Front. Electr. Electron. Eng. China 2010, 5(3): 391–404
DOI 10.1007/s11460-010-0109-8

RESEARCH ARTICLE

Runsheng CHEN, Geir SKOGERBØ

Bioinformatics — Mining the genome for information

Abstract Since the launching of the human genome sequencing project in the 1990s, genomic research has already achieved definite results. At the beginning of the present century, the complete genomes of several model organisms have already been sequenced, including a number of prokaryote microorganisms and the eukaryotes yeast (*Saccharomyces cerevisiae*), nematode (*C. elegans*), fruit fly (*Drosophila melanogaster*) and thale cress (*Arabidopsis thaliana*) as well as the major part of the human genome. These achievements signified that a new era of data mining and analysis on the human genome had commenced. The language of human genetics would gradually be read and understood, and the genetic information underlying metabolism, development, differentiation and evolution would progressively become known to mankind. Large amounts of data are already accumulating, but at present many of the rules that should guide the understanding of this information are yet unknown. Bioinformatics research is thus not only becoming more important, but is also faced with severe challenges as well as great opportunities.

Keywords bioinformatics, genome, algorithm, noncoding RNA, biological network

1 What is bioinformatics?

Bioinformatics is a recent and developing interdisciplinary field of study. Its research object is the biological data derived from genome studies, whereas its research tools are mathematical and computational.

In the 1950s, the elucidation of the DNA double helix greatly stimulated the development of molecular biology, and life science research entered a stage of unforeseen rapid development. The number of accumulated publications in genetics and molecular biology rose rapidly from close to one hundred thousand in the mid-1960s to more than two hundred thousand by the end of that decade, that is, a doubling within a span of 3–4 years. Thereafter, the figure reached three hundred thousand by the mid-1980s, signifying a yearly production of 6–7 thousand papers, and tipped four hundred thousand by the mid-1990s, as the average yearly increase reached 10 thousand papers. A second rapid growth in publication occurred in the five years leading up to the end of the millennium, during which another hundred thousand publications were produced. In 2005, approximately two thousand papers were published every day, and in 2007 the number of papers in these fields that were registered by the PubMed database (http://www.ncbi.nlm.nih.gov) increased by 670 thousand within this single year. At the same time, DNA sequence data increased rapidly. During the three years from 1979 to 1982 when the American nucleic acid database 'GenBank' was established [1], approximately one hundred sequences were entered, amounting to some hundred thousand base pairs of total sequences. By the end of the 1990s, this figure had risen to several billion base pairs, including a large number (several million) of human expressed sequence tags (ESTs), cDNAs, single gene sequences and genomic sequences. Reaching the end of year 2000, the 10 million DNA sequences registered in international databases exceeded a total of 10 billion base pairs sequence. This is to say, within the very short period of 18 years, the amount of DNA sequence data increased a hundred thousand fold[1)]. The company 454 Life Sciences launched its 454 FLX pyrosequencing platform in 2005 (454 Life Sciences was acquired by Roche in 2007), the US company Illumina made available the Solexa Genome Analyzer Platform in 2006, and ABI presented its SOLiD sequencer in 2007. With the development of these so-called 'second generation' sequencers, nearly any laboratory may produce

Received March 31, 2010; accepted April 25, 2010

Runsheng CHEN (✉), Geir SKOGERBØ
National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
E-mail: crs@sun5.ibp.ac.cn

1) Obtained from http://www.ddbj.nig.ac.jp/ddbjnew/statistics-e.html and http://www.ebi.ac.uk/genomes/mot/

over and above 100 million base pairs of sequences on a daily basis. Presently, the third generation of sequencing technology is under development and will almost certainly further improve the speed and quality of DNA sequencing. In response to this, validation and analysis of gene and protein structural data have shaken off its former slow pace, and registered data have reached a biannual doubling rate[2)]. It should be noted that due to the completion of the human genome draft and the depth and accuracy of its annotation, the detection and cloning of human disease-related genes have been greatly accelerated. Simultaneously, technologies related to DNA sequencing, such as EST analysis, single nucleotide polymorphism (SNP) analysis, microarray analysis and their corresponding data assemblies have increased dramatically in importance and amount.

The above description illustrates the challenges posed by the immensity and complexity of present day biological data. The complexity of biological data stems on one hand from the intricate structural and functional features of the living organism with all its inherent and diverse processes constituting life. On the other hand, the actual generation of experimental biological data, which commonly occur without strict rules regulating its applied semantics and syntax, further aggravates the analytical complexity of the data.

How to obtain knowledge from the immensely complex genomic data has already become the crux of genomic research. Over the last decades, electronic computers have increased the speed of numerical operations, so that today large computers have been developed to the level of handling trillions and even tens of trillions of computations per second. Only by applying this technology to the study of genomics may the enormous amounts of biological data be efficiently analyzed and its informational significance be expounded, thus shaping the new field of bioinformatics.

Generally speaking, bioinformatics is engaged in the extraction, processing, storing, analysis and interpretation of biological data, its application forging a synthesis of applied mathematics, computational science and biology in the pursuit of the biological information contained in the data.

In practical terms, bioinformatics uses genomic DNA sequence analysis as its information source. This implies discovering protein and RNA genes embedded in the sequences, elucidating the informational content of non-protein coding sequence, deciphering the grammatical code of the genetic language contained in the DNA sequence, while at the same time inferring, ordering and releasing the information of the genetic language and its RNA and protein profiles, thereby learning the rules governing metabolism, development, differentiation and evolution. Bioinformatics integrates genetic information with that of large scale protein structure and interaction to model and predict protein spatial conformation and function. Subsequently, this information will be combined with information on the specific organisms and its biochemical and physiological processes to unravel their molecular mechanisms, leading up to molecular drug design aimed at individually oriented medical care. Consequently, the bioinformatics of the genomics era should at least include three important fields of application: genomics, protein structure computation and modeling, and design of pharmaceuticals. These three closely related topics all center on the central dogma of genetic information transmission and are thus necessarily organically connected.

The central aim of bioinformatics is to bring to light "the complexity of the structure of genomic information and the basic rules of the genetic language". This implies the close integration of the three important concepts of modern natural science and technology, namely "genome", "information structure" and "complexity". Development of bioinformatics is a tool to know the language of genetics and read the complete human DNA sequence, thus ultimately leading to a better understanding of this being we call human. However, it will also inevitably enlighten our understanding of the profundity of the "information structure" and "complexity" concepts. In turn this will enable a comprehensive view of the connections between genetics, development and evolution which will greatly enhance the developments of the theory and methodology of a number of related fields such as biology, physiology, chemistry, mathematics, computational science, informatics, and systems science. The result of this will be to invigorate and increase the influence of these fields of multidisciplinary research as well as promote their further development. As of present, bioinformatics has already become one of the most attractive and prospering novel high technologies.

2 Bioinformatics research topics

The field of bioinformatics includes topics ranging from sequencing, assembly and analysis of whole genomes to the study of the rules governing biological processes of metabolism, development, differentiation and evolution. Its main aspects are:

2.1 Large scale genome sequencing, assembly and analysis

Growth, development, differentiation and metabolism of an organism are all dependent on the storage,

2) Obtained from http://www.rcsb.org/pdb/holdings.html

expression, processing and transmission of genetic information. The bulk of genetic information is stored in the genomic sequence of the four bases A, C, G and T, and represents a form of 1-dimensional digital data. Large scale sequencing is the experimental process employed to retrieve these data, and is thus the basic task of genomic research, its every step closely linked to information analysis. Figure 1 shows the outline of a typical large scale sequencing process. It is clear from the figure that from detection and analysis of light intensity to the base read-out, vector labeling and removal, sequence assembly, sequence gap filling, repeat marking, reading frame orientation and gene annotation, every step is strongly dependent on computational software and database. Among these, sequence assembly and gene annotation are the main keys to efficient genome bioinformatics analysis. Despite the changes in computational technology required by the emergence of second generation sequencing, the essential computational concepts are unaltered.

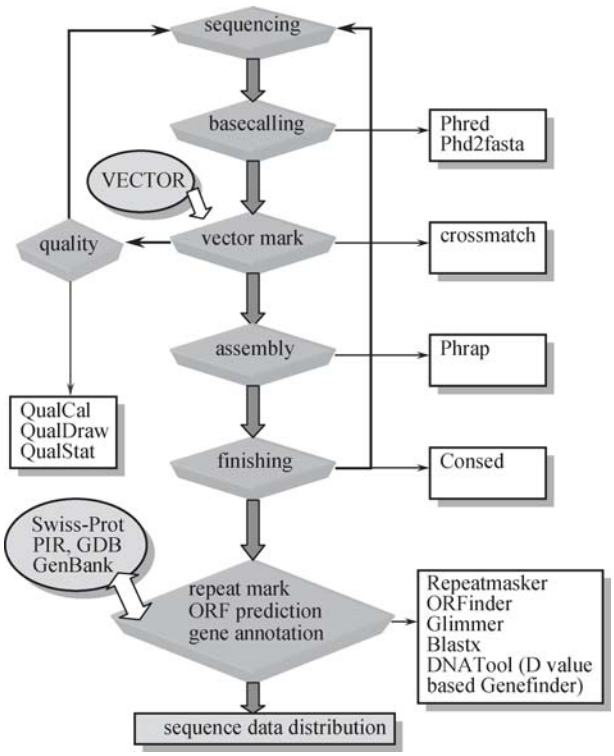

Fig. 1 Outline of a typical large scale sequencing process

The joining and assembly of sequences in a large scale sequencing project is based on the similarity of single sequence reads obtained in the experimental sequencing process, and involves linking up sequence read that represent the same genomic DNA sequence. The central computational problem is comparing the similarity of different sequence readouts. Rigorous sequence comparison involves dynamic programming, and owing to its large computational burden, applied algorithms will always utilize simplifications such as the well-known Smith-Waterman algorithm. A number of software programs applied in genome research are freely available, such as the Phred-Phrap and the GigAssembler software packages. The Phred-Phrap package was developed by Phil Green and Brent Ewing [2,3] and is composed of two main elements, the base calling software Phred and the sequence assembling software Phrap. Phrap is used for joining and assembling raw readouts from shotgun sequencing into contiguous sequences (contigs), its core being the Smith-Waterman dynamic algorithm. The GigAssembler software was created by W. James Kent and David Haussler [4] based on a greedy algorithm [5], and operates by aligning and orientating overlapping fingerprint clone contigs to assemble the overall genome sequence.

Even though a number of different software packages may in principle solve the problems involved in linking up and filling the gaps in enormous sequence data, the problems created by the large numbers of repeat sequences in the genomes are hard to conquer. As a consequence, various sub-cloning approaches have been applied, such as the BAC-by-BAC sequencing strategy, which strongly reduces the number of repeats within a single BAC sequence.

Subsequent to genome sequence assembly the main task is sequence analysis, the main problems of which being the determination of reading frame and gene annotation. The essence of this work is to distinguish various sequence characteristics and identify protein coding sequence in the enormous amount of genomic sequences. To this end, a number of specific signal sequences are utilized, these being start and stop codons, splicing and branch sites, promoters and various protein binding sites. It is also possible to utilize complex characteristics, thereby developing a variety of different methodologies, such as artificial neural networks and pattern recognition software, linguistic analysis methods, and Hidden Markov Models. Examples of such software are GRAIL (http://compbio.ornl.gov/Grail) [6,7], GeneParser (http://beagle.colorado.edu/~eesnyder/GeneParser.html) [8], GENEID (http://www.imim.es/GeneIdentification/Geneid/geneid_input.html) [9], Genlang (http://cbil.humgen.upenn.edu/~sdong/genlang_home.html) [10], Genie (http://www.cse.ucsc.edu/~dkulp/cgi-bin/genie) [11], and HMMgene (http://www.cbs.dtu.dk/services/HMMgene/) [12]. Again, other methods have been based on non-linear programming of fractal dimensions [13] and the mathematics of complexity [14]. Our research group has applied cryptological methods to discern coding sequences, based on the coincident index (IC) of a DNA sequence, calculated by the following equation:

$$\mathrm{IC} = \left(1 - \frac{64 * \mathrm{STP}}{I * L}\right) * \mathrm{IC}'. \qquad (1)$$

STP denoting the number of stop codons, and I and

L being the number of words and characters, respectively, in the query sequence [15]. Applied to the AIDS-associated ARV-2 the method produced good results, enabling efficient determination of three coding regions within the viral sequence. Currently, the often used software packages for identification of coding genes in genomic sequence data are GeneFinder, GENSCAN and others.

2.2 Utilizing EST databases to identify novel genes and SNPs

Discovery of novel genes is a central issue of genomic research. Gene prediction solely based on computational methods is applicable to prokaryotes, since their genes rarely contain introns, and also to small eukaryote genomes with few introns, such as the 13 Mb yeast genome containing only about six thousand genes, of which 60% were identified by computational analysis. However, the situation is quite another for the human genome, as human genes are composed of relatively short exons (on average 50 codons or 150 bp long) and long introns, the largest of which may exceed 10 kb. Consequently, in addition to computational methods, cDNA and EST data along with comparative genomics are employed to predict genes in the human genome.

Expressed sequence tags (ESTs) are short cDNA sequences representing fragments of expressed genes. As of March 2010 the GenBank EST data (dbEST release 031910) contained more than 8.3 million human EST sequences, covering 90% of all known human genes. How to utilize the information contained in the ESTs to identify new genes has become an important research topic in recent years. This is basically carried by aligning and joining ESTs, thereby obtaining new gene candidates. Although this route is feasible, the programming is complex and the computational load is high. Crucial steps include removal of erroneous sequence, such as primers, 3' and 5' end noncoding sequence and non-human EST contamination, and creation of special databases such as seed sequence libraries. A test carried out by NCBI suggested that approximately 1.5% of the sequences in the EST database are erroneous (http://www.ncbi.nlm.nih.gov/dbEST/synopsis_detailsR.html), a fact worth noting for researchers engaged in this field. Internationally, several lists of genes based on EST data have already emerged, such as UniGene (ftp://ncbi.nlm.nih.gov/pub/schuler/unigene) [16], Merck-Gene index (http://genome.wustl.edu/est/esthmpg.html) [17], GenExpress-index (http://www.cshl.org) [18]. These gene indices provide a genetic framework which greatly facilitates related research.

Research on single nucleotide polymorphisms (SNPs) has followed the intensification of the genomics research in recent years, and as the speed of DNA sequencing has increased, SNP studies have attracted extensive attention. Although the SNP frequency is quite low, a small number of frequently occurring polymorphisms explain a large amount of the observed heterozygosity. Moreover, human genetic differences do not only appear as individual polymorphisms, but more commonly occur in the form of a series of interlinked alleles (haplotypes). Through such concentrated SNPs it is therefore possible to define common haplotypes and their association to human disease. In 1998, an international program for discovery SNP was establish, mainly based on the principle that if EST data from a single gene could be aligned to several, slightly different transcripts, the differences between these might represent SNPs. Later, the International SNP Map Working Group has also utilized large scale sequencing data, combined the multi-ethnic test panel with the method of comparing the large overlapping fragment, thereby identifying 1.4 million SNP (one SNP for every 1.9 kb of the human genome), and carried out preliminary analysis on the SNP material.

2.3 Informational and structural analysis of noncoding sequence

Just as scientists were getting a grasp of large scale genome data, they found that the regions of DNA encoding protein (i.e., what is generally known as 'genes') only occupied a small fraction, not exceeding 3% of the genome. The remaining 97% or so of the DNA sequence was mostly without any clear function. At first, researchers were wont to label this DNA as 'noncoding DNA' or simply as 'junk DNA'. After whole genome comparisons were carried out it was discovered that lower organisms such as viruses and bacteria only contained small amounts of 'junk DNA', whereas in higher animals and plants, 'junk DNA' might even make up the major part of the genome. This is to say, as one move from 'lower' to 'higher' organisms, from simple to complex, from cells with low information content to organisms with high information content, the amount of noncoding DNA in the genomes increases. This implies that 'junk DNA' may contain information related to the complexity of organisms. Whole genome comparisons have shown that the number of genes (i.e., genes encoding proteins) in the fly and nematode (14000 to 20000) are only 2–3 times that of yeast (approximately 6000), and only slightly lower than the gene number in human and mouse (approximately 24000). Thus, the increase in gene numbers does not reflect the increase in biological complexity.

The results of large scale transcriptional analysis in recent years suggest that sequences in the noncoding parts of the genome may be expressed in the form of

noncoding RNAs, and accumulating evidence suggest that noncoding RNA has important functions. The research on microRNA is the most prominent example. Carrying out research on the nematode *C. elegans* at Dartmouth Hospital, Victor Ambros' group used gene gun technology to investigate the functions of certain genes involved in nematode development. The nematode commonly undergoes four larval stages before maturity, and Ambros' group found a gene that when mutated caused the nematode larvae to remain in the first larval stage. What surprised people were that this gene did not encode a protein coding gene, but rather a very short RNA, later named 'microRNA'. Subsequent research showed that this RNA gene could also be found in the fly, in mollusks, in fish and even in the human genome. Presently, researchers have found microRNA genes in fungi and plants, in both mammals and nonvertebrate animals, and even in prokaryote organisms, and have demonstrated that these genes have important biological functions. In addition to the 21–24 nucleotides long microRNAs, a number of other types of small RNA have been discovered. In July 2006 two independent research groups published data on a novel group of testis-specific small RNAs that interacted with Piwi-class proteins of the Argonaute family. These RNAs were about 26–31 nucleotides long, slightly longer than the previously described microRNAs, and were labeled Piwi-interacting RNAs, or piRNAs [19].

Our laboratory has also discovered two new classes of short noncoding RNAs [20]. The genomic sequences corresponding to functional noncoding RNAs are named 'noncoding RNA genes'. The recent years have not only seen the discovery of noncoding RNAs in the size range of some hundred nucleotides, but also very long noncoding transcripts of several thousand nucleotides have been found. One long noncoding RNA named Xist is of high importance for reproduction and development in mammals through its function in the dose-compensation phenomenon. The expressed region of the Xist gene is measured in kilo bases, and the Xist transcript can alter the structure of the mammalian X chromosome, thereby silencing the entire chromosome [21]. This mechanism maintains balanced expression levels of genes on the X chromosome between males and females. Noncoding RNAs do not only participate in a multitude of fundamental biological processes, but are also implicated in a variety of animal and human pathologies. The noncoding RNA PCGEM1 was discovered in 2000, and four years later reported to be involved in prostate cancer development [22]. His-1 is a noncoding RNA gene that in 1999 was reported to be connected with mouse leukemia and shown to participate in oncogenesis and control of the cell cycles [23]. MALAT-1 is a noncoding transcript of more than 8000 nucleotides, which in 2003 was related to non-small cell lung cancer [24]. The emergences of noncoding RNAs have aroused new interest in 'the RNA world' and 'RNA as the original molecule of life'.

The ENCODE (Encyclopedia of DNA Elements) project has provided a deeper and more comprehensive understanding of the genomic complexity. The ENCODE project is the first systematic attempt to describe the position and organization of all categories of functional elements in the human genome. The ENCODE research topics include protein coding genes, non-protein coding genes, regulatory regions, and DNA elements involved in chromosome stability and replication. Analysis of the 200 most recent data sets revealed that 93% of the DNA is transcribed into RNA under some condition, a large number of these transcripts are noncoding, and some transcripts are apparently 'fusion transcripts' that may be a result of molecular interactions. A large number of novel transcription start sites were detected, many of which show various histone modification. The data also showed that modifications on distal regulatory elements are different from those of proximal promoters. In short, these new achievements indicate that histone modifications, DNase sensitive sites and transcription as well as replication are extensively interconnected. This strongly supports the hypothesis that the genome has a higher level of functional organization. Consequently, the human genome itself is an extremely complex network, and the amount of non-functional DNA (or so-called junk DNA) is in reality very low. In this picture, protein coding genes are no more than one of several specific functional DNA sequence elements. The ENCODE project challenges the common perception that the genome is composed of isolated genes in a vast amount of "useless DNA fragments", and suggests the alternative view of the genome as a complex network system.

Noncoding sequence, noncoding genes and noncoding RNA have presented bioinformatics research with unprecedented favorable opportunities and numerous challenges. Many years of work on coding gene prediction and protein simulation has produced a series of computational methods that are not applicable to noncoding research. Since presently known noncoding RNAs are mostly quite short (20–200 nt) and have no three-letter codons, it is difficult to use statistical methods to identify their characteristics. Another problem that urgently needs a solution is more and better methods for prediction of the noncoding RNA spatial structure.

2.4 Origin of genetic code and evolutionary research

Since the publication of Darwin's *The Origin of Species* in 1859, the theory of evolution has been a major influence on the development of human natural science and natural philosophy. The core of evolutionary research is

the history of biological evolution (the phylogenetic tree) and the exploration of evolutionary mechanisms. Along with continuous development of molecular biology since the middle of the last century, evolutionary research has entered the molecular level [25]. At present, research on molecular evolution is already its most important tool, and evolutionary research has already developed a set of computational methods based on sequence information from nucleic acids, protein and whole genomes. A comprehensive computational analysis need to include the following steps:

Comparison of sequence similarity. This involves carrying out comparisons between the sequence of interest and sequences in DNA and protein databases, in order to determine the biological features of the sequence, as well as identifying the molecules with highest similarity to the sequence of interest. To this end, it is necessary to employ pair-wise sequence alignment methods. Commonly used algorithms include BLAST, FASTA and others [26,27];

Analysis of sequence homology. This involves comparing the sequence of interest to multiple homologous sequences from different organisms, in order to determine its degree of identity to these sequences. It is the most important step in the computational analysis, and requires multi-sequence comparison algorithms, of which the most commonly used is CLUSTAL [28];

Construction of phylogenetic trees. In order to reconstruct the genetic or ancestral relationship between different organisms according to the result of the sequence homology analysis, a number of different algorithms has been developed, such as PHYLIP, MEGA and others [29];

Test of robustness. To know whether the reconstructed phylogenetic tree is reliable, it is necessary to test for stability or robustness. This is often done by repeating the construction of the phylogenetic tree many times. The phylogenetic tree being accepted only when a majority (70% or more) of the branches emerge. The Bootstrap algorithm [30] is commonly used for this purpose, and corresponding software is often included in software packages used for construction of phylogenetic trees.

When discussing molecular evolution, it is necessary to point out the distinction between 'similarity' and 'homology'. 'Similarity' only reflects the resemblance between two entities, and does not imply any suggestion of an evolutionary relationship. 'Homology', on the other hand, is similarity based on a common ancestry. Thus, sequence similarity is an insufficient basis for reconstruction of evolutionary relationship, and commonly leads to errors. When it comes to research on extinct species there are additional factors that merit attention. For example, how shall protein or nucleic acid sequences that are extracted from many thousand year old fossils be compared to current data? It should also be considered how sequences may have been altered in the long period of fossilization?

Biological molecules have different replacement rates. For example, in the protein fibrin one amino acid residue is changed every 2 million years, whereas in histones, on average 30 million years pass for each amino acid replacement. In noncoding sequence, on the other hand, a nucleotide residue may be replaced every 10 to 20 thousand years [31]. Because of their differing rates of changes, different molecules can be used as time standards in evolutionary research, similar to the hour, minute and second of a clock. When used appropriately, different sequences can be used to measure varying evolutionary time scales [32].

Accompanying the enormous increase in sequence data, disputes over the relationship between sequence differences and evolution has grown hotter in recent years. Quite a few results do not support the molecular clock hypothesis. Since a phylogenetic tree that is reconstructed on the basis of a given sequence only will reflect the development of this molecule, it is not at all certain to represent the real evolutionary relationship among studied organisms. Thus, there may exist differences between the gene tree and species tree [25]. Meanwhile, the discussion on the relationship between vertical evolution (leading to orthologous sequences) and horizontal evolution (leading to paralogous sequences) is gradually attracting more attention [33].

2.5 Large scale gene expression profile analysis

As the human genome sequencing approached completion, one question naturally arose: Even though we now have a complete atlas of all human genes, to what extent can we now explain human biological functions? A series of questions were put forward which could not be answered by the genomic data, such as the timing and level of gene expression, the extent and timing of gene product modification, the effects of gene knock-out and over/expression, and so forth. In essence, knowing the DNA sequence and its genes does not imply that we know how the genes exert their functions in time and space. Much experimental evidence indicates that there are large variations in the number of genes that are expressed in different tissues. The brain shows the highest number of expressed genes, giving rise to 30–40 thousand different transcripts, whereas in other tissues only some tens or some hundreds of genes are expressed. Without precise knowledge of the number and extent to which genes are expressed in any tissue, there is no way of understanding its functional activity at the molecular level. Research work also shows that the same tissue may express different sets and numbers genes under

specific stages of development, as some genes may be expressed at a younger stage, others during middle stages, and again others may only be expressed towards mature or older stages. If the developmental stage and its state of gene expression are not taken into consideration, there is no way to accurately explain a biological process. Upon completion of the human genome, there was thus widespread anticipation that genomics research would enter a correspondingly richer and more profound era. The central issue of this era would be obtaining functional gene expression profiles, and to this end new technologies were specifically developed, both on the nucleic acid and the protein level. On the nucleic acid level this technology was the DNA microarray and on the protein level this was ion exchange chromatography and sequence determination technology, the latter also known as protein mass spectrometry or simply 'proteomics'.

The DNA microarray is one form of 'biochip'. The DNA microarray consists of a large number of DNA probes fixed to a small slide or chip of silicon, glass or metal in a certain way. The DNA microarray is used to assay and research the mRNA level in cells and tissues, commonly referred to as transcriptome research. The proteome is the complete protein output of the genome. Presently, a combination of 2-dimensional (2D) gel electrophoresis and mass spectrometry is used to analyze the functional gene expression at the protein level. With the ever-increasing amounts of data and refinement of experimental techniques for functional genomics, databases have become an inevitable component of support for these technologies, such as Swiss-Prot (protein sequences), GenBank (nucleic acid sequences), PROSITE (protein domains), PDB (protein 3D structure), SWISS-2DPAGE (2D gel electrophoresis data), O-GLYCBASE (post-translational protein modification), OMIM (human genes and genomics), KEGG (metabolism), and others. Without the material in these databases, the new technology would be of little practical use. Large scale expression profile analysis also give rise to additional methodological problems. Expression profile data and protein profile data does not only consist of numerical data, but also include graphical data and even data projected in the multi-dimensional time and space continuum. From a mathematical point of view, these are not simple NP problems, dynamic system problems or uncertainty problems, but are rather a feature of the gene expression network [34], thus requiring the development of new methods and tools. Also microarray design needs additional theoretical and software support. Thus, both the developments of biochips and proteomics technology are becoming even more dependent on bioinformatics theory, technology and data resources.

2.6 Complex biological networks and systems biology

Complex biological networks are an important foundation for a true understanding of biological structures and functions, but even if we knew every gene network and every regulatory or metabolic pathway, this might not be sufficient to explain how biological structures like cells and organisms are composed and function. It will be required to link up all related elements and unite and integrate all systems. This is the idea which systems biology is gradually working towards. One might also say that we move from sequencing the genome to functional genomics, from gene networks to network coupling and integration, and eventually to systems biology. This stepwise process is in short the present trend of genomics research.

In a commentary article in Cell in 2001, Marc Vidal expounded on the work design of systems biology [35]. The paper assumed that previous research has mostly been carried out at one given level, such as at the gene expression level, at the protein folding level, at the enzymatics level, at the level of determining protein cellular locations, and so forth. At present, the research objectives are increasingly becoming more complex, moving from the genome to the proteome and the transcriptome, and further on to the interactome, the localizome, the foldome, the metabolome and so on, each of these in reality being a functional module. In order to provide a more comprehensive idea of what constitutes a 'cell', an 'organ', a 'tissue', systems biology will inevitably have to integrate all these aspects of molecular structure and function.

In conventional molecular biology a gene is expressed as a protein, which has a given structure to accomplish a particular function, or in other words, the thinking follows a line from sequence to structure to function. Presently, accumulating evidence suggests that the expression of a single gene is often not sufficient to elicit a biological event, and that the occurrence of a biological event requires the simultaneous expression of a group of genes, thus being the result of the concerted action of several proteins. Consequently, a more realistic view of biological function would be a group of correlated genes and their interacting proteins acting in a molecular network. This picture of biological function arising from a constantly interacting macromolecular network describes the changes in theory and research practice brought about by systems biology. Having entered the era of systems biology, genomics will shift towards the integration of biological data from different sources in order to simulate the complex and dynamic biological processes. Systems biology would thus emphasize data integration and modeling of complex processes, and pay particular attention to system kinetics and the influence of the environment on the system dynamic behavior.

The theory and practice of conventional biological research is essentially multi-information integration and system modelling. Systems biology will by no means replace other disciplines, but will clarify results from studies on the gene and protein level. It will do what molecular biologist and geneticist normally do not do, that is, to connect and integrate data from these two fields in every possible way. Thus, systems biologists strive to link together data from different levels and systems of research, by seeking out any relationship or interconnected activity between an event occurring between different systems or levels. Systems biologists are thus looking for a particular interaction between different systems or levels of data. To achieve this, it is necessary to integrate all information from the gene, RNA and protein levels. This is essentially the research theory of systems biology. Systems biology research should include three parts: The first would be to assemble and integrate all theoretical work and experimental data on different levels of biological systems. Based on these data, the second step would be to suggest a model that can describe the coordination of these systems at various levels. The final step would be to use this model to predict what future events would occur in the system.

To exemplify this, we will describe a few models suggested by our research group.

2.6.1 Spectral analysis method applied to a protein interaction network [36]

Spectral analysis is an efficient method for revealing the deep structure of large scale complex data. A famous example is the outstanding study by David Gibson, Jon Kleinberg and Prabhakar Raghavan on the extraction of information from the link structure of the Internet [37,38]. We applied the spectral analysis method to distinguish topological structures in the complex protein-protein interaction network. With this method, the network is represented by a non-directional graph $G(V, E)$, that is, a vertex set including every protein as a vertex $V = \{P_1, P_2, \ldots, P_n\}$, and an edge set defined as $E = \{P_i, P_j\}$, where there is an interaction between protein P_i and P_j. The symmetrical $n \times n$ adjacent matrix is defined as $A = (a_{ij})$,

$$a_{ij} = \begin{cases} 1, (P_i, P_j) \in E, \\ 0, (P_i, P_j) \notin E. \end{cases} \tag{2}$$

The spectrum of the adjacency matrix is then an important estimate of the characteristics of vertices that are propagated through the interactions. Consider giving each vertex a score X indicating its 'importance'. Through its interactions, a vertex with a high score would then increase the score of its neighbors, and the scores of two interacting vertices would thus be mutually strengthened, leading through several cycles of calculation to the determination of their scores:

$$\Delta X_i = \sum_{j=1}^{n} a_{ij} \times X_j. \tag{3}$$

The iterative method by Gibbs et al. was applied to break the cycle. An interesting point is that X_i will approach a given value independent of its starting value. It can be shown that this value is an eigenvector of matrix A, and thus an inherent property of protein interactions. To this comes that since A is a symmetrical matrix, all its eigenvectors will be orthogonal to each other, and their properties may also be orthogonal. In other words, each eigenvector may display a particular property not belonging to any other eigenvector. From a topological point of view, the graph spectrum helps elucidate the hidden structure of the complex interaction network. We found that for an eigenvector with a positive eigenvalue, components with relatively high absolute values tended to form a quasi-clique (i.e., positive and negative extremes separately tend to form an internally connected clique), whereas for an eigenvector with a negative eigenvalue, its component proteins tend to quasi-bipartites (i.e., the protein in two opposite sets are not connected within each set, but tend to form connections with protein in the other set).

We applied the spectral analysis to a yeast protein interaction network consisting of 2617 proteins connected by 11855 interactions of middle to high confidence, and calculated the eigenvalues and eigenvectors of its corresponding adjacent matrix. By applying the following criteria to the eigenvectors with highly positive eigenvalues we obtained a set of quasi-cliques: (i) All proteins were ranked according to the absolute value of their eigenvalues, and the upper 10% were selected for further analysis; (ii) The proteins were included according to their rank, and each entered protein must interact with at least 20% of the previously entered proteins. The CC-values were used to estimate connectivity, and the parameters were adjusted to maintain clique quality; (iii) Each quasi-clique should contain at least 10 proteins. Applying these criteria, we obtained 48 quasi-cliques, the largest and smallest of which included 109 and 10 proteins, respectively, with an average of 26.6 protein per quasi-clique (allowing a protein to appear in more than one clique). A similar analysis was applied to eigenvectors with negative eigenvalues, obtaining six quasi-bipartites. These two topological structures displayed different interaction profiles. In the quasi-cliques proteins tend to interact among themselves, whereas in the quasi-bipartites most interactions occur between the two sets, whereas there are no interactions within each set. Distinguishing between these two models is not only applicable to the demonstration of the complexity of interaction networks, but more importantly provides a tool

for feasible exploration of complex networks in general. A single quasi-clique may represent several biological functions. The P-value method can be regarded as a standard to endow a quasi-clique with a main function. Hypergeometric distribution can be applied to calculate the likelihood of protein versus function in a clique, that is, the likelihood of finding k proteins in n cliques will have a given function, if C protein in the entire proteome G has this function. The P-value of randomly selected cliques is

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}. \quad (4)$$

The above criteria describe the probability that a randomly selected protein clique would be enriched for a certain functional category. If the P-value approaches zero, this signifies the likelihood that the quasi-clique has a random composition of proteins is very low. The functional categories with the lowest P-value were regarded as the main functional category of the clique. We calculated P-values for the MIPS (Munich Information Center for Protein Sequences) hierarchical functional categories, and were able to assign a single functional category to 43 of the 48 quasi-cliques, the five remaining cliques being assigned two or more functions. Analysis of the functional annotations of individual proteins showed that there was a tendency that most proteins in a quasi-clique belong to the same functional category, such as ribosome biogenesis, rRNA and tRNA synthesis, processing, transcriptional control, mRNA splicing, and others. Only a minority of the proteins had no annotated function, or was annotated with a function not coherent with the main functional category of the clique.

2.6.2 Analysis of motif preferences in transcription regulatory networks [39]

Transcription regulatory networks (TRNs) can be depicted as directed graphs. In such graphs, transcription factors (TFs) and the genes they regulate (transcription target genes, TGs) make up the nodes of the graph, and the regulatory activities exerted by TFs on TGs are indicated as an edge pointing from the TF to TG. The regulatory activity between TFs and their respective TGs is thus shown as multi-node graph. Some subgraphs have been extensively researched owing to their topological and biological meaningfulness [40,41], such as feed-forward loops (FFLs), feedback loops (FBLs), single input motifs (SIMs), and multiple input motifs (MIMs). These subgraphs are not isolated functional entities unrelated to other parts of the network. In fact, these subgraphs tend to accumulate around highly connected TFs, these often being members of several different subgraphs. The topological structure of the network indicates that the number of TGs controlled by a TF follows a power law distribution, implying that the network contains a small number of TFs that regulates a large number of genes. These highly connected TFs have been named 'transcriptional hubs (THubs)', and evidence indicates that the THubs are commonly essential genes for an organism [42,43]. The TRN can be regarded as a signal transduction structure through which indicators of external nutrition or environmental stress is transmitted, and one would expect differences in the behavior in signal transmission among the TFs [44]. Under different external conditions or developmental stages, the frequency of different subgraphs (or motifs) in the TRN has been found to vary [45]. However, there is no strong understanding of what influences these changes in motif frequency on a genome-wide scale, and no methodology for measuring the behavior of TFs towards their downstream TGs had been published. To solve this problem, we designed a set of methods to estimate the preference of TFs for particular TRN subgraphs.

It was first necessary to determine the aggregation of TFs and motifs. We concentrated mainly on the THubs. Network hubs are often determined applying a threshold to the node distribution curve, regarding as hubs those nodes whose connectivity surpasses the threshold. With respect to basic subgraphs, we selected circular ('closed') and tree-like ('open') motifs. A regulatory motif was called open if and only if it topologically forms a single closed circle, regardless of regulatory orientation. Similarly, a motif was called a 'tree' (or 'open') if and only if it did not contain any closed circle as part of the subgraph. We only considered three and four node rings or trees, as motifs with two nodes are respectively either trivial or nearly non-existing, and motifs with more than four nodes were not included due to limitations on computational capacity. These motifs covered all basic motif forms, that is, all other motifs can be composed from this selection of motifs. After selecting TF and motif sets, the preference (A_w) of a TF (H) for a given motif (P) was determined as follows:

$$A_w(H,P) = \sum_{sg \in \mathrm{SG}(H,P)} \sum_{k \in \mathrm{N}(sg)} \frac{1}{\left(d(H,k)+1\right)^2}, \quad (5)$$

where $\mathrm{SG}(H,P)$ denoting the set of all motifs (P) downstream of a TF (H), sg being a member of the set $\mathrm{SG}(H,P)$, and $\mathrm{N}(sg)$ the set of nodes in sg. $d(H,k)$ is the shortest path from the TF (H) to a node (k) in the network, and the weight factor $1/(d(H,k)+1)^2$ was introduced to quantify the reducing effect of increasing distance from the TF (H) on its 'signal strength'. For

each TF-motif pair, the preference of the TF for the motif was calculated. However, since the number of regulatory motifs downstream of a TF is variable, and similarly, and the overall frequency of different motifs in the TRN differed greatly, the preference values cannot be directly compared. To remove the influence of this variation and thus enable direct comparison of TF preferences, the following equation was applied:

$$A_N(H_i, P_j) = \frac{A_w(H_i, P_j)/\sum_{j\in \mathrm{SG}} A_w(H_i, P_j)}{\sum_{i\in \mathrm{THubs}} A_w(H_i, P_j)/\sum_{i\in \mathrm{THubs}}\sum_{j\in \mathrm{SG}} A_w(H_i, P_j)}, \quad (6)$$

where $A_N(H, P)$ is the normalized preference of transcription factor H for motif P, THubs denotes the TF set (all being THubs), and SG denoting all regulatory motifs (all 3- and 4-node rings and trees). This equation calculates the normalized preferences of all THubs for all motifs. The preferences of one TF for all subgraphs were called the 'subgraph preference profile (SPP)' of this TF, and constitute the vector of the normalized abundances of all subgraphs in the motif set (SG). The matrix of all subgraph preferences of all TFs was subsequently referred to as the 'subgraph preference landscape'.

To provide a measure of the statistical significance of difference in subgraph preferences, we created a number of random networks with the same number of inbound and outbound edges, and observed the difference between the true network and the random networks. By repeating the above calculations, we obtained a measure of the sub graph preferences in the random networks, and could also observe the robustness and stability of the TRN.

Applying the above method, we investigated the relationship between THubs and their downstream subgraph preferences in TRNs of brewer's yeast. The analysis included TRNs constructed based on data from 5 specific growth conditions (cell cycle, sporulation, diauxic shift, DNA damage and stress response). A sixth TRN was constructed based on the complete data set, labelled the static TRN[3)]. The analysis showed clearly that different TFs tended to employ different regulatory subgraphs, indicating that the 'subgraph preferences' reflected some important tendencies in the behaviour of THubs under different growth and cellular conditions. The significance of this study lies in the realisation that the variations in the subgraph preference landscape suggest changes in TFs under different external conditions, and hints at the biological mechanisms involved in these changes.

2.6.3 Noncoding RNAs are participants in complex biological network [46]

Much evidence suggest the existence of a large number of noncoding RNAs in higher organisms, and results from the ENCODE project indicates that 93% of the human genome is transcribed as RNA, of which the far larger part is noncoding [47]. Consequently, the introduction of noncoding RNAs to the biological network will increase its size many fold. Even more important is that the inclusion of noncoding RNA alters the mechanisms of interaction in the network on both the transcriptional and post-transcriptional level. Through their effects on mRNA stability and translational inhibition, microRNAs exemplify these important regulatory activities [48]. Realizing the strong influence of noncoding RNAs on biological networks, research has been initiated to investigate their impact on network complexity. The inclusion of noncoding RNAs has greatly enriched our understanding of biological networks. By interacting with certain proteins, noncoding RNAs can form a variety of RNA-protein complexes through which they exert their functions. Examples include the five snRNAs U1–U2 and U4–U6 which together with more than 75 different proteins form the spliceosome complex are responsible for pre-mRNA splicing [49], and the mouse noncoding RNA NRON which combines with 11 proteins to control the intracellular transport of the transcription factor NFAT [50]. Noncoding RNAs may act through sequence similarity to identify their RNA target sequences and recruit interacting proteins to exert their function on the target. C/D box snoRNAs act in this manner to identify methylation sites on the rRNAs, and recruits the necessary protein to carry out the modifications [51]. MicroRNA seed sequences identify target mRNAs through sequence similarity to the 3'UTRs, while the effects on mRNA stability and translation are mediated by the associated RISC complex [48]. Including noncoding RNAs into the molecular networks may also facilitate research on specific noncoding RNA functions. Similar to the protein interaction network, a large number of noncoding RNAs may form an RNA-RNA interaction network which takes part in the regulation of cellular activities, being the equivalent of the cosmological light and dark matter. In the past, biological function has mainly been attributed to proteins, and proteins have thus been the focus of most functional and regulatory research, revealing large networks involved in protein synthesis, regulation and metabolism. These discoveries have made an enormous contribution to our understanding of the nature of biological activity. The

3) The network data was obtained from http://sandy.topnet.gersteinlab.org/, excluding self-regulatory interactions

coming years will see increasing research efforts on noncoding RNAs, as more types of noncoding RNAs will be detected, their cellular functions being investigated and their interactions with proteins and roles in cellular pathways in networks being explored. If in the past we assumed that biological networks were structured with protein modules, the future biological network will be composed of both protein and noncoding RNA, forming a 'bicolor' network. We therefore believe that the concept of a 'bicolor network' will inevitably be included in network and systems biology research.

With respect to coding gene research, network clustering to identify functional modules and the employment of network neighbors to predict protein have become the new weaponry of biological research. The demonstrated success of these analytical methods should enable an even better research on the 'bicolor network' constituted by protein and RNA interactions.

Prediction of microRNA targets shows that nearly 1/3 of human genes may be under microRNA control at the post-transcriptional level, each microRNA on average regulating several hundred coding genes [48]. A number of researchers have created post-transcriptional regulatory networks composed of microRNAs and their targets. We proposed a new hypothesis: miRNAs may also regulate a particular group of noncoding transcripts called mRNA-like RNAs (mlRNAs) at the post-transcriptional level, thereby creating a noncoding RNA regulatory network. Experimental evidence has shown that eukaryote cells contain large numbers of long noncoding transcripts. Some of these resemble protein coding mRNAs in various ways; they are long, are transcribed by RNA polymerase II, are often spliced, and are furnished with both a cap and polyA tail. They, however, do not contain any prolonged or conserved open reading frame, and have thus been called mRNA-like noncoding RNAs [52]. The function of a few mRNA-like noncoding genes has already been demonstrated, but the large majority is still functionally unknown. Since mlRNAs are similar to mRNAs in sequence and structure characteristics, we surmised that mlRNAs might also be targeted by miRNAs. To show this, we borrowed a method for the analysis of the influence of miRNA expression on the stability of mRNAs [53]. The method utilizes mRNA microarray data to estimate the reliability of miRNA target prediction. Our work made use of the 34000 mlRNA collected by the FANTOM database. Among these 34000 sequences there are approximately 11000 sequences that are included in a set of microarray data derived from 20 different mouse tissues. We selected 8 miRNAs with verified tissue specific expression as our miRNA set. As miRNAs can significantly influence the stability of their target mRNAs, a tissue specific miRNA should down-regulate its targets in the tissue of miRNA expression. We therefore used the software miRanda to predict targets of the 8 tissue specific miRNAs among the 11000 mlRNAs. Applying the Wilcoxon's rank sum test on the expression profiles of the predicted targets, we found that three different miRNAs significantly reduced the expression level of the targets in the four tissues where these miRNA were expressed. The significance level was similar to that found for mRNA target analysis under the same conditions. This result extends greatly the involvement of miRNAs in regulatory networks. An even more interesting aspect is that among the mlRNAs there are a number of miRNA-encoding transcripts. Thus, these miRNA primary transcripts themselves may be under miRNA control, potentially giving rise to a complex network of mutual miRNA-miRNA regulation.

2.7 Protein structure modeling and molecular design

Along with the rapid development of genomics research, the rate of primary protein sequence determination has also increased. However, the amino acid sequence is insufficient to understand the function of a protein, and it is necessary to obtain its 3-dimensional structure. At the same time, pharmaceutical drug design also requires an understanding of the 3-dimensional structure of the receptor protein target. This is the imminent challenge put before the scientist. Presently, X-ray crystallography and nuclear magnetic resonance (NMR), 2-dimensional electron diffraction and 3-dimensional image restructuring provide efficient tools for the determination of the protein spatial structure. However, although being able to obtain the spatial structure of several macromolecules in a single day is certainly an improvement, the rate with which 3-dimensional structures are produced cannot keep up with the speed of primary sequence acquisition. Besides, these methods are still hampered by a number of limitations, and a number protein structures may not be applicable to experimental determination in the foreseeable future. Consequently, computational modeling and structure prediction are of great importance. Computational research will not only provide structural information, but will also provide information of the electronic structure of the protein, such as energy levels, surface charge, and molecular track interactions, as well as information on dynamic behavior, such as energy change of biochemical reactions, electric charge transfer, structural changes and so forth. Such information is difficult to obtain directly from experimental approaches. The modeling results are of great significance for understanding the basic processes of biological phenomena at the macromolecular, submolecular and electronic structure levels. They also provide a basis for conformational modifications in natural

macromolecules and for drug design modeled on receptor structure. In other words, they provide a theoretical basis for protein engineering. Protein engineering is a new and developing field that was initiated in the 1980s based on the relationship between protein function and the rules guiding protein structure. Employing molecular design, regulatory gene modifications and gene synthesis, protein engineering aims at making intended changes in existing proteins in order to create proteins with optimized characteristics which better serve human needs. At the same time it provides science with new technological possibilities and efficient new tools that greatly facilitate the solution of important problems in molecular research.

The computational development in recent years has endowed applied physics with the theoretical possibility of handling macromolecular systems composed of more than ten thousand atoms. In 1987 the American scientist W. F. DeGrado and colleagues at DuPont utilized accumulated rules for protein spatial structure to meticulously design and produce a 74 amino acid long artificial protein from peptides [54]. The protein contained all the four characteristic α-helices anticipated by the designer, and was called the first milestone of protein molecular design. Later, protein structure simulation and drug design has developed rapidly, and present protein structure simulation employs such methods as homology modeling, protein structure threading, and *ab initio* design using molecular dynamics modeling or Monte Carlo simulations. For drug design one may use the 3D-QSAR method and flexible atom receptor methods. These methods are already in extensive use, producing good results. Among the better known examples of successful drug design may be mentioned such as antibodies [55,56], AIDS viral compounds [57,58] and others. At the end of the last century there also emerged servers offering protein design services to customers. With respect to protein structure simulation it appears that protein folding configurations are limited, presently estimated to number from a few hundred to a few thousand [59,60]. This is far less than the number of freedom degrees in a protein structure. At the same time, protein folding configurations are correlated with their components and primary sequence, and it should be possible to determine the ultimate protein structure from its sequence information. If biological information is added to modeling systems it will certainly produce even better methods for protein simulation.

Bioinformatics is a new field in rapid development linking biology and informatics. Bioinformatics is a valuable discipline that nobody doing genomics research can do without. Given the rapid development in this field in recent years it seems inevitable that bioinformatics will produce brilliant results also in the future.

References

1. Benson D A, Boguski M S, Lipman D J, Ostell J, Ouellette B F. GenBank. Nucleic Acids Research, 1998, 26(1): 1–7
2. Ewing B, Hillier L, Wendl M C, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research, 1998, 8(3): 175–185
3. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research, 1998, 8(3): 186–194
4. Kent W J, Haussler D. GigAssembler: An algorithm for the initial assembly of the human genome working draft. Technical Report, UCSC-CRL-00-17, 2000
5. Cormen T H, Leiserson C E, Rivest R L. Introduction to Algorithms. MIT Press, 1990
6. Uberbacher E C, Xu Y, Mural R J. Discovering and understanding genes in human DNA sequence using GRAIL. Methods in Enzymology, 1996, 266: 259–281
7. Uberbacher E C, Mural R J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proceedings of the National Academy of Sciences of the United States of America, 1991, 88(24): 11261–11265
8. Synder E E, Stormo G D. Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. Nucleic Acids Research, 1993, 21(3): 607–613
9. Guigó R, Knudsen S, Drake N, Smith T. Prediction of gene structure. Journal of Molecular Biology, 1992, 226(1): 141–157
10. Pesole G, Attimonelli M, Saccone C. Linguistic analysis of nucleotide sequences: Algorithms for pattern recognition and analysis of codon strategy. Method in Enzymology, 1996, 266: 281–294
11. Girbal L, Soucaille P. Regulation of solvent production in *Clostridium acetobutylicum*. Trends in Biotechnology, 1998, 16(1): 11–16
12. Henderson J, Salzberg S, Fasman K H. Finding genes in DNA with a hidden Markov model. Journal of Computational Biology, 1997, 4(2): 127–141
13. Xiao Y, Chen R S, Shen R Q, Sun J, Xu J. Fractal dimension of exon and intron sequences. Journal of Theoretical Biology, 1995, 175(1): 23–26
14. Shen R Q, Chen R S, Ling L J, Sun J, Xiao Y, Xu J. The complexity of different regions of protein coding genes. Chinese Science Bulletin, 1993, 38(21): 1995–1997 (in Chinese)
15. Xu J, Chen R S, Ling L J, Shen R, Sun J. Coincident indices of exons and introns. Computers in Biology and Medicine, 1993, 23(4): 333–343
16. Miller G, Fuchs R, Lai E. IMAGE cDNA clones, UniGene clustering, and ACeDB: An integrated resource for expressed sequence information. Genome Research, 1997, 7(10): 1027–1032
17. Eckman B A, Aaronson J S, Borkowski J A, Bailey W J, Elliston K O, Williamson A R, Blevins R A. The Merck

Gene Index browser: An extensible data integration system for gene finding, gene characterization and EST data mining. Bioinformatics, 1998, 14(1): 2–13
18. Houlgatte R, Mariage-Samson R, Duprat S, Tessier A, Bentolila S, Lamy B, Auffray C. The genexpress index: A resource for gene discovery and the genic map of the human genome. Genome Research, 1995, 5(3): 272–304
19. Girard A, Sachidanandam R, Hannon G J, Carmell M A. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature, 2006, 442(7099): 199–202
20. Deng W, Zhu X P, Skogerbø G, Zhao Y, Fu Z, Wang Y D, He H S, Cai L, Sun H, Liu C N, Li B, Bai B Y, Wang J, Jia D, Sun S W, He H, Cui Y, Wang Y, Bu D B, Chen R S. Organization of the *Caenorhabditis elegans* small non-coding transcriptome: Genomic features, biogenesis and expression. Genome Research, 2006, 16(1): 20–29
21. Chureau C, Prissette M, Bourdet A, Barbe B, Cattolico L, Jones L, Eggen A, Avner P, Duret L. Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. Genome Research, 2002, 12(6): 894–908
22. Petrovics G, Zhang W, Makarem M, Street J P, Connelly R, Sun L, Sesterhenn I A, Srikantan V, Moul J W, Srivastava S. Elevated expression of *PCGEM1*, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. Oncogene, 2004, 23(2): 605–611
23. Xu F, McFarland M, Askew D S. His-1: A noncoding RNA implicated in mouse leukemogenesis. Current Science, 1999, 77(4): 545–549
24. Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider P M, Tidow N, Brandt B, Buerger H, Bulk E, Thomas M, Berdel W E, Serve H, Müller-Tidow C. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene, 2003, 22(39): 8031–8041
25. Li W H, Graur D. Fundamentals of Molecular Evolution. Sinauer Associates, 1991
26. Pearon W R. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in Enzymology, 1990, 183: 63–98
27. Altschul S F, Madden T L, Schäffer A A, Zhang J, Zhang Z, Miller W, Lipman D J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 1997, 25(17): 3389–3402
28. Higgins D G, Bleasby A J, Fuchs R. CLUSTAL V: Improved software for multiple sequence alignment. Computer Applications in the Biosciences, 1992, 8(2): 189–191
29. Kumar S, Tamura K, Nei M. MEGA: Molecular evolutionary genetic analysis. University Park: Pennsylvania State University, 1993
30. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. Evolution, 1985, 39(4): 783–791
31. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press, 1983
32. Wang N, Chen R S. Comparison between phylogeny of introns and exons in primates. Chinese Science Bulletin, 1999, 44(21): 1940–1946
33. Koonin E V, Tatusov R L, Galperin M Y. Beyond complete genomes: From sequence to structure and function. Current Opinion Structural Biology, 1998, 8(3): 355–363
34. Somogyi R, Sniegoski C. Modeling the complexity of gene networks: Understanding multigenic and pleiotropic regulation. Complexity, 1996, 1(6): 45–63
35. Vidal M. A biological atlas of functional maps. Cell, 2001, 104(1): 333–339
36. Bu D B, Zhao Y, Cai L, Xue H, Zhu X P, Lu H C, Zhang J F, Sun S W, Ling L J, Zhang N, Li G J, Chen R S. Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Research, 2003, 31(9): 2443–2450
37. Gibson D, Kleinberg J, Raghavan P. Inferring Web communities from link topology. In: Preceedings of the 9th ACM Conference on Hypertext and Hypermedia. 1998, 225–234
38. Kleinberg J M. Authoritative sources in a hyper-linked environment. In: Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms. 1998, 668–677
39. Zhang Z H, Liu C N, Skogerbø G, Zhu X P, Lu H C, Chen L, Shi B C, Zhang Y, Wang J, Wu T, Chen R S. Dynamic changes in subgraph preference profiles of crucial transcription factors. PLoS Computational Biology, 2006, 2(5): e47
40. Lee T I, Rinaldi N J, Robert F, Odom D T, Bar-Joseph Z, Gerber G K, Hannett N M, Harbison C T, Thompson C M, Simon I, Zeitlinger J, Jennings E G, Murray H L, Gordon D B, Ren B, Wyrick J J, Tagne J B, Volkert T L, Fraenkel E, Gifford D K, Young R A. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science, 2002, 298(5594): 799–804
41. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chclovskii D, Alon U. Network motifs: Simple building blocks of complex networks. Science, 2002, 298(5594): 824–827
42. Vázquez A, Dobrin R, Sergi D, Eckmann J P, Oltvai Z N, Barabási A L. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proceedings of the National Academy of Sciences of United States of America, 2004, 101(52): 17940–17945
43. Guelzim N, Bottani S, Bourgnie P, Képès F. Topological and causal structure of theyeast transcriptional regulatory network. Nature Genetics, 2002, 31(1): 60–63
44. Bray D. Protein molecules as computational elements in living cells. Nature, 1995, 376(6538): 307–312
45. Luscombe N M, Babu M M, Yu H, Snyder M, Teichmann S A, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature, 2004, 431(7006): 308–312
46. Zhao Y, He S M, Liu C N, Ru S W, Zhao H T, Yang Z, Yang P C, Yuan X Y, Sun S W, Bu D B, Huang J F, Skogerbø G, Chen R S. MicroRNA regulation of messenger-like noncoding RNAs: A network of mutual microRNA control. Trends in Genetics, 2008, 24(7): 323–327
47. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel

S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey D K, Ganesh M, Ghosh S, Bell I, Gerhard D S, Gingeras T R. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science, 2005, 308(5725): 1149–1154

48. Bartel D P. MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell, 2004, 116(2): 281–297
49. Will C L, Lührmann R. Spliceosomal UsnRNP biogenesis, structure and function. Current Opinion in Cell Biology, 2001, 13(3): 290–301
50. Willingham A T, Orth A P, Batalov S, Peters E C, Wen B G, Aza-Blanc P, Hogenesch J B, Schultz P G. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science, 2005, 309(5740): 1570–1573
51. Chen C L, Liang D, Zhou H, Zhuo M, Chen Y Q, Qu L H. The high diversity of snoRNAs in plants: Identification and comparative study of 120 snoRNA genes from *Oryza sativa*. Nucleic Acids Research, 2003, 31(10): 2601–2613
52. Numata K, Kanai A, Saito R, Kondo S, Adachi J, Wilming L G, Hume D A, RIKEN GER Group, GSL Members, Hayashizaki Y, Tomita M. Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. Genome Research, 2003, 13(6B): 1301–1306
53. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N. Cell-type-specific signatures of microRNAs on target mRNA expression. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(8): 2746–2751
54. Ho S P, DeGrado W F. Design of a 4-helix bundle protein: Synthesis of peptides which self-associate into a helical protein. Journal of the American Chemical Society, 1987, 109(22): 6751–6758
55. Riechmann L, Clark M, Waldmann H, Winter G. Reshaping human antibodies for therapy. Nature, 1988, 332(6162): 323–327
56. Liu X F, Xiao S, Gu Z, Wang Y, Chen A, Lin Q, Zhang W G, Huang H L, Sun J, Chen R S, Shen B F, Chen X. The expression of CD3 single chain and reshaping single-domain antibody. Science in China (Series C), 1996, 26(5): 428–435 (in Chinese)
57. Greer J, Erickson J W, Baldwin J J, Varney M D. Application of the three-dimensional structures of protein target molecules in structure-based drug design. Journal of Medicinal Chemistry, 1994, 37(8): 1035–1054
58. Lam P Y, Jadhav P K, Eyermann C J, Hodge C N, Ru Y, Bacheler L T, Meek J L, Otto M J, Rayner M M, Wong Y N, Chang C H, Weber P C, Jackson D A, Sharpe T R, Erickson-Viitanen S. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. Science, 1994, 263(5145): 380–384
59. Blundell T L, Johnson M S. Catching a common fold. Protein Science, 1993, 2(6): 877–883
60. Orengo C A, Jones D T, Thornton J M. Protein superfamilies and domain superfolds. Nature, 1994, 372(6507): 631–634

Professor Runsheng Chen is now Professor in Systems Biology Research Center and National Laboratory of Biomacromolecules at the Institute of Biophysics, Chinese Academy of Sciences. He is also a member of Human Genome Organization (HUGO), and a member of the biomacromolecule group of The Committee on Data for Science and Technology (CODATA). From 1992 to 1996 he was member of the Biophysics Professional Committee of the International Union of Pure and Applied Physics (IUPAP), and was ever the General-Secretary and Vice President of Chinese Society of Biophysics. He graduated in 1964 from the Department of Biophysics of the University of Science and Technology of China. From 1985 to 1987 he studied the electronic structure of biomacromolecules at the University of Erlangen-Nürnberg, as a fellow of the Alexander von Humboldt Foundation. After that he has been engaged in research cooperation with The Hong Kong University of Science and Technology, The Chinese University of Hong Kong, Osaka University, University of Erlangen-Nürnberg, University of California, Los Angeles, and Harvard University. In October 1996, Prof. Chen was invited to give a lecture called "From DNA sequence database to protein three-dimensional structure" at the 15th International CODATA Conference, and won the "Kotani Prize". In 2007, he was elected as Member of the Chinese Academy of Sciences. Professor Chen was awarded "Ho Leung Ho Lee Prize" in 2008.

Prof. Chen has occupied with studies in bioinformatics over a number of years. He was the first in China to accomplish the assembly and gene annotation of a complete bacterial genome. He has further established statistical DNA sequence analysis, fractal dimension analysis, and work on neural networks, complexity, local area degeneracy factor analysis, cryptology and other methodologies. Among these, Prof. Chen set up cryptology studies in China for the first time. He also took part in the sequencing of 1% of the human genome and computer analysis of the rice genome draft. For 20 years, Prof. Chen has taken a systematic study in the field of bioinformatics, and published more than 120 SCI papers; besides, he was invited to give a report at international academic conference many times.