Postprint of Complex Biological Networks and Two-Color Networks Involving Non-Coding RNA

Authors: Liu Changning, Sun Shiwei, Yi Zhao, Bu Dongbo

Date: 2017-03-09T00:00:00+00:00

Abstract

Noncoding RNA (ncRNA) and complex biological networks represent two rapidly advancing directions in contemporary biological research, particularly within the genomics domain, and constitute hot topics in bioinformatics research. This paper respectively introduces the relevant background and latest research progress of noncoding RNA studies and complex biological network studies, especially regarding the applications of bioinformatics therein, and further discusses the significant importance of incorporating noncoding RNA into complex biological network research as well as possible directions for future work.

Full Text

Abstract

Non-coding ribonucleic acid (ncRNA) and complex biological networks represent two rapidly developing directions in biological research, particularly in genomics, and have become focal points in bioinformatics research. This paper introduces the relevant background and latest research progress in both ncRNA studies and complex biological network analysis, with special emphasis on the application of bioinformatics in these fields. We further discuss the significance of introducing ncRNA into complex biological network research and propose possible directions for future work.

Keywords: bioinformatics; protein interaction network; non-coding RNA genes; complex network analysis

1 Introduction

Following the completion of the *Haemophilus influenzae* genome sequencing in 1995 [1], the genomes of a series of organisms have been determined. The near-completion of human genome sequencing in 2004 [2] particularly marked the

arrival of the post-genomic era in biological research. As carriers of genetic information, whole genome sequences provide us with rich information but also raise new questions. Traditional concepts hold that biological complexity is determined by the scale of protein-coding genes. However, through comparative analysis of genome sequences across species, we find that although humans and nematodes differ dramatically in morphology and complexity, the number of protein-coding genes is not substantially different [3]. What then accounts for this enormous difference? Research indicates that living organisms can be represented as complex, dynamically changing networks where nodes are various biomolecules such as DNA, RNA, and proteins, and their associations are represented as edges (undirected edges, such as protein interactions) or arrows (directed edges, such as gene regulatory relationships) [4,5]. Systems biology posits that such network relationships are the source of life' s complexity, with various complex biological phenomena arising from different combinations and dynamic changes of interactions and regulatory relationships among nodes in the network. Increased biological complexity can be achieved through two distinct approaches: first, by increasing the number of nodes in the network, i.e., adding more proteins and RNAs to expand network scale; more importantly, by enhancing the dynamic adjustment capabilities of interactions between nodes and introducing new interaction mechanisms. Therefore, the study of life requires a holistic approach, investigating these complex interactions and regulatory relationships within biological networks rather than merely studying individual nodes in isolation [6].

Non-coding RNA genes are transcribed but not translated into proteins; they function directly as RNA molecules [7,8]. Non-coding RNAs play functional roles in multiple critical processes including chromatin epigenetic modification, mRNA transcription and degradation, protein transport, and RNA processing and modification [9-12], and are closely associated with various diseases and tumorigenesis [13-16]. With the successive discovery of numerous non-coding RNAs in multiple model organisms, research on non-coding RNAs has become a key focus and hotspot in biological studies [17-19]. The incorporation of non-coding RNAs into biological networks has substantially increased life' s complexity: first, by adding tens of thousands of nodes to expand network scale; more importantly, by introducing various new interaction mechanisms. For example, microRNA (miRNA), discovered only in recent years, introduces an entirely new post-transcriptional regulatory layer in biological complex networks through its complementary inhibition of messenger RNAs [10]. Recognizing the potential significant impact of non-coding RNAs on biological networks, research on non-coding genes has expanded from discovering new non-coding genes and exploring their functions to studying and constructing mixed networks of non-coding and coding genes. Although research on mixed networks has just begun, it will undoubtedly become a new hotspot in non-coding gene research. In this paper, we combine our recent work on non-coding RNAs and complex biological networks to introduce existing work and future research directions in non-coding RNA functional studies, complex biological network analysis, and

the construction of biological networks involving non-coding RNAs.

2 Non-coding RNA and RNAomics

In higher organisms and humans, non-coding regions constitute the majority of the genome sequence. For instance, protein-coding sequences account for only about 3-5% of the human and mouse genomes, with the remaining 95-97% being non-coding regions [20,21]. These regions were once considered "junk DNA" without any function. However, from an evolutionary perspective, the increasing proportion of non-coding sequences with the functional perfection and complexity of organisms indicates that non-coding regions must have important biological functions. In recent years, domestic and international scholars have conducted increasingly in-depth research on large-scale transcriptomes. Massive experimental data demonstrate that genomic non-coding regions not only participate in transcriptional regulation as binding sites but also transcribe numerous non-coding RNA products. Related research includes: (1) large-scale complementary DNA (cDNA) annotation studies, such as the RIKEN consortium's 2003 cloning analysis of mouse full-length cDNAs, which identified nearly 4,280 full-length cDNAs lacking protein-coding open reading frames as non-coding RNA genes [17,22]; (2) gene chip studies, such as Affymetrix' s 2005 high-density oligonucleotide chip study of the transcriptome of 10 human chromosomes, which confirmed the existence of numerous non-coding RNA genes [23]; and (3) experimental RNomics, such as the 2006 study by Chen Runsheng' s laboratory at the Institute of Biophysics, Chinese Academy of Sciences, on *C. elegans* microRNAs, which discovered large numbers of new non-coding RNAs, including two new classes: small nuclear-like RNA (snlRNA) and stem-bulge RNA (sbRNA) [24]. Numerous similar works cannot be enumerated here. To date, scientists worldwide have discovered large numbers of non-coding RNAs in various organisms including mice, fruit flies, *Arabidopsis*, rice, archaea, and even *E. coli* [18,19,25-28].

Existing research has found that these non-coding RNAs of varying lengths and structures perform diverse functions in organisms: small nuclear RNAs (snRNA) participate in mRNA splicing [29]; small nucleolar RNAs (snoRNA) participate in ribosomal RNA (rRNA) methylation and pseudouridylation [12]; guide RNAs (gRNA) participate in RNA editing [30]; Signal Recognition Particle RNA (SRP-RNA) participates in protein cellular localization [11]; telomerase RNA participates in DNA telomere synthesis and affects cell lifespan [31]; transfer-messenger RNA (tmRNA) participates in terminating protein synthesis from damaged mRNAs [32]; Xist induces X chromosome inactivation [33]; and piRNAs participate in regulating chromosomal epigenetic modifications [9]. Additionally, recent medical research on various diseases and tumors has identified numerous disease- and tumor-specifically expressed non-coding genes, such as the highly expressed non-coding RNA gene MALAT-1 in non-small cell lung cancer [16] and the abnormally expressed non-coding RNA gene PCGEM1 in prostate cancer [15]. In contrast to non-coding RNAs with known functions,

we know virtually nothing about the functions of the vast majority of non-coding RNAs. How to study the regulation and function of these non-coding RNAs has become a new challenge in biological research. Chinese and foreign scientists have both recognized the RNAome issue with this as the research object. As early as 1998, Chinese scientist Jin Youxin proposed the "Functional RNAome Research Plan" at the 109th Xiangshan Science Conference. Large-scale experimental and computational RNAome research began abroad around 2000, with five major discoveries in this field being selected as top ten scientific breakthroughs of the year by *Science* between 2001 and 2006. RNAomics, with non-coding RNAs as the research subject, has become a hotspot in both experimental biology and bioinformatics.

Piwi-interacting RNAs are a class of small RNA molecules approximately 29-30 nucleotides in length, expressed only in mammalian testes, and can bind with Piwi proteins to form piRNA complexes (piRCs). They are related to RNA silencing.

2.1 Establishment of the NONCODE Database

With increasing attention to non-coding genes and the deepening of related research, more and more new members and new classes of non-coding genes have been discovered, and databases collecting and organizing non-coding gene-related information have emerged. Some databases focus only on specific classes of non-coding genes, such as SRP RNA, tmRNA, and RNase P RNA, while others collect various non-coding gene data, such as "Small RNA Database," "Non-coding RNA Database," and "Rfam Database" [34-39]. However, these databases have several problems. First, because they collect data manually from literature, the collected non-coding gene data has many omissions in both quantity and type. A more serious problem is that they lack a unified system for classifying and annotating non-coding genes, which causes even more troublesome issues. NONCODE was developed against this background. On one hand, NONCODE adopts a workflow of computer-automated filtering of GenBank [40] data followed by manual inspection and confirmation. This improves both the comprehensiveness and accuracy of data collection while ensuring efficiency. On the other hand, to address the lack of a unified classification system for non-coding genes, we propose a new, unified classification system called the "Process-Function" classification system, which is based on the cellular biochemical processes in which non-coding genes participate and the functions they perform. In the first version of the NONCODE database, we collected a total of 5,339 non-redundant records of all types of non-coding genes except transfer RNA (tRNA) and ribosomal RNA, involving 861 species across eubacteria, archaea, and eukaryotes [41].

To efficiently and comprehensively collect non-coding gene data, we designed a computer-automated analysis workflow with manual confirmation starting from PubMed (see Figure 1 [Figure 1: see original paper]). PubMed is an Internet biomedical information retrieval system developed by the National Center for

Biotechnology Information of the U.S. National Library of Medicine, covering abstracts and partial full texts from over 4,300 major biomedical journals in more than 70 countries worldwide. We search PubMed using a keyword table, and the retrieved literature is manually inspected to confirm its relevance to non-coding genes. By reading these non-coding gene-related literature, we obtain new non-coding gene keywords. Based on these new keywords, we update the keyword table and then use the new keyword table to automatically filter GB format files in GenBank. GenBank, established and maintained by the National Center for Biotechnology Information, contains all known nucleic acid and protein sequences, along with related literature and biological annotations. Each GB format file contains a brief description of the sequence, scientific nomenclature, taxonomic name, references, a sequence feature table, and the sequence itself. The sequence feature table includes annotations of biological features such as coding regions, transcription units, repeat regions, mutation sites, or modification sites. Based on these annotations in GB files and our non-coding gene-related keyword table, we can preliminarily screen potential non-coding genes and perform initial classification.

All GB files are divided into 16 categories including bacteria, viruses, primates, rodents, EST data, genome sequencing data, and large-scale genome sequence data. Our search mainly focuses on several categories in the nucleic acid database: eukaryotes, prokaryotes, bacteria, viruses, and viroids. The searched data are imported into a MySQL database awaiting manual inspection and confirmation. After manual confirmation as authentic non-coding gene data, a series of annotation procedures are performed. Similarly, the entire annotation process is basically completed automatically by computer, with only a few special cases requiring manual confirmation. Finally, based on this database, we established a user-friendly, comprehensive web interface (www.noncode.org) providing services including data browsing, keyword searching, online Blast sequence queries, and data downloads.

In existing non-coding gene nomenclature, some non-coding genes are named according to their cellular localization, such as small nuclear RNA (in the nucleus) and small nucleolar RNA (in the nucleolus) [29,42]; some are named according to function, such as pRNA (package RNA) and guide RNA [43,44]; and some are even named directly by sedimentation coefficient, such as 6S RNA and 5.3S RNA [45]. These different naming methods result in the same class of non-coding genes often having multiple names from different laboratories, and many names being identical for functionally unrelated non-coding genes. Based on the cellular biochemical processes in which non-coding genes participate and the functions they perform, we have developed a unified classification system. We hope this classification will avoid previous confusion and enable researchers to understand the function of a class of non-coding genes directly from its classification. In the NONCODE database's "Process-Function" classification system, cellular processes refer to biological reactions with DNA, RNA, and proteins as substrates, such as DNA replication and modification, RNA alternative splicing and methylation modification, protein transport and degradation, etc. Each

non-coding gene is named according to its function in a cellular process. The entire name is given by two to three levels of keywords connected by underscores. The first-level keyword is DNA, RNA, or Protein, representing which molecular type is the key component in a cellular process. The second keyword describes a specific process. If this process has more detailed branches, a third keyword is used to further explain the specific function. For example, the non-coding gene snRNA U1 participates in the mRNA splicing process, with RNA as the main molecule, the process being RNA processing, and the more detailed specific process being splicing. Therefore, snRNA U1 is assigned to the RNA_{{{processing}}*{{splicing}}} class. RNase P RNA participates in the 5' end maturation process of tRNA, cleaving the 5' end of tRNA precursors, and is therefore assigned to RNA*{{{processing}}_{{cleavage}}}. The "Process-Function" classification system is the first attempt to integrate the processes and functions in which non-coding genes participate into a single classification system. As our understanding of non-coding genes deepens, the NONCODE database content will be further expanded, and this classification system will be further improved to enable full utilization of the database. Details of the "Process-Function" classification system are shown in Table 1 .

2.2 Prediction and Validation of miRNA-encoding Non-coding Genes

In recent years, whole-genome microarray experiments and full-length complementary DNA (cDNA) library construction for several important model organisms have revealed numerous long non-coding transcripts in genomes. They share some similarities with protein-coding mRNAs: both are long, transcribed by RNA polymerase II, and undergo splicing, capping, and polyadenylation after transcription, but lack protein-coding open reading frames. Therefore, they are called "mRNA-like non-coding genes" [46-51]. The number of discovered mRNA-like non-coding genes is astonishing. For instance, approximately 4,000 full-length cDNAs lacking protein-coding open reading frames were found to be mRNA-like non-coding genes in the FANTOM mouse full-length cDNA library [52]; similarly, nearly 5,800 mRNA-like non-coding genes were discovered in the human full-length cDNA library [53]. The functions of a few mRNA-like non-coding genes have been confirmed. For example, Marahrens et al. found that knockout of the mRNA-like non-coding gene Xist on female mice affects X chromosome inactivation [54]; Young et al. found that microRNA interference of the mRNA-like non-coding gene TUG1 in newborn mouse retinal cells causes eye developmental abnormalities [55]; and Willingham et al. found that the mouse mRNA-like non-coding gene NRON is an inhibitor of the transcription factor NFAT [56]. However, the functions and mechanisms of the vast majority of mRNA-like non-coding genes remain unknown. MicroRNAs are tiny non-coding genes widely present in higher animals and plants that play important regulatory roles in life activities by controlling mRNA stability or inhibiting mRNA translation [57,58]. Based on genomic location, microRNAs can be divided into three categories: (1) microRNAs located in introns of protein-coding transcripts; (2) microRNAs located in introns of non-coding transcripts; and

(3) microRNAs located in exons of non-coding transcripts. We hypothesize that more mRNA-like non-coding RNAs encode microRNAs in their exons. They constitute a special class of non-coding RNAs, which we call "microRNA-encoding ncRNAs" (me-ncRNAs). In this paper, by analyzing 20 known me-ncRNAs encoding known microRNAs in the mouse genome, we designed a novel prediction method (PriMir) and used it to predict 65 new candidate me-ncRNAs among 34,030 microRNA-like ncRNAs in the FANTOM3 database, of which 24 were experimentally validated. We further analyzed these known and predicted me-ncRNAs and found they all contain conserved motifs. The me-ncRNAs discovered in our work represent a new class of non-coding RNAs. We also provide new interpretations for some mRNA-like non-coding RNAs with unknown functions.

We also used PriMir to search for me-ncRNAs from all mRNA-like non-coding RNAs. PriMir filters out stem-loop structures matching known microRNA precursors (pre-miRNAs) in sequence length and base pairing by scanning secondary structures of all mRNA-like non-coding RNAs; then filters out all stem-loop structures conserved between mouse and rat through conservation analysis; finally predicts all possible microRNA precursors and their corresponding me-ncRNAs through the PriMir score matrix (PMS matrix). Figure 2 [Figure 2: see original paper] shows a flowchart of the PriMir method. To establish the training set, we analyzed 270 pre-miRNAs from miRBase 8.0, filtering out cases where the stem-loop structure was too short (less than 45 nt) or the mature microRNA sequence position was special (on the loop of the stem-loop structure), yielding a training set of 220 known pre-miRNA sequences. We also needed to establish a background set composed of non-pre-miRNA stem-loop structures. We used RNAfold to predict secondary structures of all 34,030 mRNA-like non-coding RNAs in FANTOM3, and PriMir extracted stem-loop structures satisfying two conditions: (1) stem-loop sequence length greater than 45 nt; (2) number of paired bases in the stem-loop structure greater than 28. This yielded 184,000 stem-loop structures. These 184,000 stem-loop structures contain real unknown microRNA precursors that we need to identify and discover, but most are certainly non-microRNA precursor stem-loop structures. Therefore, we use these 184,000 stem-loop structures as the background set. After determining the training and background sets, we can establish the PMS matrix by analyzing differences in the values of 11 characteristic parameters between the training and background sets (see Figure 2). We further filtered the 184,000 stem-loop structures extracted from secondary structures of all 34,030 mRNA-like non-coding RNAs based on sequence conservation between mouse and rat. To determine the conservation threshold, we compared 220 known mouse pre-miRNAs from the training set with the rat genome using BLASTN. The results showed that 160 pre-miRNAs satisfied two criteria: (1) aligned sequence length exceeding 50 nt; (2) alignment identity value greater than or equal to 98%. Therefore, PriMir further filtered the 184,000 stem-loop structures according to these two standards, obtaining 4,463 stem-loops conserved between mouse and rat, including 18 known pre-miRNAs. PriMir then scored

the 4,463 conserved stem-loops using the PMS matrix. To reduce false positives, a PriMir score of "7" was used as the cutoff value. This is a strict criterion, as only 73% of the 220 known mouse pre-miRNAs in the training set scored above this value. MicroRNA precursors with PriMir scores greater than or equal to 7 were identified as candidate microRNA precursors. This yielded 84 candidate microRNA precursor genes from the 4,463 conserved stem-loops, corresponding to 80 possible microRNA precursors. Among them, 15 microRNA precursors belong to known microRNA precursors included in miRBase 8.0, corresponding to 15 known me-ncRNAs. Therefore, we call the remaining 69 stem-loops and their corresponding 65 mRNA-like non-coding RNAs candidate microRNA precursors and me-ncRNAs.

The sensitivity of the PriMir method can be assessed from its prediction performance on 20 known me-ncRNAs. Since we predicted 15 of them, the sensitivity should be above 75%. Because mRNA-like non-coding RNAs are often expressed in a tissue-specific or developmental stage-specific manner, and their expression levels are often low, estimating the specificity of predictions is more difficult. To estimate the specificity of PriMir predictions, we designed a microarray with 168 26-nt probes corresponding to both arms of the stems of the 84 predicted pre-microRNAs. To prevent interference from long RNAs during hybridization, we filtered out RNA molecules longer than 200 nt from the extracted total RNA. Microarray signals were hybridized with RNA extracted from newborn mouse brain and thymus tissues, 2-month-old male adult mouse brain, and 15-day mouse embryos. Signal analysis showed 46 probes with significant signals, corresponding to 46 different microRNAs, 40 different pre-microRNAs, and 39 me-ncRNAs (including 15 known me-ncRNAs). Significant signals were detected from both arms for 6 pre-microRNAs. Among the 15 microRNAs included in miRBase 8.0, 14 showed significant signals, indicating our microarray worked well. For the 32 new microRNAs detected by microarray, we searched miRBase 9.0 and found 5 had been newly included. We further verified the remaining 27 new microRNAs using Stem-loop RT-PCR plus sequencing, and results showed all new microRNAs were authentic. Thus, 24 of the 65 me-ncRNA candidates passed our strict experimental verification. While our work was in progress, another 10 microRNAs were discovered by other laboratories. These microRNAs corresponded to 5 me-ncRNA candidates and 4 known me-ncRNA genes in our work. One me-ncRNA candidate passed our PriMir prediction but was not detected by our microarray and RT-PCR plus sequencing methods. Therefore, if we count both our own microarray and RT-PCR plus sequencing experimental validation and support from published literature by other laboratories, the specificity of the PriMir method should be 50% ((39+1)/80). Of course, if we examined more mouse tissues and developmental stage samples in our experiments, we believe we would obtain higher specificity scores.

Furthermore, we analyzed sequence conservation and sequence motifs of me-ncRNAs. To measure me-ncRNA conservation, we scored each base of me-ncRNAs based on mouse genome PhastCons scores relative to 17 vertebrates, then used the average of all base scores for an me-ncRNA (average PhastCons

scores, APCSs) as its conservation score. The APCSs of 65 candidate me-ncRNAs averaged 41%, significantly higher than the APCS mean of 20 known me-ncRNAs (26%). The low conservation of these 20 known me-ncRNAs may be due to statistical fluctuations because 20 sequences are too few. Another possible reason is that me-ncRNAs do not require conservation across the entire sequence. Therefore, we further analyzed the sequence conservation of the microRNA precursor portion of me-ncRNAs. The results showed that the average APCSs of microRNA precursor portions of known and predicted me-ncRNAs were 88% and 72% respectively, both substantially higher than the conservation scores of entire me-ncRNA sequences. By analyzing known and predicted me-ncRNAs, we identified an internal motif IM1 within me-ncRNAs with the conserved sequence CNCTUNCTU (see Figure 4 Figure 4: see original paper). We constructed a positional weight matrix (PWM) based on IM1 and used this matrix to search all mRNA-like non-coding gene sequences. The matrix score threshold was set to ensure IM1 appeared on 50% of confirmed me-ncRNAs (20 known plus 24 experimentally confirmed). The results showed IM1 was present in 65% of known me-ncRNAs and 42% of predicted me-ncRNAs, while only 23% of all mRNA-like non-coding genes had IM1 (see Figure 4(b)). Because the proportion of internal motif IM1 in mRNA-like non-coding genes is relatively high, we further analyzed the relationship between IM1 frequency in sequences and PriMir scores. We used the highest PriMir score among all stem-loop structures on an mRNA-like non-coding RNA as the score for that RNA. Analysis of all mRNA-like non-coding RNAs revealed a strong correlation between the number of IM1 occurrences in a sequence and its PriMir score (R-squared = 0.91, p-value = 2.2e-16) (see Figure 4(b)), meaning the higher the likelihood of encoding microRNAs (PriMir score), the higher the likelihood and number of IM1 occurrences. On the other hand, more conserved stem-loop structures are more likely to be authentic microRNAs. Therefore, we reasoned that if IM1 is indeed related to whether a sequence encodes microRNAs, the correlation between IM1 and PriMir scores would decrease after adding conservation constraints, because under strict conservation requirements, low-scoring stem-loop structures by PriMir might still encode microRNAs and thus still contain many IM1 motifs. Further analysis results were consistent with our prediction. In the subset of 3,670 mRNA-like ncRNAs with conserved stem-loop structures, the correlation between IM1 and PriMir scores decreased significantly (R-squared = 0.1, p-value = 0.03) (see Figure 4(b)), while the proportion of sequences containing IM1 motifs was higher than in the entire mRNA-like non-coding RNA set. Based on the above analysis, we conclude that the IM1 motif is indeed related to me-ncRNA encoding of microRNAs. The discovery of IM1 will facilitate further prediction of me-ncRNAs, while the function of IM1 in the process of me-ncRNA encoding microRNAs requires further in-depth study.

3 Complex Biological Networks and Systems Biology

Systems biology views living organisms as complex, dynamically changing networks. The study of life requires a holistic approach, investigating these com-

plex interactions and regulatory relationships within biological networks rather than merely studying individual nodes in isolation [59]. Initially, due to experimental technology limitations, systems biology research mainly remained at the theoretical level of computer simulation systems. With the determination of whole genomes of a series of organisms and the emergence of various high-throughput experimental technologies such as whole-genome microarrays, yeast two-hybrid, and chromatin immunoprecipitation microarrays [60-62], global observation of biological network topology and quantitative changes of nodes in networks has become possible: by analyzing whole-genome sequences and predicting protein-coding and non-coding genes, we can rapidly identify most nodes in biological networks; through yeast two-hybrid technology for large-scale detection of protein-protein interactions, we can find undirected edges in biological networks; through chromatin immunoprecipitation microarray technology for detecting transcription factor-specific binding sites on chromatin, we can identify directed edges; and finally, through whole-genome microarrays and other gene chip technologies, we can dynamically and quantitatively observe expression levels of nodes in biological networks. Although these high-throughput experimental technologies inevitably contain various noise and measurement biases, they still provide a foundation for systems biology research. Systems biology has now become a new research paradigm in life sciences. Starting from multi-data source integration and based on network analysis, it uses various means including statistics, informatics, and artificial intelligence to make predictions about various life phenomena and guide traditional biological experiments to verify these predictions. This new paradigm has greatly promoted progress in life sciences research, ushering post-genomic era systems biology research into a new period of rapid development.

The completion of the Human Genome Project is a major step forward in life science development. The next step is functional genomics to study the functions of deciphered genes and control them, ultimately serving humanity' s conquest of nature and defeat of diseases. As Robert Tepper of Millennium Pharmaceutical said, "We know what' s in the dictionary; now we need to know what each word means." Although 99% of gene sequences have been deciphered, only 10% of gene functions are known. How to obtain functions of more genes has become the main topic of functional genomics. For a long time, gene function research has been conducted on single genes following the "sequence → structure → function" paradigm, assuming one gene expresses one protein, one protein has one structure, and one structure performs one function. Relative to post-genomic era functional genomics goals, this "one gene at a time" research model is not only completely inadequate in efficiency but more importantly cannot reveal the complexity and essence of life activities. Increasing evidence shows that expression of a single gene alone often cannot dominate the occurrence of a biological event. Biological functions are generally achieved through simultaneous expression of a group of genes and synergistic action of a group of proteins. In a biological event, complex gene transcriptional regulatory networks control simultaneous expression of related genes, and various protein-protein interac-

tion networks exist, even including RNA interactions. Therefore, changing the original "one gene at a time" research approach and "sequence → structure → function" thinking, adopting a systems biology perspective with the new "interaction → network → function" paradigm to integrate different aspects and levels of information about genes and proteins for gene function analysis has become the new direction of functional genomics research. In coding gene research, studies based on protein interaction networks and gene transcriptional regulatory networks have demonstrated the great power of network research: finding functional modules through network clustering, predicting protein functions based on network neighbor nodes, and studying the topological structure and signal transduction characteristics of network motifs. These network-based studies have become new tools in biological research.

3.1 Spectral Analysis Methods for Protein Interaction Networks

A major challenge in the post-genomic era is understanding how gene information leads to synergistic interactions among gene products and how they perform biological functions in time and space to ultimately interact and form an organism. Therefore, developing reliable proteomics methods to better understand protein function is crucial. Genomics methods have been used to predict functions of numerous genes based on sequence features. However, proteins rarely work alone at the biochemical level but interact with other proteins to form complexes to accomplish specific cellular tasks. System functions are richer than the functions manifested by their individual parts. Traditionally, protein interaction research studies some proteins at a certain moment from genetic, biochemical, and physiological perspectives. We now realize that this jigsaw puzzle-like research of genetic and biochemical pathways in cells has hindered further understanding of cells as integrated biological processes. Protein complexes, cellular pathways, and protein interactions are the basic components that are decisive for protein function. Therefore, it is certain that all biological processes are more precisely manifested through protein interactions. In the past three years, high-throughput interaction detection methods have been developed, such as yeast two-hybrid systems, mass spectrometry-based protein purification methods, expression profile analysis with related information, genetic interaction network methods, and other interaction prediction methods based on gene correlation computational models (gene fusion and fission, gene neighborhood, and co-occurring genes). These have generated considerable large-scale protein interaction data in several organisms (such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Helicobacter pylori*) [63-75]. These high-throughput large-scale data have opened a door to more comprehensive understanding of genetic and biochemical phenomena in cells. Subsequently, several methods have been successfully applied in this research area. For example, Schwikowski et al. and Hishigaki et al. successfully used interaction neighbors to predict functions of unknown proteins. Ge et al. first provided evidence that proteins with similar expression profiles tend to have protein interactions. Fraser et al. revealed that proteins with conserved interactions show negative correlation with

their mutation rates [76-79]. All these studies suggest that the interaction network of *S. cerevisiae* may have properties different from other complex networks. Topological patterns of interactions are important starting points for studying protein biological function information, so we need to develop methods to mine and understand interaction networks. Here we apply spectral analysis methods, which have been successfully used in other fields, to proteomics research to identify topological patterns in protein interaction networks, namely quasi-cliques and quasi-bipartite graphs. Interestingly, we found that proteins in the same group have similar protein functions. More importantly, for nearly one-third of proteins with unknown functions in *S. cerevisiae*, this method provides a means to predict protein functions based on protein structure.

Spectral analysis is an effective method for revealing deep structures of massive complex data relationships. As a famous example, David Gibson, Jon Kleinberg, and Prabhakar Raghavan did excellent work in information mining of World Wide Web link structures [80,81]. The World Wide Web consists of an increasing number of web pages linked to other pages through hyperlinks. In addition to the complexity of web structure, spectral analysis methods have successfully discovered data information sources such as "authoritative points" and "hubs." We apply spectral analysis methods to complex protein-protein interaction networks to identify interesting topological structures. In this method, the network is represented as an undirected graph G(V,E), where the node set contains each protein as a node: V = {v_i}, and the edge set is defined as E = {(v_i, v_j)} for interacting proteins v_i and v_j. A symmetric n×n adjacency matrix can be defined as A = (a_{ij}). The spectrum of the adjacency matrix A is essentially an important measure of node attributes that can be transmitted through interactions. Each node can be given a score x_i to represent its "importance." A node with a high score will increase the scores of its linked nodes through interactions, meaning the scores of two interacting nodes reinforce each other, which can be defined cyclically. Gibson et al.' s iterative algorithm introduces a method to interrupt this cycle. Interestingly, regardless of initial values, x_i will converge to a specific point. It can be proven that this point is an eigenvector of matrix A. This proves that its properties are an essential attribute of interactions. Moreover, because matrix A is symmetric, all eigenvectors are orthogonal. This means the corresponding attributes are also orthogonal. In other words, each eigenvector may represent a special property that other vectors do not have. From a topological perspective, the graph spectrum helps reveal hidden structures of complex interaction networks. We found that for each eigenvector corresponding to a positive eigenvalue, components with larger absolute values tend to form quasi-cliques (i.e., positive and negative ends each form a group tending to have internal links) (see Figure 5a [Figure 5: see original paper]), while for each eigenvector with negative eigenvalues, such proteins tend to form quasi-bipartite graphs (i.e., proteins that are not connected internally between positive and negative ends form a structure tending to be tightly linked) (Figure 5b).

There are many false positives in high-throughput interaction data. To measure

the accuracy of these data and determine their bias, von Mering et al. evaluated over 80,000 interactions among more than 5,400 proteins from published literature, giving each interaction a confidence score [82]. To reduce false positives in data, we focused our analysis on medium- and high-confidence data, including 11,855 interactions among 2,617 proteins. To analyze interaction data, we first applied spectral analysis methods to compute all eigenvalues and eigenvectors of the adjacency matrix of the corresponding network. Quasi-cliques were generated on eigenvectors with large positive eigenvalues using the following criteria: (1) all proteins were sorted by the absolute value of their eigenvalues, and the top 10% of eigenvectors were selected for analysis; (2) proteins were added in sorted order, with newly added proteins required to have interactions with at least 20% of existing proteins. Here we use the clustering coefficient CC to measure the degree of connectivity between nodes, adjusting parameters to ensure quasi-clique properties; (3) quasi-cliques must contain at least 10 proteins. Using these criteria, we obtained 48 quasi-cliques. The largest contained 109 proteins, while the smallest contained 10 proteins, with an average of 26.6 proteins (a protein can appear in multiple cliques). Similar analysis on eigenvectors with negative values yielded 6 quasi-bipartite graphs. These two topological patterns show different interaction spectra. In quasi-cliques, proteins tend to interact with themselves (see Figure 5a), while in quasi-bipartite graphs, two sets tend to have interactions between them but no interactions within each set (see Figure 5b). Identification of these two patterns not only makes representation of complex interaction networks more orderly but more importantly provides a more convenient means to analyze complex networks. An isolated quasi-clique includes different biological functions. The P-value method can serve as a criterion for assigning primary functions to quasi-cliques. The hypergeometric distribution can calculate the probability of a protein group having a certain function. For a clique with n proteins containing k proteins with a certain function, assuming the protein group has G proteins and the functional class has C proteins, the P-value of random occurrence of such a clique is:

$$P = \Sigma_{i=k}^{min(n,C)} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}$$

This standard describes the abundance of a specific functional class in protein quasi-cliques compared to random occurrence. If the P-value abundance approaches zero, it indicates the probability of randomly selecting such proteins in the quasi-clique would be very low. Here we take the function with the lowest P-value among all functional classes in each quasi-clique as the primary function of that clique. For each of the 48 quasi-cliques, we annotated them using hierarchical functional annotations from the Munich Information Center for Protein Sequences (MIPS) and calculated P-values for functional annotations. In the P-value calculation process, MIPS annotations allow a protein to have more than one function. The results showed that 43 quasi-cliques could be assigned one function, while the other 5 could be assigned several functions. Functional

analysis of individual proteins in quasi-cliques found that most proteins tend to share a common function, such as ribosome biogenesis, ribosomal RNA and transfer RNA synthesis, processing, transcription control, and mRNA splicing. Only a small portion of proteins are either unannotated or have functions conflicting with the primary function in the quasi-clique (see Figure 6 [Figure 6: see original paper]).

The separated quasi-cliques provide good clues for predicting functions of unannotated proteins. In the original data of 2,617 proteins, 555 proteins had no annotation in MIPS hierarchical functional classification. Among the 48 quasi-cliques, 76 unannotated proteins were included. We predicted the functions of each such protein using the primary function of the clique they belonged to. If a protein fell into multiple cliques, we used the clique with the smallest P-value for prediction; if multiple cliques had the smallest P-value and the cliques had multiple functions, we assigned multiple functions to the unknown protein. Among them, 43 proteins were related to rRNA processing; 7 were related to pre-RNA processing; 11 proteins were related to ribosome biogenesis; and the other 15 proteins were related to energy, metabolism, cytoskeleton, and transcription regulation, respectively.

We used the computational P-value method of Lani F. Wu et al. to evaluate functional annotation methods. As a control, we generated and analyzed random network data with the same degree distribution as the original network. The results showed that more than 87.5% of functional class annotations in the 48 cliques from our experimental data analysis were meaningful (i.e., $P < 0.01/(cN)$, where cN is the total number of functional classes), while in the control group of quasi-cliques generated from random networks, only 2.1% of cliques met this standard. This means that a considerable portion of the separated quasi-cliques may have biological significance. Some of our predictions have been confirmed by recent experimental evidence. Among all these quasi-cliques, five are occupied by uncharacterized proteins (i.e., unknown proteins account for at least 50% of all proteins). This suggests that these unknown proteins under the same quasi-clique may form a complex related to a specific cellular process. As shown in Figure 7 [Figure 7: see original paper], in our predicted quasi-clique, most proteins are related to rRNA processing according to our prediction, which partially matches recent experimental results.

3.2 Analysis of Regulatory Pattern Preferences in Transcriptional Regulatory Networks

Transcriptional regulatory networks control the expression levels of all genes in cells, and their study is an important issue in the post-genomic era. With rapid advances in related experimental technologies, transcriptional regulatory networks of multiple model organisms have been experimentally determined [83,84]. We can simply view transcriptional regulatory networks as directed graphs where transcription factors (TFs) and transcription target genes (TGs) are represented as nodes. The regulatory effect of a transcription factor on its

regulated gene is represented as a directed edge from the transcription factor to the regulated gene. The regulatory relationships between transcription factors and their target genes appear as subgraphs with multiple nodes in the graph. Some subgraphs have been widely studied due to their obvious biological implications in topology [83-85], such as feed-forward loops, feedback loops, single input motifs, and multi-input motifs (see Figure 8 [Figure 8: see original paper]). These subgraphs, or regulatory patterns, may contain specific regulatory capabilities. For example, single input motifs may regulate a group of functionally related genes, while feed-forward loops may play a role in temporal control in certain biological processes. However, these subgraphs are not independent functional units unconnected to other parts of the network. In fact, these subgraphs tend to cluster around network hub transcription factors with high connectivity. Thus, a transcription factor often becomes a member of multiple different pattern subgraphs. Globally, topological analysis of regulatory networks shows that the number of genes regulated by transcription factors follows a power-law distribution. This means a small portion of transcription factors regulate most genes in the transcriptional regulatory network. These highly connected transcription factors are called network hub transcription factors (THubs). Research shows these network hub transcription factors are usually essential key genes in organisms [86-88]. Transcriptional regulatory networks can be viewed as network structures for signal transmission (such as external nutrients, environmental stress), and different transcription factors should exhibit different behaviors when transmitting signals in the network [89]. Similar situations were recently found in mammalian signal transduction networks, where three important ligands at different network layers used different subgraphs to function [90]. On the other hand, in transcriptional regulatory subnetworks under different external conditions or developmental stages, the overall density of different regulatory patterns in the network also varies [91]. However, so far, no one has understood at the genomic scale how this different pattern density affects each transcription factor; nor are there methods to measure transcription factor behavior when regulating their downstream target genes. Therefore, we hope to design a set of methods to measure and represent transcription factor preferences for different regulatory patterns in regulatory networks.

To calculate the network subgraph preference spectrum, we first need to determine our studied transcription factor set and subgraph set. Our studied transcription factors mainly focus on the hub transcription factor set. For complex networks with power-law degree distributions like transcriptional regulatory networks, most nodes are connected to only a few other nodes, while a small number of nodes are connected to many other nodes. Such nodes are called hub nodes. In transcriptional regulatory networks, we call these hub nodes hub transcription factors, which regulate significantly more genes than most other transcription factors in the network. These hub transcription factors are usually key genes in organisms. Hub nodes are typically determined by taking inflection points in node distribution curves as thresholds, with nodes having connectivity greater than this threshold defined as hub nodes. For basic regulatory patterns,

we selected two categories: ring patterns and tree patterns (see Figure 9 [Figure 9: see original paper]).

We determine the type of a regulatory pattern based on whether its topology is open or closed. Regardless of regulatory direction, if a regulatory pattern topologically forms a single closed loop, it is called a "ring." If a regulatory pattern contains no loops as subgraphs, it is called a "tree." We only consider rings and trees with three and four nodes because at the two-node level, T2-1 is a trivial regulatory structure, while R2-1 is a very rare pattern in networks. For regulatory patterns with more than four nodes, they are not included in our study due to computational limitations. We selected these as basic regulatory patterns not only because their mutual similarity is sufficiently small but also because these regulatory patterns cover all basic patterns, meaning all other patterns can be generated from combinations of these basic patterns. After selecting transcription factor sets and subgraph sets, for each given transcription factor H and subgraph P, we define the usage preference w_A of this transcription factor for this subgraph as:

$$w_A(H,P) = \Sigma_{sg \in SG(H,P)} \Sigma_{k \in N(sg)} \frac{1}{(d(H,k)+1)}$$

where SG(H,P) is the set of all instances of subgraph P downstream of transcription factor H, N(sg) is all nodes in subgraph instance sg, and d(H,k) is the shortest distance between transcription factor H and its downstream gene k in the network. The weighting factor 1/(d(H,k)+1) appears in the formula to quantify the characteristic that the influence of a transcription factor on downstream genes gradually decreases with increasing distance.

For selected transcription factor sets and subgraph sets, usage preferences can be calculated for each pair of transcription factor and subgraph. However, because different transcription factors regulate downstream regions of different scales and different subgraphs have vastly different overall abundances in the network, these usage preference values between different transcription factors and subgraphs cannot be directly compared. To make these usage preference values comparable, we use the following formula to eliminate the influence of these two factors:

$$A'(H,P) = \frac{A(H,P) - \mu_A(H,P)}{\sigma_A(H,P)}$$

where A' (H,P) is the normalized usage preference of transcription factor H for subgraph P, _A(H,P) is the mean of A(H,P) over all transcription factors and subgraphs, and σ_A(H,P) is the standard deviation. This yields the "normalized usage preference" of all hub transcription factors for all subgraphs. We define the "subgraph preference profile (SPP)" of a transcriptional regulatory factor as a vector composed of normalized usage preference values of this transcription

factor for all subgraphs. We define the matrix composed of subgraph preference profiles of all studied transcription factors as the "subgraph preference landscape (SPL)" of this network.

We can visualize the subgraph preference landscape as a grayscale map (see Figure 10 [Figure 10: see original paper]). Points that appear very black in the map indicate that transcription factors have preferences for these regulatory patterns different from other patterns. However, because studies have shown that global network structure and some local patterns are mutually determined, we cannot simply determine from this grayscale whether transcription factors truly preferentially use a regulatory pattern or whether it is just a common property of networks with this abundance distribution. To provide significance determination for network subgraph preferences, for a specific network with a particular abundance distribution, we examine a cluster of random networks generated from this network, making the generated random networks have the same out-degree and in-degree distributions as the real network. By calculating the subgraph preference landscape in the random network cluster, we can obtain the distribution of regulatory pattern preferences in the random network cluster. Because transcriptional regulatory network subgraph preference values roughly follow a power-law distribution, we use a z-score-based algorithm to determine the significance threshold for preferential use of a subgraph: from the random network cluster, we remove data with subgraph preference z-scores greater than or equal to 2, then based on the distribution of remaining random network cluster subgraph preference values, we take z-score greater than or equal to 2 as the threshold for determining significant preference. Regulatory patterns with subgraph preferences greater than this threshold are considered to be significantly preferentially used by the corresponding transcription factor. To provide statistical significance for this preference, we compare the real network with 1,000 randomly generated networks. The p-value for the preference of each transcription factor and network subgraph pair is given by the percentage of this preference value being smaller than the corresponding preference value in the 1,000 random networks.

For a network-based quantitative computational analysis method, we need to consider not only whether results are significant relative to random networks but also whether results have a certain degree of robustness and stability. This issue is particularly important considering the special nature of transcriptional regulatory networks and the relatively large noise from high-throughput experiments. For robustness analysis, we examine three aspects:

1. **Robustness to network node scale**: We randomly knock out and add some nodes in the network, then observe the impact of node addition/deletion on the subgraph preference landscape.
2. **Robustness to noise**: We add certain connection noise to the network, including randomly adding, deleting, and swapping some connections, then observe the impact of noise on the subgraph preference landscape.

3. **Robustness to downstream gene number**: We change the original threshold to obtain a new set of hub transcription factors, then observe the impact on the subgraph preference landscape.

Additionally, in our studied network, hub transcription factors are identified using inflection points in their degree distribution as thresholds. This threshold selection has no absolute standard, so we need to examine the influence of threshold selection on the subgraph preference landscape. For each network, we increased and decreased the selected threshold by 1 to obtain two new sets of hub transcription factors. Using these different hub transcription factor sets, we recalculated their subgraph preference landscapes. For the above three situations, we generated reference network clusters relative to the original network and calculated subgraph preference landscapes for these networks. For each given original network and its corresponding network cluster, we determined robustness to corresponding operations by comparing their subgraph preference landscapes. The comparison method is: for each pair of subgraph preference landscapes (original network and a network in the corresponding network cluster), we calculate the Euclidean distance between corresponding vectors in all network subgraph preference profiles to obtain a distribution of distances between network subgraph preference profiles. Based on the distribution of distances between subgraph preference profiles within the reference network cluster, we can obtain thresholds for determining whether there are significant differences between network subgraph preference profiles. Based on these thresholds, we can determine the p-value for significant differences between the original network subgraph preference landscape and the corresponding network cluster subgraph preference landscape. Through the above robustness analysis, we found that the analysis results of subgraph preference profiles and landscapes have good stability for various node and edge noise and changes in transcription hub factor definition thresholds.

Furthermore, we examined the relationship between hub transcription factors and their downstream regulatory pattern preferences in the *S. cerevisiae* transcriptional regulatory network. Our analysis included six conditions: static, cell cycle, sporulation, diauxic shift, DNA damage, and stress response. The static condition network is the full network, while the other five networks are subnetworks under various conditions. Network data came from http://sandy.topnet.gersteinlab.org/, with self-interaction edges removed from the network. Similar to previous findings in *E. coli*, the yeast transcriptional regulatory network is also a multi-layer hierarchical structure. The yeast static regulatory network has 14 layers, while the subnetworks under cell cycle, sporulation, diauxic shift, DNA damage, and stress response conditions have 13, 14, 9, 9, and 7 layers, respectively. When we arranged hub transcription factors in the subgraph preference landscape according to their order in the hierarchical structure, we observed a general trend in all subgraph preference landscapes: transcription factors in upper parts of the network have more complex subgraph preference profiles than those in lower parts (Figure 10).

Furthermore, we first analyzed characteristics of subgraph preference landscapes under various conditions, then performed comparative analysis of subgraph preference landscapes under different conditions. Figure 10 shows the regulatory pattern preference landscape for each network, where grayscale in each square reflects preference degree, with darker indicating higher preference, and boxes indicating significant parts. We can clearly see that different transcription factors in regulatory networks preferentially use different regulatory patterns. We also note that although single input motifs (T3-3 and T4-7) are the only regulatory patterns appearing in all networks and at all levels, no transcription factor prefers to use this regulatory pattern in any of the 6 transcriptional regulatory networks. In contrast, for feed-forward loops (R3-1), there are transcription factors that preferentially use this regulatory pattern at various levels in the network. In cell cycle and sporulation subnetworks, feedback loops (R3-2, R4-1) are preferentially used by high-level transcription factors in the network, while in the other three subnetworks (diauxic shift, DNA damage, and stress response), feedback loops are more preferentially used by lower-level transcription factors. Some regulatory patterns have high relative abundance in the network that cannot explain transcription factors' preferential use of these patterns. For example, studies in regulatory networks have found that feed-forward loops, feedback loops, and single input motifs are relatively high-frequency regulatory patterns, called network motifs. However, as mentioned above, single input motifs are not preferentially used by any transcription factor in any network we examined. Conversely, some patterns that are not significantly high-frequency, such as T3-1 and T3-2, are preferentially used by some transcription factors. High significant use of a regulatory pattern by a transcription factor may be the result of clustering of that pattern around the factor. For example, in the cell cycle regulatory subnetwork, transcription factor YLR013W highly preferentially uses feed-forward loops. Careful examination of YLR013W' s upstream and downstream regions in the regulatory network shows 4 feed-forward loops forming a symmetric grid pattern (see Figure 11 Figure 11: see original paper). However, not all high preferential use can be explained by clustering. In the integrated network, we detected 3 cases of feedback loops (R3-2). These feedback loops not only cluster together but also connect to form a large feedback loop, and all nodes in this large feedback loop are hub transcription factors (see Figure 11(b)). Nevertheless, there are still hub transcription factors in this large feedback loop that do not preferentially use feedback loops (such as YGL073W), and only one hub transcription factor, YBR049C, preferentially uses feedback loops in all networks. Since transcription factors' preferential use of regulatory patterns cannot be fully explained by high abundance of these patterns nor by high clustering of these patterns in local regions, we believe this suggests that our defined "preferential use" reflects some important behavioral preferences of these hub transcription factors in transcriptional regulatory networks under certain growth or cellular conditions.

When we compared subgraph preference landscapes of five subnetworks under different conditions, we observed dynamic changes in preferences for regulatory

patterns across different networks. First, we used the Kolmogorov-Smirnov test to test whether regulatory pattern preference values in subgraph preference landscapes of subnetworks under different conditions came from the same distribution. In diauxic shift, DNA damage, and stress response subnetworks (also called exogenous networks), the three-node subgraph preference landscapes clearly differed from endogenous networks (cell cycle and sporulation) (see Table 2). However, the sporulation subnetwork and diauxic shift condition subnetwork were somewhat similar, perhaps reflecting some exogenous characteristics during sporulation. Among exogenous networks, their subgraph preference landscapes basically came from the same distribution. In contrast, between the two endogenous networks, the distributions of preference p-values in their regulatory pattern preference landscapes were significantly different. We observed similar situations in four-node subgraph preference profiles. Second, by comparing Euclidean distances between subgraph preference profiles of transcription factors within the same layer, we found that similarities between subgraph preference profiles within the same layer differ across different regulatory networks. For three-node regulatory patterns, transcription factors in the same layer in integrated, cell cycle, and stress response networks tend to have more similar network subgraph preference profiles (see Table 3), while transcription factors in the same layer in the other three networks tend to have different network subgraph preference profiles. At the four-node level, transcription factors in the same layer in integrated, cell cycle, and diauxic shift subnetworks tend to have more similar pattern preference profiles, while transcription factors in the same layer in the other three networks tend to have more different preference profiles (see Table 4). Because preference profiles of bottom-layer transcription factors are simple and obvious, containing only single input motifs, to remove the influence of this obvious bias, we excluded bottom-layer transcription factors when analyzing intra-layer pattern preference similarities. Finally, we examined dynamic characteristics of nine transcription factors identified as hub transcription factors in subnetworks under five conditions (see Figure 12 [Figure 12: see original paper]). For three-node regulatory patterns, among these nine transcription factors, only YLR013W' s preference profile changed significantly between the two endogenous networks, although the distributions of preferences in the two endogenous networks were completely different. In contrast, among the three exogenous networks, four hub transcription factors (YMR043W, YJR060W, YKL043W, YLR013W) had significantly changed preference profiles, although the distributions of preferences in the three exogenous networks were similar. At the four-node level, we observed more dynamic changes in preference profiles. These dynamic changes may reflect dynamic transformations of biological functions performed by transcription factors during changes in external environment or growth status. For example, in cell cycle and stress response transcriptional regulatory networks, YJR060W required for nucleosome function preferentially uses three-node and four-node feedback loops [92], while in DNA damage transcriptional regulatory networks, YJR060W switches to preferentially using feedforward loops [93]. As gene regulatory patterns similar to clocks and oscillators, feedback loops may regulate cell growth rates. Therefore, the preferential use

of feedback loops by this transcription factor in cell cycle and stress conditions suggests its role may be a clock controller or frequency regulator, while in DNA damage processes, YJR060W may function as a signal amplifier. This example demonstrates that comparing changes in subgraph preference landscapes under different conditions can reveal hidden biological functions that transcription factors perform in different external environments.

4 Complex Biological Networks Involving Non-coding RNA

The vast majority of life activities involve complex interactions among many biological molecules (including genes, proteins, and other biomolecules). These interactions actually manifest as network relationships, with biological functions emerging from these network interactions. In other words, the complexity of organisms originates not only from the enormous number and types of macromolecules but more importantly from the intricate relationships between macromolecules. Complex networks manifest as interaction relationships at three levels: first, gene transcriptional regulatory networks, manifested as interactions between transcription factors and their binding sites upstream of regulated genes during gene transcription; second, protein interaction networks; and third, post-transcriptional regulatory networks, particularly interactions between microRNAs and their regulatory sequences. The addition of large numbers of non-coding RNAs has a huge impact on the complexity of biological networks. First, studies have shown that non-coding RNAs are abundant in higher organisms. ENCODE project results indicate that 93% of the human genome is transcribed into RNA, with more than half being non-coding RNA [94]. Therefore, the addition of non-coding RNAs will expand network scale exponentially. More importantly, the addition of non-coding RNAs introduces various new interaction mechanisms, playing important roles in gene transcriptional regulation, post-transcriptional regulation, and modification. For example, microRNAs, tiny non-coding RNAs widely present in higher animals and plants, play important regulatory roles in life activities by controlling mRNA stability or inhibiting mRNA translation [10]. Recognizing the significant impact of non-coding RNAs on biological networks, research on complex biological networks involving non-coding RNAs has also been initiated. Incorporating non-coding RNAs into biological network research helps us better understand biological networks and greatly enriches our understanding of entire biological networks. Non-coding RNAs can interact with specific proteins to form various complexes, functioning as RNA-protein complexes. For example, snRNAs U1, U2, U4, U5, and U6 form splicing complexes with up to 75 proteins to responsible for pre-mRNA splicing [29]; mouse NRON RNA binds with 11 proteins to control NFAT protein transport [95]. Non-coding RNAs can also locate targets by matching target RNA sequences and further recruit functional proteins to perform functions. For example, C/D box snoRNA locates sites on ribosomal RNA through complementary sequences and recruits proteins to methylate ribosomal RNA [25]; microRNAs locate specific mRNA 3' UTR regions through their seed sequences and control mRNA stability or inhibit mRNA translation

through recruited RNA-induced silencing complex (RISC) proteins [10]. Incorporating non-coding RNAs into network research also helps us study the functions of non-coding RNAs themselves. In coding gene research, studies on protein interaction networks and gene transcriptional regulatory networks have demonstrated the great power of network research: finding functional modules through network clustering and predicting protein functions based on network neighbor nodes. These network-based studies have become new tools in biological research. We believe these analysis methods that have proven very successful in network research will certainly help us better predict functions of non-coding RNAs.

MicroRNAs are tiny non-coding genes widely present in higher animals and plants that play important regulatory roles in life activities by controlling mRNA stability or inhibiting mRNA translation. MicroRNA primary transcripts (primiRNA) transcribed from chromosomes are processed into pre-microRNAs by RNase Drosha in the nucleus. Pre-microRNAs are then transported to the cytoplasm by Exportin-5/Ran-GTP, where they are further cleaved by RNase Dicer into approximately 22 base pair microRNA duplexes. This duplex is unwound by an RNase, and one strand combines with proteins to form the RNA-induced silencing complex (RISC), which inhibits protein synthesis or degrades target genes through partial complementary pairing with mRNA 3' UTR. Several research groups have conducted large-scale predictions of microRNA targets based on known microRNA regulatory features, with results showing that nearly one-third of human genome genes are regulated at the post-transcriptional level by microRNAs, with each microRNA regulating hundreds of coding genes on average [10]. Numerous researchers are working to establish post-transcriptional regulatory networks through microRNAs and their corresponding target genes and to study the biological functions of microRNAs and their target genes through microRNA post-transcriptional regulatory networks. MicroRNA target genes discovered in existing studies are all protein-coding genes, i.e., mRNAs. We hypothesize that microRNAs may regulate a special class of non-coding RNAs —mRNA-like non-coding RNAs—at the transcriptional level, forming a post-transcriptional regulatory network for non-coding RNAs (see Figure 13 [Figure 13: see original paper]). Most known non-coding genes are relatively short, but several important whole-genome microarray experiments and full-length cDNA library constructions in recent years have found large numbers of long non-coding transcripts in genomes. They share some similarities with protein-coding mRNAs: both are long, transcribed by RNA polymerase II, and undergo splicing, capping, and polyadenylation, but lack protein-coding open reading frames, and are therefore called mRNA-like non-coding genes [17]. The functions of a few mRNA-like non-coding genes have been confirmed, but the functions and mechanisms of the vast majority remain unknown. Due to sequence and structural similarities between mRNA-like non-coding RNAs (mlRNA) and mRNAs, we believe mRNA-like non-coding RNAs are also likely targets of microRNAs.

To verify our hypothesis, we borrowed methods used to verify microRNA regulation of mRNAs [96]. Recent studies have shown that microRNAs can accelerate

target RNA degradation, so microRNA target gene RNA levels can be detected by gene chips to evaluate prediction reliability. We selected 34,000 mRNA-like non-coding RNAs collected in the FANTOM database as our research objects. About 11,000 of these 34,000 sequences have gene expression profile data in 20 tissues. We selected 8 microRNAs with confirmed tissue-specific expression as our microRNA research set (see Table 5). Because microRNAs can significantly downregulate target gene RNA levels, for tissue-specifically expressed microRNAs, expression levels of their target genes should be significantly lower in their specifically expressed tissues than in other tissues. For the 8 microRNAs we analyzed, we predicted their target genes on the set of 11,000 mRNA-like non-coding RNAs based on miRanda prediction results and target site sequence conservation. We then performed Wilcoxon' s rank sum test on expression profiles of predicted target genes, with results shown in Table 5. The results showed that 3 microRNAs had significantly downregulated target gene expression levels in 4 specifically expressed tissues (see Figure 14 [Figure 14: see original paper]), with significance levels comparable to microRNA regulation levels on mRNAs, verifying our hypothesis that microRNAs can regulate mRNA-like non-coding RNAs. Our results greatly expand the post-transcriptional regulatory network involving microRNAs. More interestingly, our previous research results show that large numbers of microRNA-encoding RNAs exist among mRNA-like non-coding RNAs, and microRNAs also have regulatory relationships with these own primary transcripts, thus forming a complex inter-microRNA regulatory network (see Figure 15 [Figure 15: see original paper]).

5 Conclusion

This paper introduces our research group' s work in recent years on non-coding RNAs and complex biological networks, covering non-coding RNA functional studies, complex biological network analysis, and construction of biological networks involving non-coding RNAs. Introducing non-coding RNAs into network research is a frontier topic at the intersection of two hotspots in bioinformatics: non-coding RNA research and complex biological network research. Although some achievements have been made in this area in recent years, many problems remain to be solved. Existing biological studies show that the impact of non-coding RNAs on biological networks is global, participating in all levels of networks including gene transcriptional regulation, protein interaction, and post-transcriptional regulation. Our work is currently limited to the level of post-transcriptional regulatory networks involving non-coding RNAs. How to effectively study the impact of non-coding RNAs on other levels of complex biological networks remains a problem. On the other hand, compared with traditional networks composed of coding genes and their corresponding proteins, the new network will be a mixed "two-color network" with two-color nodes (coding genes, non-coding genes). Therefore, we urgently need to construct a theoretical framework for analyzing such two-color networks involving non-coding RNAs to guide our computational work. Introducing non-coding RNAs into complex biological network analysis allows us to study biology from a new

perspective and will inevitably bring various new problems and challenges. We believe that as these new problems and challenges are solved, our understanding of biology will enter a new stage.

References

[1] Fleischmann, R.D., et al., Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 1995. 269(5223): p. 496-512.

[2] Finishing the euchromatic sequence of the human genome. *Nature*, 2004. 431(7011): p. 931-45.

[3] Lander, E.S., et al., Initial sequencing and analysis of the human genome. *Nature*, 2001. 409(6822): p. 860-921.

[4] Strogatz, S.H., Exploring complex networks. *Nature*, 2001. 410(6825): p. 268-76.

[5] Alon, U., Biological networks: the tinkerer as an engineer. *Science*, 2003. 301(5641): p. 1866-7.

[6] http://www.systemsbiology.org/*{{{the}}}*{{21st}}}*{{{Century}}}*{{Science}}}*{{{Intro}}}*{{to}}}*{{{ISB}}}{

[7] Mattick, J.S. & Makunin, I.V. Non-coding RNA. *Hum Mol Genet* 15 Spec No 1, R17-29 (2006).

[8] Liu, C. et al. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res* 33, D112-5 (2005).

[9] Yin, H. & Lin, H. An epigenetic activation role of Piwi and a Piwi-associated piRNA in *Drosophila melanogaster*. *Nature* 450, 304-8 (2007).

[10] Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281-97 (2004).

[11] Lutcke, H. Signal recognition particle (SRP), a ubiquitous initiator of protein translocation. *Eur J Biochem* 228, 531-50 (1995).

[12] Kiss, T. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *Embo J* 20, 3617-22 (2001).

[13] Nemes, J.P., Benzow, K.A., Moseley, M.L., Ranum, L.P. & Koob, M.D. The SCA8 transcript is an antisense RNA to a brain-specific transcript encoding a novel actin-binding protein (KLHL1). *Hum Mol Genet* 9, 1543-51 (2000).

[14] Smilinich, N.J. et al. A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. *Proc Natl Acad Sci U S A* 96, 8064-9 (1999).

[15] Petrovics, G. et al. Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene* 23, 605-11 (2004).

[16] Ji, P. et al. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031-41 (2003).

[17] Numata, K. et al. Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res* 13, 1301-6 (2003).

[18] Marker, C. et al. Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Curr Biol* 12, 2002-13 (2002).

[19] Huttenhofer, A. et al. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *Embo J* 20, 2943-53 (2001).

[20] Rubin, G.M. The draft sequences. Comparing species. *Nature* 409, 820-1 (2001).

[21] Nadeau, J.H. et al. Sequence interpretation. Functional annotation of mouse genome sequences. *Science* 291, 1251-5 (2001).

[22] Okazaki, Y. et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563-73 (2002).

[23] Cheng, J. et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149-54 (2005).

[24] Deng, W. et al. Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome Res* 16, 20-9 (2006).

[25] Chen, C.L. et al. The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. *Nucleic Acids Res* 31, 2601-13 (2003).

[26] Yuan, G., Klambt, C., Bachellerie, J.P., Brosius, J. & Huttenhofer, A. RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res* 31, 2495-507 (2003).

[27] Tang, T.H. et al. Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc Natl Acad Sci U S A* 99, 7536-41 (2002).

[28] Rivas, E., Klein, R.J., Jones, T.A. & Eddy, S.R. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr Biol* 11, 1369-73 (2001).

[29] Will, C.L. & Luhrmann, R. Spliceosomal UsnRNP biogenesis, structure and function. *Curr Opin Cell Biol* 13, 290-301 (2001).

[30] Hinz, S. & Goringer, H.U. The guide RNA database (3.0). *Nucleic Acids Res* 27, 168 (1999).

[31] Podlevsky, J.D., Bley, C.J., Omana, R.V., Qi, X. & Chen, J.J. The telomerase database. *Nucleic Acids Res* 36, D339-43 (2008).

[32] Zwieb, C., Gorodkin, J., Knudsen, B., Burks, J. & Wower, J. tmRDB (tmRNA database). *Nucleic Acids Res* 31, 446-7 (2003).

[33] Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A. & Panning, B. Xist RNA and the mechanism of X chromosome inactivation. *Annu Rev Genet* 36, 233-78 (2002).

[34] Rosenblad, M.A., Gorodkin, J., Knudsen, B., Zwieb, C. and Samuelsson, T. (2003) SRPDB: Signal Recognition Particle Database. *Nucleic Acids Res.*, 31, 363-364.

[35] Zwieb, C., Gorodkin, J., Knudsen, B., Burks, J. and Wower, J. (2003) tmRDB (tmRNA database). *Nucleic Acids Res.*, 31, 446-447.

[36] Brown, J.W. (1999) The Ribonuclease P Database. *Nucleic Acids Res.*, 27, 314.

[37] Gu, J., Chen, Y. and Reddy, R.(1998) Small RNA database. *Nucleic Acids Res.*, 26,160-162.

[38] Szymanski, M., Erdmann, V.A. and Barciszewski, J. (2003) Noncoding regulatory RNAs database. *Nucleic Acids Res.*, 31, 429-431.

[39] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, 31, 439-441.

[40] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2004) GenBank: update. *Nucleic Acids Res.*, 32, D23-D26.

[41] Changning Liu, Baoyan Bai, et al. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, 33, D112-D115

[42] Kiss, T. (2001) Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.*, 20, 3617-3622.

[43] Hendrix, R.W. (1998) Bacteriophage DNA packaging: RNA gears in a DNA transport machine. *Cell*, 94, 147-150.

[44] Sugisaki, H. and Takanami, M. (1993) The 5' terminal region of the apocytochrome b transcript in *Crithidia fasciculata* is successively edited by two guide RNAs in the 3' to 5' direction. *J. Biol. Chem.*, 268, 887-891.

[45] Zhanybekova, S.S.h., Polimbetova, N.S., Nakisbekov, N.O. and Iskakov, B.K. (1996) Detection of a new small RNA, induced by heat shock, in wheat seed ribosomes. *Biokhimiia*, 61, 862-870.

[46] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. *Science* 309: 1559-1563.

[47] Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* 36: 40-45.

[48] Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, et al. (2006) Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet* 2: e62.

[49] Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* 306: 2242-2246.

[50] Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, et al. (2006) Characterization of the piRNA complex from rat testes. *Science* 313: 363-367.

[51] Erdmann VA, Szymanski M, Hochberg A, de Groot N, Barciszewski J (1999) Collection of mRNA-like non-coding RNAs. *Nucleic Acids Res* 27: 192-195.

[52] Okazaki, Y., et al., Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 2002. 420(6915): p. 563-73.

[53] Ota, T., et al., Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet*, 2004. 36(1): p. 40-5.

[54] Marahrens Y, Loring J, Jaenisch R (1998) Role of the Xist gene in X chromosome choosing. *Cell* 92:

[55] Young TL, Matsuda T, Cepko CL (2005) The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr Biol* 15: 501-512.

[56] Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, et al. (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* 309: 1570-1573.

[57] Bartel, D.P., MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 2004. 116(2): p. 281-97.

[58] Ambros, V., The functions of animal microRNAs. *Nature*, 2004. 431(7006): p. 350-5.

[59] Strogatz, S.H. Exploring complex networks. *Nature* 410, 268-76 (2001).

[60] Bertone, P., et al., Global identification of human transcribed sequences with genome tiling arrays. *Science*, 2004. 306(5705): p. 2242-6.

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

[62] Lee, T.I., et al., Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 2002. 298(5594): p. 799-804.

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

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

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

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

[67] Cho, R.J., M.J. Campbell, et al. (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." *Mol Cell* 2(1): 65-73.

[68] Hughes, T.R., M.J. Marton, et al. (2000). "Functional discovery via a compendium of expression profiles." *Cell* 102(1): 109-26.

[69] Tong, A.H., M. Evangelista, et al. (2001). "Systematic genetic analysis with ordered arrays of yeast deletion mutants." *Science* 294(5550): 2364-8.

[70] Mewes, H.W., D. Frishman, et al. (2002). "MIPS: a database for genomes and protein sequences." *Nucleic Acids Res* 30(1): 31-4.

[71] Enright, A.J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." *Nature* 402(6757): 86-90.

[72] Marcotte, E.M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." *Science* 285(5428): 751-3.

[73] Dandekar, T., B. Snel, et al. (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." *Trends Biochem Sci* 23(9): 324-8.

[74] Gerdes, S.Y., M.D. Scholle, et al. (2003). "Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655." *J Bacteriol* 185(19): 5673-84.

[75] Huynen, M.A., B. Snel, et al. (2003). "Function prediction and protein networks." *Curr Opin Cell Biol* 15(2): 191-8.

[76] Schwikowski, B., P. Uetz, et al. (2000). "A network of protein-protein interactions in yeast." *Nat Biotechnol* 18(12): 1257-61.

[77] Hishigaki, H., K. Nakai, et al. (2001). "Assessment of prediction accuracy of protein function from protein–protein interaction data." *Yeast* 18(6): 523-31.

[78] Maslov, S. and K. Sneppen (2002). "Specificity and stability in topology of protein networks." *Science* 296(5569): 910-3.

[79] Ge, H., Z. Liu, et al. (2001). "Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*." *Nat Genet* 29(4): 482-6.

[80] Gibson, D., J. Kleinberg, et al. (1998). "Inferring Web communities from link topology." *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia* ACM Press, New York, NY.

[81] Kleinberg, J. (1998). "Authoritative sources in a hyper-linked environment." *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia* ACM Press, New York, NY.

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

[83] Shen-Orr SS, Milo R, Mangan S, Alon U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 31: 64-68.

[84] Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298: 799-804.

[85] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: Simple building blocks of complex networks. *Science* 298: 824-827.

[86] Vazquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, et al. (2004) The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc Natl Acad Sci U S A* 101: 17940-17945.

[87] Guelzim N, Bottani S, Bourgnie P, Kepes F (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* 31: 60-63.

[88] Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) Genomic analysis of essentiality within protein networks. *Trends Genet* 20: 227-231.

[89] Bray D (1995) Protein molecules as computational elements in living cells. *Nature* 376: 307-312.

[90] Ma' ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, et al. (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network. *Science* 309: 1078-1083.

[91] Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, et al. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431: 308-312.

[92] Thomas D, Jacquemin I, Surdin-Kerjan Y (1992) MET4, a leucine zipper protein, and centromere-binding factor 1 are both required for transcriptional activation of sulfur metabolism in *Saccharomyces cerevisiae*. *Mol Cell Biol* 12: 1719-1727.

[93] Wolf DM, Arkin AP (2003) Motifs, modules, and games in bacteria. *Curr Opin Microbiol* 6: 125.

[94] Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 2005.

[95] Willingham, A.T. et al. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* 309, 1570-3 (2005).

[96] Sood, P., et al. (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc Natl Acad Sci U S A* 103, 2746-2751.

Authors:

Liu Changning: Forward-looking Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences

Sun Shiwei: Forward-looking Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences

Zhao Yi: Forward-looking Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, biozy@ict.ac.cn

Bu Dongbo: Forward-looking Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences

Note: Figure translations are in progress. See original paper for figures.