Chapter 12
Advanced Topics in Bioinformatics and Computational Biology

Bailin Hao, Chunting Zhang, Yixue Li, Hao Li, Liping Wei, Minoru Kanehisa, Luhua Lai, Runsheng Chen, Nikolaus Rajewsky, Michael Q. Zhang, Jingdong Han, Rui Jiang, Xuegong Zhang, and Yanda Li

12.1 Prokaryote Phylogeny Meets Taxonomy

Bailin Hao
Fudan University, Shanghai, China
The Santa Fe Institute, Santa Fe, NM, USA

Phylogeny defined as the context of evolutionary biology is the connections between all groups of organisms as understood by ancestor/descendant relationships. Since many groups of organisms are now extinct, we can’t have as clear a picture
of how modern life is interrelated without their fossils. Phylogenetics, the science of phylogeny, is a useful tool severed as one part of the larger field of systematic including taxonomy which is a practice and science of naming and classifying the diversity of organisms.

All living organisms on the earth are divided into prokaryotes and eukaryotes. Prokaryotes are unicellular organisms that do not have a nucleus in the cell, and DNA molecules encoding the genetic information just float in the cells. Prokaryotes are the most abundant organisms on earth and have been thriving for more than 3.7 billion years. They shaped most of the ecological and even geochemical environments for all living organisms. Yet our understanding of prokaryotes, particularly their taxonomy and phylogeny, is quite limited. It was the Swedish naturalist Carolus Linnaeus (1707–1778) who firstly introduced the taxonomic hierarchy made of kingdom, phylum, class, order, family, genus, and species. In 1965, Zukerkandl and Pauling suggested that evolutionary information may be extracted from comparison of homologous protein sequences in related species, thus opening the field of molecular phylogeny. A breakthrough in molecular phylogeny of prokaryotes was made by Carl Woese and coworkers in the mid-1970s. They compared the much conserved RNA molecule in the tiny cellular machines that make proteins, the so-called small-subunit ribosomal RNAs (SSU rRNAs), to infer the distance between species. This method has led to a reasonable phylogeny among many prokaryote species by the alignment of the symbolic sequences of about 1,500 letters long. The modern prokaryotic taxonomy as reflected in the new edition of Bergey’s Manual of Systematic Bacteriology is now largely based on 16S rRNA analysis.

Recently proposed CVTree approach, using entirely different input data and methodology, supports most of the 16S rRNA results and may put the prokaryotic
branch of the tree of life on a secure footing. The CVTree approach was first announced in 2002 and has been described in several follow-up works. A web server has been installed for public access. In brief, the input to CVTree is a collection of all translated amino acid sequences from the genome of an organism downloaded from NCBI database. Then the number of K-peptides is counted by using a sliding window, shifting one letter at a time along all protein sequences. These counts are kept in a fixed lexicographic order of amino acid letters to form a vector with $20^k$ components. A key procedure leading to the final composition vector is the subtraction of a background caused mainly by neutral mutations in order to highlight the shaping role of natural selection. As mutations occur randomly at molecular level, this is done by using a (K-2)th-order Markovian prediction based on the number of (K-2)- and (K-1)-peptides from the same genome. A distance matrix is calculated from these composition vectors, and the standard neighbor-joining program from the PHYLIP package is used to generate the final CVTrees.

CVTree method has many new characteristics compared with more traditional methods. It is an alignment-free method as each organism is represented by a composition vector with $20^k$ components determined by the number of distinct K-peptides in the collection of all translated protein sequences. This strategy overcomes the huge computational complexity caused by the difference in genome size and gene number of prokaryotes in sequence alignment method. Moreover, the CVTree provides a parameter-free method that takes the collection of all proteins of the organisms under study as input and generates a distance matrix as output. In fact, the CVTree method has shown rather high resolution to elucidate the evolutionary relationship among different strains of one and the same species. It does not require the selection of RNA or protein-coding gene(s) as all translated protein products in a genome are used. Since there may be large number of gene transfers in species, this phenomenon may show more affection on original methods than CVTree.

The CVTree results are verified by direct comparison with systematic bacteriology. The CVTrees constructed from many organisms bear a stable topology in major branching patterns from phyla down to species and strains. As compared to some traditional phylogenetic tree construction methods, the CVTree approach gets a nice feature of “the more genomes the better agreement” with Bergey’s taxonomy. The high-resolution power of the CVTrees provides a means to elucidate evolutionary relationships among different strains of one and the same species when the 16S rRNA analysis may not be strong enough to resolve too closely related strains.

While the 16S rRNA analysis cannot be applied to the phylogeny of viruses as the latter do not possess a ribosome, the CVTree method has been successfully used to construct phylogeny of coronaviruses including human SARS virus and double-strand DNA viruses and chloroplasts as well.

Many experimental results have supported CVTree method can output very high precision of phylogenetic tree construction agreeing well with the taxonomy which is based more and more on the 16S rRNA analysis. The CVTree approach and the 16S rRNA analysis use orthogonal data from the genome and utilize different methodology to infer phylogenetic information. Yet they support each other in
an overwhelming majority of branchings and clusterings of taxa, thus providing a reliable framework to demarcate the natural boundaries among prokaryote species. It would be useful to see the CVTree method applied to such eukaryotes as fungi. This method also gives new calculation of phylogenetic distance between different species and can be further used in many other bioinformatics problems.

12.2 Z-Curve Method and Its Applications in Analyzing Eukaryotic and Prokaryotic Genomes

Chunting Zhang (✉)
Tianjin University, Tianjin, China

At the stage of post-genomics, ever-increasing genomic sequences across various prokaryotic and eukaryotic species are available for the mechanism exploration of biological systems. The computer-aided visualization of the long sequences with complex structures which could provide us the intuitive clues at understanding biological mechanisms is urgently needed for both biologists and scientists from other disciplines. One representative tool of visualizing methods is the Z-curve, which makes us capable of analyzing the genomic sequences in terms of geometric approach as the complement of algebraic analysis. The Z-curve is a three-dimensional curve which is a unique representation for a given DNA sequence in the sense that each can be uniquely reconstructed given the other. The Z-curve has been proven that it is the generalized formula of several proposed visualization methods, for example, the H-curve, the game representation, the W-curve, and the two-dimensional DNA walk. The Z-curve has been used in several biological applications, including gene recognition and Isochore prediction, furthermore, succeeded at helping identifying novel biological mechanisms. The Z-curve database for the available sequences of archaea, bacteria, eukaryote, organelles, phages, plasmids, viroids, and viruses is established and open for biological researches. This is a review of the Z-curve and the successful applications for both prokaryotic and eukaryotic species.

The Z-curve is a unique three-dimensional curve representation for a given DNA sequence in the sense that each can be uniquely reconstructed given the other. The Z-curve is composed of a series of nodes \( P_1, P_2, P_3, \ldots, P_N \), whose coordinates \( x_n, y_n \) and \( z_n \) (\( n = 1, 2, \ldots, N \), where \( N \) is the length of the DNA sequence being studied) are uniquely determined by the Z-transform of DNA sequence:

\[
\begin{align*}
  x_n &= (A_n + G_n) - (C_n + T_n) \\
  y_n &= (A_n + C_n) - (G_n + T_n) \\
  z_n &= (A_n + T_n) - (C_n + G_n)
\end{align*}
\]

\( x_n, y_n, z_n \in [-N, N] \), \( n = 0, 1, 2, \ldots, N \)
where $A_n$, $C_n$, $G_n$, and $T_n$ are the cumulative occurrence numbers of A, C, G, and T, respectively, in the subsequence from 1st base to the $n$th base in the sequence. The Z-curve is defined as the connection of the nodes $P_1, P_2, P_3, \ldots, P_N$ one by one sequentially with straight lines starting from the origin of the three-dimensional coordinate system. Once the coordinates $x_n, y_n,$ and $z_n$ ($n = 1, 2, \ldots, N$) of a Z-curve are given, the corresponding DNA sequence can be reconstructed from the inverse Z-transform. In terms of biology, the three components of the Z-curve make sense that $x_n, y_n,$ and $z_n$ represent the distributions of purine/pyrimidine (R/Y), amino/keto (M/K), and strong H bond/weak H bond (S/W) bases along the sequence, respectively. The three components of the Z-curve uniquely describe the DNA sequence being studied and contain all the information in the original sequences.

The perceivable form of the Z-curve provides an intuitive insight to the researches of genomic sequences. The database of the Z-curves for archaea, bacteria, eukaryote, organelles, phages, plasmids, viroids, and viruses is established and contains pre-calculated coordinates of more than 1,000 genomes. As the complement of GenBank/EMBL/DBJ, the Z-curve database provides a variant resolution geometric insight into some features of the nucleotide composition of genomes, ranging from the local scale to the global scale. It has been shown that the visualization and the complete information of the Z-curve offer benefits to Bioinformatics community. The joint effect of three components of the Z-curve has been identified at recognizing genes in available genomes across various prokaryotic species, including $S.\, cerevisiae$, bacteria, archaea, coronavirus, and phages. The respective software services are established in the Z-curve database. The linear combinations of the $x_n$ and $y_n$ components of the Z-curve are defined as a family of disparity curves, which could be used to analyze the local deviations from Chargaff Parity Rule 2 showing that globally both $\%A \approx \%T$ and $\%G \approx \%C$ are valid for each of the two DNA strands. The AT- and GC-disparity curves calculated by the Z-curves have been applied to predict the replication origins and terminations of some bacterial and archaeal genomes. The individual $z_n$ component has a significant advantage at the calculation of the $G + C$ content. With the Z-curve, the calculation of the $G + C$ content of genomes is windowless in contrast to the previous methods and could be performed at any resolution. The $z'_n$-curve, a transform of the Z-curve, termed GC profile, is defined to describe the distribution of the $G + C$ content along the DNA sequences. Intuitively, a jump in the $z'_n$-curve indicates an A + T-rich region, whereas a drop means a G + C-rich region. A sudden change in the $z'_n$-curve might imply a transfer of foreign DNA sequence from other species. A typical example is the $z'_n$-curve for the smaller chromosome of $Vibrio\, cholera$, where the position of the integron island is precisely identified by observing a sudden jump in the $z'_n$-curve. Recent studies have shown that the windowless calculation and analysis of the $G + C$ content of the eukaryotic genomes obtain several impressive achievements at the Isochore edge determination, the Isochore structure exploration, and the Isochore predictions.
12.3 Insights into the Coupling of Duplication Events and Macroevolution from an Age Profile of Transmembrane Gene Families

Yixue Li (✉)
Shanghai Center for Bioinformatics Technology, Shanghai, China

This study stemmed from another project focusing on the evolution of transmembrane proteins. We noticed the noncontinuous distribution of duplication events over evolutionary time, and the pattern to some extent overlapped with fossil evidences of macroevolution, which led us to think about the relationship between molecular evolution and macroevolution. The neutral evolution theory for molecular level and Darwin’s macroevolution theory conflict with each other, but both are supported by evidences at different levels. What’s the connection?

We tried to answer this question by studying the duplication of transmembrane proteins. The evolution of new gene families subsequent to gene duplication may be coupled to the fluctuation and dramatically alternations of population and environment variables. By using the transmembrane gene family, which is a key component for information exchange between cells and the environment, it is possible to find the cycles and patterns in the gene duplication event records and the relationship between the evolutionary patterns on the molecular level and the species level.

We started by building transmembrane gene family. First, we predicted transmembrane proteins from 12 eukaryotes by including proteins with at least one predicted transmembrane helix. Then we developed a pipeline to build gene families by integrating strategy of COG, HOBACGEN, and other additional steps. After a manual check we had 863 homology families of eukaryote transmembrane proteins. We then constructed the phylogenetic tree for these families by neighbor-joining method. The molecular time scale of the inferred tree was calibrated and adjusted by both fossil data and molecular data with known molecular age. Finally we were able to detect 1,651 duplication events in the final dataset with 786 gene families. All of the identified duplication events were recorded with the corresponding ages.

The overall age distribution was determined on the basis of 1,620 transmembrane gene duplication events. Similar to previous report, this distribution clearly shows three peaks (0.13 billion years [Gyr], 0.46 Gyr, and 0.75 Gyr ago approximately).

We next examined the relationship of the apparent disturbances of the age distribution with oxidation event records reconstructed from geochemical and fossil research. Interestingly enough, the time point at which the density of the duplicates increases distinctly under the baseline distribution is completely consistent with the reliable minimum age for the advent of oxygenic photosynthesis (2.75 Gyr ago) and the age of the domain Eucarya concluded from molecular fossils (about 2.7 Gyr ago). Our findings imply the linkage between the oxygen level and the transmembrane gene duplicates.
We performed decomposition in Fourier series of the detrended density trace of the duplicates in the Phanerozoic phase. We identified three potential peaks, which are 60.92-Myr, 27.29-Myr, and 10.32-Myr cycles. The 60.92-Myr cycle has the strongest cyclicity in the density trace of the transmembrane gene duplicates. Consistent with the opinion that the macroevolutionary time series have characteristics of a random walk, because these cycles are not statistically significant, they cannot reject the null hypothesis of a random walk. The existence of the most potential cycle of 60.92 Myr in the age distribution of transmembrane gene families is a very interesting discovery, because it is not indistinguishable from the 62 ± 3-Myr cycle that is the most statistically significant cycle detected in biodiversity recently reported.

We had clearly shown that the duplication events of transmembrane genes are coupled with the macroevolution measurement and asynchronous with the animal biodiversity. The evolution history is a coevolution process of the environment and life. The overall shape of the age distribution is driven by the oxygen level in the atmosphere, while the waves of the distribution might be driven by some rhythmic external force. Furthermore, we proposed a plausible evolutionary scenario to explain these findings based on the factors finally determining the fate of the duplicates, which implies that the environment alternation would induce the redundancy of the existent genome system that is beneficial for survival in a rigorous condition. In addition, we presented a methodology to provide a unique, temporally detailed understanding of the interaction of the transmembrane gene duplication events and the environment variables. Since the sequence data are thoroughly independent from the fossil record and more readily attainable, this methodology may give us a new strategy to validate patterns such as the 62-Myr cycle, which was detected from fossil or other geophysical records. Further studies using this method may offer important insights into the interplay of the microevolution and macroevolution factors.

12.4 Evolution of Combinatorial Transcriptional Circuits in the Fungal Lineage

Hao Li
University of California, San Francisco, CA, USA

Now it’s a good time to study basic mechanism of evolution, because the whole-genome sequences of an increasing large number of species are available. Bioinformatics provides important tools for research on evolution. We study the evolution of combinatorial control in yeast transcriptional regulation network.

Transcription network can response to external environmental changes, and the regulation often involves combinatorial control of multiple transcriptional factors (TFs). One can view the transcriptional network as a black box, with the activity of TFs as input and the transcript levels of all genes as the output.
We are interested in two questions: first, how to construct the transcriptional network? And second, why should it be like this? What are the functional and evolution constraints? Here we focus on the later, that is, how transcriptional networks evolve? What are the basic patterns or steps? Here we will first show a study on the transcriptional circuits controlling yeast mating types.

Mating type in the yeasts *Saccharomyces cerevisiae* and *Candida albicans* is controlled by the MAT locus, which has two versions: MATa and MATα. Cells that express only the MATa- or MATα-encoded proteins are a-cells and α-cells, respectively. The a-cells express a-specific genes (asgs), which are required for a-cells to mate with α-cells. On the other hand, α-cells express the α-specific genes (αsgs). In *S. cerevisiae*, the αsgs are on by default and are repressed in other cells by protein α2 encoded by MATα. In *C. albicans*, however, the asgs are off by default and are activated in a-cells by protein α2 encoded by MATα. Both molecular mechanisms give the same logical output: αsgs are expressed only in a-cells. By comparative genomics analysis, we show that α2-activation most likely represents the ancestral state and the α2 gene was recently lost in the *S. cerevisiae*, now using the α2-repressing mode of asg regulation. In the promoters of several asgs, we found a regulatory element with several distinctive features. First, the sequence contains a region that closely resembles the binding site of Mcm1, a MADS box sequence-specific DNA-binding protein that is expressed equally in all three mating types, and is required for the regulation of both asgs and αsgs in *S. cerevisiae*. We also show that the Mcm1 residues that contact DNA are fully conserved between *C. albicans* and *S. cerevisiae*, strongly implicating this region of the element as a binding site for Mcm1 in *C. albicans*. Second, the putative Mcm1 site in *C. albicans* asg promoters lies next to a motif of the consensus sequence CATTGTC. The spacing between this motif and the Mcm1 site is always 4 bp. This motif is similar to demonstrated binding sites for α2 orthologues in *Schizosaccharomyces pombe* and *Neurospora crassa* and to the α2 monomer site of *S. cerevisiae*.

These evidences suggest the following changes in cis- and trans-elements can lead to a profound evolutionary change in the wiring of a combinatorial circuit: (1) “tuning up” of a binding site for a ubiquitous activator, making gene expression independent of a cell-type-specific activator; (2) a small change in an existing DNA-binding site, converting its recognition from one protein to that of an unrelated protein; and (3) a small change in the amino acid sequence of a sequence-specific DNA-binding protein, allowing it to bind DNA cooperatively with a second protein.

In a second study, we center on Mcm1. Besides the regulatory role in mating, it also involves in many other processes such as cell cycle. By comparing data from *Saccharomyces cerevisiae*, *Kluyveromyces lactis*, and *Candida albicans*, we find that the Mcm1 combinatorial circuits undergone substantial changes. This massive rewiring of the Mcm1 circuitry has involved both substantial gain and loss of targets in ancient combinatorial circuits as well as the formation of new combinatorial interactions. We have dissected the gains and losses on the global level into subsets of functionally and temporally related changes. One particularly dramatic change is the acquisition of Mcm1 binding sites in close proximity to Rap1 binding sites at 70 ribosomal protein genes in the *K. lactis* lineage. Another intriguing and very
recent gain occurs in the *C. albicans* lineage, where Mcm1 is found to bind in combination with the regulator Wor1 at many genes that function in processes associated with adaptation to the human host, including the white-opaque epigenetic switch.

### 12.5 Can a Non-synonymous Single-Nucleotide Polymorphism (nsSNP) Affect Protein Function?

**Analysis from Sequence, Structure, and Enzymatic Assay**

Liping Wei (✉)

Peking University, Beijing, China

After the completion of the human genome project, increasing attention has focused on the identification of human genomic variations, especially single-nucleotide polymorphisms (SNPs). It is estimated that the world population contains a total of ten million SNP sites, resulting in an average density of one variant per 300 bases. SNPs in coding and regulatory regions may play a direct role in diseases or differing phenotypes. Among them, the single amino acid polymorphisms (SAPs, conventionally known as non-synonymous SNPs or nsSNPs), which cause amino acid substitutions in the protein product, are of major interest because they account for about 50% of the gene lesions known to be related to genetic diseases. Through large-scale efforts such as the HapMap project (http://www.hapmap.org), The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov), and whole-genome association studies, available SAP data is accumulating rapidly in databases such as dbSNP, HGVbase, Swiss-Prot variant page, and many allele-specific databases. However, because of the high-throughput nature of these efforts, many SAPs could not be experimentally characterized in terms of their possible disease association. Furthermore, the underlying mechanisms that explain why a SAP may be associated with disease and have deleterious functional effect are not yet fully understood.

In the past 5 years, several bioinformatics methods have been developed to use sequence and structural attributes to predict possible disease association or functional effect of a given SAP. A popular sequence-based method is SIFT, which predicts whether an amino acid substitution is deleterious or tolerated based on the evolutionary conservation of the SAP site from multiple sequence alignment. More recent methods incorporate both sequence and structural attributes and use a range of classifiers such as rule-based, decision trees, support vector machines (SVMs), neural networks, random forests, and Bayesian networks to annotate SAPs. Zhi-Qiang Ye et al. recently employed machine learning method, named SAPRED (http://sapred.cbi.pku.edu.cn/), and obtained better performance (82.6%) than early methods. SAPRED first constructed a relatively balanced dataset from the Swiss-Prot variant pages, then investigated the most complete set of structural and sequence attributes to date, and identified a number of biologically informative new
attributes that could explain why a SAP may be associated with disease. Finally, the method incorporated these attributes into an SVM-based machine learning classifier.

SAPRED investigated a large set of structural and sequence attributes including both commonly used ones such as residue frequency and solvent accessibility and new ones that are novel to this study. These attributes include residue frequency and conservation, solvent accessibilities, structural neighbor profiles, nearby functional sites, structure model energy, hydrogen bond, disulfide bond, disordered region, aggregation properties, and HLA family. They accessed these attributes and get many new findings. They confirmed that residue frequencies provided the best discrimination reported as early researches. Previous studies found solvent accessibilities to be the type of attributes with the second most predictive power. However, their study identified two new types of attributes that showed higher predictive power than solvent accessibilities.

The new attributes studied in this work may further the understanding of the biological mechanism underlying the functional effect and disease association of SAPs. At the same time, SAPRED also contribute to the increase in accuracies of predicting the disease association of SAPs. In particular, the predictive power of structural neighbor profile is almost as high as that of residue frequencies, highlighting the importance of the microenvironment around a SAP. In addition, the predictive power of nearby functional sites is higher than solvent accessibilities, the second most powerful type of attributes in previous studies. By considering residues both at and near functional sites in terms of both sequence and structure, SAPRED significantly enlarged the coverage and overcame the limitations in previous work that used only the functional site residues themselves. The other new attributes, such as disordered regions and aggregation properties, also provided direct biological insights into the study of SAPs and contributed to the overall accuracy of prediction.

After prediction, further biological experiments can be used to describe the association between SAP and disease. An excellent example in SAP study is R41Q associated with certain severe adverse reactions to oseltamivir. The use of oseltamivir, widely stockpiled as one of the drugs for use in a possible avian influenza pandemic, has been reported to be associated with neuropsychiatric disorders and severe skin reactions, primarily in Japan. R41Q, near the enzymatic active site of human cytosolic sialidase, a homologue of virus neuraminidase is the target of oseltamivir. This SNP occurred in 9.29% of Asian population and none of European and African American population. Structural analyses by SAPRED and Ki measurements using in vitro sialidase assays indicated that this SNP could increase the unintended binding affinity of human sialidase to oseltamivir carboxylate, the active form of oseltamivir, thus reducing sialidase activity. In addition, this SNP itself results in an enzyme with an intrinsically lower sialidase activity, as shown by its increased K_m and decreased V_max values. Theoretically administration of oseltamivir to people with this SNP might further reduce their sialidase activity. The reported neuropsychiatric side effects of oseltamivir and the known symptoms of human sialidase-related disorders are correlated. This Asian-enriched sialidase variation caused by the SNP, likely in homozygous form, may be associated with certain severe adverse reactions to oseltamivir.
The “sequence → structure → function” model for describing protein activity states that the amino acid sequence determines the higher structures of a protein molecule, including its secondary and tertiary conformations, as well as quaternary complexes and further states that the formation of a definite ordered structure represents the foundation for the function of the protein. Different attributes including sequence and structure can give more contribution to the final function prediction and show the association between disease and SAP. However, to better predict the possible disease association of SAPs, existing methods still need to be improved in several aspects. First, more biologically informative structural and sequence attributes need to be investigated to further understand the underlying mechanism of how a SAP may be associated with a disease. Second, several studies used imbalanced datasets which impeded the performance of their classifiers. Third, by using more biologically informative attributes and a better dataset, the overall accuracy of the prediction can be improved.

12.6 Bioinformatics Methods to Integrate Genomic and Chemical Information

Minoru Kanehisa
Kyoto University, Kyoto, Japan

Although the comprehensive genome sequence has only recently been revealed, biologists have been characterizing the roles played by specific proteins in specific processes for nearly a century. This information spans a considerable breadth of knowledge and is sometimes exquisitely detailed and stored as primary literature, review articles, and human memories. Recent efforts have established databases of published kinetic models of biologic processes ranging in complexity from glycolysis to regulation of the cell cycle. These chemical information and genomic information can be found in many databases such as NCBI, KEGG, PID, and Reactome, which allow researchers to browse and visualize pathway models and, in some cases, to run simulations for comparison with experimental data.

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a knowledge base for systematic analysis of gene functions, linking genomic information with higher-order function information. KEGG mainly consists of the PATHWAY database for the computerized knowledge on molecular interaction networks such as pathways and complexes, the GENES database for the information about genes and proteins generated by genome sequencing projects, and the LIGAN database for the information about chemical compounds and chemical relations that are relevant to cellular processes. KEGG BRITE is a collection of hierarchies and binary relations with two interrelated objectives corresponding to the two types of graphs: to automate functional interpretations associated with the KEGG pathway reconstruction and to assist discovery of empirical rules involving genome-environment interactions. In addition to these main databases, there are several other databases including EX-
PRESSION, SSDB, DRUG, and KO. KEGG can be considered as a complementary resource to the existing database on sequences and three-dimensional structures, focusing on higher-level information about interaction and relations of genes and proteins.

The KEGG databases are highly integrated. In fact, KEGG should be viewed as a computer representation of the biological system, where biological objects and their relationships at the molecular, cellular, and organism levels are computerized as separate database entries. Cellular functions result from intricate networks of molecular interactions, which involve not only proteins and nucleic acids but also small chemical compounds. The genomic and chemical information in database can be integrated by bioinformatics methods which are summarized as two procedures: decomposition into building blocks and reconstruction of interaction networks. First, chemical structure can be decomposed into small building blocks by comparison of bit-represented vectors (fingerprints) or comparison of graph objects. Conserved substructures can be viewed as building blocks of compounds and variable substructures as building blocks of reactions. Here small building blocks are divided into two categories. One is metabolic compounds that are subject to enzyme-catalyzed reactions that maintain the biological system. The other category is regulatory compounds that interact with proteins, DNA, RNA, and other endogenous molecules to regulate or perturb the biological system. In the same way, the proteins can be decomposed into domains or conserved motifs considered as small building blocks. Second, using the existed pathway information (metabolic pathways) and genomic information (such as operon structure in bacterial), interaction networks (or network modules) can be reconstructed.

Extracting information from the chemical structures of these small molecules by considering the interactions and reactions involving proteins and other biological macromolecules, a knowledge-based approach for understanding reactivity and metabolic fate in enzyme-catalyzed reactions in a given organism or group was presented by Mina Oh et al. They first constructed the KEGG RPAIR database containing chemical structure alignments and structure transformation patterns, called RDM patterns, for 7,091 reactant pairs (substrate-product pairs) in 5,734 known enzyme-catalyzed reactions. A total of 2,205 RDM patterns were then categorized based on the KEGG PATHWAY database. The majority of RDM patterns were uniquely or preferentially found in specific classes of pathways, although some RDM patterns, such as those involving phosphorylation, were ubiquitous. The xenobiotics biodegradation pathways contained the most distinct RDM patterns, and a scheme was developed to predict bacterial biodegradation pathways given chemical structures of, for example, environmental compounds.

If the chemical structure is treated as a graph consisting of atoms as nodes and covalent bonds as edges, two chemical structures of compounds can be compared by using related graph algorithms. On the basis of the concept of functional groups, 68 atom types (node types) are defined for carbon, nitrogen, oxygen, and other atomic species with different environments, which has enabled detection of biochemically meaningful features. Maximal common subgraphs of two graphs can be found by searching for maximal cliques (simply connected common subgraphs) in the
association graph. The procedure was applied to the comparison and clustering of 9,383 compounds, mostly metabolic compounds, in the KEGG/LIGAND database. The largest clusters of similar compounds were related to carbohydrates, and the clusters corresponded well to the categorization of pathways as represented by the KEGG pathway map numbers. When each pathway map was examined in more detail, finer clusters could be identified corresponding to subpathways or pathway modules containing continuous sets of reaction steps. Furthermore, it was found that the pathway modules identified by similar compound structures sometimes overlap with the pathway modules identified by genomic contexts, namely, by operon structures of enzyme genes.

With increasing activity of metabolomics, chemical genomics, and chemical biology in which the biological functions of a large number of small molecules are uncovered at the molecular, cellular, and organism levels, new bioinformatics methods have to be developed to extract the information encoded in the small molecular structures and to understand the information in the context of molecular interactions and reactions involving proteins and other biomolecules. As above shown results, the integrated analysis of genomic (protein and nucleic acid) and chemical (small molecule) information can give more effective results in inferring gene functions or cellular processes. Furthermore, combined with more other database information such as microarray data, more results and higher performance can be got.

12.7 From Structure-Based to System-Based Drug Design

Luhua Lai (✉)
Peking University, Beijing, China

For the purpose of better understanding the structure-based and the later system-based drug design, we first introduce the general procedure of drug discovery. This procedure, cost 10–12 years averagely, is composed of lead compound discovery, lead compound optimization, activity test, preclinical research, clinical research, and medicaments listing. Though the drug design usually refers to the first two steps, currently there is a very important trend that the subsequent steps should be considered as from the beginning. That is, from the early stage of the drug research and design, we may consider the absorption, metabolism, and toxicity of a drug, which can reduce the waste as much as possible.

Targeted drug design needs to study the targets. Most of our known drug targets are proteins, and the combination of drugs and targets follows a “lock and key” model, which is also a basic hypothesis of the structure-based drug design, meaning that the drug design is actually finding targets interacted with specific proteins. This structure-based drug design method is facing two main difficulties now: conformational change in the interaction and the accurate and fast calculation of binding free energies. Other design may also include the mechanism-based drug
design, but it is less successfully applied compared with that by the structure-based drug design.

When doing molecular recognition, the following issues need to be taken into account, including how to identify the interaction targets among so many molecules, faraway attraction, conformational changes and induced fit, energy interactions, desolvation interactions, dipole-dipole interactions, and so on. As is known that drug design can be regarded as a molecular recognition process, the main effects that influence the molecular recognition include the van der Waals interaction, the Coulomb law, the hydrogen bond interaction, as well as the hydrophobic interaction.

The drug design process is classified into two cases, based on whether the three-dimensional structure of the receptor is known. If the structure is known, the relevant researches include finding new types of lead compounds according to the structures of the receptors, studying the interactions between receptors and known compounds as well as optimizing the lead compounds. Otherwise if the structure is unknown, the relevant researches include studying the QSAR (quantitative structure-activity relationship) in series of compounds and analyzing the pharmacophore model. Specifically, the new lead compounds can be found through database algorithms, segment connecting method, and de novo design approach. Even though it is universally regarded that the structure-based drug design has achieved great success on the aspects of lead compound finding by means of database screening, there also exist big limitations; for example, if a newly discovered protein cannot be found in the existing database, it will fail to find a molecule to combine. Under this circumstance, though there does not exist any complete molecule which is suitable for being combined with the protein, there may exist several already synthesized molecular segments. In some manner, we can connect these molecular segments with each other and obtain a complete molecule, and this method is called segment connecting method. The disadvantage of this method is that in reality the segment combination is not always the conformation of the lowest energy, rendering the combination hard to be connected in reality. For people who are studying the drug design approaches, the most challengeable approach may be the de novo design approach, which is expected to be independent of existing compounds and databases. By knowing only the protein’s shape at the combination site, we are able to grow a new molecular, which can fit the shape very well.

In the computation of binding free energy, we need a scoring function, and it is usually given in the forms of empirical or semiempirical formula, for example, the force field-based scoring functions (D-Score, G-Score, GOLD, Autodock, and DOCK), the empirical scoring function (LUDI, F-Score, ChemScore, Fresno, SCORE, and X-SCORE), and the knowledge-based scoring function (PMF, DrugScore, and SMoG).

The network-based drug design approach is distinguished from the “one target, one molecule” approach in the structure-based drug design. In the network-based drug design, the target lies in a complex network of certain cell of certain organ of certain human bodies, which is much more complicated than a single molecule. However, with our gradually deep understanding toward the complex biological systems, some analysis can be done at the molecular network level, including
the mathematical modeling and the dynamical property simulation of the disease-related networks, finding the key nodes in use of the multiparameter analysis, implementing regulations using multiple nodes in network control, as well as simulating the effect of multi-node control. Last but not least, the network-based drug design does not mean that we do not need the structure-based drug design; instead, the latter is an indispensable step in no matter what kinds of drug design. It is appropriate to say that the network-based drug design provides a fresh idea for the structure-based drug design.

12.8 Progress in the Study of Noncoding RNAs in *C. elegans*

Runsheng Chen (✉)
Chinese Academy of Sciences, Beijing, China

With the completion of human genome project, more and more genomes are sequenced and more secret of life is discovered. Important properties of human genome are that it has lower protein number (~25,000) than expected and more regions in human genome are noncoding sequences. Noncoding sequences are segments of DNA that does not comprise a gene and thus does not code for a protein, which are interspersed throughout DNA. There are only 15% noncoding sequences in *E. coli*, 71% in *C. elegans*, and 82% in *Drosophila*. However, the noncoding sequences in human genome are 98%, while only 2% are coding regions. It is obvious the higher organisms have a relatively stable proteome and a relatively static number of protein-coding genes, which is not only much lower than expected but also varies by less than 30% between the simple nematode worm *C. elegans* (which has only $10^3$ cells) and humans ($10^{14}$ cells), which have far greater developmental and physiological complexity. Moreover, only a minority of the genomes of multicellular organisms is occupied by protein-coding sequences, the proportion of which declines with increasing complexity, with a concomitant increase in the amount of noncoding intergenic and intronic sequences. Now it is widely believed that there are many units called noncoding genes in noncoding sequences which translate in noncoding RNA but not protein. Thus, there seems to be a progressive shift in transcriptional output between microorganisms and multicellular organisms from mainly protein-coding mRNAs to mainly noncoding RNAs, including intronic RNAs.

Noncoding RNA genes include highly abundant and functionally important RNA families such as transferRNA (tRNA) and ribosomal RNA (rRNA), as well as RNAs such as snoRNAs, microRNAs, siRNAs, piRNAs, and lastly long ncRNAs. Recent transcriptomic and bioinformatic studies suggest the existence of thousands of ncRNAs encoded within human genome. Numerous researches proved that noncoding RNAs have very important functions. For example, SINE elements serve as recombination hot spots allowing the exchange of genetic material between unrelated sequences and also can act as tissue-specific enhancers or silencers of the
adjacent genes. The Xist gene lies within the X-inactivation center and is required to initiate X chromosome inactivation. Xist encodes a large, spliced, polyadenylated, and noncoding RNA that is expressed exclusively from the otherwise inactive X chromosome. NcRNAs, especially the miRNAs, have also been implicated in many diseases, including various cancers and neurological diseases. There are many estimated noncoding genes in sequenced genomes which may be much more than coding genes.

In 2004, the *C. elegans* genome was reported approximately 1,300 genes known to produce functional ncRNA transcripts including about 590 tRNAs, 275 rRNAs, 140 trans-spliced leader RNA genes, 120 miRNA genes, 70 spliceosomal RNA genes, and 30 snoRNA genes. Recently, with hundreds of noncoding RNA found, the number is increased drastically. Applying a novel cloning strategy, Wei Deng et al. have cloned 100 new and 61 known ncRNAs in *C. elegans*. Studying of genomic environment and transcriptional characteristics has shown that two-thirds of all ncRNAs, including many intronic snoRNAs, are independently transcribed under the control of ncRNA-specific upstream promoter elements. Furthermore, the percent of the transcription levels of the ncRNAs varying with developmental stages is at least 60%. This work also found two new classes of ncRNAs, stem-bulge RNAs (sbRNAs) and snRNA-like RNAs (slnRNAs). They are all identified and featured distinct internal motifs, upstream elements, secondary structures, and high and developmentally variable expression. Most of the novel ncRNAs are conserved in *C. briggsae*, but only one homolog was found outside the nematodes. Preliminary estimates indicate the transcriptome may contain ~2,700 small noncoding RNAs, which are potentially acted as regulatory elements in nematode development. This estimation highly increased the number of noncoding RNA in *C. elegans*. Furthermore, combined microarray is designed to analyze the relationship between ncRNA and host gene expression. Results show that the expression of intronic ncRNA loci with conserved upstream motifs was not correlated to (and much higher than) expression levels of their host genes. Promoter-less intronic ncRNAs, which even show a clear correlation to host gene expression, also have a surprising amount of “expressional freedom” compared to host gene function. Taken together, the microarray analysis presents a more complete and detailed picture of a noncoding transcriptome than hitherto has been presented for any other multicellular organism.

By using a whole-genome tiling microarray, the *C. elegans* noncoding transcriptome is mapped. Three samples are designed and individually produced 108669, 97548, and 5738 transfrags which after removal of redundancies suggested the presence of at least 146,249 stably expressed regions with an average and median length of 156 and 103 nt, respectively. After combining overlapping transcripts, it is estimated the total transcription in *C. elegans* is at least 70%. More tests show the experiments are very high precise and 90% transcriptions are further confirmed by added experiments. These new findings are the same as other conservative summation of the mammalian sequences (human, mouse), which indicates that (at least) 60–70% of the mammalian genome is transcribed on one or both strands.

In the past few years, considerable number of noncoding RNAs (ncRNAs) has been detected by using experiments and computations. Although the functions of the many recently identified ncRNAs remain mostly unknown, increasing evidence
stands in support of the notion that ncRNAs represent a diverse and important functional output of most genomes. To fusion these information together, NONCODE presents an integrated knowledge database dedicated to ncRNAs and has many distinctive features. First, the ncRNAs in NONCODE include almost all the types of ncRNAs, except tRNAs and rRNAs. Second, all added ncRNA sequences and their related information have been confirmed manually by consulting relevant literature, and more than 80% of the entries are based on experimental data. Third, based on the cellular process and function, which a given ncRNA is involved in, NONCODE introduced a novel classification system, labeled process function class, to integrate existing classification systems. In addition, ncRNAs have been grouped into nine other classes according to whether they are specific to gender or tissue or associated with tumors and diseases. NONCODE database is very powerful and gives much help in noncoding research.

There are rapidly research problems for ncRNAs, especially as the unknown function and mode of noncoding RNA, their complex structures, and so on. The functional genomics of ncRNAs will be a daunting task which may be equal or greater challenge than that we already face in working out the biochemical functions and biological roles of all of the known and predicted proteins. Most of the ncRNAs identified in genomic transcriptome studies have not been systematically studied and have yet to be ascribed as any functions. RNAs (including those derived from introns) appear to comprise a hidden layer of internal signals that control various levels of gene expression in physiology and development, including transcription, RNA splicing, editing, translation, and turnover. RNA regulatory networks may determine most of our complex characteristics and play a significant role in disease and constitute an unexplored world of genetic variation both within and between species. New methods and experiments will be designed to these new studies. In this area, bioinformatics will be a key, as it should be possible to use more information to identify transmitters and their receivers in RNA regulatory networks.

12.9 Identifying MicroRNAs and Their Targets

Nikolaus Rajewsky (✉)
Max-Delbrück-Center for Molecular Medicine, Berlin, Germany

MicroRNAs (miRNAs) are endogenous 22nt RNAs that can play important regulatory roles in animals and plants by targeting mRNAs for cleavage or translational repression. They are often conserved and with hairpin structures. The early idea about RNA started in the early 1960s when Rosalind Lee and Rhonda Feinbaum discovered that lin-4 does not code for a protein but instead produces a pair of small RNAs. The shorter lin-4 RNA is now recognized as the founding member of an abundant class of tiny regulatory RNAs called miRNAs. Hundreds of distinct miRNA genes are now known to exist and to be differentially expressed during development and across tissue types. Several detailed studies have shown that miRNA families can expand in plants and animals by the same processes.
of tandem, segmental, and whole-genome duplication as protein-coding genes. Many confidently identified miRNA genes are likely to regulate large number of protein-coding genes in these human and animals. These numbers will undoubtedly increase as high-throughput sequencing continues to be applied both to miRNA discovery and the validation of some of the many additional candidates proposed. The breadth and importance of miRNA-directed gene regulation are coming into focus as more miRNAs and their regulatory targets and functions are discovered.

Computational approaches have been developed to complement experimental approaches to miRNA gene identification. Homology searches are the early used methods and have revealed orthologs and paralogs of known miRNA genes. Gene-finding approaches that do not depend on homology or proximity to known genes have also been developed and applied to entire genomes. The two most sensitive computational scoring tools are MiRscan, which has been systematically applied to nematode and vertebrate candidates, and miRseeker which has been systematically applied to insect candidates. Both MiRscan and MiRseeker have identified dozens of genes that were subsequently (or concurrently) verified experimentally. They typically start by identifying conserved genomic segments that both fall outside of predicted protein-coding regions and potentially could form stem loops and then score these candidate miRNA stem loops for the patterns of conservation and pairing that characterize known miRNA genes.

Obviously, methods that rely on phylogenetic conservation of the structure and sequence of a miRNA cannot predict nonconserved genes. To overcome this problem, several groups have developed ab initio approaches to miRNA prediction that use only intrinsic structural features of miRNAs and not external information. With these ab initio prediction methods used, many nonconserved miRNAs have been discovered and experimentally verified in viruses and human. Very recently, by using deep sequencing technology, new method was developed to find miRNA effectively. Deep sequencing is a new biotechnology which may displace microarrays in the future and have already used for DNA sequencing, disease mapping, expressing profiling, and binding sites mapping. But analysis of output is highly nontrivial which needs new generation of computing power. MiRDeep is the first algorithm extract miRNA from deep sequencing and uses a probabilistic model of miRNA biogenesis to score compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor. MiRDeep can find overall miRNAs with a signal to noise ratio of at least 11:1 and sensitivity at least 90%. MiRDeep finds 17 new miRNAs in *C. elegans* (signal over noise) at least of 3:1, and numerous novel miRNA genes are found in *C. elegans*, *planaria*, etc.

To understand biological miRNA function, it may be important to search for combinations of miRNA binding sites for sets of coexpressed miRNAs. The current predictions by TargetScan, PicTar, EMBL, and EILMo all require stringent seed pairing and have a high degree of overlap. PicTar is a miRNA target-finding algorithm, which uses a probabilistic model to compute the likelihood that sequences are miRNA target sites when compared to the 3′ UTR background. This computational approach successfully identifies not only microRNA target genes for single microRNAs but also targets that are likely to be regulated by
microRNAs that are coexpressed or act in a common pathway. Massive sequence comparisons using previously unavailable genome-wide alignments across eight vertebrate species strongly decreased the false positive rates of microRNA target predictions, allowed PicTar to predict (above noise), on average, ~200 targeted transcripts per microRNA. PicTar has been used to predict targets of vertebrate and *Drosophila* miRNAs. Vivo experimental validation suggests a high degree of accuracy (80–90 %) and sensitivity (60–70 %) for the PicTar algorithm in flies.

The most pressing question to arise from the discovery of the hundreds of different miRNAs is what all these tiny noncoding RNAs are doing. MiRNA is highly involved in posttranscriptional control. Standard Affymetrix analysis before and after knockdown with “antagomirs” results mRNA roughly regulated 300 gene go up and 300 down. For example, B cell numbers were reduced in the spleen of mice ectopically expressing *miR150*, just as the gene *cMyb* does. Is *cMyb* a direct target of *miR150*? It is really true. Experiments show ectopically expressed *miR150* downregulates *cMyb* protein. Further mutation experiments were done on *miR150* binding sites which were described by PicTar, and results show that some special sites are essential for downregulation of *cMyb* by *miR150*. There are many SNP databases available which give much useful information about evolution and disease and so on. SNP is at least 1 out of 100 bp in human genome. By population genetics methods to study function and evolution of short *cis*-regulatory sites using SNP, the binding sites of MiRNA can be further analyzed to find more evolution information. Many researches show ectopic, moderate experiment of a single miRNA can profoundly interfere with a developmental process by moderately repressing protein levels of a transcription factor. In evolution content, there may be broad rewiring of posttranscriptional control by miRNAs during metazoan evolution. Since many miRNAs and recognition motifs are deeply conserved but regulatory relationships only weekly conserved, it is interesting to further research whether lots of miRNA target rewiring during organism evolution. Many studies have suggested that TF is more conversed and their binding sites are low conserved, while miRNA’s targets are more conserved. In this direction, the relationship and difference in evolution between TF and miRNA is also another charming topic. Although much progress is done in miRNA finding, target recognition, and special function description, there is much space to improve and lots of work to be continued.

### 12.10 Topics in Computational Epigenomics

Michael Q. Zhang

Department of Molecular and Cell Biology, The University of Texas at Dallas,
800 West Campbell Rd, RL11, Richardson, TX, USA 75080

Tsinghua National Laboratory for Information Science and Technology,
Tsinghua University, Beijing 100084, China

This talk is divided into two parts: (1) introduction about epigenetics and epigenomics and (2) genome-wide study on CTCF binding sites.
Epigenetics refers to the inheritable changes in gene expression that cannot be attributed to changes in DNA sequences. There are two main mechanisms, RNAi and histone modification. Epigenomics is the genome-wide approach to studying epigenetics. The central goal is to define the DNA sequence features that direct epigenetic processes.

There are a number of epigenetic examples. In yeast, these include yeast silent mating-type loci and the silence of loci near telomeres and ribosomal DNA; in fly, the position-effect variegation (PEV) for eye color determination; and in plant, the first discovery of RNAi. In mammal, the well-known example is genomic imprinting, that is, epigenetic modification leading to differential expression of the two alleles of gene in somatic cells of the offspring.

The structure of chromatin plays important roles in epigenetic control. Two important histone modifications for transcriptional control are acetylation and methylation. The latter is mostly involved in the silencing of expression. Different combination of histone modification marker is referred to as the histone code.

The relationship between DNA methylation and histone modification is unclear. During development, methylations change. Different methylation patterns may determine cell types. Methylation pattern is largely fixed in post-development stage.

DNA methylation in CpG island and promoters also play roles in gene regulation. Now we show some of our works. In one study, we have experimentally identified whole-genome methylation markers and computationally identified sequence patterns associated with markers. Our studies show that unmethylated regions contain many CpG islands, more likely to be in promoter region and more conserved. Young Alu elements often localize close to unmethylated region. We also found that genes with methylated promoters are under-expressed. We also found sequence motif near unmethylated region. We have tried to build classifiers to predicted methylated and unmethylated regions and developed methods HMDFinder and MethCGI.

Recently, in collaboration with several labs, we tried to identify insulator (CTCF) binding sites in the human genome. CTCF is a DNA-binding domain zinc finger family, which accounts about one third of human TFs. CTCF is a ubiquitous, 11-zinc finger DNA-binding protein involved in transcriptional regulation, reading of imprinted sites, X chromosome inactivation, and enhancer blocking/insulator function. In cancer, CTCF appears to function as a tumor suppressor gene. Genome is organized into discrete functional and structural domains. Insulators are defined by function, block the action of distal enhancer when inserted between the enhancer and promoter, and serve as boundaries between heterochromatin and euchromatin. Mutations in insulators are implicated in human diseases. We performed whole-genome CTCF ChIP-chip experiment to identify CTCF binding sequences. We then validated by real-time quantitative PCR, showing a specificity >95 % and a sensitivity >80 %. We show CTCF sites are conserved, and CTCF binding correlates with gene number but not with chromosome length. On average CTCF binding sites are 8 kb away from the nearest 5' end, which is consistent with insulator function.

We also found two distinct patterns of CTCF distribution: (1) CTCF binding sites are depleted in superclusters of related genes, such as zinc finger protein clusters and olfactory receptor clusters. (2) CTCF binding sites are highly enriched in genes
with a larger number of alternate transcripts, such as immunoglobulin locus. A vast majority of CTCF binding sites are characterized by a specific 20mer motif, present in >75% of all CTCF binding sites. The motif is similar to previously characterized motif and extends and refines at several key nucleotide positions.

In conclusion, we have defined a large number of CTCF binding sites in the human genome that exhibit unique distribution among sequence-specific transcriptional factors. CTCF binding sites can function as enhancer blockers and chromatin boundary elements. These CTCF binding sites are uniformly characterized by a unique 20mer motif, which are conserved and display interesting model of change throughout the vertebrate genomes, suggesting.

12.11 Understanding Biological Functions Through Molecular Networks

Jingdong Han (✉)
Chinese Academy of Sciences, Beijing, China

In this talk I’ll show several studies based on molecular networks. Networks consist of nodes and edges connecting them. Some important concepts include degree, the number of edges of a node and hub, node with high degree, and characteristic path length, the average length of path between any two nodes in the network. There are mainly three types of molecular networks: protein network, regulatory network, and miRNA–mRNA network. We have been focused on protein networks.

Our first example is to map the structure of interactome networks of C. elegans. We use yeast two-hybrid systems. The resulted network shows some surprising features. For example, yeast orthologs are almost homogeneously distributed in the network, instead of as a core module. The network also shows a power-law degree distribution.

The next step is to annotate the mapped networks by data integration. Our first example is to extract molecular machines that function in C. elegans early embryogenesis, by integrating expression profiling, binary interactome mapping, and phenotypic profiling. We searched for fully connected subnetworks (cliques) and derived methods to prediction protein functions by completing incomplete cliques. Our second example is to generate breast cancer network using a similar approach. Based on this network, we predict new proteins involved in breast cancer. Proteins interacted with four seed proteins (known as breast cancer genes) are pulled out to generate a breast cancer gene network. A significant proportion of the network is not covered by existed literature. We also predicted a large number of high-quality interactions by Bayesian analysis. We integrated 27 omics datasets and predicted about 300,000 interactions among 11,000 human proteins. All data are deposited into the IntNetDB database.

With high-confidence network available, one can study other advanced topics. For example, we have shown that hubs can be further divided into two types: date
hub and party hub, by the coexpression level hubs and their partners. In silico simulation of node removal shows that date hub is more important to the main CPL.

Recently network motifs are introduced to the study of molecular networks. One typical motif is feedback, including both positive feedback and negative feedback. Another interesting type is toggle switch. Please refer to the review of on Nature, 2002.

Another question we tried to answer is about aging, a complex process that involves many seemly unrelated biological processes. Could this reflect different aspects of concerted changes of the aging network? Is this network modularly designed? We proposed an algorithm to identify modules with an aging network and show existence of proliferation/differentiation switch at systems level. We also show that these modules are regulated like a toggle switch, that is, mutual inhibition.

The fifth question we have tried to answer is “can we infer the complete network from incomplete network?” The current interactome maps cover only a small fraction of the total interactome (3–15 %). Most networks show scale-free property. Will it be an artifact due to incomplete mapping? Our simulation studies show that limited sampling of networks of various topologies give rise to scale-free networks. Thus, at current coverage level, the scale-free topology of the maps cannot be extrapolated to the complete interactome network.

The next step of network study is to make prediction by network modeling at different levels, including differential equation model, Boolean model, Bayesian model, and statistical correlation models. Here we show a study using network statistics to predict disease-associated genes. The authors tried to find out the disease gene within a linkage loci containing more than 100 genes. The method first pulled out direct neighbors of the gene under study and then check whether these neighbors are involved in similar diseases. Genes with many partners involved in similar diseases are ranked higher.

### 12.12 Identification of Network Motifs in Random Networks

Rui Jiang (✉)
Tsinghua University, Beijing, China

The numerous persuasive evidences from many experimental investigations have proved that networks across many scientific disciplines share global statistical properties. The “small world” of short paths between nodes and highly clustered connections is one characteristic of natural networks. It has also been shown that many networks are “scale-free” networks, in which the node degrees follow a power-law distribution. However, recent studies have shown that the local different structures, termed “network motifs,” are widespread in natural networks. Network motifs are these patterns that occur in networks at significantly higher frequency than the average in randomized networks and have been found in a wide variety of networks, ranging from the World Wide Web to the electronic circuits, from
the transcriptional regulatory networks of *Escherichia coli* to the neural network of *Caenorhabditis elegans*. Therefore, network motifs are considered as the basic building blocks of most complex networks.

As in most previous studies, the networks and network motifs have been represented as deterministic graphs in which the connections between nodes include the presence and absence states. Nevertheless, most biological processes in organisms are in a dynamical equilibrium adapted to the rivalship between the internal conditional variations and the external environmental perturbations, rather than static. For example, in a living cell, DNA-binding proteins are believed to be in a balance between the bound and unbound states, thus introducing uncertainties in protein-DNA interactions. The evolution of regulatory pathways in cells is itself a stochastic process, and some protein-DNA interactions could change without affecting the functionality of the pathways. Functionally related network motifs are therefore not necessarily topologically identical. Therefore, networks and network motifs are intrinsic uncertain. Additionally, incomplete and/or incorrect observations due to experimental resolutions, systematic errors, and random noises also introduce considerable uncertainties into the observations. This situation prevails in biological interaction networks constructed by using data collected by high-throughput techniques such as the yeast two-hybrid assays and the chromatin immunoprecipitation method. With the intrinsic and experimental uncertainties, it is more suitable to describe the biological networks and network motifs as stochastic graphs in which connections between nodes are associated with probabilities and discuss the network motif identification problem in the circumstance of stochastic networks.

A stochastic network with *N* nodes could be represented by a probability matrix \( P = (\pi_{ij})_{N \times N}, 0 \leq \pi_{ij} \leq 1 \). \( \pi_{ij} \) is the probability that the connection between node *i* and node *j* occurs in the stochastic network. By the probabilistic representation, a family of deterministic networks could be randomly drawn with respect to a stochastic network. The deterministic network is described by using the adjacency matrix \( A = (a_{ij})_{N \times N} \) which charactering the flow directions on the network. For directed graphs, \( a_{ij} = 1 \) if there is a directed edge pointing from node *i* to node *j*. For undirected graphs, \( a_{ij} = 1 \) if an undirected edge connects node *i* and node *j*. Network motifs are a small set of subgraphs embedded in stochastic networks with significant occurrences at number and could be also described by using probability matrices. A subgraph could be found in networks as various isomorphic structures with the same topology but dissimilar node orderings. These isomorphic structures could be mapped to each other by permuting their node labels. In the circumstance of stochastic networks, the network motif could be represented by a parametric probabilistic model, \( \Theta_f = (\theta_{ij})_{n \times n}, 0 \leq \theta_{ij} \leq 1 \), which could be used to calculate the occurrence probability of a given *n*-node subgraph isomorphic structure *P* in the foreground. Another parametric probabilistic model is established to describe the suitable random ensemble drawn from the background and is characterized by a three-tuple \( \Theta_b = \{I, O, M\} \), where *I*, *O*, and *M* are the distribution of the in, out, and mutual degrees for the stochastic network, respectively. Given the foreground and background mixture probabilistic model, a given *n*-node subgraph isomorphic structure *P* could be sampled from the
population in the following procedure: randomly choose a generation distribution; for the foreground with the probability \( \lambda_f = \lambda \), sample a subgraph according to the probability matrix \( \Theta_f \); for the background with the probability \( \lambda_b = 1 - \lambda \), generate a subgraph under the parametric model \( \Theta_b \); and then select one isomorphic structure of the subgraph at random. Therefore, mathematically, the network motif identification problem could be formulated as searching for the mode of the likelihood function which describing the co-occurrence probability of the observations. Expectation maximization (EM) is an efficient statistical inference method to estimate the parameters which maximizing the likelihood of the mixture model through introducing some appropriate latent variables.

In experiments, this method has been applied to a wide range of available biological networks, including the transcriptional regulatory networks of *Escherichia coli* and *Saccharomyces cerevisiae*, as well as the protein-protein interaction networks of seven species [*E. coli*, *S. cerevisiae* (core), *C. elegans*, *Helicobacter pylori*, *Mus musculus*, *Drosophila melanogaster*, and *Homo sapiens*], and identifies several stochastic network motifs that are consistent with the current biological knowledge. Additionally, the ChIP-chip datasets for the *S. cerevisiae* regulatory network are also involved in the experiments. In the 3-node case, the feed-forward loop motifs are identified in the transcriptional regulatory networks for both species. Recent studies have shown that the feed-forward loop serves as a sensitive delay element in regulatory networks and could speed up the response time of the target gene’s expression following stimulus steps in one direction (e.g., off to on) but not in the other direction (on to off). In 4-node case, the patterns with the solid biological evidences in regulatory networks are the stochastic bi-fan motifs in which the interaction between two regulators is of two states. Undirected stochastic motifs are also identified in protein-protein interaction networks via the available datasets.

The approach with the mixture probabilistic model and EM solver could be used to identify the stochastic network motifs embedded in the stochastic networks. The proposed unified parametric probabilistic model takes the intrinsic and experimental uncertainties into considerations, could capture the stochastic characteristics of network motifs, and could be applied to other types of networks conveniently.

### 12.13 Examples of Pattern Recognition Applications in Bioinformatics

Xuegong Zhang (✉)
Tsinghua University, Beijing, China

Concerning about the pattern recognition applications in bioinformatics, we give some examples on the basis of our major research interests. These interests are mainly surrounding the central dogma and focused on the molecular regulation systems in eukaryotes, especially in human beings. Most of the relevant interests are concentrated on the aspect of genome, including transcriptional regulation of genes,
splicing regulation, regulation by and of noncoding RNAs, epigenomics, interaction of regulatory factors, and posttranslational modifications. Other interests include population genetics and traditional Chinese medicine.

First, on the issue of alternative splicing, two examples are mentioned. One of them is about how to identify the factors that make a splicing site alternative. From a pattern recognition point of view, this issue can be regarded as a classification problem to differentiate alternative splicing sites from constitutive splicing sites, and by training a machine to imitate the real splicing mechanism, the reasonableness of our supposed mechanism can be judged by the performances of the classification. The other problem is about the VSAS (very short alternative splicing events), and we feel interested to know what role the alternative splicing plays in this process. In this problem, the observed VSAS events can be classified into two groups depending on whether they insert new structure domains in the proteins, and they might be of different evolutionary status.

Next, we talk about RNA regulation, including computational recognition of noncoding RNAs especially microRNAs, study of their regulatory roles and target genes, mechanism of their transcription and regulation, and RNA editing. Most of the relevant problem can also be described or transformed into a pattern recognition problem; for example, the identification of microRNA can be solved as a classification problem of microRNA vs. pseudo microRNA.

Another point need to be noticed is the feature selection problem. Consider that we are doing a classification problem; for example, classifying the microarray data, there are two things that may need to be done, which are feature selection and classifier construction. There exist two kinds of methods to do feature gene selection, one kind of which is called two-step procedures (filtering methods), while the other kind of which is called recursive procedures (wrapper methods). For the filtering procedures, some criteria are first designed to select differentially expressed genes with certain stand-alone methods, and then classification is implemented using the selected genes. For the wrapper procedures however, the gene selection and classifier construction are done synchronously; that is, the entire genes are utilized to construct classifiers, and after that, useless genes are deserted, while useful genes are left. This gene selection process is recursive, and support vector machine (SVM) can be used in the recursive gene selection procedures. There is also an example about a story between SVM-RFE and R-SVM. The idea of the two methods is almost the same, both using the entire genes to train SVM (we just use linear SVM) and then selecting a subset of genes that gives the best performances. The advantage of the wrapper methods lies in that the following selection criteria are consistent with the classification criteria and possible collaborative effects of genes are considered. The main disadvantage lies in that the rankings are not comparable and there is no cutoff for selection. As a result, the problem existing in the wrapping methods is that it cannot evaluate the significance of rankings. In the application of the traditional statistical test method, the comparison between two samples is first transformed to the comparison of mean difference, and then the difference is ranked and measured by $p$-value, based on which we can draw conclusion accordingly. While using the machine learning method, the comparison between two samples is
first transformed to the contribution in SVM, and then the contribution is ranked and the conclusion is drawn, but lacking the measures of the ranks, some \( p \)-value-like statistics as in the statistical test method.

On the issue of gene expression and transcriptional regulation, the main topics include discovery and recognition of transcription factor binding sites (TFBSs), prediction of transcription starting points (TSSs), tissue- and developmental stage-specific regulations, as well as analysis of regulatory pathways and networks. And some problem can also be described as a pattern recognition problem.

On the issue of epigenomics, we mainly do the following research: predicting the methylation status of CpG islands, genomic factors that decide on the methylation of CpG islands, and so on. As is well known, CpG islands are areas in the genome within which some Cs are methylated, while others are not. So the pattern recognition task in front of us is to classify the methylated from the unmethylated, and our classifier is constructed and trained using the known methylated and unmethylated CpG islands.

12.14 Considerations in Bioinformatics

Yanda Li (✉)
Tsinghua University, Beijing, China

On the issue of how to understand bioinformatics, different understandings may lead to different direction choice, focus, working methods, as well as final results of the study. Some people regard bioinformatics as serviceable and auxiliary means for biologists, and some regard it as applications of computer science on information processing or technique of pattern recognition and machine learning on molecular biology data analysis. From Prof. Yanda Li’s point of view, however, bioinformatics plays an essential and central role in terms of the molecular biology or life sciences, the reason of which is explained by doing an analogy from understanding bioinformatics to understanding characteristics of a machine.

Since everything in this world is interconnected with each other, we may start from considering how we understand the characteristics of a machine or device. As is well known, the characteristics of a device are determined by the whole, but not the part. And though the whole is complicated on many aspects such as component structure, materials, driving force, and peripherals, its characteristics are mainly determined by the interaction of the components. The method of describing this interaction, which needs to omit minor factors, highlighting the dominating ones, is called system modeling. Through the mathematical (or computing) analysis toward the model, we can understand the response of the system to the outside world and also understand the overall characteristics of the system (static, dynamic) as well as various local features. However, no matter what type the system is (mathematical or of other type), the interaction of its components is actually characterized by relationship between informations. Therefore, information system modeling is the
basic method to understand the overall characteristics of a system. It is even believed that understanding toward machines and that toward living creatures is substantially the same. Of course as an open problem as yet, some people oppose the above analogy between machines and living creatures, whereas Wiener and many posterior scholars hold similar views on behalf of the above analogy. This issue is then elevated to a philosophical height, say, recognizing living creatures can be understood in the same manner as machines means life is nothing but a physical interaction at an advanced stage, which has no insurmountable difference from low-level interactions in nature. There is no God and the soul. Hence, the fundamental task of bioinformatics is to find the internal information system model of the living creatures, that is, to understand the living creatures on the whole from an information system perspective.

Bioinformatics is calling for system theory and cybernetics. Since the biological system itself is a very complex system, its overall characteristics mainly come from interactions, and cybernetics is just a theory studying interactions. These interactions, to put it further, are actually characterized by information interactions. Therefore, studying bioinformatics from a perspective of information, rather than separately analyze the chemical composition, is a more promising way to simplify a complex system.

The development of bioinformatics can be traced from data management, sequence alignment, and biological molecule identification to various molecular interactions as well as biological molecular network modeling and analysis. Now the analysis of biological molecular network will be an important issue in front of bioinformatics. Analysis of the network includes structural analysis (simplification, clustering, and modularization) and the dynamic performance analysis (evolution and the response to the disturbance), by means of evolution and analysis of random networks.

Currently in the post-genome era, functional genomics and systems biology will be new studying directions we are facing. The gradually close integration of biological science, information science, controlling theory, and systems science enables the information analysis of complex systems become a core content in the post-genome era and bioinformatics. Besides, bioinformatics has a closer integration with diseases and medicine, and the internal coding system of living creatures may also meet breakthroughs, including the coding of regulation and cerebral neurobehavioral cognition. Specifically, the major challenges come from eight research area:

- Identification and characterization of pathogenic gene in complex diseases
- Noncoding DNA analysis and data mining
- Information analysis of oriented-differentiation problem in the stem cell research
- Analysis of protein-protein interaction networks
- Identification and prediction of foreign gene control (epigenetic gene control)
- Information analysis of brain-computer interface
- Analysis and formula optimization toward compound prescriptions of TCM by means of the complex system theory
- Character representation