was observed in Log-Log scale between P/T and T (see Figure 2.4), and was especially strong at large divergence times.

The analysis above indicated that the negative linear relationship between the rewiring rate and time in real networks could be universal and reflect underlying principles in evolution. This intuitively corresponds to the saturation of percentage change. For nucleotide sequences, as divergence becomes larger, the percentage of sequence change saturates at 0.75 according to the Jukes-Cantor model. New nucleotide changes happen on top of previous changes, which have little effect on percentage difference. Our analysis showed that the same is true for networks.

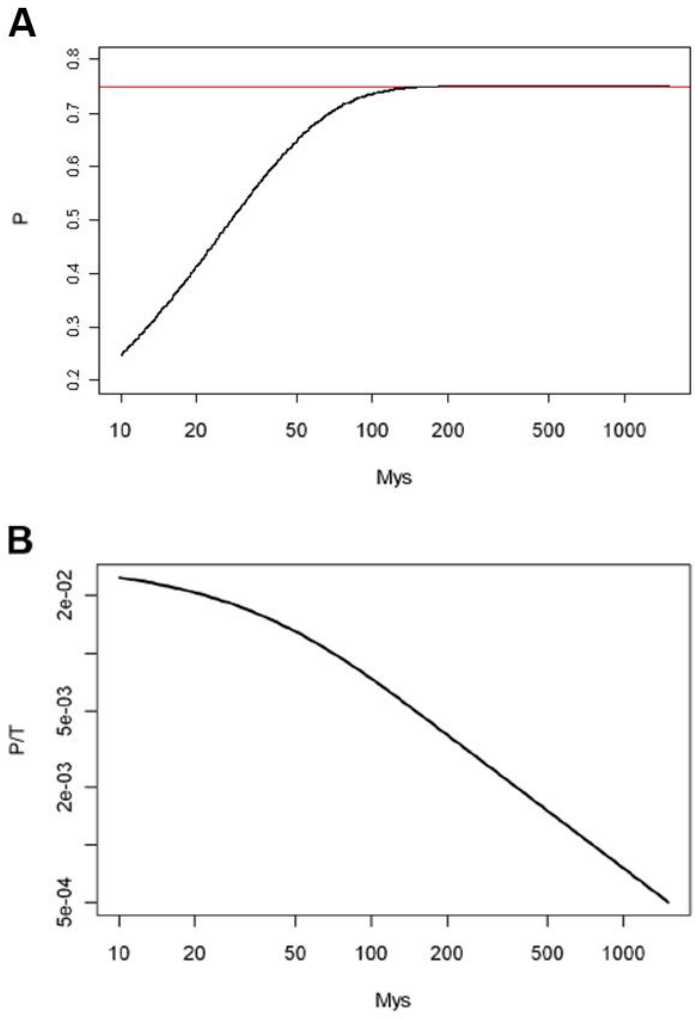

Figure 2.4: Saturation of nucleotide substitution in Jukes-Cantor Model.
(A) Saturation of percentage nucleotide substitution (P) at 75% (red horizontal line) as a function of divergence time in Log scale. (B) Relationship on Log-Log scale between sequence evolution rate, as number of nucleotide change per nucleotide per million years according to the Jukes-Cantor model, and divergence time.

Estimated Divergence Time (Mys)	~25	~75	~270	~800	~1500	Fitted 800
Metabolic Pathway Network	7.4E-6	3.1E-6	4.1E-6	5.4E-7	3.7E-7	5.7E-7
Protein Interaction Network	-	-	-	1.1E-6	1.1E-6	2.2E-6
Genetic Interaction Network	-	-	-	1.3E-5	4.0E-6	8.3E-6
Metabolic Enzyme Network	4.8E-4	1.5E-4	-	1.7E-5	8.4E-6	1.6E-5
miRNA Regulatory Network	8.6E-4	3.3E-5	-	4.1E-6	-	3.1E-6
Kinase Phosphorylation Network	-	-	3.5E-4	-	2.2E-5	6.5E-5
Transcription Factor Regulatory Network	2.3E-2	-	3.5E-3	2.1E-4	4.4E-5	2.4E-4

Table 2.3: Rewiring rate spectrum of eukaryotic biological networks
Using estimated divergence time between species pairs (see Table 2.1), we calculate rewiring rates for multiple time divergence of each type of biological networks, and show a subset of results here. 'Fitted 800' column is the fitted rewiring rate from linear regression at 800 Mys divergence time (see Figure 2.3). Network data is unavailable for rewiring rate calculation for blank cells. Rewiring rate is measured as rewiring per edge per Mys.

2.2.3 Rewiring rate as a discriminating characteristic of networks

We used the fitted rates from linear models for each type at 800 Mys divergence, roughly half the time of eukaryotic history (see Table 2.3). The "banding" of networks on the plot into characteristic groups with order of magnitude rate differences between them indicates the robustness of the rewiring rate calculation and the actual rate difference between networks.

In fact, the above described rewiring rate is an "average" rate rather than "instantaneous" rate for networks. As the Jukes-Cantor model shows for sequences, evolutionary rate (α) could only be approximately measured using instantaneous rate (dP/dT) between closely related species (dT is small), where α is proportional to dP/dT. When the divergence gets large, the approximation of instantaneous rate with the average rate is poor and the relationship between α and dP/dT

becomes non-linear. The logic is directly applicable to our case for networks.

Ideally, instantaneous rewiring rate should be measured using networks between closely related species. However, little network data are available for close species, which inhibits the calculation of instantaneous rewiring rates. The disadvantage of using the average rates described above is that at large evolutionary distance, network rewiring approaches saturation and is hard to compare. And the limited number of species network comparisons does not allow accurate estimations of instantaneous rates by the linear model at less than 10Mys divergence (see Table 2.2).

Another idea of comparing rewiring of biological networks is to use networks for a given divergence of the same species pairs. Since networks are of the same divergence, we use the percentage of edge changes among total possible changes, which is R/C, to measure the extent of rewiring (see Table 2.4). This method circumvents the disadvantages of average rewiring rate and limited species comparisons of networks, while it maintains the ability to distinguish the extent of network rewiring. For each of the 11 species comparisons listed in Table 2.4, biological networks are ordered according to their percentage of rewiring. We then count the number of cases where one type of biological network is observed to rewire more or less than another (see Table 2.5). Thus for each comparison between species (at a given level of divergence), we get an ordering of network rewiring (e.g. transcription regulatory > phosphorylation regulatory > protein interaction > metabolic pathway). We found that the ordering is consistent amongst all the 11 comparisons in this study. This result further supports the differences found in network rewiring using averaged rates.

Species Pair	Estimated Divergence Time (Mys)	Metabolic Pathway	Protein Interaction	Genetic Interaction	Metabolic Enzyme	miRNA Regulatory	Kinase Phosphorylation	Transcription Factor Regulatory
S. cer, S. mik	10	0.015 %	-	-	-	-	-	43%
S. cer, S. bay	20	0.015 %	-	-	-	-	-	46%
H. sap, M. mul	25	0.013 %	-	-	1.2%	-	-	-
C. ele, C. bri	30	0.025 %	-	-	-	2.6%	-	-
H. sap, M. mus	75	0.006 %	-	-	1.1%	0.25%	-	-
S. cer, K. lac	150	0.032 %	-	-	-	-	-	87%
S. cer, C. alb	270	0.11%	-	-	-	-	9.5%	95%
S. cer, S. pom	420	0.033 %	0.37%	0.67%	-	-	9.2%	-
D. mel, C. ele	600	-	-	-	-	-	-	13%
H. sap, D. mel	800	0.033 %	0.088%	1.04%	1.36%	0.32%	-	-
H. sap, C. ele	800	0.043 %	0.088%	0.42%	1.36%	0.33%	-	-
S. cer, D. mel	1500	-	-	-	-	-	-	6.5%
S. cer, H. sap	1500	0.056 %	0.17%	0.6%	1.26%	-	3.3%	-

Table 2.4: Percentage of rewired edges of eukaryotic biological networks

	TF regulatory (T)	Kinase phosphorylation (K)	Metabolic enzyme (E)	Genetic interaction (G)	miRNA regulatory (M)	Protein interaction (I)	Metabolic pathway (P)
T							
K	T > K: 1/1						
E	-	K > E: 1/1					
G	-	K > G: 2/2	E > G: 3/3				
M	-	-	E > M: 3/3	G > M: 2/2			
I	-	K > I: 2/2	E > I: 3/3	G > I: 4/4	M > I: 2/2		
P	T > P: 4/4	K > P: 3/3	E > P: 5/5	G > I: 4/4	M > P: 4/4	I > P: 4/4	

Table 2.5: Consistency of species comparison cases of network rewiring
The percentages of network rewiring calculated in Table 2.4 are compared for the extent of rewiring and summarized. '>' denotes the argument of greater rewiring extent of the column type of biological network over the row type. Network types are abbreviated using capital letters in rows. Only the lower triangle of this symmetric table is filled. The ratio denotes the number of cases supporting the argument out of the total number cases compared. All arguments are supported with full consistency of species pair comparisons.

2.2.4 Application of rewiring rate measurement to commonplace networks

The formalism of network rewiring was also applicable to non-biological networks to get some intuition for fast or slow rewiring processes (see Table 2.6). Three different representative commonplace networks with very different divergences were constructed, including co-authorship networks, family trees and Linux kernel design networks (see Figure 2.5). The three types of non-biological networks showed differential rewiring rates in the order of magnitudes (see Table 2.6). Consistent with our intuition, for example, family trees have less rewiring than co-authorship networks. Contrary to popular opinion of frequent computer software updates, Linux kernel design network in fact evolves approximately one order of magnitude

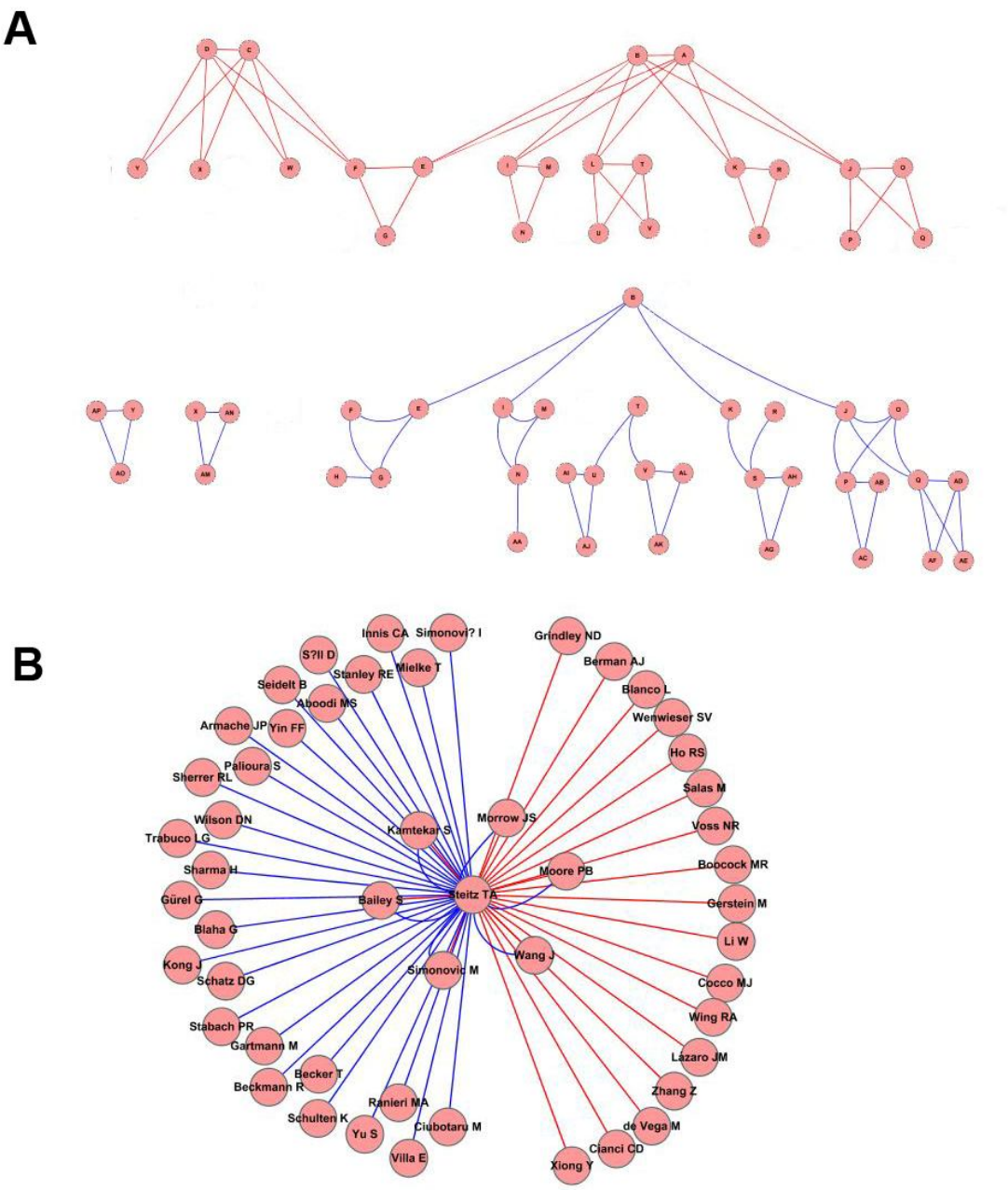

Figure 2.5: Visualization of types of social networks.
(A) A typical family tree in 1983 (red edges) and in 2009 (blue edges), with unchanged nodes aligned. (B) Dr. Steitz Lab co-authorship network in 2006 (red edges) and in 2009 (blue edges).

	Years of Change	Rewiring Rate (per edge per year)
Linux Kernel Design Network	2	1.7E-4
Family Tree	26	9.5E-4
Lab Co-authorship Network	3	2.9E-1

Table 2.6: Rewiring rates of selected commonplace network
Rewiring rates are calculated using the same method as for biological networks (see Materials and Methods). Notice that rewiring rate for social networks is measured in per year unit, as compared to per Mys unit in biological networks.

slower than a typical family tree (more family samples needed for statistically significant arguments). It is clear that rewiring rate could help us understand the nature of edge relationship in networks, thus can be used for direct comparisons among all kinds of biological and social networks.

2.2.5 Network rewiring and gene content turnover

Rewiring of biological networks consist of two sources: edge change between conserved nodes, and edge change from node gain and loss. We observed that a large fraction and in many cases the majority of network rewiring is attributed to the gain and loss of nodes (see Table 2.7). In fact, gene content turnover of two species contributes to the gain and loss of nodes in networks. Some studies have suggested differential gene content turnover of gene families, such as transcription factors and metabolic enzymes, in completely sequenced genomes [69-71]. Therefore, it is important to assess the impact of gene family evolution on the extent of their respective network rewiring.

Network Type	Species Pair (reference, compared)	Divergence (Mys)	Shared edges	Edge change from Edge Gain	Edge change from Edge Loss	Edge change from Node Gain	Edge change from Node Loss	Total possible edges	Percentage of rewiring by gene content turnover	Total Rate	Edge Gain Rate	Edge Loss Rate	Node Gain Rate	Node Loss Rate	Common Nodes	Gain Nodes	Loss Nodes
TF	*D. melanogaster, S. cerevisiae*	1500	3	80	80	12733	76543	1368376	0.06535923	4.36E-05	2.64E-05	2.64E-05	1.24E-05	7.47E-05	1011	3508	10867
TF	*D. melanogaster, C. elegans*	600	0	0	0	33752	76626	857710	0.1286892	2.14E-04	0	0	3.27E-04	1.86E-04	1833	6014	10045
TF	*S. cerevisiae, C. albicans*	270	54	NA	NA	677	193	924	NA	3.50E-03	NA	NA	3.40E-03	2.90E-03	55	193	677
TF	*S. cerevisiae, K. lactis*	150	95	NA	NA	519	152	766	NA	5.80E-03	NA	NA	5.60E-03	4.10E-03	96	152	519
TF	*S. cerevisiae, S. bayanus*	20	288	26	53	60	306	986	0.82247191	2.30E-02	3.10E-03	6.30E-03	2.80E-02	3.30E-02	213	53	229
TF	*S. cerevisiae, S. mikatae*	10	328	29	41	70	278	972	0.83253589	4.30E-02	6.00E-03	8.50E-03	7.60E-02	7.00E-02	242	46	200
PPI	*S. cerevisiae, H. sapiens*	1500	448	4189	554	48668	29693	48819933	0.94292693	1.10E-06	6.90E-06	9.80E-07	2.40E-06	5.70E-07	915	4407	7460
PPI	*H. sapiens, C. elegans*	800	45	289	408	4250	30242	38870600	0.98019267	1.10E-06	2.20E-06	3.10E-06	1.40E-06	1.10E-06	583	2233	7792
PPI	*H. sapiens, D. melanogaster*	800	113	1044	1778	21451	28804	58913366	0.94683196	1.10E-06	1.40E-06	2.30E-06	1.10E-06	1.10E-06	1405	5641	6970
PPI	*S. cerevisiae, S. pombe*	420	1093	654	4403	552	48405	14596540	0.90637612	8.80E-06	5.80E-06	3.90E-05	8.50E-06	8.10E-06	734	186	4640

Genetic	*S. cerevisiae, H. sapiens*	1500	0	2	1	57851	280	9666463	0.9999484	4.00E-06	5.80E-06	2.90E-06	4.00E-06	4.50E-06	21	4366	267
Genetic	*H. sapiens, C. elegans*	800	0	1	0	2104	281	570486	0.99958089	5.20E-06	3.50E-05	0	5.00E-06	8.40E-06	8	1020	280
Genetic	*H. sapiens, D. melanogaster*	800	0	18	6	5206	275	520765	0.99564033	1.30E-05	4.00E-05	1.30E-05	1.40E-05	8.40E-06	33	946	255
Genetic	*S. cerevisiae, S. pombe*	420	1261	5925	10191	3287	46441	10007207	0.75523966	1.60E-05	3.80E-05	6.60E-05	3.30E-05	1.20E-05	858	242	3563
Phosphorylation	*H. sapiens, S. cerevisiae*	1500	0	87	114	3981	27806	933247	0.99371639	2.20E-05	2.10E-05	2.70E-05	2.40E-05	2.30E-05	123	1244	2413
Phosphorylation	*S. cerevisiae, S. pombe*	420	226	299	277	1817	3053	58391	0.8942343	2.20E-04	1.30E-04	1.20E-04	2.30E-04	2.10E-04	154	325	551
Phosphorylation	*S. cerevisiae, C. albicans*	270	385	474	383	4260	2788	84823	0.8915876	3.50E-04	2.20E-04	1.80E-04	3.50E-04	3.20E-04	192	737	513
miRNA	*H. sapiens, C. elegans*	800	3	2	0	122	5679	1785268	0.99965535	4.10E-06	8.90E-05	0	3.30E-05	4.00E-06	11	133	4199
miRNA	*H. sapiens, D. melanogaster*	800	20	9	1	99	5661	1784155	0.9982669	4.00E-06	2.80E-05	3.10E-06	3.50E-05	4.00E-06	43	104	4167
miRNA	*H. sapiens, D. rerio*	450	300	198	168	914	5214	1875715	0.94364028	7.70E-06	1.10E-05	9.70E-06	2.10E-05	6.70E-06	468	742	3742
miRNA	*H. sapiens, M. musculus*	75	2138	410	477	3178	3067	2850669	0.87563096	3.30E-05	7.80E-06	9.10E-06	4.00E-05	3.80E-05	1987	2275	2214
miRNA	*C. elegans, C. briggsae*	30	12	1	0	44	115	6212	0.99375	8.60E-04	1.20E-04	0	9.40E-04	8.80E-04	35	51	109
Metabolic Pathway	*H. sapiens, S. cerevisiae*	1500	1099	64	158	208	652	1940402	0.7948244	3.70E-07	7.10E-08	1.70E-07	5.60E-07	4.00E-07	778	145	524
Metabolic	*H. sapiens, C.*	800	1191	13	192	47	520	1783038	0.73445596	5.40E-	2.00	2.90E	6.60	7.40E	905	48	397

Pathway	*elegans*									07	E-08	-07	E-07	-07			
Metabolic	*H. sapiens, D.*	800	1400	43	160	102	340	1922026	0.68527132	4.20E-	5.20	2.00E	5.60	6.40E	1013	107	289
Pathway	*melanogaster*									07	E-08	-07	E-07	-07			
Metabolic	*S. cerevisiae, S.*	420	1133	19	87	78	154	1031728	0.68639053	7.80E-	7.50	3.50E	1.00	1.50E	775	109	148
Pathway	*pombe*									07	E-08	-07	E-06	-06			
Metabolic	*S. cerevisiae, C.*	270	463	18	179	19	743	869746	0.79457769	4.10E-	3.20	3.20E	3.80	4.30E	459	20	464
Pathway	*albicans*									06	E-07	-06	E-06	-06			
Metabolic	*S. cerevisiae, D.*	270	1196	19	80	196	97	1323634	0.74744898	1.10E-	9.80	4.10E	1.50	2.70E	847	244	78
Pathway	*hansenii*									06	E-08	-07	E-06	-06			
Metabolic	*S. cerevisiae, K.*	150	1146	7	100	102	128	1097612	0.68249258	2.10E-	6.90	9.80E	2.80	5.00E	825	138	98
Pathway	*lactis*									06	E-08	-07	E-06	-06			
Metabolic	*S. cerevisiae, C.*	80	1204	10	36	56	133	955560	0.80425532	3.10E-	1.80	6.60E	6.70	9.90E	827	61	96
Pathway	*glabrata*									06	E-07	-07	E-06	-06			
Metabolic	*H. sapiens, M.*	75	1831	11	24	17	45	1744282	0.63917526	7.40E-	9.40	2.00E	4.50	4.50E	1250	20	52
Pathway	*musculus*									07	E-08	-07	E-06	-06			
Metabolic	*D. melanogaster,*	50	1199	24	109	42	238	1336356	0.6779661	6.20E-	5.40	2.40E	1.00	1.30E	945	43	175
Pathway	*D. pseudoobscura*									06	E-07	-06	E-05	-05			
Metabolic	*C. elegans, C.*	30	1196	73	25	190	31	1282064	0.69278997	8.30E-	2.80	9.70E	1.70	2.10E	927	184	26
Pathway	*briggsae*									06	E-06	-07	E-05	-05			
Metabolic	*H. sapiens, M.*	25	1706	12	90	10	106	1728384	0.53211009	5.00E-	3.20	2.40E	1.20	2.20E	1225	14	77
Pathway	*mulatta*									06	E-07	-06	E-05	-05			
Metabolic	*S. cerevisiae, S.*	20	1299	23	40	48	32	963026	0.55944056	7.40E-	1.40	2.50E	2.10	4.60E	904	60	19
Pathway	*bayanus*									06	E-06	-06	E-05	-05			
Metabolic	*S. cerevisiae, S.*	10	1303	16	36	61	32	994114	0.64137931	1.50E-	2.00	4.40E	4.30	9.20E	904	76	19
Pathway	*mikatae*									05	E-06	-06	E-05	-05			
Metabolic	*S. cerevisiae, S.*	10	1305	6	32	39	34	947466	0.65765766	1.20E-	7.40	3.90E	4.00	8.90E	902	52	21
Pathway	*paradoxus*									05	E-07	-06	E-05	-05			

Metabolic Enzyme	*H. sapiens, S. cerevisiae*	150 0	503	54	52	5301	15147	1634182	0.99484285	8.40E-06	1.10 E-06	1.10E -06	9.10 E-06	8.30E -06	182	467	935
Metabolic Enzyme	*H. sapiens, C. elegans*	800	506	44	12	4431	15185	1490166	0.99715331	1.70E-05	1.70 E-06	4.80E -07	2.30 E-05	1.60E -05	178	347	939
Metabolic Enzyme	*H. sapiens, D. melanogaster*	800	586	132	38	21080	15064	2644584	0.99531861	1.70E-05	3.30 E-06	9.40E -07	1.90 E-05	1.60E -05	225	979	892
Metabolic Enzyme	*H. sapiens, M. musculus*	75	2699	116	104	11863	12744	2146602	0.99113868	1.50E-04	6.10 E-06	5.40E -06	1.80 E-04	1.80E -04	505	570	612
Metabolic Enzyme	*H. sapiens, M. mulatta*	25	1263	16	0	6947	14368	1762542	0.99924992	4.80E-04	4.80 E-06	0	5.40 E-04	5.20E -04	365	441	752
Linux	*V4, V15*	2 yrs	11072	877	3189	11981	2696	55228321	0.78306568	1.70E-04	2.50 E-05	9.00E -05	1.90 E-04	2.20E -04	8498	5585	1334
Linux	*V4, V27*	4.5 yrs	7111	1156	4213	25451	5633	10776233 6	0.85271445	7.50E-05	2.00 E-05	7.20E -05	6.70 E-05	1.10E -04	7286	12506	2546
Family	*1983, 2009*	26 yrs	19	0	1	25	19	1821	0.97777778	9.50E-04	0	2.20E -04	8.30 E-04	1.50E -03	19	33	18
Co-author ship	*2006, 2009*	3 yrs	8	0	1	38	38	445	0.98701299	5.80E-02	0	1.60E -02	5.00 E-02	7.50E -02	7	17	13

Table 2.7: Detailed rewiring rates for networks and species pairs

Detailed information of rewiring rate results for all networks and species-pairs studied. Numbers of common nodes, gain nodes and loss nodes are provided. Four types of rewired edges (gain edge between common nodes, loss edge between common nodes, gain edge involving gain/loss nodes, loss edge involving gain/loss nodes) are also distinguished for separate rewiring rates. Note for biological networks, rewiring rates are measured by per edge per Mys, while for commonplace networks by per edge per year.

		H. sapiens – M. musculus	C. elegans – C. briggsae	S. cerevisiae – K. lactis
Transcription factor activity	Non-conserved genes	99	165	19
	Total genes	1063	504	235
	Content turnover	9%	33%	8%
Kinase activity	Non-conserved genes	39	100	7
	Total genes	767	460	250
	Content turnover	5%	22%	3%
Metabolic process	Non-conserved genes	129	200	68
	Total genes	2015	1072	1172
	Content turnover	6%	19%	6%

Table 2.8: Gene content turnover of 3 GO categories
Genes in *H. sapiens*, *C. elegans* and *S. cerevisiae* from 3 GO categories are identified from annotations. In the counter species (*M. musculus*, *C. briggsae* and *K. lactis*) their orthologous counterparties are mapped. Gene content turnover for the species pair is measured as the number of non-conserved genes over the total number of genes in the GO category.

In order to examine whether the turnover of a specific set of genes, such as kinases and TFs, have impact on their corresponding network rewiring, we examined the gene content turnover of 3 GO categories using 3 species pairs (see Table 2.8). The 3 GO categories (transcription factor activity, kinase activity, and metabolic process) are selected to be compared with TF-target regulatory network, kinase-substrate phosphorylation network, and metabolic enzyme network, respectively. For the 3 categories of proteins, we did not observe a clear pattern in which some categories had faster turnover than others. This suggests that differences in network rewiring across networks may not come from the gene content turnover of corresponding GO category proteins. The rewiring of networks should mostly reflect the characteristic of biological

relationships rather than specific GO category molecules themselves.

2.2.6 Biological networks evolve in rates comparable to protein sequences

Cellular molecules, as nodes in biological networks, are under differentiated selection pressure, and therefore evolve at different rates. Genomic analyses from model organisms have shown the spectrum of sequence conservation among types of genomic annotations, in which protein coding exon sequences are the most conserved, intron sequences are the least conserved, and regulatory cis/trans elements are somewhere in between [72]. Proteins as the products of DNA coding sequences are generally thought to be under great constraint. Another special product from DNA sequences is ribosomal RNA, which is considered the most conserved locus in the genome [73].

We asked whether the edge rewiring rates in biological networks were in the range of node changes. Since there is no analogous concept of "total possible edges between nodes" in sequence comparisons, a naïve sequence/network identity-based method was used to measure the percentage change between two sequences/networks for consistency. Here, only edge changes in networks are counted to compare with nucleotide change in sequences. Sequence identity is calculated as the percentage of the number of unchanged nucleotides or amino acids in global alignment per length of the alignment. Similarly, network identity is calculated as the percentage of the number of unchanged edges out of total number of edges in two networks. Then, the rate can be calculated as (1- percentage identity)/(divergence time) for both sequence and network.

This equates one edge change with one nucleotide or amino acid change. We realized this might not be the best, but a default to start with.

Using this definition, we observed that biological networks evolve in a range comparable to that of protein sequences in both species cases (see Figure 2.6). Transcription factor-target regulatory networks, the fastest rewiring biological networks, were comparable to the top 0.1% and 4% of the fastest evolving protein sequences in *Homo sapiens* and *Sacchromyces cerevisiae*, respectively. The slowest rewiring metabolic pathway network was comparable to the bottom 23% and 36% of the slowest evolving protein sequences. The density distribution of protein coding DNA sequence rates had a similar peak position but a smaller standard deviation than protein sequence rates, because an amino acid change does not necessarily result from changes of all its three codon positions. Therefore the evolutionary rate distinction between protein coding sequences and biological networks became more significant: with 0.5% and 4% of sequences slower than metabolic pathway networks in human and yeast, respectively, and 0% and 4% of sequences faster than transcription factor-target regulatory networks. The 18S rRNA sequences evolved slower than all biological networks analyzed here: approximately 60% rate of the slowest rewiring metabolic pathway network in human and 1% of the rate in yeast.

2.2.7 Permanent protein interactions rewire slower than transient interactions

Since rewiring rates are capable of distinguishing different network types, we attempted to use rewiring rates to study different subtypes of edges within protein interaction networks.

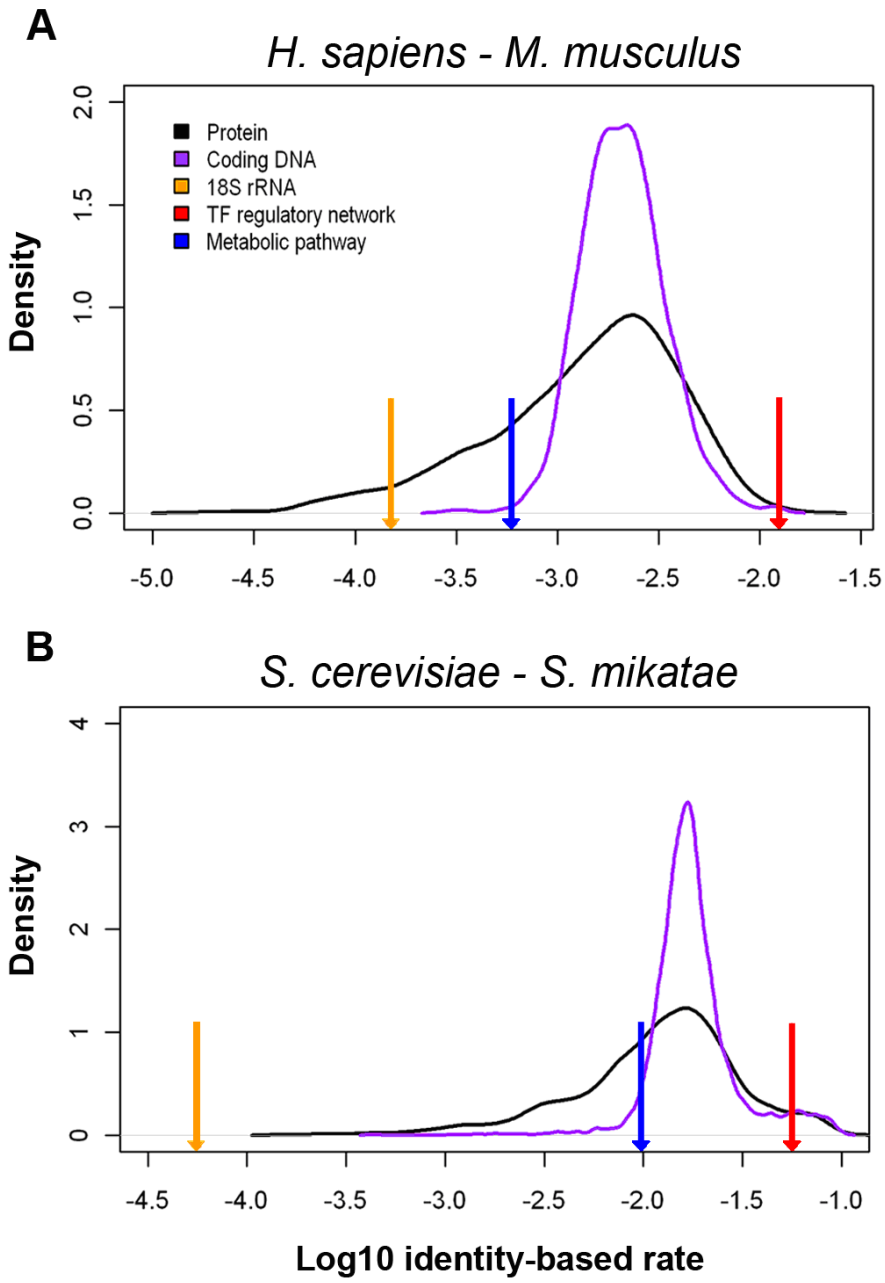

Figure 2.6: Network rewiring rates is comparable to molecular sequence change.
(A) Network rewiring evolution is compared to molecular sequence evolution using *H. sapiens* and *M. musculus* data, and (B) using *S. cerevisiae* and *S. mikatae* data. Two density distributions of identity-based evolutionary rate are shown as for protein sequences (black line) and protein coding DNA sequences (purple line). 18S rRNA rate (orange arrow), transcription factor regulatory network rate (red arrow) and metabolic pathway network rate (blue arrow) are also shown for relative positions to sequence rate distributions.

Edge Type	Human Permanent	Human Transient	Yeast Permanent	Yeast Transient
Conserved	8	8	38	66
Non-Conserved	1088	2874	318	1106
Total	1096	2882	356	1172

Table 2.9: Permanent protein interactions rewires slower than transient interactions
We distinguish permanent and transient edges for protein physical interactions. Fisher's Exact Test is performed to test conservation difference between permanent and transient edges, with P-value=0.05 for human and P-value=0.002 for yeast. Human network edges are compared to *D. melanogaster* for conservation, and yeast *S. cerevisiae* network is compared to *S. pombe*.

Relating protein 3-D structures to protein interaction networks helped us to distinguish simultaneously possible (permanent) interactions from mutually exclusive (transient) interactions [35]. The difference between the two types of interactions is whether an interaction between two proteins has competition from a third potential interacting protein for the same interacting site. It has long been hypothesized that protein pairs of permanent interactions tend to co-evolve during evolution [74]. The co-evolutionary effect could help to maintain the stability of permanent interactions.

Structural interaction networks (SINs) for both human and yeast were constructed using updated and coherent datasets. Permanent and transient interactions were identified through interacting site regions in proteins and number of interacting partners for each site. Conservation of permanent and transient interactions was measured by their presence in another reference species network (see Table 2.9). Significant conservation distinction was observed for permanent and transient interactions in both yeast (p-value=0.001) and human networks (p-value=0.05)

using Fisher's Exact Test. Stronger conservation of permanent protein interactions indicated that the interacting sites within two proteins were more constrained to maintain the interaction via co-evolution of interacting sites.

2.2.8 Paralogs rewire at a close pace in protein interaction networks

The results above showed that the rewiring rate of network edges reflects the biological nature of edge types. It is also plausible that proteins with different characteristics might have different rewiring rates than their network partners. Here, we used protein interaction networks to investigate how protein paralogs behave during evolution in terms of changing their interacting partners. We collected all paralog pairs present in human and yeast interaction networks and calculated the rewiring rate difference between each pair. The distribution of the rate difference was then compared with a background distribution calculated for all protein pairs in the networks (see Figure 2.7).

In both human and yeast networks, the paralog pairs had rate difference distribution shifted to zero compared to background (Wilcoxon test p-value $< e^{-15}$ in yeast, p-value = 0.004 in human). The result suggested that paralog pairs tend to have a smaller rewiring rate difference, demonstrating a closer evolutionary rate of network change. In fact, as paralogs emerge from the event of gene duplication in ancestral species, they share sequence similarities [75]. Here, we showed that paralogs also shared network similarities as the network rewiring rates of paralogs were similar. After the gene duplication events which lead to their formation in ancestral species,

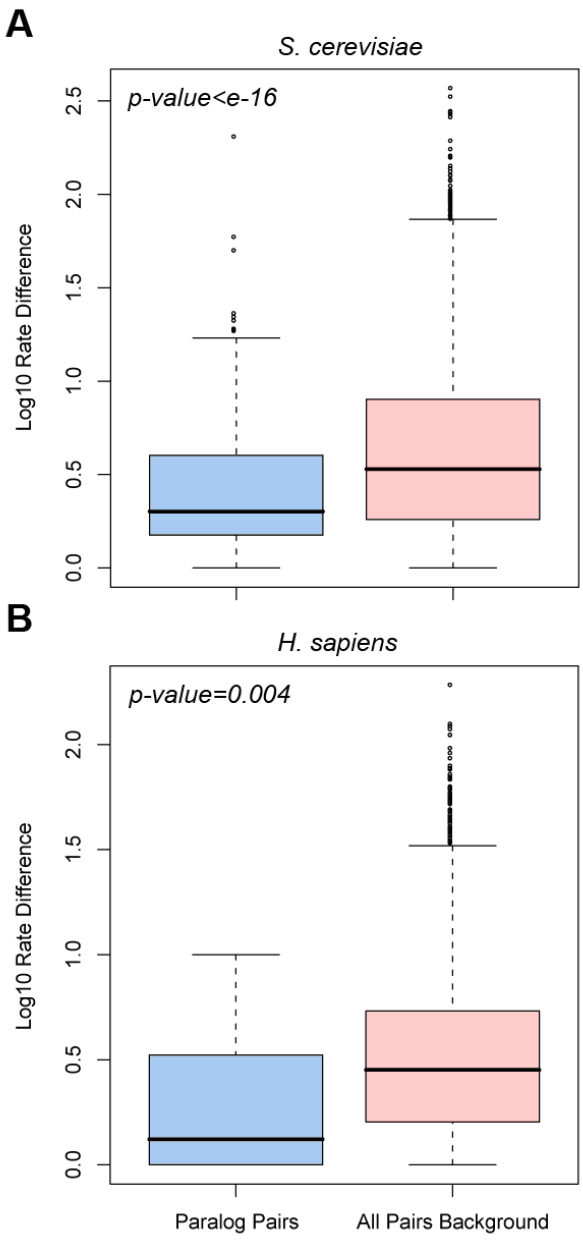

Figure 2.7: Rewiring rate difference of paralog pairs in protein interaction networks.
(A) Boxplot of rewiring rate difference in yeast and (B) human protein interaction networks between paralog pairs (blue) and between all node pairs as background (pink). Paralog pairs tend to have smaller rewiring rate difference than expected.

paralogs are likely to have similar constraint on sequences and network partners due to their shorter evolutionary history than random protein pairs.

2.3 Discussion

King and Wilson proposed [76] and Bourman et al. [38] then demonstrated that fast changing regulatory relationships in transcription factor-target networks could account for the species differences, which could hardly be explained by the highly conserved protein and DNA coding sequences. Following that study, small- and large-scale evidence has been presented to support the view that after the divergence of two species, fast change in regulatory relationships may have a critical role in speciation [39, 41]. As we have shown above, transcription factor-target regulatory networks and kinase-substrate phosphorylation networks are two major types of regulatory networks that have the fastest evolutionary changing rates among networks and protein sequences, confirming the importance of regulation in species evolution.

2.3.1 Collaborative networks and regulatory networks

Biological networks are characterized by their functional relationships: protein binding, expression regulation, phosphorylation, etc. We introduce another way to categorize biological networks into collaborative and regulatory networks by the reversibility of edges to help understand rewiring rate distinction among network types. Collaborative networks are the biological networks with reversible edges—either the edges are undirected or directed but

reversible. By reversibility we mean that a reversed edge is biologically possible between a pair of nodes. Regulatory networks have irreversible edges: a reversed edge may not be biologically possible. By this definition, transcription factor-target regulatory networks, miRNA-target regulatory networks, and kinase-substrate phosphorylation networks fall into the regulatory network group; and protein interaction networks, genetic interaction networks, and metabolic networks fall into the collaborative network group.

Our network rewiring analysis shows that in general, regulatory networks tend to rewire faster than collaborative networks (see Table 2.3). Two of the regulatory networks, transcription factor-target regulatory networks and kinase-substrate phosphorylation networks, are the fastest rewiring biological networks in this study. Transcriptional regulation of gene expression by transcription factors is carried out by transcription factor binding to the transcription start site commonly upstream of a gene. Recognition of a binding site is often specific to a sequence pattern buried in the site [77]. Post-translational modification of protein substrate by kinases also involves recognition of sequence patterns in substrate's phosphorylation site [78]. Sequence pattern matching as a major factor in establishing regulatory relationships could be an important reason of fast rewiring. A single nucleotide/amino acid change in the target's binding-recognition sites, could lead to a "digital" recognition site change. Besides, a number of studies have showed that both transposable element insertion and genomic rearrangement led to considerable indel changes at transcription factor binding sites [79-84]. The digital and indel changes in binding-recognition sites greatly contribute to the large turnover of transcription factor-target regulatory network.

Collaborative networks show slower rewiring rates than regulatory networks. Contrary to "digital" or "indel" changes in regulatory networks, changes tend to be "structurally continuous" in collaborative networks. Here, we generally refer to the globular interactions which are the majority in physical interaction networks. On the other hand, the general collaborative physical interaction network in this study still includes interactions mediated by kinases and domains such as SH3 which are in fact regulatory relationships. In fact, protein functions gradually change as sequence changes, and most proteins do not change their functions radically with their sequences conserved. As a natural implication of the sequence-function paradigm, it is not surprising that collaborative protein networks rewire as protein sequences evolve. In this study we include two representations of metabolic networks. Metabolic enzyme networks are constructed using enzymes as nodes and edges connect two nodes if the product of one serves as the substrate of the other. The rewiring rates of metabolic enzyme networks are similar to other collaborative networks (see Table 2.4). On the other hand, metabolic pathway networks that are constructed using chemical compounds as nodes and reactions as edges rewire the slowest. For example, the biosynthesis metabolic pathway of acetyl-CoA from pyruvate is identical in human and yeast, but the corresponding metabolic enzyme network rewires (see Figure 2.8). In fact, metabolic reactions process chemical compounds into energy and nutrition, and are mostly essential for living. Our results suggest that the essentiality is partly reflected in the slower rewiring rate of metabolic pathway networks than that of other types of biological networks and protein sequences. Based on these results, we think that enzymes for reactions are less constrained to change while the underlying reactions remain highly conserved.

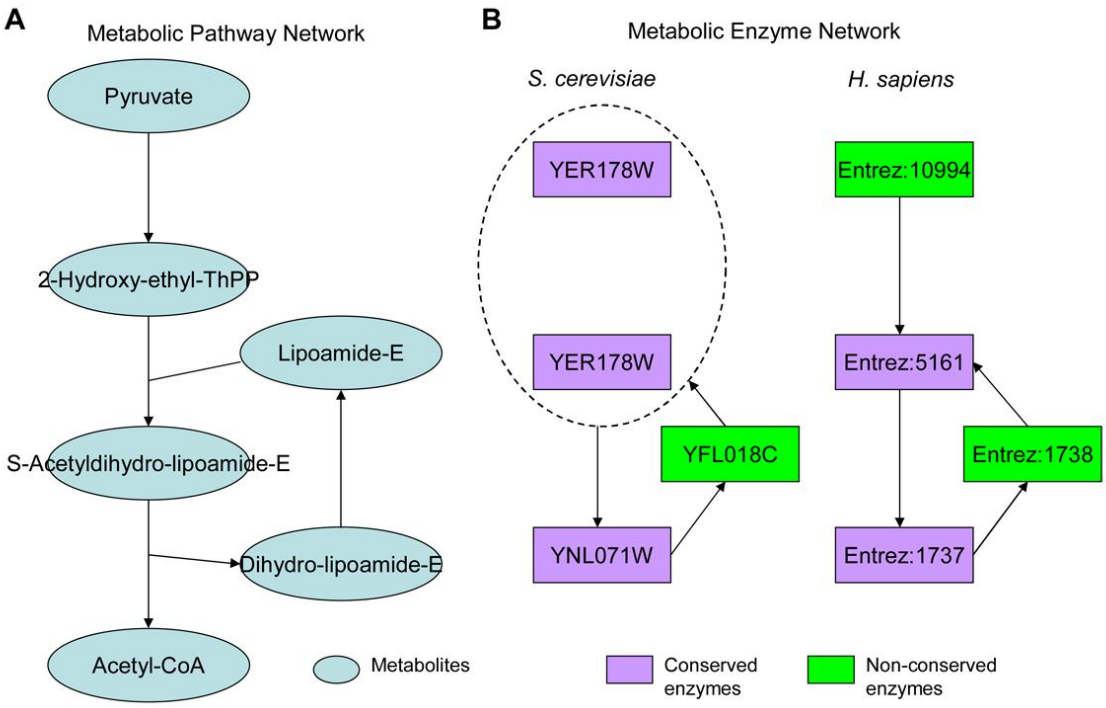

Figure 2.8: Example rewiring of metabolic pathway network and metabolic enzyme network. (A) The biosynthesis pathway of acetyl-CoA from pyruvate showing metabolites (circles) and reactions (arrows). The pathway is identical in human and yeast. (B) The corresponding metabolic enzyme networks in yeast and human showing enzymes (rectangles) and product-substrate relationships (arrows). Each enzyme corresponds to a reaction in (A). Purple rectangles represent orthologous enzymes from two species, while green rectangles represent non-orthologous enzymes. The dashed circle shows one yeast enzyme coded by YER178W catalyzes two consecutive reactions, but different enzymes catalyze each reaction in human.

2.3.2 Network rewiring as an important aspect of cellular system evolution

We now know that there are two layers of cellular evolution, individual molecules and organizations of molecules. Therefore, it is our ultimate goal to understand how individual molecule changes affect cells and their organization and collaboration.

Some factors may also influence and shape the landscape of biological networks (see Figure 2.9). It has been shown that external environment can influence the conservation of regulatory relationship and network motifs in prokaryotic transcription factor-target networks [85, 86]. Relationships tend to be conserved in organisms living in similar environmental niches, despite large evolutionary distance. Whole-genome duplication events rapidly reorganized transcription regulatory networks through the survived duplicates and their functional divergence afterwards [87-91]. And the regulatory networks, in a feedback way, could affect the survival of duplicated genes [92].

This study attempts to systematically investigate the evolutionary rate of all known types of biological networks in terms of rewiring. According to our results, it is possible that small changes of molecular sequences lead to large network re-organizations and this augmentation effect makes small molecular changes more detectable by natural selection. This is especially true for regulatory networks with the greatest augmentation effects caused by minor changes of regulators. If the above assumptions are true, network rewiring should be an essential tool to understand the differences between closely related species such as human and chimpanzee, because their molecular sequences are nearly identical. More importantly, intra-species network

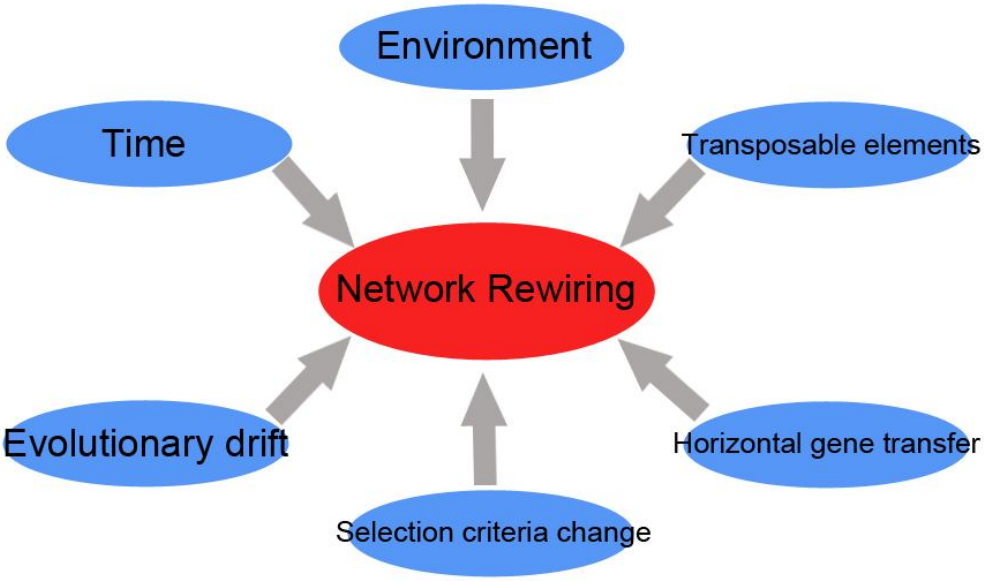

Figure 2.9: Factors shaping network rewiring.

rewiring variations will help at an individual level beyond SNPs and structural variations.

2.3.3 Future directions of network rewiring analysis

In the future, we foresee additional calculations and analyses that could be performed when accurate and more complete network data becomes available for more species. Analogous to sequence analysis, we can build species trees comparing biological networks and infer branch lengths using rewiring rates. From this study, we know that types of biological networks and molecular sequences evolve at different rates, but it is still unclear whether network rewiring "speeds up" in some species and "slows down" in others. We can use benchmark rates and develop comparative ratios to measure this. This is actually quite similar to using dN/dS ratio

(non-synonymous changes versus synonymous changes) to measure selection pressure on coding sequences. Building the tree is important to understanding biological system evolution compared to traditional molecular evolution.

Network hubs and bottlenecks are of general interest in biological research due to their topological importance. Both hub and bottleneck proteins in human and yeast protein interaction networks tend to rewire their edges faster than non-hub non-bottleneck proteins (see Figure 2.10). One reason for this is that hubs with large degrees tend to have more rewired edges, and therefore faster rewiring rates. Further detailed analysis is needed to understand the rewiring rates of bottleneck proteins.

It is also interesting to look for rewiring "hotspots" and "coldspots" within biological networks. Subnetworks and motifs that are enriched in fast or slow rewiring edges may have biological function implications. Immune response, transport and localization associated genes in human protein interaction networks have been found to change interacting partners relatively quickly [61]. The analysis could also be applied to other types of biological networks.

Further network rewiring analysis will possibly investigate factors affecting network rewiring (see Figure 2.9). These efforts will greatly increase our understanding of cellular system evolution, intra-species variation, and speciation.

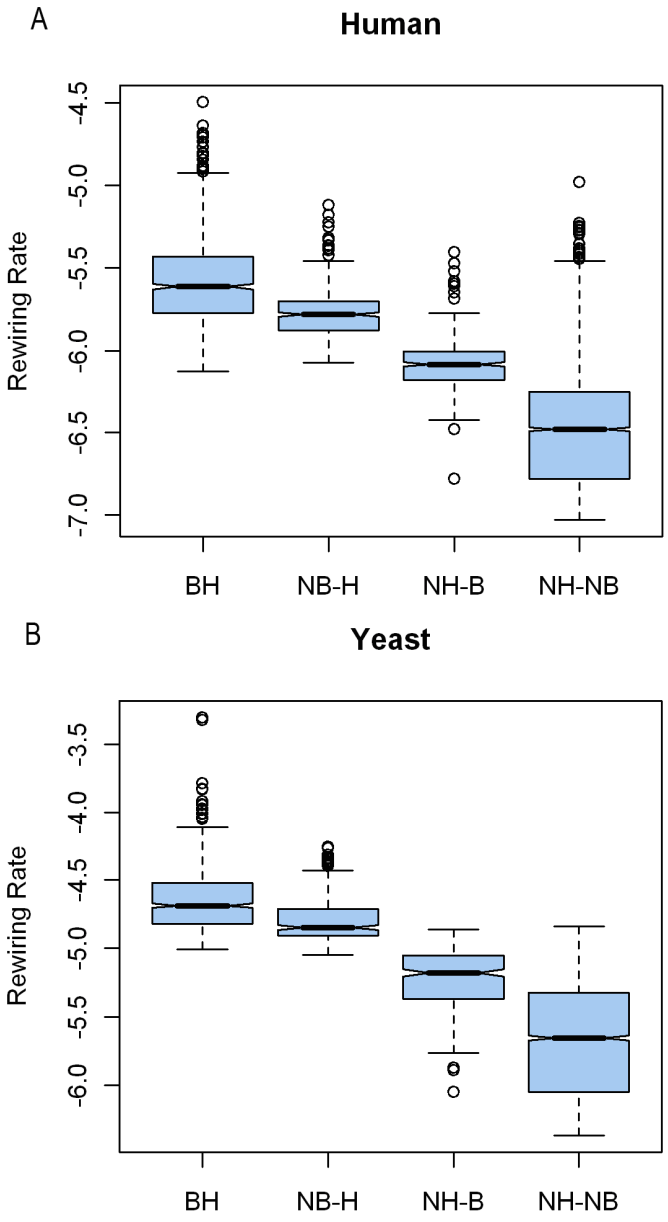

Figure 2.10: Rewiring rate of hubs and bottlenecks in protein interaction networks.
Rewiring rates are calculated for all proteins in (A) human and (B) yeast protein interaction networks. Hubs are defined as top 20% proteins ranked by their degree, and bottlenecks as top 20% ranked by betweenness. Proteins are grouped into 4 categories: Bottleneck hubs (BH), Non-bottleneck hubs (NB-H), Non-hub bottlenecks (NH-B) and Non-hub non-bottlenecks (NH-NB). Either hubs or bottlenecks are found to have faster rewiring rates than NH-NBs (Wilcoxon p-val<e-15).

2.4 Methods

2.4.1 Datasets of networks, sequences and homologs

For different types of biological networks, we gathered data from multiple sources. Binary protein physical interaction networks and genetic interaction networks were extracted from BioGRID database v2.0.55 (http://thebiogrid.org/) for 5 species: *H. sapiens*, *C. elegans*, *D. melanogaster*, *S. pombe* and *S. cerevisiae* [93]. Metabolic pathway networks of compound reactions were obtained from KEGG database (http://www.genome.jp/kegg/) for 16 species: *H. sapiens, M. mulatta, M. musculus, C. elegans, C. briggsae, D. melanogaster, D. pseudoobscura, S. pombe, D. hansenii, C. albicans, K. lactis, C. glabrata, S. bayanus, S. mikatae, S. paradoxus* and *S. cerevisiae* [94]. Metabolic enzyme networks were constructed from the pathway networks for 7 species: *H. sapiens, M. mulatta, M. musculus, C. elegans, D. melanogaster, D. hansenii,* and *S. cerevisiae*, by establishing directed edges from upstream reaction enzymes to downstream reaction enzymes. miRNA-target regulatory networks were constructed from miRBase (http://www.mirbase.org/) predictions with edges pointing from miRNAs to target genes in 5 species: *H. sapiens*, *M. musculus, D. rerio, C. elegans* and *D. melanogaster* [95]. Transcription factor-target regulatory networks were extracted from various sources: *S. cerevisiae, C. elegans* and *D. melanogaster* networks from large-scale ChIP-Chip and ChIP-Seq experiments [21, 22, 40], *C. albicans, K. lactis, S. bayanus, S. mikatae* networks from recent small-scale experiments [38, 39]. Kinase-substrate phosphorylation network for *S. cerevisiae* was obtained from

large-scale protein chip experiments [26]. Phosphorylation networks of yeast species *S. cerevisiae, C. albicans* and *S. pombe* were constructed in two steps. We first obtained phosphorylation sites identified by MassSpec [41], and also obtained kinase binding specificity data from kinase binding specificity experiments [96]; then used MOTIPS analysis pipeline to identify responsible kinases for each phosphorylation site by matching position weight matrices (PWMs) [97]. Structural Interaction Networks (SINs) for *H. sapiens* and *S. cerevisiae* were constructed in a similar way as the first version of yeast SIN [35], using protein domain interaction data from iPfam database Release 20.0 (http://ipfam.sanger.ac.uk/) [98].

For social co-authorship network, we parsed the co-author lists of 2009 Nobel Prize Winner Thomas A. Steitz's 2009 and 2006 publications from PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), and constructed co-authorship networks for Dr. Steitz. For social family tree network, we obtained data from a typical family with its trees in 1983 and 2009 (see Figure 2.5). Edges in family trees stand for either marriage or child/parent relationship. Linux kernel design networks are obtained for 3 versions, v2.6.4, v2.6.15 and v2.6.27. From v2.6.4 to v2.6.15 and from v2.6.15 to v2.6.24, the time separations are around 2 years and 2.5 years, respectively [99]. One edge in Linux kernel design networks represents one function calling or using another function.

Protein sequences and protein coding DNA sequences for *H. sapiens*, *M. musculus* and *S. cerevisiae* were downloaded from BioMart database (http://www.biomart.org/) [100], and from SGD (http://www.yeastgenome.org/) for *S. mikatae*. 18S ribosome RNA sequences for all 4 species were extracted from Entrez database (http://www.ncbi.nlm.nih.gov/Entrez/) [101].

Orthologous sequences in *H. sapiens-M. musculus* and *S. cerevisiae-S. mikatae* pairs were then aligned using MUSCLE software v4.0 (http://www.drive5.com/muscle/) [102] for calculations of sequence identity.

Sequence orthology for non-fungi species pairs used in this study was downloaded from InParanoid database v7.0 (http://inparanoid.sbc.su.se/cgi-bin/index.cgi) [103]. Orthology for fungi species pairs was obtained from Fungal Orthogroups Repository v1.1 (http://www.broadinstitute.org/regev/orthogroups/) [104]. Paralog pairs in *H. sapiens* and *S. cerevisiae* were extracted from HomoloGene database Release 64 (http://www.ncbi.nlm.nih.gov/homologene) [101].

2.4.2 Calculating network rewiring rates

We used a consistent method to calculate rewiring rates comparing two networks for all network types. First, orthology relationships between nodes from the same network type in two species were established. Second, three sets of nodes were distinguished. Common Node (CN) set includes nodes having orthologous counterparts present in both networks. Loss Node (LN) set includes nodes present in the reference network but absent of orthologous counterparts in the compared network. And Gain Node (GN) set includes nodes present in the compared network but not having orthologous counterparts present in the reference network. Third, we counted the total number of rewired edges (R) between two networks. Rewired edges between two networks were defined as the union of edges between pairs of CNs that only present in one network and all edges

involving LNs and GNs. Fourth, we counted the total number of possible edges (C) in the two networks. This was basically the number of non-redundant edges if two networks are both fully connected. Finally, the following equation was used to calculate the rewiring rate for a pair of networks:

$$\text{Rewiring rate} = \frac{\text{R}}{\text{C} \times \text{Time divergence}}$$

The time divergence is either estimated evolutionary divergence time (in Mys) between two species in biological networks or passed period of time (in years, and then coerced to Mys) in commonplace networks (see Table 2.1). Thus, the rewiring rate was measured as the number of rewired edges per edge per Mys. It can be interpreted as the averaged fraction of rewired edges among all possible edges in a period of one million years.

However, total number of possible edges was calculated differently among network types. Calculation for collaborative networks, including social networks, is simpler because their edges are reversible (see Figure 2.2):

$$\text{Collabortive network C} = \frac{\text{CNs} \times (\text{CNs} - 1) + \text{GNs} \times (\text{GNs} - 1) + \text{LNs} \times (\text{LNs} - 1)}{2} + \text{CNs} \times (\text{GNs} + \text{LNs})$$

Note that here we did not allow self interactions and only allowed one edge between two nodes. For metabolic networks that allow two reciprocal edges between two nodes (for directional reactions), we just multiplied the above calculated result by 2. For regulatory networks involving irreversible edges, we further separated nodes into regulators (Regs) and targets (Tars) and only allowed edges from Regs to Tars, but not from Tars back to Regs. In addition, regulators in transcription factor-target regulatory network and kinase-substrate phosphorylation network could

themselves be targets of other regulators, but not in miRNA-target regulatory network. Considering all these factors (see Figure 2.2),

$$\text{TF or Kinase network C} = \frac{\text{Reg CNs} \times (\text{Reg CNs -1}) + \text{Reg GNs} \times (\text{Reg GNs -1}) + \text{Reg LNs} \times (\text{Reg LNs -1})}{2} + \text{Reg CNs} \times \text{Tar CNs} + \text{Reg GNs} \times \text{Tar GNs} + \text{Reg LNs} \times \text{Tar LNs} + \text{Reg CNs} \times (\text{Tar GNs} + \text{Tar LNs}) + \text{Tar CNs} \times (\text{Reg GNs} + \text{Reg LNs})$$

and

$$\text{microRNA network C} = \text{Reg CNs} \times \text{Tar CNs} + \text{Reg GNs} \times \text{Tar GNs} + \text{Reg LNs} \times \text{Tar LNs} + \text{Reg CNs} \times (\text{Tar GNs} + \text{Tar LNs}) + \text{Tar CNs} \times (\text{Reg GNs} + \text{Reg LNs})$$

2.4.3 Calculating evolutionary rates in network and sequence comparisons

The rewiring rate calculation described above was not directly comparable to sequence evolution rate calculation, as there is no equivalent to the 'total number of possible edges' as in networks. Therefore, we used identity-based evolutionary rate measures instead to compare networks and sequences as:

$$\text{Sequence identity (\%)} = \frac{\text{Number of unchanged nucleotide/amino acid positions in the alignment}}{\text{Total length of sequence alignment}} \times 100\%$$

$$\text{Network identity (\%)} = \frac{\text{Number of common edges between orthologous nodes present in both networks}}{\text{Total number of edges in two networks}} \times 100\%$$

The evolutionary rate calculated based on identity was:

$$\text{Identity based evolutionary rate} = \frac{\text{1 - Identity(\%)}}{\text{Time divergence}}$$

2.4.4 Calculating rewiring rate difference for paralog pairs in protein interaction networks

Rewiring rates for all individual nodes were calculated for *H. sapiens* and *S. cerevisiae*

protein interaction networks by comparing them to *D. melanogaster* and *S. pombe* networks, respectively. Number of rewired edges for each node was counted as the number of gained or lost edges involving this node. This number was then divided by network size and by divergence time to get rewiring rate for a node. Network size is difference for CNs, GNs and LNs. For CNs, network size is the sum of the number of CNs, GNs and LNs from the two networks; for GNs, network size is the sum of CNs and GNs; and for LNs, network size is the sum of CNs and LNs. Rewiring rate difference was then calculated for all node pairs including all paralog pairs.

Chapter 3

Computational simulation of network rewiring

3.1 Introduction

Current biological networks built from large-scale experimental data are generally thought to have some extent of false positive and false negative interactions. False positive interactions stem from non-perfect-specificity experiments, and false negative interactions result from non-perfect-sensitivity experiments. For example, protein interaction networks from Y2H method have been suggested by many computational studies to have higher number of false-positives than literature-curated or mass spectrometry data sets [105, 106]. ChIP based methods for TF regulatory networks are also susceptible from false interactions due to incompetence of target-calling methods that process binding peak data.

Researchers have tried to build higher quality biological networks based on large-scale data

sets by using small-scale data sets as gold-standard positives [107]. However, there are two major problems of filtering high-quality interactions for all biological networks in all species. First, many biological networks lack experimental or computational methods to define gold-standard interactions. Current TF regulatory networks are defined as the TF-target interactions based on ChIP experiments. And there is no another gold-standard way to verify the "real" targets of a TF. Second, gold-standard positives are not always available for all species. The quality of large-scale experiments is usually assessed by small-scale gold-standard data sets in limited number of model organisms. Thus it is difficult to construct high-quality networks based on large-scale data sets in all species.

To complement the evolutionary rewiring rate analysis of biological networks presented in chapter 2, network quality issue has to be addressed that how our calculation is affected by the imperfectness of data sets. Interactions could be randomly removed from the current data sets as if they were false positives, and random interactions could be added to the data sets as if they were false negatives.

Besides coping with the data quality issue, simulated network rewiring also helps to model the evolutionary rewiring process. By rewiring the "ancestral" network, derived "current" networks could be obtained with certain rewiring steps. This is exactly an analogy of ancestral species and present species with certain time divergence. We can therefore simulate the evolutionary network rewiring and analyze the potential mechanisms underlying.

Since almost all biological networks are scale-free, we would like to maintain the scale-free characteristic of networks when rewired. Preferential attachment has been proposed in modeling

the growth of social and computer networks [108]. It is a growing process such that the probability of gaining a new interaction is proportional to the number of interactions one already has. Preferential attachment will generate power law degree distributions and scale-free networks.

Sensitivity analysis is often used to assess the relative importance of parameters to the output of models. There are two major types of sensitivity analysis: local and global analysis [109]. Local sensitivity analysis studies the behavior of the system response locally around a chosen point for static systems with small perturbations. Global sensitivity analysis determines all of the system's critical points, such as bifurcations, turning points, response maxima, minima, and/or saddle points, in the combined space formed by the parameters and output variables.

Many sensitivity analysis methods are proposed in the scientific and engineering research fields. Statistical methods explore the parameter space by random sampling, and are especially suited for cases with large parameter space [110].

3.2 Results

3.2.1 Assessing network data quality to rewiring rates

Unlike sequence data that one is essentially sure of every base, network data either generated from experiments or computational predictions are currently subject to high number of false positives and false negatives. Because many distinct experimental approaches are used to generate network data, different biological networks may have varied systematic bias during their construction. It is inevitable that our results presented in chapter 2 might be subject to change

when new network data become available.

For each type of biological networks, we used consistent data source and method to build networks for species, which ensures the uniform definition of edges and facilitates comparison between species.

Instead of trying to build high quality networks for all biological networks in multiple species, which is difficult due to lack of gold-standard positives and negatives, we applied a general method to assess the influence of false positives and negatives to rewiring rate calculation for all biological networks. Beltrao et al. have used a sampling-based sensitivity analysis to assess the robustness of rewiring rate relative to the amount of protein interaction data used [62]. Here, we applied a similar method to six representative types of biological networks used in this study. The effects of false negatives and false positives are simulated by random sampling. That is, we randomly add and remove a fraction of edges of the two compared real networks, forming simulated "corrected" networks, and then calculate rewiring rates. A series of disruption fractions of random edges are used to simulate false positive and negative rates from low to high (see Figure 3.1).

Rewiring rates of most of the biological networks are robust to network size change and disruption, especially when the disruption fraction is lower than 50%. However, the rates of metabolic pathway networks have shown clear deviations at large disruption levels. The observed one order of magnitude difference between metabolic pathway networks and protein interaction networks (10^{-5} for protein interaction network, 10^{-6} for metabolic pathway) disappears at approximately 70% disruption level. We conclude from these results that the network rewiring

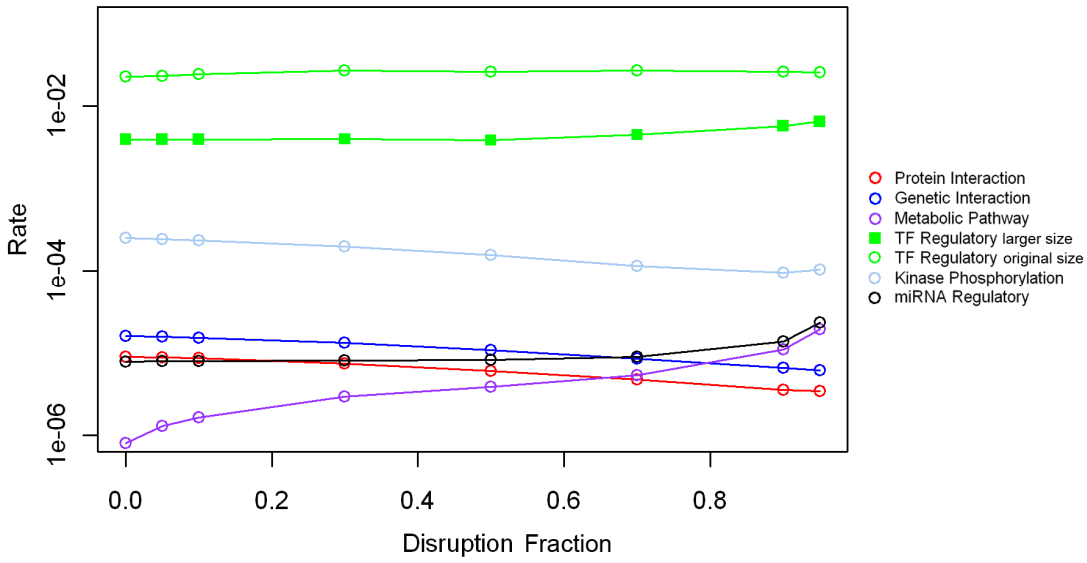

Figure 3.1: Sensitivity analysis of false positive and false negative rates to rewiring rate.
We sampled biological networks in order to test the false positive and false negative rates to rewiring rate calculation. Six biological networks are included here: protein interaction network, genetic interaction network, miRNA-target regulatory network, kinase-substrate phosphorylation network, metabolic pathway network (*S. cerevisiae* compared to *S. pombe*) and transcription factor target regulatory network (*S. cerevisiae* compared to *S. bayanus*). For each type of network, we randomly delete and add edges from the original network as a simulation of false positives and false negatives, with each a series of percentage disruptions. For transcription factor target regulatory network, we also tested rewiring rate sensitivity to network size by using a larger original network for *S. cerevisiae*.

rate is only slightly affected by network size, and is especially robust at sampling levels above 50%. The results of this study should still hold when new network data arrives.

We also investigated the potential size effect of fungi TF-target regulatory networks used in our study. These networks were constructed using binding sites from ChIP-chip experiments of one or two TFs, which results in relatively small networks. Besides the simulated disruption described previously on these small networks, edges were added to the *S. cerevisiae* network from another ChIP-chip study between the existence nodes to generate a larger network [111]. The same disruption analysis was performed on the larger network. Rewiring rates calculated from the larger network decreased about half order of magnitude than from the original small network (see Figure 3.1). This is largely due to the increase of total possible edge changes in our calculation. As a result, the current subnetwork of TF-target regulatory network might lead to a bias of faster rewiring rate.

A comprehensive simulation analysis was further performed to assess the effects of both network size and network quality. Two simulated scale-free networks were constructed with some common edges, and sub-samples of both networks were taken for comparison. Random rewiring of both sub-network were performed to mimic false positives and negatives. Percentage of edge change (R/C) was calculated for each sub-sampling fraction. As the size of the compared sub-networks decreases, percentage of rewiring increases (see Table 3.1). The upward bias of percentage of rewiring is approximately one order of magnitude corresponding to 1% sub-sampling fraction. Because the fungi TF regulatory network used in this study is approximately 20-100 times smaller than the complete networks estimated by the number of

Sub-sampling fraction	Rewiring percentage	Rewired edges	Shared edges	Total possible edge changes	Shared nodes	Unique node in network A	Unique node in network B
100%	4.3e-4±7e-6	11660±130	6054±156	2.7e7±3e5	6867±24	505±52	1±0
95%	4.5e-4±6e-6	11690±117	5576±146	2.6e7±3e5	6494±38	594±66	101±7
90%	4.8e-4±5e-6	11596±134	5129±112	2.4e7±2e5	6126±23	651±39	190±15
70%	5.8e-4±7e-6	10911±136	3377±35	1.9e7±9e4	4898±34	742±20	566±21
50%	6.2e-4±7e-6	9577±130	1875±62	1.6e7±7e4	3855±25	1013±22	864±15
30%	6.8e-4±6e-6	7187±64	767±19	1.0e7±9e4	2589±24	1227±27	1017±10
10%	9.4e-4±1e-5	3136±48	95±5	3.3e6±3e4	1030±15	1008±18	857±16
5%	1.3e-3±2e-5	1801±14	29±2	1.4e6±3e4	545±16	756±27	642±11
3%	1.7e-3±4e-5	1167±29	12±2	6.7e5±3e4	306±15	580±27	497±14
1%	3.9e-3±8e-5	431±21	2±0.5	1.1e5±7e3	75±4	290±14	236±11

For each sub-sampling fraction, we performed 10 simulations and calculated 95% confidence intervals for resulting numbers.

Table 3.1: Simulation of network size, false positives, and false negatives to rewiring rate
Based on two simulated scale-free networks, sub-networks are sampled to mimic the fact that data of many biological networks used in this study are not complete, such as the fungi TF regulatory networks. Extra random rewiring by adding and removing edges and nodes is performed to mimic the false positives and negatives in the current network data. Percentage of network rewiring is then calculated to assess the effects of those perturbations.

edges and the number of TFs [111]. We thus estimated that the true rate of fungi TF regulatory network could be half to one order of magnitude slower than we calculated. Considering the above estimation of network size effect on rewiring measurement, fungi TF regulatory network

should still rewire faster than or in a similar pace as kinase phosphorylation network, and much faster than other types of biological networks (see Table 2.3).

miRNA regulatory networks were constructed using a consistent miRNA target prediction method [95]. In the current stage of miRNA research, most miRNAs are found or predicted using sequence conservation, and regulatory relationship is predicted mainly by searching for complementary sequence in 3' UTRs [43-45]. Therefore, the turnover of miRNAs is small with lack of species-specific miRNAs and their corresponding targets. For example, a total of 459 conserved miRNAs are present in the networks comparing human and mouse. However, only 18 and 9 miRNAs are human-specific and mouse-specific, respectively. The mere gene content turnover of only 6% for miRNAs is much less than 67% and 74% for TFs and kinases (see Table 2.8). This ascertainment bias could result in under-estimation of rewiring rates.

To estimate the effect of novel miRNAs to our rewiring measurements, we randomly added a series numbers of hypothetical novel miRNAs to actual human and mouse miRNA regulatory networks. The targets of those hypothetical miRNAs are also randomly selected with degree distribution maintained. Rewiring rates calculated from these simulations showed that discovering potential species-specific miRNAs could result in an increase of rewiring rate (see Table 3.2). With the advance of miRNA research from novel miRNA discovery to better target prediction methods, it is possible that the current rewiring rates of miRNA regulatory networks will be adjusted higher.

miRs added	Rewiring rate	Rewired edges	Shared edges	Total possible edge changes	Shared miRs	Unique miRs in network A	Unique miRs in network B
0	3.3e-5	7132	2138	2.9e6	459	18	9
50	6.9e-5±1e-5	16800±3174	2138	3.2e6	459	68	59
100	1.7e-4±2e-5	44640±6281	2138	3.6e6	459	118	109
200	3.1e-4±4e-5	102110±12889	2138	4.4e6	459	218	209
400	4.9e-4±6e-5	215559±25995	2138	5.9e6	459	418	409
600	7.2e-4±7e-5	396853±38672	2138	7.4e6	459	618	609
800	9.7e-4±8e-5	647465±50143	2138	8.9e6	459	818	809

For each number of added miRNAs, we performed 10 simulations by assigning different targets to the added miRs, and calculated 95% confidence intervals for resulting numbers.

Table 3.2: Simulation analysis of the effect of novel miRNAs to miRNA regulatory network
Based on current miRNA regulatory networks for human and mouse, simulated novel miRNAs are added to both networks with their target randomly sampled, while maintaining the power-law distribution of target number distribution. Statistics are calculated comparing the simulated networks.

3.2.2 Simulation of network rewiring model

In section 2.2.2, we observed linear relationship between rewiring rate or nucleotide substitution rate and divergence time on Log-Log scale. Here, we used simulated networks to determine whether the observed relationship is specific to real biological networks. A simulation-based network rewiring model was built based on four parameters, corresponding to node changes, edge changes, and preferential attachment to rewiring networks while maintaining scale-free topology. As a simulation of evolutionary divergence, two branches of networks were compared after the same number of rewiring steps and rewiring rates calculated, as the red

double-ended arrows shown in Figure 3.2. The rewiring rate calculated from the simulation model also shows a negative linear relationship in Log-Log plot with number of rewiring steps (see Figure 3.3). This is consistent with the results from real biological networks and molecular sequence models, indicating a universal saturation effect of biological system evolution.

We also investigated whether the rewiring rates calculated from comparing two offspring networks, or comparing ancestral and offspring networks are different with a factor of two. As the green double-ended arrows shown in Figure 3.2, rewiring rates could also be calculated comparing the ancestral network with its offspring networks. The divergence in two methods is the same as the number of steps from the ancestral network. Because the distance between two offspring networks is two times the distance between the ancestral network and one of its offspring network, we suspect that the rewiring rates calculated from two comparisons may also reflect the factor of two. However, our calculation shows that the rewiring rates calculated from the two methods are largely the same, without a factor of two (see Figure 3.4). This suggests that our rewiring rate measure inherently deals with the issue that renders consistent results comparing networks with the same extent of divergence.

3.2.3 Sensitivity analysis of network rewiring model

For all types of biological networks and simulated networks we observe a negative linear relationship between rewiring rate and divergence time (see Figure 2.3 and 3.3). Generally speaking, the average rewiring rate calculated comparing distant species networks tends to be

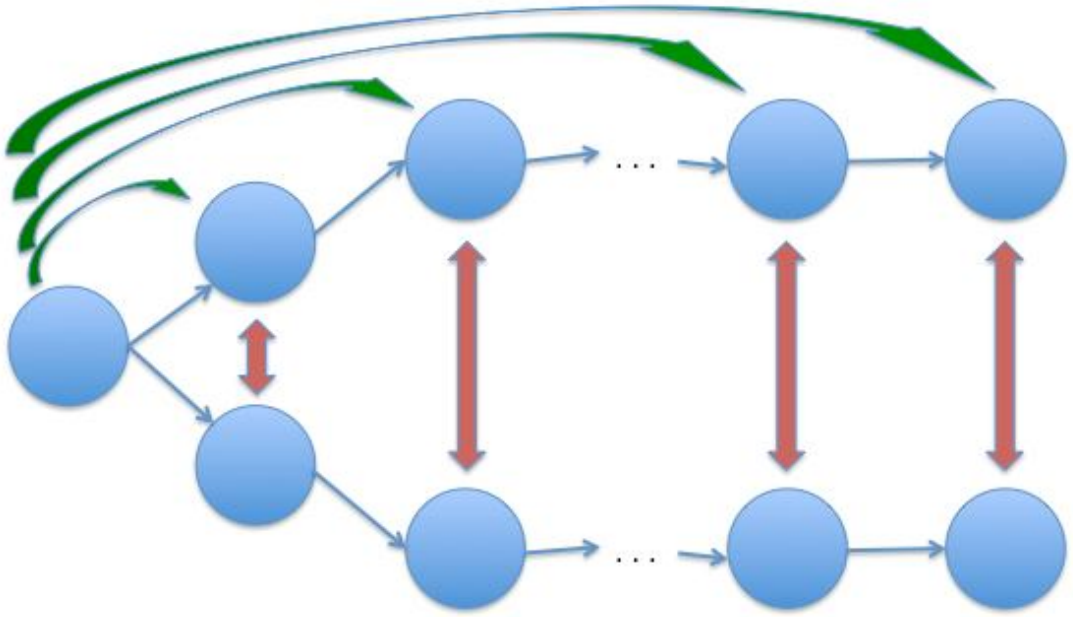

Figure 3.2: Simulation of network rewiring and rewiring rate calculation.
Simulation of network rewiring started from a seed network, and had two independent branches of simulation. Networks are schematically shown in blue circles. Each branch had 1000 rewiring steps, and snapshots of rewired networks were taken every 50 steps. For each rewiring step, the starting network was rewired to generate the next network according to the same parameter set. Rewiring rate was calculated comparing two independently rewired networks from two branches with the same number of steps, e.g. 50, 100, 150, 200 and all the way to 1000, showing in red double-ended arrows. The other way of calculating rewiring rate after certain number of rewired steps is comparing the resulting network with the original seed network, showing in green double-ended arrows.

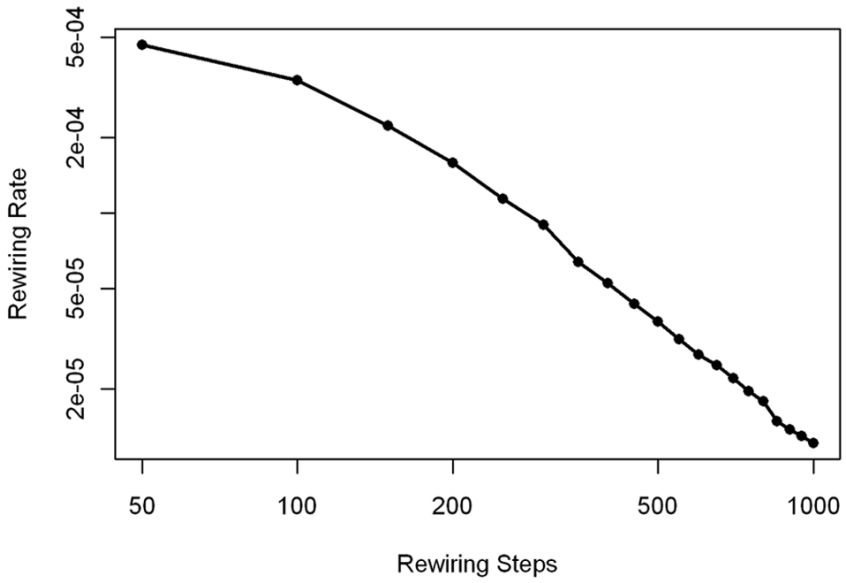

Figure 3.3: Linear relationship between rewiring rate and rewired steps on Log-Log scale. Rewiring rate is calculated comparing two independently generated offspring networks based on the same ancestral network, as the red double-ended arrows in Figure 3.2.

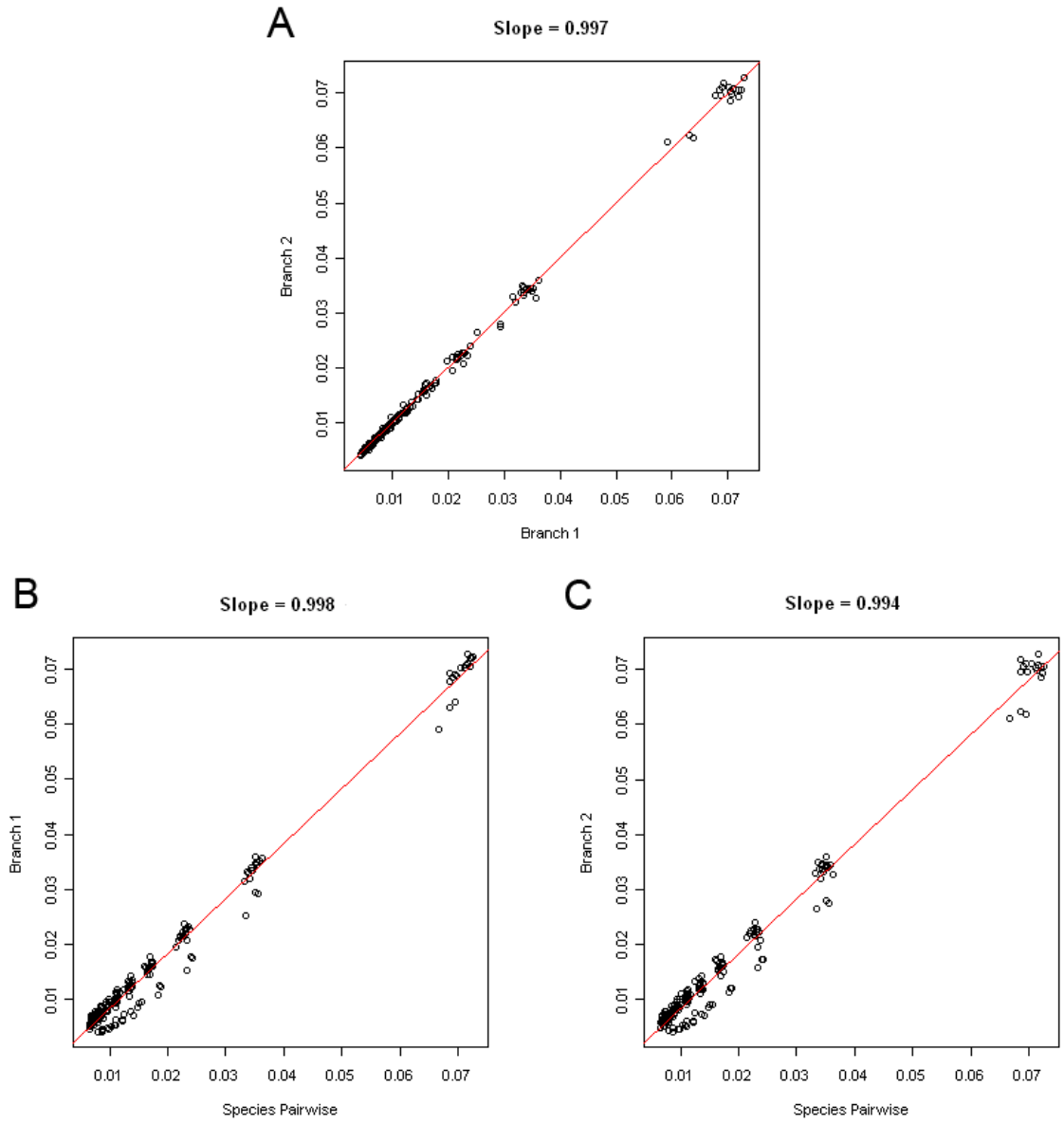

Figure 3.4: Consistent rewiring rates from two different comparisons.
Branch 1 and 2 are the results from comparing ancestral network and offspring networks from the upper and lower branches, respectively, as the green double-ended arrows shown in Figure 3.2. Species pairwise results are calculated from comparing two offspring networks with the same number of rewired steps from the ancestral network, as the red double-ended arrows shown in Figure 3.2. The consistency between the two methods is observed from simulated results.

smaller than the instantaneous rate comparing close species networks. For networks from two distant species, overlap of their nodes becomes smaller due to loss of conservation. As a result, the total number of possible edges C increases and rewiring rate decreases correspondingly. In conclusion, a larger difference between node sets of two distant species networks might be the main reason for this bias.

The major effect of node gain and loss on rewiring rate was further confirmed by a sensitivity analysis based on network rewiring simulation model. Global sensitivity analysis is applied to explore the entire parameter space. Simple uniform random sampling is used without stratification. Each of four independent parameters in our model was tested for its relative importance to model output—rewiring rate. We do not observe critical turning points in the parameter space (see Figure 3.5).

Not surprisingly, we found that some parameters are more significant to the model than others. Removal of node has the strongest effect (negative linear) on rewiring rate, because rewired edges associated with a node are removed along with the node, which decreases the total number of rewired edges. Adding node also has some effect (positive linear) on rewiring rate, because of the increased number of total rewired edges associated with the node. Nevertheless, removing and adding edges have only small effects on rewiring rate (see Figure 3.5). It is reasonable that removing and adding nodes has a major influence on rewiring rate as it affects *all* edges associated with nodes rather than individual edges.

It is also possible that there are "cores" for each type of networks that slow down the rewiring process when it approaches the cores. The cores are partial networks that are the most constrained

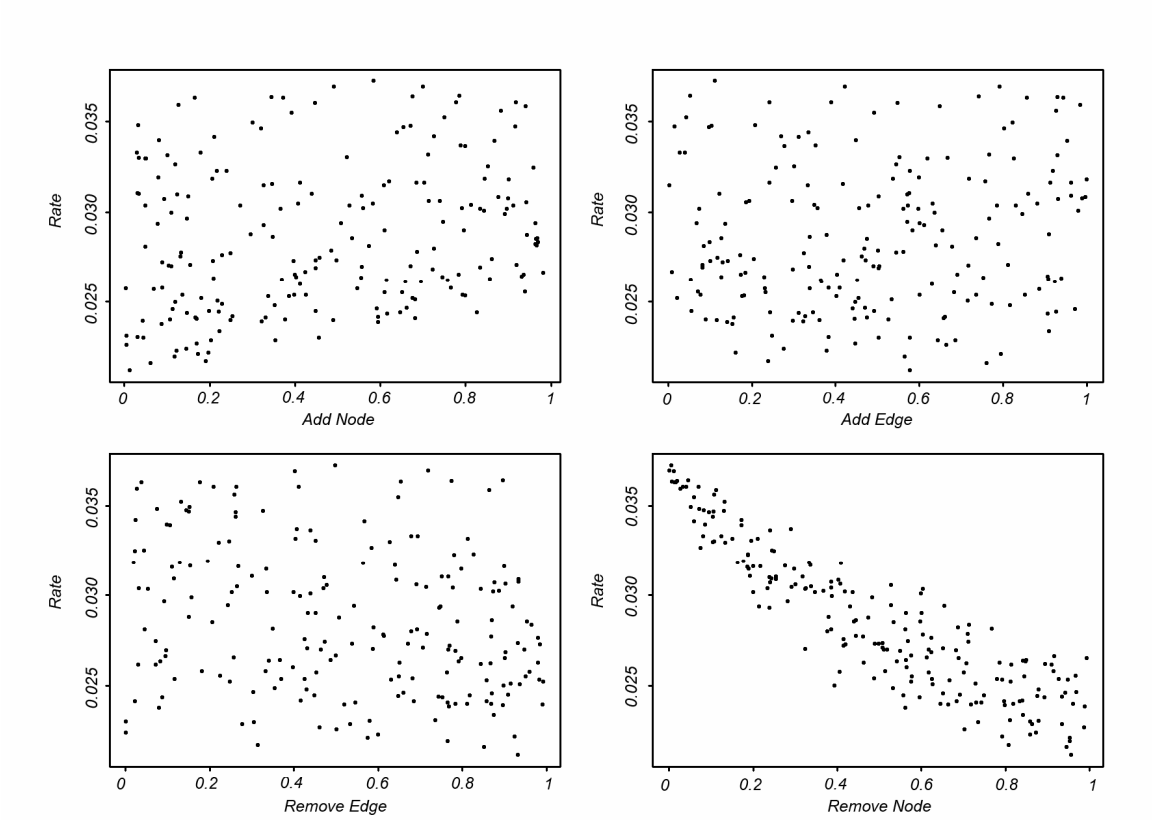

Figure 3.5: Sensitivity analysis of four network rewiring parameters to rewiring rate.
Four parameters in our rewiring simulation model - probabilities of adding a node, removing a node, adding an edge and removing an edge, are analyzed for their importance to calculated network rewiring rate. Removing node probability has the greatest negative effect on rewiring rate calculation.

and conserved during evolution, possibly reflecting their functional importance. Therefore, network types with a smaller ratio of rewiring rate changes and divergence time (flat lines) might have larger cores, because of greater resistance to rewire the cores; while network types with a larger ratio (steep lines) might have smaller cores (see Figure 2.3).

3.3 Discussion

Many different mathematical and statistical models are developed to describe the growth principles of scale-free networks, such as the Exponential Random Graph Model (ERGM) [112]. However, our work is not intended to develop a new model or to improve existing models. Rather, it is a simple complementary study to our rewiring analysis of biological networks.

The disruption analysis by random simulation of network edges confirms the stability of network rewiring measure to certain levels of data noise. It seems surprising that for some networks, rewiring rate measure is only marginally changed under extremely large disruptions levels (see Figure 3.1). The underlying mechanism of the stability of rewiring rate measure is not clear. More detailed and well designed analyses are needed in the future to address this issue. It is possible that the stability is originated from the general backbone structure of scale-free networks. The topological characteristic of biological networks may largely stay unchanged even under extreme disruptions.

Our model of network rewiring is rather simple compared to other existing network growth models. However, the major contribution of this study is its application to biological networks,

which considers all four sources of edge rewiring. Previous studies have focused on the expansion of social networks, or the fitness of models to real networks. Our study tries to model the dynamic rewiring process and reveals the relative importance of rewiring sources in the model. Further improvement of our model may include the incorporation of more complex models and relating the model with network rewiring measure formulism.

3.4 Methods

3.4.1 Simulation of network size, false positive and false negative rates

Two simulated scale-free networks were built with some common edges for comparison. To simulate the size effect of rewiring rate calculation, the pair of networks was sub-sampled of their edges to a series of fractions, from 95% to 1%. To assess the amount of false positives and false negatives in network data to rewiring rate calculation, we further perturbed the compared network pair (either real biological networks or simulated networks) by randomly adding and removing edges on both networks. Edges were added using preferential attachment, since probability of having false negative edges is proportional to the degree of the node. Edges were deleted randomly from existing edges. Nodes were added to the network, and only one edge was added to it. Nodes were removed from the network randomly, with all its edges removed with it. A series of perturbation percentages were used to simulate levels of false positive and negative rates.

3.4.2 Simulation model of network rewiring

The model had four parameters: probabilities of adding a node (adding one edge with that node using preferentially attachment), removing a node (randomly for all existing nodes and all edges with that node), adding an edge (using preferentially attachment) and removing an edge (randomly for all existing edges). Preferential attachment mechanism maintains the scale-free topology of networks. To begin with, a small scale-free network was used as a seed to the model. For each rewiring step, nodes and edges were added and removed according to the probability parameters, and the resulting network was recorded for the next step.

For the relationship analysis of rewiring rate and rewiring steps, two independent rewiring branches were simulated with each 1000 steps (see Figure 3.2). The networks from the two branches were compared after every 50 steps and rewiring rate was calculated.

For parameter sensitivity analysis, 200 parameter-set samples were generated, with the four probability parameters randomly generated from a uniform distribution on the interval [0,1]. The same seed network was used for all 200 simulations using the 200 random parameter-sets. All simulations were stopped after 100 steps and rewiring rate was calculated corresponding to each of the 200 parameter-sets.

Chapter 4

Genome-wide analysis of histone modification profiles and TF target gene prediction in yeast

4.1 Introduction

Transcription factors (TFs) regulate target gene expression through binding to specific genomic regions. In *Saccharomyces cerevisiae*, transcription factor binding sites (TFBSs) are often adjacent to and upstream of target loci due to the compact nature of the yeast genome [113, 114]. Upon binding, TFs interact with RNA polymerase II to activate or repress transcription. TFs also recruit chromatin modification enzymes to induce chromatin structure changes, which in turn affect the accessibility of factors to genomic DNA regions [115, 116]. The target genes of a TF change according to developmental, physiological and extra-cellular environmental conditions [111]. In addition, TFs interact with each other through combinatorial binding [117]. Uncovering

TF target genes and inter-relationships between TFs for all different conditions is thus important for understanding gene expression regulation, but it is also a difficult task due to the scale of the problem.

Several different experimental methods have been developed to identify TFBSs. Chromatin immuno-precipitation followed by tiling array (ChIP-chip) has been widely used to identify TFBSs in the genomic scale [118-120]. More recently, high-throughput sequencing after chromatin immuno-precipitation (ChIP-Seq) has been shown promising in getting higher resolution of the TFBSs [121, 122]. With these methods, an increasing amount of TFBS data have been accumulated for different TFs in different species, cell types, conditions, and so on, which have started to unravel the global picture of gene expression regulation. In yeast, the TFBSs and target genes for an almost complete set of TFs have been mapped in common YPD medium using ChIP-chip [111]. However, resources are still too limited to support a complete exploration of TF binding for all the combinations of cell types and conditions.

Many computational methods have also been proposed to predict TFBSs [123-129]. These methods are mostly based on the idea that the binding of a TF is mediated by the recognition of its binding motif represented as a position specific scoring matrix (PSSM). PSSMs are usually discovered as those enriched motifs from TFBSs in ChIP-chip or ChIP-seq experiments, or *de novo* from non-coding genomic sequences [111, 130]. Scanning and matching PSSMs in the genome constitute the core of these methods, which are then improved by incorporating information of motif conservation and TFBSs co-localization. Nevertheless, these methods often lead to considerably high rate of false positives. Furthermore, most of these methods are not

condition specific and thus could not reflect the dynamic nature of TF binding under different conditions.

Chromatin modifications could modulate the accessibility of DNA regions and affect the recruitment of TFs [115, 116, 131]. Both functions directly relate to transcription regulation by TFs. Genomic mapping of chromatin modifications in yeast using ChIP-chip has provided the opportunity to investigate their underlying relationships with TFBSs [132, 133]. Many chromatin modifications have been shown to be associated with transcription activation and repression [115, 116]. Recent studies have shown that incorporating histone modification data improves prediction of TFBS in mouse and human [134, 135]. In these models, chromatin modifications generally provide non-TF-specific chromatin accessibility, while PSSMs determine TF-specific bindings.

Here we propose a new method that integrates PSSMs and chromatin modifications to improve TF target gene predictions in yeast. Specifically, we trained individual support vector machine (SVM) models [136] for 203 yeast TFs using 2 types of features: the existence of PSSM upstream of genes and chromatin modifications adjacent to the ATG start codons. The models were trained and tested using TF target genes from ChIP-chip experiments. Furthermore, we used the model to investigate condition specificity and TF-specificity of chromatin modifications as well as TF-TF co-operation. Our analysis helps understand the mechanism of gene expression regulation by TFs and chromatin modifications.

4.2 Results

4.2.1 Differential histone modifications between functional and non-functional TFBSs

In order to examine whether chromatin modifications are predictive features for functional TF binding sites, we first investigated chromatin modification signals at functional and non-functional TFBSs defined as follows. Based on previous TFBS prediction models, we denoted the TFBSs of a factor as the local genomic sequences that match its PSSM. We then used ChIP-chip experimental data to distinguish functional and non-functional TFBSs based on the existence of actual binding peak signals. Although both functional and non-functional TFBSs contain TF binding PSSMs, they were found to have differential chromatin modification signals. Here, we use the factor Swi4, a component of the SBF complex regulating cell cycle gene expressions, as an example (see Figure 4.1). We observed that individual histone modifications varied significantly between functional and non-functional TFBSs. Among the 14 different histone modifications under 2 conditions (YPD and H_2O_2), 11 were significantly different (p-value<0.01) in their signals between the functional and non-functional TFBSs of SWI4 (see Figure 4.1). Among them, H3 and H4 signatures were particularly strong features for distinguishing between the two types of TFBSs, as they showed significantly lower signal in functional TFBSs than in non-functional TFBSs (p-value<10^{-20}). Consistent with previous studies, this indicates that functional binding sites of factors in regulatory regions are typically depleted of

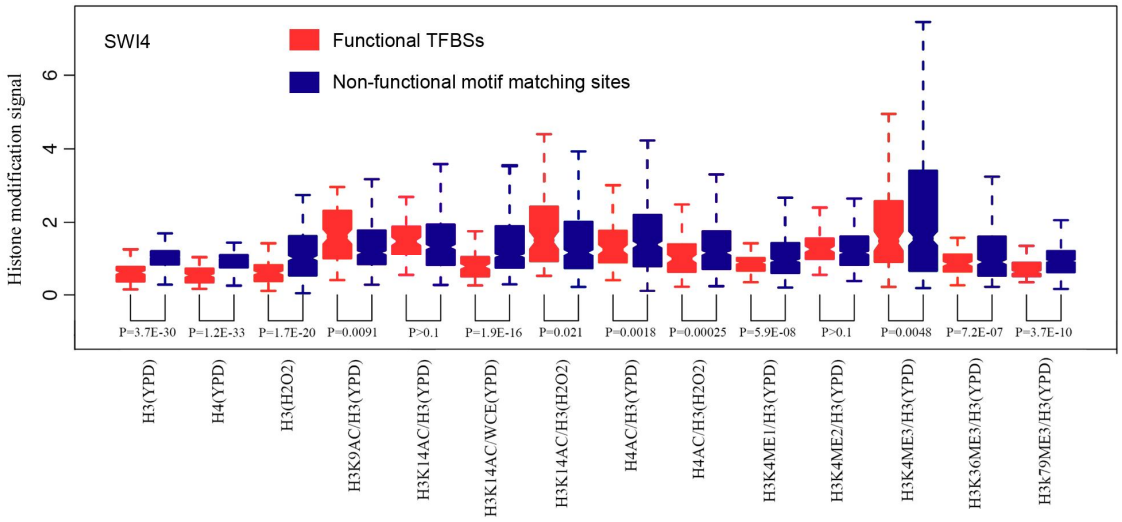

Figure 4.1: Differential histone occupation and modifications between functional TFBSs and non-functional motif matching sites.

Showing SWI4 as an example, most histone modifications (in different colors) are significantly different between functional TFBSs (left boxes), which have binding motifs and are bound by TFs in ChIP-chip experiments, and non-functional motif matching sites (right boxes), which have matching motifs but are not bound by TFs.

nucleosomes [137-140]. Encouraged by the observed differential histone modifications between functional and non-functional TFBSs, we then constructed a model that combines histone features with binding motif information for target gene prediction in yeast.

4.2.2 Improving target gene prediction by combining histone modifications and PSSMs

Since the *S. cerevisiae* genome is quite compact with respect to other higher eukaryotic species, it is reasonable to define the target genes of a TF as those with one or more upstream TFBSs. We combined chromatin modification and PSSM data and used them as input features to a SVM learning model for predicting TF target genes. The prediction accuracy of the model was tested using a gold standard dataset from ChIP-chip experiments, which provided target genes of 203 yeast TFs [111]. Specifically, we choose 0.01 as the P-value cutoff for target gene calling from ChIP-chip, which provides us with enough number of high confidence positive target genes for model training (see Figure 4.3). The data set was separated into training and testing data, and the performance of the model was assessed by cross-validation.

For chromatin modifications, we used 11 histone modifications that covered most yeast ORF regions from ChIP-chip experiments [133]. Since TFs bind to the upstream sequence of ORFs, we focused on histone modification signals 1kb flanking translation start sites (ATGs), because TFBSs were enriched in these regions. For TF PSSMs, 2 independent sets were obtained from previous studies. One set of PSSMs were discovered using sequence analysis based method,

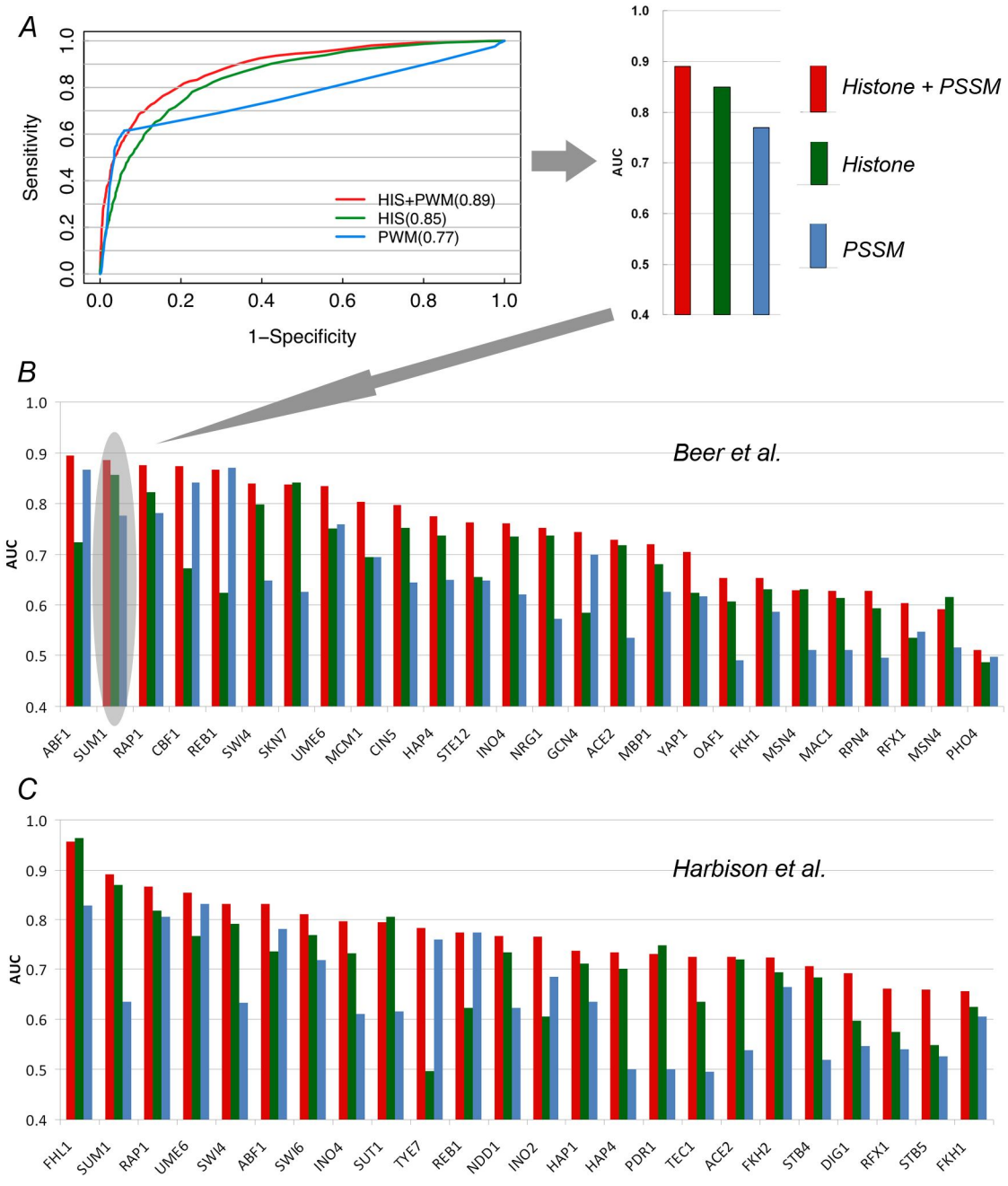

Figure 4.2: Chromatin modifications substantially improve TF target gene predictions.

(A) ROC curves show improved TF target gene predictions using histone modifications. (B) Performance of prediction models for individual TFs, with PSSMs from Beer et al and (C) from Harbison et al. TFs are sorted by prediction performance using histone modifications and PSSMs (red bars).

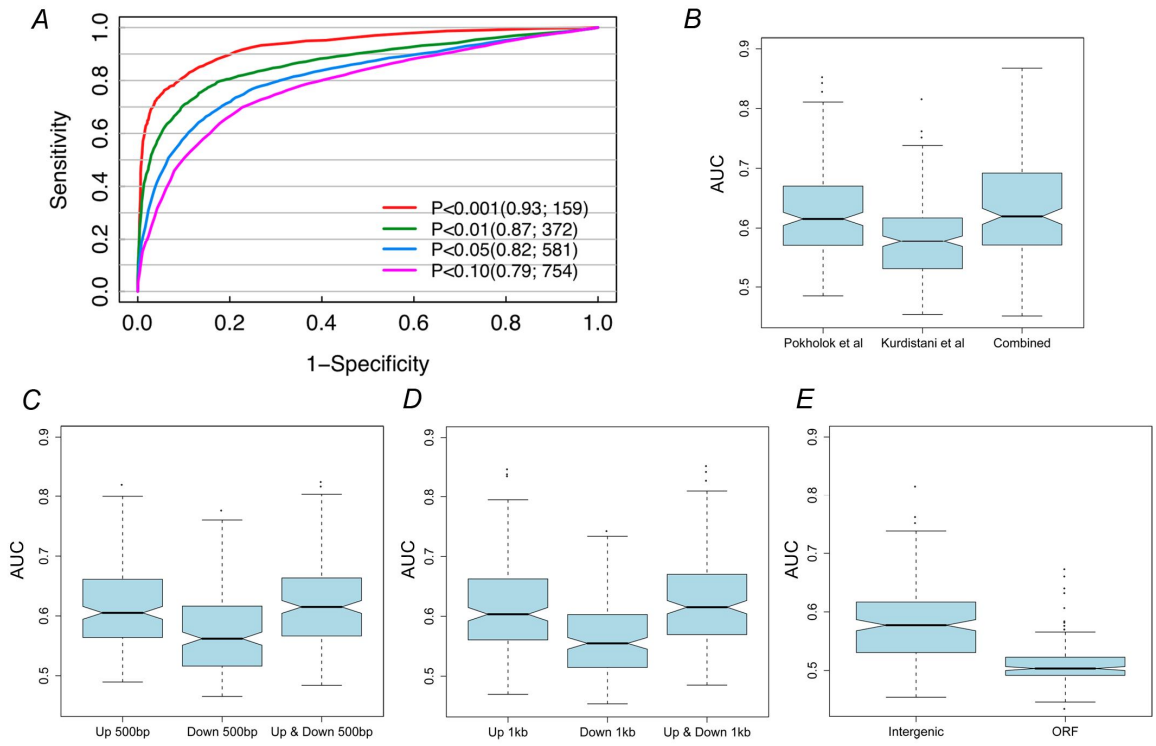

Figure 4.3: Model parameters

(A) Stricter thresholds of target gene call result in better predictions. (B) Combine independent histone modification datasets can improve target predictions. Predictions for 203 TFs are evaluated by AUC. (C) Using histone modifications within 500bp and (D) 1000bp window upstream and downstream of ATG sites of target genes achieve similar performance to using only upstream signals, and better performance to using only downstream signals. No significant performance difference between using 500bp and 1000bp window sizes. (E) Using histone modifications in intergenic regions has better predictive power than that in ORF coding regions.

basically looking for enriched motifs in the DNA regions upstream of all yeast ORFs [130]. From ~5,600 upstream sequences a total of 666 motifs have been discovered, among which 48 could be associated with known transcription factors according to literature. The other set of PSSMs were based on ChIP-chip data [111]. For each TF a target gene set was determined and then the binding motif for these TF was identified from the DNA region upstream of these genes.

Our results indicate that for almost all TFs a combination of histone modification and PSSM data had better performance (measured in AUC, area under the receiver operator characteristic curve) than using histone modification data or PSSM data alone (see Figure 4.2). For example, we obtained an AUC of 0.89 for predicting target genes of the factor SUM1 when both the histone modification and PSSM information were used. However, if only PSSM information was used, the model resulted in a much lower AUC (0.77) (see Figure 4.2). The improved performance of the combined model was observed for both of the TF PSSM sets (see Figure 4.2), indicating that the improvement does not rely on particular source or quality of PSSMs. Interestingly, for some TFs, histone-only model performed better than PSSM-only model; while for some other TFs, the opposite was observed (see Figure 4.2).

In order to examine whether we could achieve better TF target prediction when we have more chromatin modification data, we used histone modification datasets from two independent experiments, performed by Pokholok *et al* [133] and Kurdistani *et al*, respectively [132]. We found that we achieved higher prediction accuracy by using dataset from Pokholok *et al* than from Kurdistani *et al*. This is probably due to the fact that the latter contains only histone acetylation data, while the former contains both histone methylation and acetylation data, which

TF	AUC (Pokholok et al)	AUC (Pokholok et al & Kurdistani et al)
CST6	0.70	0.80
IFH1	0.53	0.64
KSS1	0.53	0.65
RDS1	0.50	0.67
SFP1	0.59	0.74
SWI5	0.62	0.74
TYE7	0.53	0.64
YKL222C	0.57	0.69
YKR064W	0.49	0.59

Table 4.1: TFs with improved prediction by including multiple histone modification datasets.

might provide complementary information for regulating chromatin structure and recruiting TF binding. We then combined the two datasets for predicting TF target genes, and found that for most TFs the performance was only slightly better than using the Pokholok dataset alone (see Figure 4.3). Nevertheless, for some TFs we observed substantial improvement by including the Kurdistani dataset (see Table 4.1). It is thus promising that we could improve our chromatin model performance by incorporating more histone modification data in the future.

We have also investigated the positional effect of histone modification signals for target gene prediction. First, we observed that signals of different types of histone modifications showed different patterns at DNA regions around the ATG, suggesting that they might affect TF binding in different manners. Second, histone modification signals from the upstream of ATG are generally more predictive than those from the downstream, as we have observed for both 500bp and 1000bp flanking region sizes (see Figure 4.3). This is somewhat expected, because TF binding

sites are more enriched in the upstream regions of ORFs for transcriptional regulation. It is also interesting to see that ATG flanking regions of 500bp and 1000bp result in almost the same performance (see Figure 4.3). Given the compact nature of the yeast genome, transcription start sites for most ORFs are located within 1000bp region upstream of ATG [8]. The comparable performance by using 500bp flanking ATG region indicates that most discriminative histone modification signals for TF binding were embedded in this region.

4.2.3 Condition specificity of the chromatin model

TFs bind to and regulate target gene expression in a complex and dynamic manner to coordinate biological processes [111, 141]. Chromatin modifications also change rapidly in response to stimulus from extra-cellular environment [115]. Therefore, chromatin modifications in one condition should match TF target binding in that condition but not other conditions.

We investigated condition specificity of our chromatin model in two conditions, YPD and H_2O_2. We tested a total of 12 TFs, for which we had the PSSM, histone modification and TF target binding data under both YPD and H_2O_2 conditions. For each of the TFs, we constructed two separate chromatin models: one model (Model A) used PSSM and histone modifications under YPD condition as features, while the other (Model B) used PSSM from YPD condition but histone modifications under H_2O_2 condition. The two models were then used to predict TF target binding under H_2O_2 conditions. It is generally believed that TFs keep their binding specificity PSSMs in different conditions and even over large evolutionary distances. Therefore, we use

PSSMs from YPD condition as a close approximation in Model B, where no PSSM information is available under H_2O_2 condition.

For TFs that are known to be functional under H_2O_2 condition, Model B achieved better performance than Model A. For example, HSF1, a heat shock TF that activates genes in response to stresses, is more active under H_2O_2 condition with 326 target genes, than in YPD medium with only 123 target genes. Using condition-matched histone modification data (Model B), our chromatin model achieved an AUC of 0.77. In contrast, using non-condition-matched data (Model A), the chromatin model only achieved 0.56 AUC (see Figure 4.4). Similar results were observed for another TF, MSN2, which is activated along with MSN4 to regulate stress response genes. MSN2 is more active under H_2O_2 condition, and our chromatin model performed better with condition-matched data (see Figure 4.4). These results indicate that histone modifications are actually dynamic and function in a condition specific manner. Thus, target genes of TFs under a certain condition can be best predicted by using histone modification data from the same condition. In practice, this enables us to predict conditional specific target genes of TFs, which cannot be achieved by using PSSM based method.

4.2.4 Relative importance of different histone modifications for target prediction

TFs bind to the upstream of target genes through recognizing their specific binding motif PSSMs. We then asked an analogous question: Do TFs have specific histone modification profiles at binding sites of their targets? To address this question, we calculated a histone modification

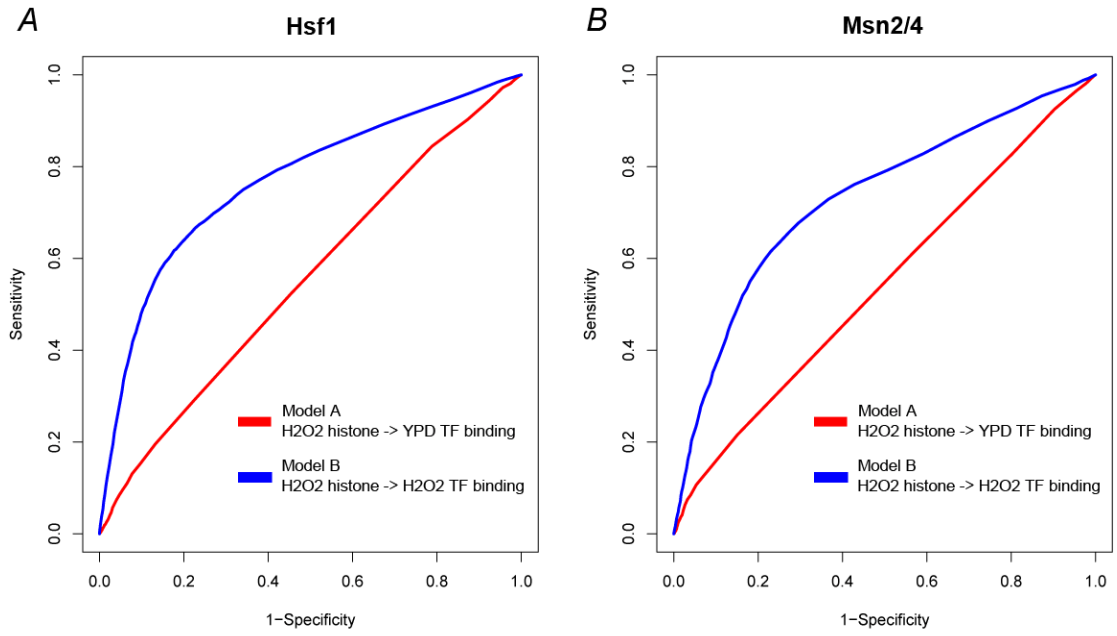

Figure 4.4: Conditional specificity of chromatin model for TF target prediction
(A) ROC curves showing performance of two chromatin models to predict target genes of HSF1 and (B) MSN2/4 complex. The model (Model B) using histone modifications in H_2O_2 condition performs better to predict target genes also in H_2O_2 condition (blue curve) than to predict target genes in YPD medium (Model A) (red curve), which indicates conditional specificity of chromatin modifications and TF target genes.

profile for each TF by averaging upstream-ATG histone modification signals over all its target genes. In our analysis, we included 25 different histone modifications from the two studies mentioned above [132, 133].

We found that different TFs have distinct target histone modification profiles. A histone modification high in one TF's profile could be low in another TF's profile (see Figure 4.5). We performed unsupervised clustering analysis for the histone modification profiles of all TFs, and detected two TF clusters (see Figure 4.5). One of the two clusters showed generally larger variations (more high and low signals) among histone modifications in the upstream of their target genes, while the signals in the other cluster are more often around the mean,($P<10^{-16}$, t-test). We thus refer the 68 TFs in the former cluster as "histone sensitive" TFs, and the 135 TFs in the latter cluster as "histone insensitive" TFs.

The correlations between pairs of histone modifications are shown in Figure 4.5C, based on their signals in histone modification profile over all TFs. Only pairs with strong correlation ($r >0.5$ or <-0.5) were connected in the form of correlation network. The dense connectivity in this network reveals strong pairwise redundancy of histone modification signals, which is also indicative of redundancy for predicting target genes.

We next examined the relative importance of each histone modification for predicting target genes of all TFs. Given a histone modification, we compared its signal difference between target and non-target genes of a TF. The signal difference was represented as the t-statistics (see "Method" section for detail), which indicated the relative importance of that histone modification for predicting target genes of a TF. A larger absolute value of t-statistic indicated more

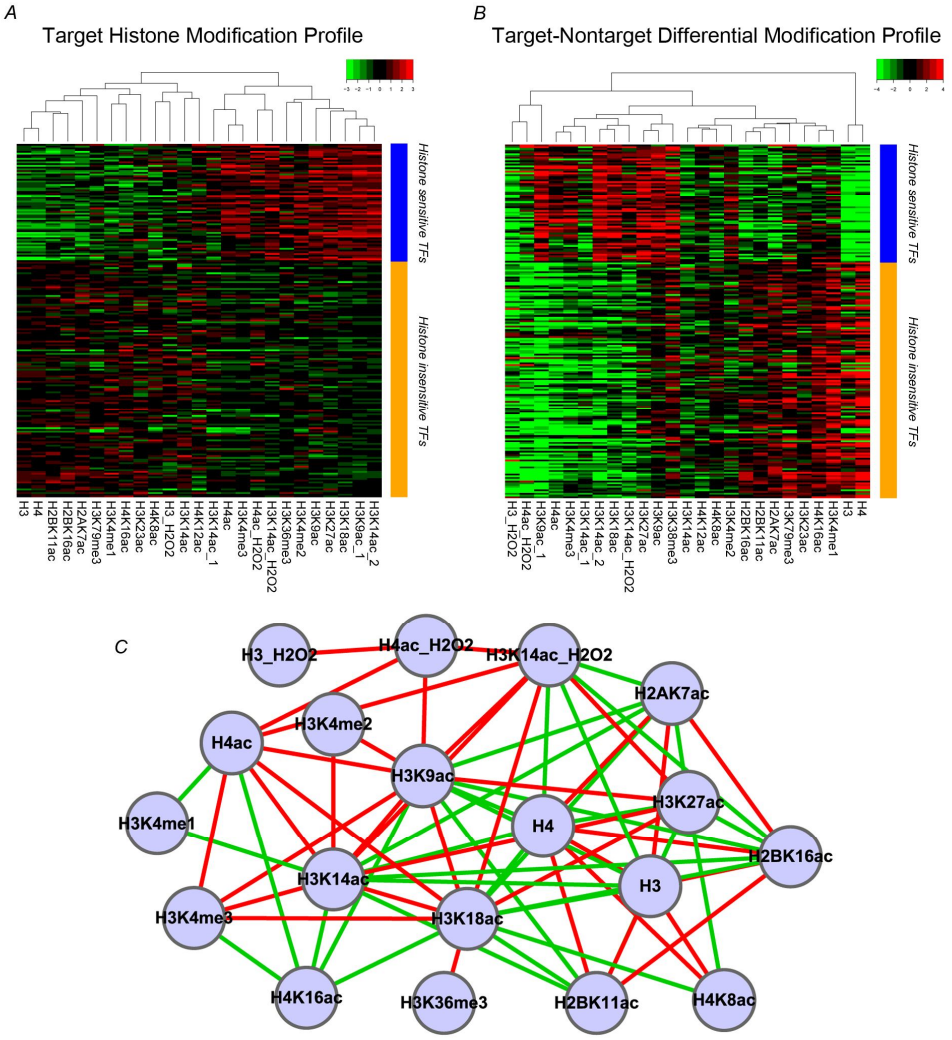

Figure 4.5: Target histone modification profiles and target-nontarget differential modification profiles of TFs

(A) Target histone modification profiles comprising different normalized histone modification signals (columns) of TFs (rows). Target histone modification profile is the averaged histone modification signals of the TF's targets. TFs are clustered into histone sensitive (blue bar) and insensitive TFs (orange bar) using their target histone modification profiles. Histone sensitive TFs have stronger histone modification signals. (B) Target-nontarget differential modification profiles of TFs showing discriminating power (t-statistic) of histone modifications to TF targets and non-targets. TFs are ordered the same as (A). Histone sensitive and insensitive TFs have distinct differential modification profiles, indicating preferential histone modifications of TFs targets. (C) Correlation network of histone modifications in terms of TF differential modification profile. Histone modification pairs with correlation coefficient larger than 0.5 (red edges) or smaller than -0.5 (green edges) are connected. The network shows high level of redundancy of histone modifications in differential modification profiles.

importance. The t-statistics for all histone modifications forms a TF specific profile, denoted as differential modification profiles of the TF. Interestingly, histone sensitive TFs and histone insensitive TFs defined based on target histone modification profiles were also distinct in their differential modification profiles (see Figure 4.5). This suggests that histone sensitive and insensitive TFs are actually robust clusters with different patterns of histone modifications in their target genes.

4.2.5 Chromatin sensitivity of transcription factors

To understand the biological nature of the histone sensitive and insensitive TFs, we explored the feature differences of these two TF classes under different biological "context". First, we observed different predictive power of histone modifications for target gene prediction between the two TF classes. As shown in Figure 4.6A, histone modifications are generally more predictive of the target genes for histone sensitive TFs than for histone insensitive TFs. This is due to the fact that target genes of histone sensitive TFs have stronger histone modification signals, which substantially improve the performance of our chromatin model.

Histone sensitive and insensitive TFs also show distinct topological characteristics in biological networks. In general, histone sensitive TFs have less target genes than histone insensitive TFs (see Figure 4.6). Yu et al has constructed a hierarchical network in yeast based on the TF-TF regulation relationships identified by the ChIP-chip [33]. We mapped the histone sensitive and insensitive TFs onto the hierarchical network, and found that histone sensitive TFs

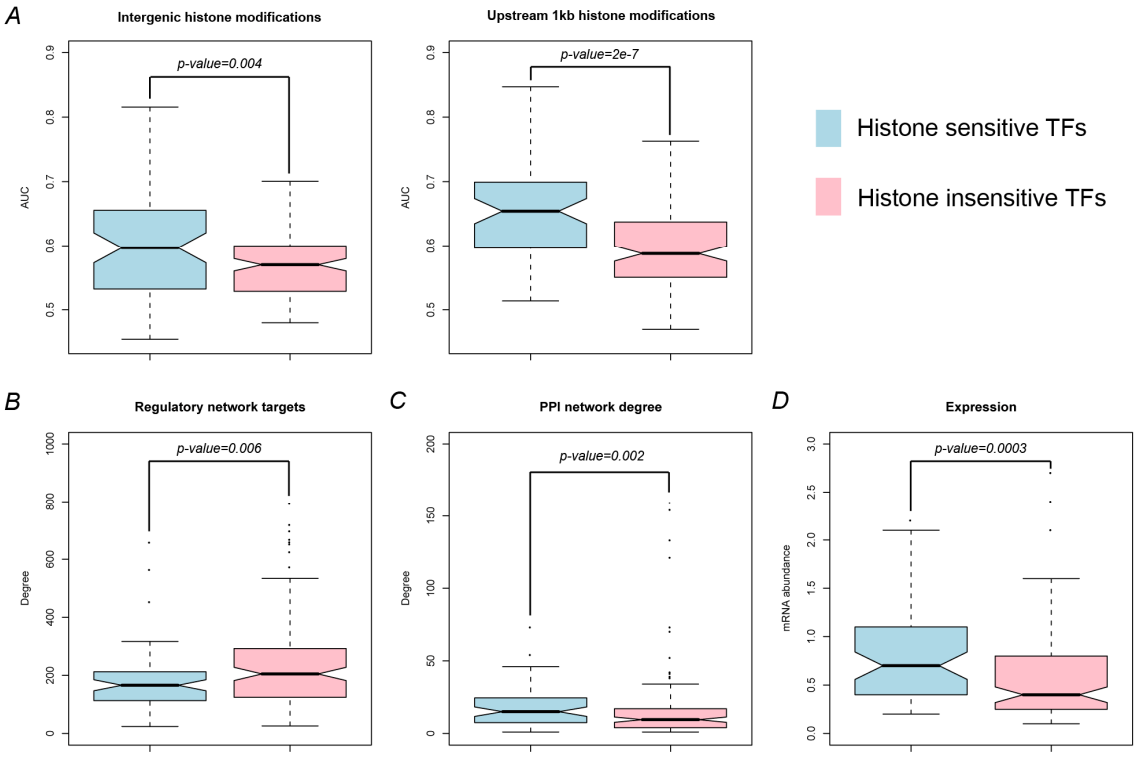

Figure 4.6: Distinctions between histone sensitive and insensitive TFs

(A) Target genes of histone sensitive TFs are better predicted using intergenic or upstream histone modifications than histone insensitive TFs. (B) Histone sensitive TFs show smaller number of target genes in regulatory network than histone insensitive TFs. (C) Histone sensitive TFs have larger number of interacting partners in protein interaction network. (D) Higher mRNA expression levels of histone sensitive TFs.

Hierarchical levels in regulatory network		Number of histone sensitive TFs		Number of histone insensitive TFs	
1	Bottom level	15	15	52	52
2		43		36	
3	Upper levels	6	51	8	48
4		2		4	

Fisher exact test p-value = 0.0007 for separate levels, p-value = 0.0002 for combined upper levels.

Table 4.2: TF histone sensitivity relates to hierarchical level in regulatory network

were enriched in the upper layers. This suggests that histone sensitive TFs are more likely to act as "managers" that regulate other TFs, while histone insensitive TFs tend to be "workers" at bottom layer in the hierarchy (see Table 4.2). We have also examined the "degrees" of these TFs in the protein-protein interaction networks [93]. Our results indicate that histone sensitive TFs tend to have more physical interacting partners than those histone insensitive TFs (see Figure 4.6). The high connectivity of histone sensitive TFs further implies their functional importance.

Interestingly, histone sensitivity of TFs also indicates distinct co-regulation relationships. Two TFs are said to co-regulate if their sets of targets significantly overlap. Among the ~14,000 possible TF pairs, we found 1,440 significant co-regulatory relationships ($P<0.05$, Fisher's exact test). Among the TFs involved in co-regulatory relationships, 64 are histone sensitive and 95 are histone insensitive. In the 1,440 significant co-regulation pairs, 447 are between two histone sensitive TFs, 437 between two histone insensitive TFs, and 556 between one histone sensitive TF and one histone insensitive TF. Fisher's Exact test showed that histone sensitive TFs are more

Cellular functions	Histone sensitive TFs	Histone insensitive TFs
Cell cycle	ACE2, ASH1, CIN5, FKH1, FKH2, MBP1, MCM1, NDD1, RLM1, STB1, STE12, STP1, SWI4, SWI5, SWI6, TEC1	CST6
Heat shock or stress conditions	GAT1, MSN4, SKN7, YAP1	GLN3, HAL9, HMS2, HSF1, MGA1, MSN2, WAR1, USV1

Table 4.3: TF histone sensitivity relates to cellular functions

likely to be involved in a co-regulatory relationship than histone insensitive TFs ($P<10^{-16}$). In summary, the histone sensitive TFs reside mostly in the upper layers of the regulatory network, tend to work and communicate with other TFs during transcriptional regulations.

Furthermore, we found that the expression levels of histone sensitive TFs were higher than those of the histone insensitive TFs (see Figure 4.6). It seems that histone sensitivity of TFs was also related to their biological functions. For example, TFs involved in cell cycle regulation predominantly belong to histone sensitive class (see Table 4.3). Out of 17 cell cycle TFs reported in a previously study [142], only CST6 was classified to be histone insensitive. We also examined TFs that were specific to certain conditions, e.g. heat shock or oxidative stress. Conditional specific TFs were classified into both histone sensitive and insensitive classes, thus it was not obvious whether the histone sensitivity and condition specificity of TFs are related.

4.2.6 PSSM predictability and cooperativity of transcription factors

TFs exhibit different PSSM predictability in that the targets of some TFs are well predicted

PSSM sensitive TFs	AUC
REB1	0.87
ABF1	0.86
CBF1	0.84
FHL1	0.83
RAP1	0.79
TYE7	0.77
SUM1	0.76
UME6	0.76
MBP1	0.72
GCN4	0.71

Table 4.4: Top 10 PSSM well-predictable TFs

using its PSSM alone but others are not. PSSM predictability reflects the extent to which TFs recognize its binding site through motif matching. Since the PSSMs are available for only 50 TFs, accounting for about a quarter of 203 TFs with histone modification profiles, it is difficult to perform systematic identification and classification to categorize them into PSSM well-predictable and weakly-predictable. However, we have identified 10 TFs with target prediction AUC>0.7 using their PSSMs alone, providing a confident subset of PSSM well-predictable TFs (see Table 4.4). To check whether the high predictability is attributed to PSSM specificity, we calculated the information content of these PSSMs. We found that the information content of the 10 well-predictable TFs' PSSMs is not different from other TFs (P=0.4, Wilcoxon test). We investigated the expression level, number of target genes, hierarchy in regulatory network of these 10 TFs, and found no significant difference from the other TFs. We will be able to make more confident conclusions when more PSSMs for TFs become available in the future.

Histone sensitive TFs	Histone insensitive TFs
ABF1, ACE2, ARG80, ARG81, ASH1, AZF1, CAD1, CBF1, CIN5, CRZ1, CUP9, ECM22, FHL1, FKH1, FKH2, GAT1, GAT3, GCN4, GCR1, GCR2, GTS1, HAP1, HAP2, HAP4, HIR1, HIR2, HIR3, HMS1, INO2, INO4, LEU3, MAC1, MBP1, MCM1, MET31, MET4, MSN1, MSN4, NDD1, OPI1, PDR1, PHO2, PUT3, RAP1, REB1, RGM1, RLM1, RME1, ROX1, RPH1, SFP1, SKN7, SMP1, SPT2, STB1, STE12, STP1, SWI4, SWI5, SWI6, TBS1, TEC1, TYE7, YAP1, YAP5, YAP6, YML081W, ZAP1	A1, ABT1, ACA1, ADR1, AFT2, ARO80, ARR1, ASK10, BAS1, BYE1, CHA4, CST6, DAL80, DAL81, DAL82, DAT1, DIG1, DOT6, EDS1, FAP7, FZF1, GAL3, GAL4, GAL80, GLN3, GZF3, HAA1, HAC1, HAL9, HAP3, HAP5, HMS2, HOG1, HSF1, IFH1, IME1, IME4, IXR1, KRE33, KSS1, MAL13, MAL33, MBF1, MDS3, MET18, MET28, MET32, MGA1, MIG1, MIG2, MIG3, MOT3, MSN2, MSS11, MTH1, NDT80, NNF2, NRG1, OAF1, PDC2, PDR3, PHD1, PHO4, PIP2, PPR1, RCO1, RCS1, RDR1, RDS1, RFX1, RGT1, RIM101, RLR1, RPI1, RPN4, RTG1, RTG3, RTS2, SFL1, SIG1, SIP3, SIP4, SKO1, SMK1, SNF1, SNT2, SOK2, SPT10, SPT23, SRD1, STB2, STB4, STB5, STB6, STP2, STP4, SUM1, SUT1, SUT2, THI2, TOS8, UGA3, UME6, UPC2, USV1, WAR1, WTM1, WTM2, XBP1, YAP3, YAP7, YBL054W, YBR239C, YBR267W, YDR026C, YDR049W, YDR266C, YDR520C, YER051W, YER130C, YER184C, YFL044C, YFL052W, YGR067C, YHP1, YJL206C, YKL222C, YLR278C, YNR063W, YOX1, YPR022C, YPR196W, YRR1, ZMS1

Table 4.5: Histone sensitive and insensitive TFs

For the TFs that are weakly-predictable using PSSM information, we hypothesized that these TFs may bind to their targets indirectly by cooperating with other TFs. If the hypothesis is true, we would expect to predict its target genes accurately by using the PSSM of its cooperative TF. We tested this by using each PSSM to predict targets of all TFs (see Table 4.5). The targets of most TFs were best predicted by their own PSSMs, but some TFs have their targets better

predicted using other TFs' PSSMs. For example, YAP1's target genes were better predicted using CAD1's PSSM, with AUC increased from 0.71 to 0.75. Similarly, using INO2's PSSM, INO4 was better predicted of its target genes with AUC increased from 0.78 to 0.81. In fact, YAP1 and CAD1 work together in stress induced transcriptional responses, and INO2 and INO4 form heteromeric complexes involved in phospholipids biosynthesis [143, 144]. We also found that the PSSMs of the cooperative TFs are actually quite similar, measured by a similarity score range from 0 to 1. The PSSMs of CAD1 and YAP1 render a similarity score of 0.72 (top 1% among all pairs), and those of INO2 and INO4 render 0.55 (top 5%). This further indicates the cooperation between the two TF pairs through indirect binding. Therefore, TF target gene prediction using cross PSSMs could help identify cooperations between TFs. On the other hand, this suggests that using TF's own PSSM may not always be the best for predicting its target genes, when there is evidence of TF cooperations.

4.2.7 Comparison with previous methods

We compared our SVM based method with several previous published approaches including Cluster-Buster [124], MCAST [145], EEL [146], Stubb [127]. We calculated the prediction accuracy of each method by applying it to 10 TFs with more than 200 target genes under YPD condition. As shown in Table 4.6, our method that integrates histone modification and PSSM data sets achieves the best prediction for most factors. For those histone-senstive TFs such as Swi4 and Swi6, including histone modification data can improve target prediction accuracy

		ROC AUC						
	Number of Target genes	HIS+PSSM	HIS alone	PSSM alone (FIMO)	Cluster -Buster	MCAST	EEL	Stubb
ABF1	549	0.830	0.736	0.781	0.776	0.676	0.807	0.893
FHL1	207	0.957	0.963	0.827	0.855	0.874	0.852	0.887
FKH1	284	0.656	0.625	0.606	0.680	0.546	0.661	0.725
FKH2	216	0.723	0.694	0.664	0.698	0.566	0.688	0.735
HAP1	215	0.738	0.711	0.635	0.675	0.624	0.676	0.663
RAP1	408	0.865	0.818	0.805	0.752	0.774	0.802	0.811
REB1	278	0.773	0.623	0.774	0.727	0.818	0.765	0.758
SWI4	252	0.831	0.790	0.634	0.680	0.626	0.664	0.651
SWI6	230	0.809	0.768	0.719	0.720	0.629	0.742	0.665
UME6	298	0.854	0.767	0.831	0.774	0.783	0.814	0.815

Table 4.6: Comparison of several computational methods for target gene prediction

substantially, and PSSM alone gives relatively poor predictions no matter what algorithms are used to search for TF binding sites.

Among those previous published methods, EEL and Stubb take advantage of conservation of TF binding motifs between related species, and as shown they achieve relatively more accurate prediction results than FIMO, Cluster-Buster and MCAST. We also tried the "Chromia" method proposed by Won et al [134]. Similar to our method, Chromia integrates histone modification and PSSM data sets but using a hidden Markov model. The method has shown impressive performance when applied to genome-wide ChIP-seq data in mouse for predicting TF binding sites. However, when applied to the yeast data in our case, it does not result in good prediction due to the low coverage of the histone modification and TF binding data sets [111, 133]. For example, the arrays used for Pokholok ChIP-chip data contains ~42,000 probes (60-mers),

representing only about 20% of the yeast genome [133]. The arrays used for identifying yeast TF binding sites are essentially promoter arrays, covering only DNA regions around the translation start site of yeast ORFs [111]. In practice, our method only requires data for interested regions, e.g. promoter regions, and thereby is more flexible and can be applied to a wide range of data sets.

4.3 Discussion

4.3.1 Condition specific gene regulation: contribution from chromatin modifications

It is widely known that transcriptional regulation is condition specific in that TFs change their binding sites under different conditions. We showed here that histone modification data is most predictive of TF target binding under the same condition. This is true especially for those TFs that are mostly active in specific stress conditions.

Because of limited resources, it is impossible to perform exhaustive experiments for every TFs, cell types, species and all possible conditions. As an alternative method, we proposed the feasibility to predict target genes of a TF under an interested condition by combining histone modification data under that condition with its PSSM. In this way, we can achieve much higher results than using PSSM alone, and more importantly the predictions are also condition specific.

On the other hand, PSSMs, the TF binding recognition motifs, are generally thought to be non-condition-specific which do not change under different conditions [115]. Similarly, it will

also be interesting to investigate condition specificity of the relative importance profiles of that histone modification of TF target binding (see Figure 4.5). However, with histone modification data mostly for YPD medium, we are currently unable to test this hypothesis. If this were in fact true, it would be sufficient to predict TF targets by matching PSSMs and relative importance histone modification profiles, both TF-specific and condition-non-specific, to DNA sequences and chromatin modifications under certain conditions, respectively. This can be very useful in not only understanding of transcription regulation of chromatin modifications, but also getting a broader picture of TF targets turnover under different conditions.

4.3.2 Histone-sensitive and insensitive TFs

For the 203 yeast TFs used in our study, we classified them into 68 histone-sensitive and 135 histone-insensitive TFs based on the upstream histone modification signals of their target genes. The two classes have generally opposite characteristics in histone modification signals, expression levels, topology in regulatory networks and other biological features.

Hitone sensitive TFs might target highly regulated genes. It is known that gene expression is regulated by specific TFs and their orchestrating chromatin modification enzymes. Thus, stronger histone modification signals upstream of the target genes of the histone sensitive TFs are indicative of more intensive transcriptional regulation. Our results showed that cell cycle TFs were mostly histone-sensitive TFs, consistent with the previous knowledge that cell cycle is highly-regulated to achieve cyclical expression of genes.

The majority of histone modification data used in this study is based on YPD medium. We found that histone sensitive TFs tend to be active under YPD medium, as indicated by larger number of target genes and higher expression levels with respect to those insensitive TFs. It is possible that histones upstream of the target genes of the histone sensitive TFs have more chance to be modified by histone modification enzymes, because they are recruited by these more active TFs.

We found that histone sensitive TFs were enriched in higher layers of the hierarchical regulatory network. This suggests that histone sensitive TFs tend to be "managers" that regulate other TFs and for such a reason their binding to target genes is highly regulated through histone modifications. In consistent with this hypothesis, we have observed stronger histone modification signals in the upstream regions of their target genes.

The two classes of TFs were also different in their histone modification profiles. Some histone modifications show opposite signal patterns between histone sensitive and insensitive TFs. For example, H3K9ac and H3K14ac modifications show higher signals in target than non-targe genes for histone sensitive TFs, while the opposite is observed for histone insensitive TFs. This might be relevant to the TF behavior as a transcription activator or repressor, since histone acetylations are generally known to be an active mark during transcription regulation [115].

Here, we have attempted to provide some biological intuition on the difference between histone sensitive and insensitive TFs. Detailed analysis is required to further understand the biology behind the classification of TFs.

4.3.3 Combinatorial interaction of TFs: direct and indirect binding

When a TF binds directly to the promoter regions of its target genes, the enriched motifs identified from its binding sites can be regarded as its own PSSM. However, TFs do not always act individually; sometimes they cooperate with (physically bind to) each other to form regulatory functional units, such as yeast cell cycle complexes SBF (SWI4-SWI6) and MBF (MBP1-SWI6) [30]. In these indirect binding cases, it is important to distinguish the TFs that are motif-recognizing and the ones that are not.

By examining the PSSM sensitivity of TFs, we were able to infer some possible combinatorial interactions between TFs. If TF A's targets are better predicted by using another TF B's PSSM instead of its own PSSM, then this is an indication of potential cooperation of the two TFs. In particular, TF B directly binds to promoter regions through its PSSM, and TF A indirectly binds to promoter regions through physical binding to TF B [147]. This is also referred to as indirect piggy-back binding [117].

PSSM sensitivity under indirect TF binding is important for our target gene prediction model. Instead of using a TF's own PSSM, PSSM of another TF through which the TF binds should be used for more accurate predictions. Therefore, identifying those cases before using our model will be necessary to achieve better results.

4.3.4 Implications on gene expression regulation

In this study, we showed that incorporating chromatin modification information could

substantially improve the prediction of TF target genes. In fact, chromatin modifications relate to gene expression regulation in two layers [115]. First, chromatin is modified to form structure as euchromatin, within which genes could be turned on and off, or heterochromatin, within which genes are silenced. Second, euchromatin is further modified by enzymes recruited by specific TFs to mark the "on and off" status of transcription. We examined the target vs. non-target differential histone modification profiles for each individual TFs, and observed TF specific chromatin modifications marked in the target genes. Therefore, we suggest that chromatin modifications might function as both non-specific euchromatin marks and TF specific regulatory marks. And our model takes advantage of the chromatin information from both of the two layers.

However, it is still not quite clear the sequential order in terms of time and causality of chromatin modification and TF binding. It is possible that one of them happens first which then drives the occurrence of the other. The other possibility is that the two events might be interactive in a feedback manner to regulate gene expression. More fine-tuned experiments in the future would be helpful for unraveling the time-dependent interaction between chromatin modification and TF binding.

4.4 Methods

4.4.1 Chromatin modification data

The yeast histone modification data sets used in this study are basically from two sources. The first data set is available from Pokholok et al [133] at

http://web.wi.mit.edu/young/nucleosome/, which contains the profiles of 14 chromatin features under YPD or H_2O_2 condition. These chromatin features include histone H3 and H4 occupation, H3K9ac, H3K14ac, H4K5ac8ac12ac16ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3, and H3K79me3. The profiles of these features are measured by ChIP-chip experiments using over 40,000 probes, which cover 85% of the yeast genome. We calculated the signal of each chromatin feature in the 1kb upstream region of each open reading frame (ORF) by averaging signals of all the probes within this region. Similarly, for each ORF the average signal of each feature in the 1kb region downstream of the start codon was also calculated. We named them as the upstream chromatin signal and downstream chromatin signal for ORFs, respectively.

The second data set is available from Kurdistani et al [132]. The data contains levels of acetylation of 11 lysines in intergenic regions (IR) as well as ORF regions. These profiles were also measured by using ChIP-chip experiments. These 11 histone acetylations are H2AK7ac, H2BK11ac, H2bK16ac, H3K9ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H4K8ac, H4K12ac and H4K16ac. We named the signal in IR and ORF as the IR chromatin signal and coding region (CR) chromatin signal for ORFs, respectively.

4.4.2 Target genes of yeast transcription factors

Target genes for 203 yeast transcription factors under various conditions (including YPD and H_2O_2) have been identified using the ChIP-chip experiment by Harbison et al [111]. For each binding interaction, a probability score (P-value) was calculated, measuring the binding potential

of a TF with the promoter region of a gene.

When TF target genes are determined according to ChIP-chip data, one needs to set a cutoff for P-values, which indicates the confidence of regulation of genes by TFs. A small (strict) P-value cutoff would result in fewer but more confident target genes, while a large (loose) P-value cutoff would do the opposite. For instance, there are 159 target genes for RAP1 using a cutoff value of 0.001, while the target gene number increases to 581 when the cutoff value 0.05 is used. We therefore tested the influence of P-value cutoff on our model performance. As shown in Figure 4.3, our results indicate that a more stringent P-value cut-off, i.e. smaller target gene set, improves the prediction accuracy of our model. Moreover, at all cutoff values the models combining histone modification and PSSM data outperform the models using either of them alone. On the other hand, a more stringent cutoff results in less target genes. To ensure enough positive target genes for model training, we decided to use 0.01 as the P-value cutoff in our analysis.

4.4.3 Position-specific scoring matrices of transcription factors

Two sets of position-specific scoring matrices (PSSMs) for yeast transcription factors have been identified previously using different strategies [10,18]. The first set was downloaded from http://genomics.princeton.edu/tavazoie/Gene%20Expression.htm, which was based on *de novo* motif finding in all yeast promoter sequences [130]. The promoter DNA sequences (from start codon of a ORF to 800bp upstream) of all yeast ORFs were analyzed to identify enriched motifs by using the AlignACE program [148]. A total of 666 motifs have been found, among which 51

can be associated with known yeast transcription factors. The occurrences and matching scores of these motifs in the promoter regions of all yeast genes were also provided by Beer et al [130].

The second set of PSSMs was available from http://fraenkel.mit.edu/Harbison/, which is based on motif analysis of target promoters identified by the ChIP-chip experiment [111]. The detail about motif discovery procedure can be found in [111]. In brief, for a transcription factor the motifs were discovered by applying a suite of motif discovery programs to the intergenic sequences identified by the binding data for this factor. The resulting motifs were subsequently clustered, filtered and selected to give rise to a single PSSM that can best represent the motif of a factor. For some factors the above procedure failed to identify their motifs and in such cases motifs were derived from literature or databases.

The information content (IC) of a PSSM is calculated as $IC = -\sum_{i,j} p_{i,j} \times \log\left(p_{i,j} / p_b\right)$, where i and j represent positions in PSSMs and four nucleotides, respectively. $p_{i,j}$ is then the weight at each PSSM position of each nucleotide, and p_b is the background nucleotide frequency of *S. cerevisiae* genome. Specifically, we use 37% as the GC content to calculate p_b for each nucleotide.

The similarity between two PSSMs is calculated as the averaged dot product at each PSSM positions, $Similarity = \frac{1}{n}\sum_{i,j} p_{1,i,j} \times p_{2,i,j}$, where n is the length of PSSM. If two PSSMs are of different lengths, we compare each possible alignment of the two PSSMs with no gap, and keep the maximum similarity from each alignment. The similarity score has the range from 0 to 1.

4.4.4 Searching promoters for known motifs

Given the list of PSSMs for transcription factors, we searched the promoters of all yeast genes for the occurrences of these motifs using FIMO of the MEME suite [148] available at http://meme.nbcr.net/meme4_3_0/downloads.html. The promoter region was defined the DNA region from the start codon to 800bp upstream of an ORF. The cumulative matching score (CMS) of the occurrences of a motif in the promoter region of a gene was calculated, which was subsequently used as features for predicting TF target genes.

4.4.5 Comparison of chromatin modifications between functional TFBSs and non-functional motif matching sites

We performed comparative analysis of chromatin modification differences between functional TFBSs and non-functional motif matching sites for TFs with available PSSMs. SWI4 was shown here as an example. A list of binding sites (TFBS) of the factor SWI4 was downloaded from the Saccharomyces Genome Database (SGD) [149] at http://www.yeastgenome.org/. This list contains 99 binding sites that were targeted by the SWI4 under YPD medium according to the ChIP-chip data. We also collected a list of non-TFBSs by selecting DNA regions that was consistent with the SWI4 motif but not targeted by SWI4 under YPD medium as indicated by the ChIP-chip results ($P>0.4$). These non-TFBSs were further filtered to ensure that there is no SWI4 TFBSs within the nearby 2kb region, which ultimately resulted in 485 non-TFBSs for SWI4. All of the TFBSs and non-TFBSs are less than 20bp in size. The levels of the chromatin features on

these TFBSs and non-TFBSs were calculated based on the intensities of probes covering them. Finally, the signal of the 14 chromatin features was compared between the TFBS and non-TFBS groups using the t-test.

4.4.6 Support vector machine model for transcription factor target prediction

For a transcription factor in each gene we have obtained the following features: cumulative matching score from motif searching, the upstream and downstream signal of 14 histone or histone modification profiles, the IR and CR signal of 11 histone acetylation profiles. All or subsets of these features were integrated using a support vector machine (SVM) model [136] for predicting target genes of a TF.

As a supervise machine learning model, the class of gene must be know to train the SVM model. We split the data into two sets, a training set and a testing set. The model was then trained using the training set and applied to the testing set to predict target genes. The prediction power of the model was estimated based on the testing set. In general, the SVM model outputs a probability indicating how likely a gene is the target of a TF. By setting different cut-off values, we can balance the sensitivity (true positive rate) and specificity (true negative rate) of predictions of the model. The plot of the sensitivity versus 1-specificity is called receiver operating characteristic (ROC), which can be used to show the classification accuracy of the SVM model. AUC, the area under the ROC curve can be used to summarize the prediction power of the model.

4.4.7 Clustering of TFs using target chromatin modification profile

For each TF, target histone modification profile was calculated by averaging histone modification signals among all its targets. 1kb upstream chromatin modifications from Pokholok et al [20] and intergenic chromatin modifications from Kurdistani et al [132][19] were used. Unsupervised k-means clustering algorithm was performed to generate two TF clusters, histone sensitive and insensitive TFs, by their target histone modification profiles.

To understand the relative importance of each chromatin modification to target prediction, target-nontarget differential histone modification profiles for TFs were calculated based on t-statistic. For each chromatin modification in differential modification profile for a TF, modification signals for target genes and non-target genes were collected and t-statistic calculated. The t-statistics in differential modification profiles indicated the directional significance of chromatin modifications to distinguish target genes.

4.4.8 Inferring interactions between transcription factors

The target genes identified by ChIP-chip experiment could be wither direct or indirect targets of a TF. For example if two transcription factors A and B that are interacted with each other, the ChIP-chip for A can potentially identify target genes of B as well. Conversely, the existence of B's motif would be informative for predicting target genes of factor A. We used the TF target prediction model with chromatin modifications and the TF' own PSSM, and then compared the model's AUC performance to the models with chromatin modifications and other TFs' PSSMs.

Models with improved AUC performances suggest better predictive power of other PSSMs than the TF's own. These cases might indicate interactions between the TFs.

4.4.9 Application of previously reported methods

We run all methods with their default parameter settings. Internal thresholding is turned off in all cases to report a full list of predictions with scores. PSSMs of 10 TFs and upstream 1kb DNA sequences for all annotated yeast *S. cerevisiae* ORFs are used as inputs to MCAST [132] and Cluster-Buster [132]. Pairwise pre-aligned upstream 1kb DNA sequences of all annotated *S. cerevisiae* and *S. paradoxus* orthologous ORFs are used instead for running EEL [132] and Stubb [132]. Predicted binding targets with respective scoring systems from the programs are collected for all 10 TFs. ROC curves and AUCs are calculated based on the same thresholding scheme for all methods.

Chapter 5

Conclusion

In this thesis, we first present a formulism of measuring evolutionary rewiring of biological networks. Network rewiring rate is measured as percentage edge change per Mys, which is a normalized metric by the size of the comparing networks. The metric shows Log-Log linear relationship with divergence time, indicating saturation effect of evolutionary changes also found in sequences. We found that different types of biological networks rewire at different rates during evolution. TF regulatory network and kinase substrate phosphorylation network are in the fast rewiring group, and metabolic pathway network is the slowest. The rewiring ordering of biological networks is consistent in all species comparisons. The formulism of measuring rewiring rate is directly applicable to other types of networks, commonly observed in our real lives. The differences in rewiring rates do not come directly from gene content turnover of specific GO categories of genes. Most biological networks evolve in similar rates as protein and coding DNA sequences. We argue that regulatory networks rewire faster than collaborative networks. And understanding of quickly evolving regulatory networks could potentially help us

unravel the differences of close species, such as human and chimpanzee.

To investigate the robustness of the measure of rewiring rate in comparing different biological networks, we then develop a computational simulation method. By randomly adding edges and nodes to and removing those from biological networks, we simulated the effect of potential false positives and negatives in measuring rewiring rates. Our simulation results show that the rewiring rate measure is robust even the data sets have large percentage errors. The simulation of network rewiring process identifies node removal parameter being the most influential to rewiring rate measure in our model.

Although it is time for evolutionary analysis of biological networks, there are still not enough high quality network data sets available today. We used a machine learning method to predict TF target genes in yeast, which contributes to future high quality data sets. Histone modification profiles have been found different in TF binding sites and non-binding sites. This information is used as features in SVM model, along with TF binding motifs. Our model shows significantly better prediction power than previously reported methods. Different in their histone modification profiles, 203 yeast TFs could be clustered into histone sensitive and insensitive ones. For histone sensitive TFs, incorporating histone modification signals could significantly improve the prediction their target genes, and they are enriched of cell cycle regulated TFs and hub TFs in higher hierarchies in networks.

Future research topics following this thesis would possibly include more detailed analysis of network rewiring rate, relating network rewiring model to experimental results, and application of TF target prediction methods to more conditions, cell lines, and species. It will be interesting to

analyze the rewiring hotspots and coldspots of networks, and to construct species trees based on their networks, which could be compared with molecular trees. More theoretical studies are needed to understand the topological mechanism of rewiring rate differences. A more complex network rewiring model bringing in previously reported network growth models may be helpful. Application of the TF target prediction model to more conditions in yeast, or more cell lines in other organisms may generate predictions when experimental data are not readily available. What is more, further analysis of histone sensitive and insensitive TFs in yeast may uncover interesting biological implications in their regulatory roles. And it is also important to check this clustering in other model organisms, such as fly, worm, and human, when more ChIP data sets become available.

Chapter 6

Appendix: Co-expression network of non-coding RNAs in *C. elegans*

6.1 Introduction

The massive amounts of data from tiling arrays and high-throughput sequencing have driven the discovery of novel transcripts [150-152]. Unlike the main transcription products mRNAs which are then translated into proteins, many transcripts are not, and hence are called non-coding RNAs (ncRNAs). ncRNAs include many well-known RNA types such as rRNA, tRNA and snoRNA, as well as small RNAs such as miRNA, siRNA, and piRNA. Some of ncRNAs, including miRNAs and siRNAs, carry expression regulatory functions in eukaryotes that increasingly proved to be important in cellular regulatory systems.

With the advent of high-throughput sequencing technologies, it is now possible to experimentally survey novel transcriptomes to find ncRNAs [153]. Lu et al. develop a

comprehensive model, *incRNA* (*in*tegrated *ncRNA* finder), which integrates sequence, structure, and expression data [154]. The large-scale expression data sets are gathered from the modENCODE consortium, which includes tiling array and deep sequencing data from different tissues and developmental stages of *C. elegans* [21, 155]. 7,237 novel ncRNA candidates (merged from 10,994 high-confidence ncRNA bins) are predicted using *incRNA* and the expression of a random sample has been experimentally validated.

Understanding the expression profiles of these novel ncRNA candidates is important for their characterization. We here present a co-expression network approach to identify distinct expression patterns of novel ncRNA candidates, in combination with other known ncRNAs and coding RNA transcripts.

6.2 Results

6.2.1 Novel ncRNA candidates and known ncRNAs

The 10,994 novel ncRNA candidate bins from our prediction were pooled with our sample set of 476 known ncRNAs, a total of 11,473 ncRNA bins, to study their expression profiles. We collected ncRNA expression level data from 11 small RNAseq experiments conducted in different developmental stages. Compared to RNAseq and tiling array experiments in measuring transcript expression levels, small RNAseq is more sensitive and accurate for short-length ncRNAs. We found that known ncRNAs generally have higher expression levels than novel ncRNA bins (Wilcoxon test pval<e-10) and smaller expression level variance (Wilcoxon test pval<e-10).

11,473 ncRNA candidate bins were then further clustered into three groups reflecting their expression patterns, 348 universal expression, 6925 differential expression, and 4200 undetectable expression ncRNA candidate bins. Among universal expression ncRNAs, 202 (58%) were known ncRNAs; however, only 167 (2%) were known ncRNAs among differential expression ncRNAs (see Figure 6.1). It indicated that known ncRNAs are enriched in universal expression cluster, while novel ncRNAs are enriched in differential expression cluster (Fisher's Exact Test pval<e-15), which contains only 2% of the known ncRNAs. Since the majority known ncRNAs are miRNAs, universal expression across developmental stages might be a reason why miRNAs are the best characterized ncRNAs. However, our predicted novel ncRNAs are largely differentially expressed in developmental stages indicating their difficulty of experimental discovery. The rest predicted novel ncRNAs bins have no detectable expression in all stages experimented. The finding that most differentially expressed ncRNAs come from the novel candidates in intriguing, suggesting their specialized roles in specific stages.

6.2.2 Novel ncRNA candidates and coding transcripts

In order to examine the differential expression of ncRNAs and coding transcripts, co-expression network of above mentioned 476 known ncRNAs, 10,994 novel ncRNAs and also 27,322 coding transcripts was constructed by calculating the Euclidean distance of pair wise expression vectors. 11 small RNAseq experiments, 6 poly-A RNAseq experiments, 29 total RNA tiling array experiments, and 12 poly-A RNA tiling array experiments are used for expression

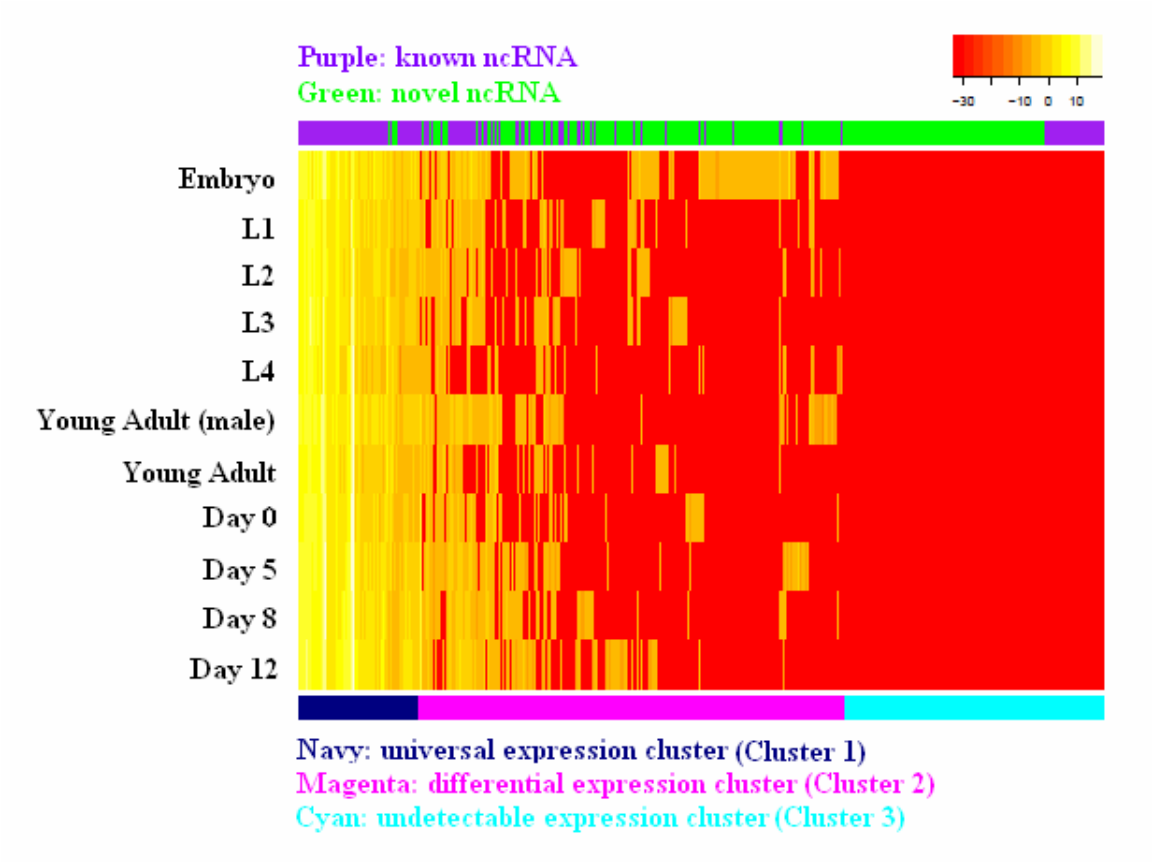

Figure 6.1: Expression profile of novel ncRNA candidate bins.

The expression of novel ncRNA bins is compared to known ncRNAs among 11 developmental stages in *C. elegans*. Novel ncRNA bins are labeled as green bars on the top row, and known ncRNAs in purple. ncRNAs are ordered according to 3 major expression profile clusters, shown on the bottom row.

levels of transcripts. Topological Overlap Matrix (TOM) distance is then calculated to measure the pair wise topological distance in the co-expression network using the WGCNA package [156]. TOM distance is defined with the following formula: $distTOM_{ij} = 1 - \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$, where a_{ij} denotes the expression vectors Euclidean distance of transcript i and j, and k_i denotes co-expression network degree of transcript i. This distance measure is better than Euclidean distance because it considers not only Euclidean distance between two objects, but also the overlap of their network neighbors. Therefore, using TOM distance in co-expression network is more capable to find co-expression clusters and modules. We observed three co-expression clusters in the network: two of them (Cluster A and Cluster B) have distinct expression patterns with small topological distances among their member transcripts, and Cluster C is generally an out-group in the network (see Figure 6.2). In fact, the largest cluster (Cluster A with 19655 transcripts) is enriched of coding transcripts (93%), while the second cluster (Cluster B with 13902 transcripts) is enriched of novel ncRNAs (70%, Fisher's Exact Test pval<e-15) (see Table 6.1). The result shows that our predicted novel ncRNAs have very different expression patterns to coding transcripts, too.

It is currently difficult to infer possible biological functions of each predicted novel ncRNA with experimental evidence. However, novel ncRNA candidates may share functional similarities with those coding transcripts that have close expression profiles. We performed Gene Ontology (GO) analysis on coding transcripts that are clustered together with novel ncRNA candidates (see Table 6.2). Five clusters, cluster 0, 1, 6, 9, and 11, are enriched of novel ncRNA candidates with a

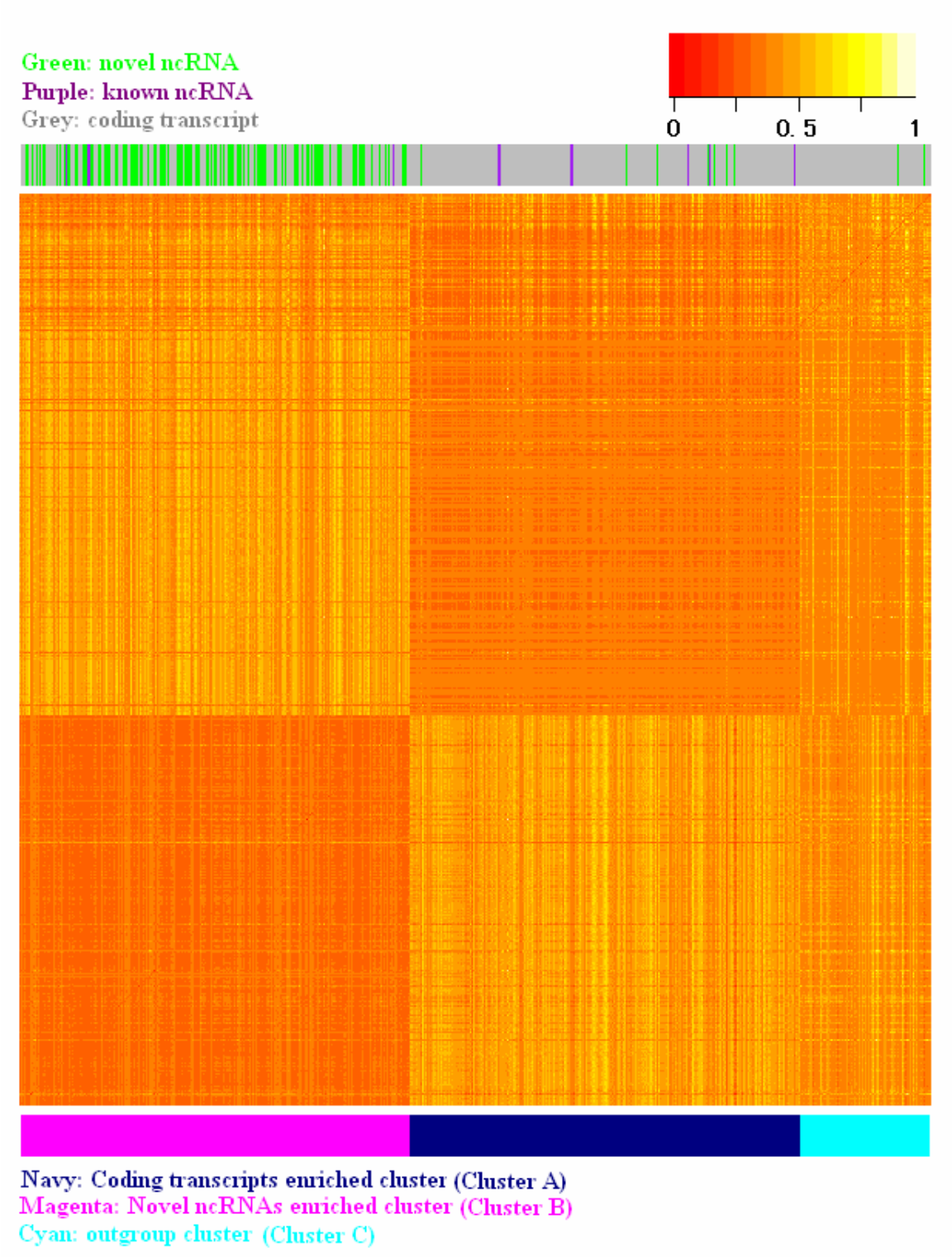

Figure 6.2: Co-expression network of novel ncRNA bins, known ncRNAs, and coding transcripts. The top row denotes the labels of transcripts. Three main co-expression clusters are found using unsupervised learning method, shown on the bottom row.

Cluster	Subcluster	Novel ncRNA bins	Coding transcripts	Known ncRNAs
1	1	133	12287	68
	2	814	792	235
	3	36	2842	3
	4	31	1318	0
	5	20	1073	3
	Subtotal	1034	18312	309
2	1	6823	711	137
	2	1441	867	10
	3	189	873	4
	4	254	492	2
	5	345	289	3
	6	279	265	2
	7	156	312	6
	8	132	143	0
	9	101	29	1
	10	31	5	0
	Subtotal	9751	3986	165
3	Subtotal	209	5024	5

Table 6.1: Main and sub clusters of *C. elegans* transcripts co-expression network.

Cluster	Coding	Non-coding	Total	nc %	Coding transcripts GO enrichment	Coding transcripts GO depletion
0	1199	1193	2392	49.87%	Transmembrane proteins, Receptors, Signal transducer activity	Protein binding, Development, Growth regulation
1	2727	4003	6730	59.48%	Chromatin assembly, DNA binding, Organelle organization	Protein binding
2	4210	604	4814	12.55%	Ion channel, Receptor, Membrane, Signal transducer actitivity, Transcription	Protein binding, Laval development, Growth regulation, Organelle, Cell cycle
3	4377	424	4801	8.83%	Development, Growth regulation, Reproduction	Receptor, Signal transducer activity
4	4274	356	4630	7.69%	Development, Cell cycle, Growth regulation, Reproduction	Receptor, Membrane, Signal transducer activity, Transcription factor activity
5	2931	362	3293	10.99%	Lipid metabolism, Sugar binding, Anion transport	Development, Receptor, Signal transducer activity, Growth regulation
6	577	1805	2382	75.78%	No significant enrichment	No significant depletion
7	2185	115	2300	5.00%	Membrane, Signal transducer activity, Receptor	Development, Growth regulation, Organelle
8	1548	149	1697	8.78%	Metabolism, Kinase	Development, Signal transducer activity, Receptor, Expression regulation
9	276	1276	1552	82.22%	No significant enrichment	No significant depletion
10	1182	75	1257	5.97%	Receptor, Membrane, Signal transducer activity, Ion channel	Development, Growth regulation, Reproduction
11	677	527	1204	43.77%	Transcription factor activity, Ion binding	No significant depletion
12	660	42	702	5.98%	Chromatin assembly, DNA binding, Organelle organization	Membrane
13	499	63	562	11.21%	Organelle	No significant depletion
Total	27322	10994	38316			

Table 6.2: GO analysis of coding transcripts in clusters with non-coding RNA candidates.

fraction >40%. Coding transcripts in cluster 0, 1, and 11 are enriched of expression regulatory genes, such as TFs and chromatin binding proteins. More interestingly, in cluster 6 and 9, which mainly contain novel ncRNA candidates (>75%), no GO term enrichment is found for coding transcripts in each cluster. This also indicates that many novel ncRNA candidates may have distinct functions that are not previously known.

6.3 Methods

To examine the expression pattern of known and novel ncRNAs, expression data from 11 small RNAseq experiments of known and novel ncRNAs were log-transformed. ncRNAs with all zero expression levels across 11 experiments were first identified and clustered as Cluster 3. Unsupervised clustering method (kmeans function in R) is then used to further split the remaining ncRNAs into two groups. The group with universal high expression levels across 11 experiments is labeled as Cluster 1, and the group with differential expression levels is labeled as Cluster 2. All 479 known ncRNAs are kept while 1,000 novel ncRNAs are randomly sampled from a total 10,994 novel ncRNAs for heatmap visualization.

Transcript co-expression network modules is detected using the WGCNA package in R. Known ncRNAs, novel ncRNAs and coding transcripts are combined with their expression levels measured in small RNAseq, poly-A RNAseq and tiling array experiments. Small RNAseq and poly-A RNAseq data is log-transformed, and expression distributions for all experiments are shifted to have the same median while maintaining their distribution shape. Adjacency matrix of

Euclidean distances was calculated for all pairs of expression profiles, and then the Topological Overlap Matrix (TOM) distances were calculated based on the adjacency matrix to reflect topological distances between transcript pairs in the co-expression network. Transcripts were first clustered using unsupervised method and then hierarchically clustered. Three network modules representing closely connected transcripts were detected and grouped using WGCNA package. 300, 300 and 100 transcripts from Cluster A, B and C are randomly sampled for heatmap visualization.

Bibliography

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

2. Sanger, F. and H. Tuppy, The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem J*. **49**(4): 481-90 (1951).

3. Sanger, F. and H. Tuppy, The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J*. **49**(4): 463-81 (1951).

4. Holley, R.W., J. Apgar, G.A. Everett, et al., Structure of a Ribonucleic Acid. *Science*. **147**: 1462-5 (1965).

5. Sanger, F., S. Nicklen, and A.R. Coulson, DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. **74**(12): 5463-7 (1977).

6. Schuster, S.C., Next-generation sequencing transforms today's biology. *Nat Methods*. **5**(1): 16-8 (2008).

7. Wang, Z., M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. **10**(1): 57-63 (2009).

8. Nagalakshmi, U., Z. Wang, K. Waern, et al., The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. **320**(5881): 1344-9 (2008).

9. Needleman, S.B. and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. **48**(3): 443-53 (1970).

10. Smith, T.F. and M.S. Waterman, Identification of common molecular subsequences. *J Mol Biol*. **147**(1): 195-7 (1981).

11. Matthews, L.R., P. Vaglio, J. Reboul, et al., Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*. **11**(12): 2120-6 (2001).

12. Uetz, P., L. Giot, G. Cagney, et al., A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*. **403**(6770): 623-7 (2000).

13. Ito, T., T. Chiba, R. Ozawa, et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*. **98**(8): 4569-74 (2001).

14. Giot, L., J.S. Bader, C. Brouwer, et al., A protein interaction map of Drosophila melanogaster. *Science*. **302**(5651): 1727-36 (2003).

15. Li, S., C.M. Armstrong, N. Bertin, et al., A map of the interactome network of the metazoan C. elegans. *Science*. **303**(5657): 540-3 (2004).

16. Rual, J.F., K. Venkatesan, T. Hao, et al., Towards a proteome-scale map of the human protein-protein interaction network. *Nature*. **437**(7062): 1173-8 (2005).

17. Gavin, A.C., P. Aloy, P. Grandi, et al., Proteome survey reveals modularity of the yeast cell machinery. *Nature*. **440**(7084): 631-6 (2006).

18. Krogan, N.J., G. Cagney, H. Yu, et al., Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*. **440**(7084): 637-43 (2006).

19. Boone, C., H. Bussey, and B.J. Andrews, Exploring genetic interactions and networks with yeast. *Nat Rev Genet*. **8**(6): 437-49 (2007).

20. Bartel, D.P., MicroRNAs: target recognition and regulatory functions. *Cell*. **136**(2): 215-33 (2009).

21. Gerstein, M.B., Z.J. Lu, E.L. Van Nostrand, et al., Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. *Science*. (2010).

22. Roy, S., J. Ernst, P.V. Kharchenko, et al., Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science*. (2010).

23. Cohen, P., The regulation of protein function by multisite phosphorylation--a 25 year

update. *Trends Biochem Sci*. **25**(12): 596-601 (2000).

24. Ficarro, S.B., M.L. McCleland, P.T. Stukenberg, et al., Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat Biotechnol*. **20**(3): 301-5 (2002).

25. Manning, G., G.D. Plowman, T. Hunter, et al., Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci*. **27**(10): 514-20 (2002).

26. Ptacek, J., G. Devgan, G. Michaud, et al., Global analysis of protein phosphorylation in yeast. *Nature*. **438**(7068): 679-84 (2005).

27. Breitkreutz, A., H. Choi, J.R. Sharom, et al., A global protein kinase and phosphatase interaction network in yeast. *Science*. **328**(5981): 1043-6 (2010).

28. Pace, N.R., The universal nature of biochemistry. *Proc Natl Acad Sci U S A*. **98**(3): 805-8 (2001).

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

30. Yu, H., D. Greenbaum, H. Xin Lu, et al., Genomic analysis of essentiality within protein networks. *Trends Genet*. **20**(6): 227-31 (2004).

31. Kim, P.M., J.O. Korbel, and M.B. Gerstein, Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A*. **104**(51): 20274-9 (2007).

32. Yu, H., P.M. Kim, E. Sprecher, et al., The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*. **3**(4): e59 (2007).

33. Yu, H. and M. Gerstein, Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci U S A*. **103**(40): 14724-31 (2006).

34. Bhardwaj, N., K.K. Yan, and M.B. Gerstein, Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. *Proc Natl Acad Sci U S A*. **107**(15): 6841-6 (2010).

35. Kim, P.M., L.J. Lu, Y. Xia, et al., Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*. **314**(5807): 1938-41 (2006).

36. Shou, C., Bhardwaj, N., Lam, H.Y.K., Yan, K.K., Kim, P.M., Snyder, M., Gerstein, M.B., Measuring the Evolutionary Rewiring of Biological Networks. *PLoS Comput Biol*. **7**(1): e1001050 (2011).

37. Cheng, C., Shou, C., Yip, K.Y., Yan, K.K., Gerstein, M.B., Genome-wide analysis of chromatin features identifies chromatin-sensitive and chromatin-insensitive classes of yeast transcription factors. *Genome Biol*. (Submitted).

38. Borneman, A.R., T.A. Gianoulis, Z.D. Zhang, et al., Divergence of transcription factor binding sites across related yeast species. *Science*. **317**(5839): 815-9 (2007).

39. Tuch, B.B., D.J. Galgoczy, A.D. Hernday, et al., The evolution of combinatorial gene regulation in fungi. *PLoS Biol*. **6**(2): e38 (2008).

40. Lee, T.I., N.J. Rinaldi, F. Robert, et al., Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*. **298**(5594): 799-804 (2002).

41. Beltrao, P., J.C. Trinidad, D. Fiedler, et al., Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol*. **7**(6): e1000134 (2009).

42. Tong, A.H., G. Lesage, G.D. Bader, et al., Global mapping of the yeast genetic interaction network. *Science*. **303**(5659): 808-13 (2004).

43. Krek, A., D. Grun, M.N. Poy, et al., Combinatorial microRNA target predictions. *Nat Genet*. **37**(5): 495-500 (2005).

44. Lewis, B.P., C.B. Burge, and D.P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. **120**(1): 15-20 (2005).

45. John, B., A.J. Enright, A. Aravin, et al., Human MicroRNA targets. *PLoS Biol*. **2**(11): e363 (2004).

46. Walhout, A.J., R. Sordella, X. Lu, et al., Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*. **287**(5450): 116-22 (2000).

47. Stelzl, U., U. Worm, M. Lalowski, et al., A human protein-protein interaction network: a resource for annotating the proteome. *Cell*. **122**(6): 957-68 (2005).

48. Ho, Y., A. Gruhler, A. Heilbut, et al., Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*. **415**(6868): 180-3 (2002).

49. Gavin, A.C., M. Bosche, R. Krause, et al., Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. **415**(6868): 141-7 (2002).

50. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. **437**(7055): 69-87 (2005).

51. Jukes, T.H., Cantor,C.R., *Evolution of protein molecules*, ed. H.N. Munro. 1969, New York: Academic Press. p. 21-132.

52. Kimura, M., *The neutral theory of molecular evolution*. 1983, Cambridge: Cambridge University Press. p. 156-67.

53. Fay, J.C., G.J. Wyckoff, and C.I. Wu, Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature*. **415**(6875): 1024-6 (2002).

54. Dujon, B., D. Sherman, G. Fischer, et al., Genome evolution in yeasts. *Nature*. **430**(6995): 35-44 (2004).

55. Kellis, M., N. Patterson, M. Endrizzi, et al., Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*. **423**(6937): 241-54 (2003).

56. Bustamante, C.D., A. Fledel-Alon, S. Williamson, et al., Natural selection on protein-coding genes in the human genome. *Nature*. **437**(7062): 1153-7 (2005).

57. Bloom, J.D., D.A. Drummond, F.H. Arnold, et al., Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol*. **23**(9): 1751-61 (2006).

58. Dunker, A.K., C.J. Brown, J.D. Lawson, et al., Intrinsic disorder and protein function. *Biochemistry*. **41**(21): 6573-82 (2002).

59. Yu, H., N.M. Luscombe, H.X. Lu, et al., Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*. **14**(6): 1107-18 (2004).

60. Kelley, B.P., R. Sharan, R.M. Karp, et al., Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A*. **100**(20): 11394-9 (2003).

61. Sharan, R., S. Suthram, R.M. Kelley, et al., Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A*. **102**(6): 1974-9 (2005).

62. Beltrao, P. and L. Serrano, Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput Biol*. **3**(2): e25 (2007).

63. Moses, A.M., D.A. Pollard, D.A. Nix, et al., Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol*. **2**(10): e130 (2006).

64. Roguev, A., S. Bandyopadhyay, M. Zofall, et al., Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*. **322**(5900): 405-10 (2008).

65. Dixon, S.J., Y. Fedyshyn, J.L. Koh, et al., Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc Natl Acad Sci U S A*. **105**(43): 16653-8 (2008).

66. van Dam, T.J. and B. Snel, Protein complex evolution does not involve extensive network rewiring. *PLoS Comput Biol*. **4**(7): e1000132 (2008).

67. Ravasi, T., H. Suzuki, C.V. Cannistraci, et al., An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*. **140**(5): 744-52 (2010).

68. Hinman, V.F., A.T. Nguyen, R.A. Cameron, et al., Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc Natl Acad Sci U S A*. **100**(23): 13356-61 (2003).

69. Koonin, E.V., N.D. Fedorova, J.D. Jackson, et al., A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*. **5**(2): R7 (2004).

70. van Nimwegen, E., Scaling laws in the functional content of genomes. *Trends Genet*. **19**(9): 479-84 (2003).

71. Ranea, J.A., A. Grant, J.M. Thornton, et al., Microeconomic principles explain an optimal genome size in bacteria. *Trends Genet*. **21**(1): 21-5 (2005).

72. Nobrega, M.A., I. Ovcharenko, V. Afzal, et al., Scanning human gene deserts for long-range enhancers. *Science*. **302**(5644): 413 (2003).

73. Woese, C.R., G.E. Fox, L. Zablen, et al., Conservation of primary structure in 16S ribosomal RNA. *Nature*. **254**(5495): 83-6 (1975).

74. Yip, K.Y., P. Patel, P.M. Kim, et al., An integrated system for studying residue

coevolution in proteins. *Bioinformatics*. **24**(2): 290-2 (2008).

75. Lynch, M. and J.S. Conery, The evolutionary fate and consequences of duplicate genes. *Science*. **290**(5494): 1151-5 (2000).

76. King, M.C. and A.C. Wilson, Evolution at two levels in humans and chimpanzees. *Science*. **188**(4184): 107-16 (1975).

77. Mitchell, P.J. and R. Tjian, Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*. **245**(4916): 371-8 (1989).

78. Kemp, B.E., D.B. Bylund, T.S. Huang, et al., Substrate specificity of the cyclic AMP-dependent protein kinase. *Proc Natl Acad Sci U S A*. **72**(9): 3448-52 (1975).

79. Jordan, I.K., I.B. Rogozin, G.V. Glazko, et al., Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet*. **19**(2): 68-72 (2003).

80. Bourque, G., B. Leong, V.B. Vega, et al., Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*. **18**(11): 1752-62 (2008).

81. Wang, J., N.J. Bowen, L. Marino-Ramirez, et al., A c-Myc regulatory subnetwork from human transposable element sequences. *Mol Biosyst*. **5**(12): 1831-9 (2009).

82. Kunarso, G., N.Y. Chia, J. Jeyakani, et al., Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*. **42**(7): 631-4 (2010).

83. Xie, D., C.C. Chen, L.M. Ptaszek, et al., Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res*. **20**(6): 804-15 (2010).

84. Schmidt, D., M.D. Wilson, B. Ballester, et al., Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*. **328**(5981): 1036-40 (2010).

85. Madan Babu, M., S.A. Teichmann, and L. Aravind, Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol*. **358**(2): 614-33 (2006).

86. Lavoie, H., H. Hogues, J. Mallick, et al., Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biol*. **8**(3): e1000329 (2010).

87. Ihmels, J., S. Bergmann, M. Gerami-Nejad, et al., Rewiring of the yeast transcriptional

network through the evolution of motif usage. *Science*. **309**(5736): 938-40 (2005).

88. Ward, J.J. and J.M. Thornton, Evolutionary models for formation of network motifs and modularity in the Saccharomyces transcription factor network. *PLoS Comput Biol*. **3**(10): 1993-2002 (2007).

89. Presser, A., M.B. Elowitz, M. Kellis, et al., The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication. *Proc Natl Acad Sci U S A*. **105**(3): 950-4 (2008).

90. Fusco, D., L. Grassi, B. Bassetti, et al., Ordered structure of the transcription network inherited from the yeast whole-genome duplication. *BMC Syst Biol*. **4**: 77 (2010).

91. Conant, G.C., Rapid reorganization of the transcriptional regulatory network after genome duplication in yeast. *Proc Biol Sci*. **277**(1683): 869-76 (2010).

92. Amoutzias, G.D., Y. He, J. Gordon, et al., Posttranslational regulation impacts the fate of duplicated genes. *Proc Natl Acad Sci U S A*. **107**(7): 2967-71 (2010).

93. Stark, C., B.J. Breitkreutz, T. Reguly, et al., BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. **34**(Database issue): D535-9 (2006).

94. Kanehisa, M., S. Goto, M. Furumichi, et al., KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*. **38**(Database issue): D355-60 (2010).

95. Griffiths-Jones, S., H.K. Saini, S. van Dongen, et al., miRBase: tools for microRNA genomics. *Nucleic Acids Res*. **36**(Database issue): D154-8 (2008).

96. Mok, J., P.M. Kim, H.Y. Lam, et al., Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Sci Signal*. **3**(109): ra12 (2010).

97. Lam, H.Y., P.M. Kim, J. Mok, et al., MOTIPS: automated motif analysis for predicting targets of modular protein domains. *BMC Bioinformatics*. **11**: 243 (2010).

98. Finn, R.D., M. Marshall, and A. Bateman, iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*. **21**(3): 410-2 (2005).

99. Yan, K.K., G. Fang, N. Bhardwaj, et al., Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc*

Natl Acad Sci U S A. **107**(20): 9186-91 (2010).

100. Haider, S., B. Ballester, D. Smedley, et al., BioMart Central Portal--unified access to biological data. *Nucleic Acids Res*. **37**(Web Server issue): W23-7 (2009).

101. Tatusova, T., Genomic databases and resources at the National Center for Biotechnology Information. *Methods Mol Biol*. **609**: 17-44 (2010).

102. Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32**(5): 1792-7 (2004).

103. O'Brien, K.P., M. Remm, and E.L. Sonnhammer, Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*. **33**(Database issue): D476-80 (2005).

104. Wapinski, I., A. Pfeffer, N. Friedman, et al., Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*. **23**(13): i549-58 (2007).

105. von Mering, C., R. Krause, B. Snel, et al., Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*. **417**(6887): 399-403 (2002).

106. Bader, J.S., A. Chaudhuri, J.M. Rothberg, et al., Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*. **22**(1): 78-85 (2004).

107. Yu, H., P. Braun, M.A. Yildirim, et al., High-quality binary protein interaction map of the yeast interactome network. *Science*. **322**(5898): 104-10 (2008).

108. Barabasi, A.L. and R. Albert, Emergence of scaling in random networks. *Science*. **286**(5439): 509-12 (1999).

109. Cacuci, D.G., M. Ionescu-Bujor, and I.M. Navon, *Sensitivity and uncertainty analysis*. 2003, Boca Raton: Chapman & Hall/CRC Press.

110. Helton, J.C., Johnson, J.D., Sallaberry, C.J., Storlie, C.B.,, Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*. **91**(10-11): 1175-1209 (2005).

111. Harbison, C.T., D.B. Gordon, T.I. Lee, et al., Transcriptional regulatory code of a eukaryotic genome. *Nature*. **431**(7004): 99-104 (2004).

112. Saul, Z.M. and V. Filkov, Exploring biological network structure using exponential random graph models. *Bioinformatics*. **23**(19): 2604-11 (2007).

113. Goffeau, A., B.G. Barrell, H. Bussey, et al., Life with 6000 genes. *Science*. **274**(5287): 546, 563-7 (1996).

114. Goffeau, A., Four years of post-genomic life with 6,000 yeast genes. *FEBS Lett*. **480**(1): 37-41 (2000).

115. Kouzarides, T., Chromatin modifications and their function. *Cell*. **128**(4): 693-705 (2007).

116. Li, B., M. Carey, and J.L. Workman, The role of chromatin during transcription. *Cell*. **128**(4): 707-19 (2007).

117. Kato, M., N. Hata, N. Banerjee, et al., Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol*. **5**(8): R56 (2004).

118. Ren, B., F. Robert, J.J. Wyrick, et al., Genome-wide location and function of DNA binding proteins. *Science*. **290**(5500): 2306-9 (2000).

119. Iyer, V.R., C.E. Horak, C.S. Scafe, et al., Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*. **409**(6819): 533-8 (2001).

120. Horak, C.E. and M. Snyder, ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol*. **350**: 469-83 (2002).

121. Robertson, G., M. Hirst, M. Bainbridge, et al., Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*. **4**(8): 651-7 (2007).

122. Johnson, D.S., A. Mortazavi, R.M. Myers, et al., Genome-wide mapping of in vivo protein-DNA interactions. *Science*. **316**(5830): 1497-502 (2007).

123. Stormo, G.D., DNA binding sites: representation and discovery. *Bioinformatics*. **16**(1): 16-23 (2000).

124. Frith, M.C., M.C. Li, and Z. Weng, Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*. **31**(13): 3666-8 (2003).

125. Zhou, Q. and W.H. Wong, CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*. **101**(33): 12114-9 (2004).

126. Gupta, M. and J.S. Liu, De novo cis-regulatory module elicitation for eukaryotic

genomes. *Proc Natl Acad Sci U S A*. **102**(20): 7079-84 (2005).

127. Sinha, S., Y. Liang, and E. Siggia, Stubb: a program for discovery and analysis of cis-regulatory modules. *Nucleic Acids Res*. **34**(Web Server issue): W555-9 (2006).

128. Van Loo, P., S. Aerts, B. Thienpont, et al., ModuleMiner - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol*. **9**(4): R66 (2008).

129. Xie, D., J. Cai, N.Y. Chia, et al., Cross-species de novo identification of cis-regulatory modules with GibbsModule: application to gene regulation in embryonic stem cells. *Genome Res*. **18**(8): 1325-35 (2008).

130. Beer, M.A. and S. Tavazoie, Predicting gene expression from sequence. *Cell*. **117**(2): 185-98 (2004).

131. Berger, S.L., The complex language of chromatin regulation during transcription. *Nature*. **447**(7143): 407-12 (2007).

132. Kurdistani, S.K., S. Tavazoie, and M. Grunstein, Mapping global histone acetylation patterns to gene expression. *Cell*. **117**(6): 721-33 (2004).

133. Pokholok, D.K., C.T. Harbison, S. Levine, et al., Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*. **122**(4): 517-27 (2005).

134. Won, K.J., B. Ren, and W. Wang, Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol*. **11**(1): R7 (2010).

135. Ernst, J., H.L. Plasterer, I. Simon, et al., Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res*. **20**(4): 526-36 (2010).

136. Cristianini, N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press. 2000.

137. Pina, B., U. Bruggemeier, and M. Beato, Nucleosome positioning modulates accessibility of regulatory proteins to the mouse mammary tumor virus promoter. *Cell*. **60**(5): 719-31 (1990).

138. Lee, C.K., Y. Shibata, B. Rao, et al., Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet*. **36**(8): 900-5 (2004).

139. Sekinger, E.A., Z. Moqtaderi, and K. Struhl, Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell*. **18**(6): 735-48 (2005).

140. Yuan, G.C., Y.J. Liu, M.F. Dion, et al., Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*. **309**(5734): 626-30 (2005).

141. Ni, L., C. Bruce, C. Hart, et al., Dynamic and complex transcription factor binding during an inducible response in yeast. *Genes Dev*. **23**(11): 1351-63 (2009).

142. Wu, W.S. and W.H. Li, Systematic identification of yeast cell cycle transcription factors using multiple data sources. *BMC Bioinformatics*. **9**: 522 (2008).

143. Hirata, D., K. Yano, and T. Miyakawa, Stress-induced transcriptional activation mediated by YAP1 and YAP2 genes that encode the Jun family of transcriptional activators in Saccharomyces cerevisiae. *Mol Gen Genet*. **242**(3): 250-6 (1994).

144. Schuller, H.J., K. Richter, B. Hoffmann, et al., DNA binding site of the yeast heteromeric Ino2p/Ino4p basic helix-loop-helix transcription factor: structural requirements as defined by saturation mutagenesis. *FEBS Lett*. **370**(1-2): 149-52 (1995).

145. Bailey, T.L. and W.S. Noble, Searching for statistically significant regulatory modules. *Bioinformatics*. **19 Suppl 2**: ii16-25 (2003).

146. Palin, K., J. Taipale, and E. Ukkonen, Locating potential enhancer elements by comparative genomics using the EEL software. *Nat Protoc*. **1**(1): 368-74 (2006).

147. Gordan, R., A.J. Hartemink, and M.L. Bulyk, Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res*. **19**(11): 2090-100 (2009).

148. Bailey, T.L. and C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*. **2**: 28-36 (1994).

149. Cherry, J.M., C. Adler, C. Ball, et al., SGD: Saccharomyces Genome Database. *Nucleic Acids Res*. **26**(1): 73-9 (1998).

150. Margulies, M., M. Egholm, W.E. Altman, et al., Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. **437**(7057): 376-80 (2005).

151. Shendure, J., G.J. Porreca, N.B. Reppas, et al., Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. **309**(5741): 1728-32 (2005).

152. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*. **457**(7232): 1028-32 (2009).

153. He, H., J. Wang, T. Liu, et al., Mapping the C. elegans noncoding transcriptome with a whole-genome tiling microarray. *Genome Res*. **17**(10): 1471-7 (2007).

154. Lu, Z.J., K.Y. Yip, G. Wang, et al., Prediction and characterization of non-coding RNAs in C. elegans by integrating conservation, secondary structure and high throughput sequencing and array data. *Genome Res*. (2010).

155. Celniker, S.E., L.A. Dillon, M.B. Gerstein, et al., Unlocking the secrets of the genome. *Nature*. **459**(7249): 927-30 (2009).

156. Langfelder, P. and S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. **9**: 559 (2008).