H. Tanaka

Department of Bioinformatics
Medical Research Institutes
Tokyo Medical and Dental University
Japan

Synopsis

Computational approach towards challenges in the post-genomic era

In this section, articles which reflect current post-genomic trends in bioinformatics are collected. Here we overview the various post-genomic challenges and, in relation to them, we briefly introduce the contents of the collected articles.

1. Bioinformatic studies in the post-genomic era

Since the human genome project is almost finished [1,2], main interests in the life science community are now moving to post-genomic challenges, such as functional genomics, comparative genomics, proteomics, metabolomics, pathway analysis, systems biology. In bioinformatics analysis, although sequence analyses have been and are still the most common tasks in the routine analyses, new topics in bioinfomatics studies have appeared to tackle post-genomic challenges. We briefly overview several study fields below.

(1) Whole genome informatics

Now that whole genome sequences of more than 100 species are finished to read, though most are prokaryotes, the study of the whole genome structure becomes possible. Comparative genomics [3] is a new branch of genome sciences which compares whole genome sequences between different species to find genome-wide common structure and its evolutionary change. One of the classical examples in comparative genomics is to infer the minimum gene set of life from the comparison among the genomes of the primitive microbial organisms [4, 5]. Other typical studies are related to the evolutionary trace of large-scaled change of genome structures, such as chromosome duplication in the course of evolution concerning the gene cluster [6].

In the whole of genome informatics, effective usage of genome databases is a prerequisite. There are many well-known databases for whole genome such as GDB (Genome DataBase), LocusLink/ReqSeq, FlyBase and WormBase. To ensure usability and accessibility of the database, it becomes important to eliminate the factors militating against the full exploitation of the genomic information.

In the paper by Coppel, a Malaria genome database (Plasmodium DB) is taken and discussed with regards to several problems related to ensuring the full exploitation of whole genome sequences. The paper presents several lessons learned which would be of use to other organism-specific databases.

(2) Transcriptome analysis and microarray data processing

Transcriptome, comprehensive information of gene expression (mRNA) at the whole cell level, can now be observed by DNA chip and cDNA microarray. Whereas the genome is a possible repertory of biological function in terms of gene set and the proteome is the currently expressed whole set of functional protein, the transcriptome reflects the current production rate of functional protein in each cell, so that it shows, so to speak, intentions of living cells under the imposed cell conditions, which are not found in the information of the genome and the proteome.

A great deal of new bioinformatics studies related to microarray data have emerged over the years. There are mainly two sub fields in microarray information processing. One is phenomenological processing of gene expression data in micro-arrays such as the classification of the expression profile through clustering methods to make groups both for genes and subjects, or to identify the differentially expressed genes under the two comparative conditions of micro-arrays [7, 8].

The other is to identify structural relations among expressions of each gene. Along this line of the study, a typical study is to identify genetic (regulation) networks from the expression profiles of micro-arrays [9,10]. Many models are used for representing the genetic network; for example the Boolean genetic network model where, though its connected path, a gene facilitates or suppresses the other's expression is a simple deterministic model of the genetic network. On the other hand, there are also probabilistic models where the interactions between connected genes are random.

In the paper by Reis et al., temporal slopes of the expression level of each gene are determined based on the sequential observations of mRNA expression pattern of *Saccharomyces cerevisiae* and correlation coefficients are calculated between these temporal slopes of all genes. By random permutation of expression data, they estimated the variation of the correlation coefficients when there is no significant correlation between genes, and determined the thresholds over which we judge the existence of significant correlation. By connecting pairs of genes having over-threshold correlation, so to speak dynamically correlated pattern ("relevance network") among the genes can be obtained. They show high association in this dynamic correlation well agrees with functional and regulatory relationships between genes.

(3) Other areas

The bioinformatics field related to proteins, called "Protein Informatics" has also drastically progressed recently. This field includes (1) structural genomics where structural prediction of protein or classification of representative structure of basic protein folds is a main topic, (2) proteomics which is related to the comprehensive observation and characterization of functional proteins in the cell and (3) functional genomics which predicts protein function from the sequences and structures such as to estimate binding sites or reaction sites.

2. Pathway analysis and systems biology

Other important new areas of bioinformatics research are those which aim to understand life as a whole functional organization from comprehensive bio-information. This area ranges from the more confined topic called "pathway analysis", to more generally proposed disciplines of "systems biology".

(1) Pathway analysis

In the pathway analysis, metabolic pathway or protein-protein networks such as signal transduction cascade are of main interests. We have already several well known pathway DBs, for example, KEGG (Kyoto encyclopedia of Genes and Genomes), an online database for metabolic and regulatory pathways, ExPAsy molecular biology server, a scanned map of Boehringer-Mannheim 'Biochemical Pathways', EcoCyc, a comprehensive database for metabolic and regulatory pathways of E. Coli. TRANSPATH and CNSDB (Cell signaling Network Database) are databases especially for signal transduction cascades.

In these pathway DBs, pathways are visualized in graph format but mostly in a static way. It would be preferable that graph representation of the pathways can be automatically and dynamically updated when new components are incorporated. In the paper by M. Becker, a new algorithm is presented to draw the metabolic pathway by combination of circular, hierarchic and force-directed graph layout. This automatic drawing of the metabolic network is of great use to promote the feasibility of pathway DBs.

(2) Systems biology

1) Concept

Although Kitano coins the word "systems biology", it is now widely accepted as a generic term. The field and purpose of systems biology [11] is not new in bioinfomatics. So far, similar disciplines have been called "integrative biology" or "biological system analysis". One of the typical and established research areas for this sort of analysis has been in the area of "metabolic control analysis" [12] where traditional control system theory is applied to the metabolic pathway to investigate the system performance or sensitivity of rate-limiting path in the metabolic system.

Prior to the post-genomic era, there had been no largely organized comprehensive biological data ("–omic" data) to be utilized in the model analysis. These studies were therefore forced to be theoretical using only simple mathematical models and few experimental data. Now, many kinds of comprehensive biological data are available and the studies dealing with the integrative behavior of life are gradually changing with regard to study style. They use large-scaled data obtained by genome-wide measurement such as whole gene expression profile by cDNA micro-array, and the modeling becomes more comprehensive and realistic in dealing with whole cells, though modeled organisms are primitive microbes. The following topics are now being studied currently in systems biology:

(1) Comprehensive simulation of large-scale pathways of metabolism, including a whole cell metabolism
(2) System analysis of the metabolic pathway or genetic regulatory network
(3) System identification of the metabolic pathway or genetic regulatory network from comprehensive experimental data such as the cDNA expression profile.
(4) System design of the artificial organism or artificial bacteria having preferable characteristics.
(5) Pathway databases and signal transaction cascade databases
(6) Tools and software to support the system biology such as the cell

simulation package and standard documentation language for cell modeling.

2) Projects

In the last several years, systems biology has become one of the nation-wide projects in the post-genomic era. In the United States, NIH/NIGMS supports the Alliance for the Cellular Signaling (AFCS) project which aims at examining the signal transduction inside cells, by analyzing mouse's G-protein coupled signaling system where 1000 proteins work cooperatively. They also support a cellular communication and cell migration consortium and many other groups conducting the quantitative analysis of complex biological systems.

The Department of Energy (DOE) is also conducting a "Genome to Life" project mainly aiming to model the microbial "virtual cell", especially its metabolic system organization and its migration. The biomedical engineering projects in NIH called physiome projects also have a strong relation to systems biology. In EU, Model of Life (MOL) projects are now being conducted.

In Japan, E-Cell, a simulator for virtual cells with minimal gene set (127 genes) was developed by Tomita which is used as a base model for a human red blood cell [13].

3) Related areas – Biological modeling, complex systems, dynamical network theory

From a slightly different perspective, another stream also attracts interests in the systems biology community. This stream is nonlinear modeling or the complex systems approach to biological systems. So far in the field of theoretical biology, complex systems approaches were adopted because biological systems are always nonlinear. In the complex systems approach, the whole system is considered more than the sum of its parts and emerging properties of biological systems are well recognized in the origins of life, biological evolution, and development process. So far, as often seen in the Kauffman theory [13], for modeling those essential phenomena of life, Erdos random network theory has been ordinarily used for the base model of biological relations. For example, random reaction networks of autocatalytic sets of biopolymers are used for modeling the origin of life, random epigenetic interaction networks among the genes are used for calculating the integrative fitness landscape of multiple genes of evolution, and random Boolean networks are utilized to describe the regulation of cell types in biological development.

But recently random networks haven't been seen as the appropriate model for real networks, instead "scale free" networks [15] in which the frequency distribution of number of edge connecting to the nodes has a long tail obeying power-law (straight line in log-log plot) or otherwise "small world" networks [16] are used. It was shown that the metabolic pathway [17] and protein-protein interaction network of Yeast Two Hybrid [18] is a scale-free network. Hence, with comprehensive biological information and a new dynamical network theory, system level organization of biological networks will drastically be clarified in theory.

In the paper by Yates, by taking the immunological system as an example, techniques and the issues in building "good" biological modeling are discussed, where nonlinear threshold effects and bifurcation or emerging phenomena in immunological response are investigated by a phase plain method about the immunological cytokine (TNF) network and T helper T cell differentiation. Monte Carlo simulation is described for stochastic simulation for cross talk for T cell receptors.

3. Clinical bioinformatics

(1) Polymorphism of the human genome

Comprehensive approaches for biological information also begin to exert important influences on clinical medicine. Various polymorphisms of the human genome sequence characterize the individual specificity of the patient genome, such as restriction enzyme polymorphism, VNTR (variable number of tandem repeat), SINE, LINE and SNP (single nucleotide polymorphism). Especially SNPs are recently the main target of comprehensive surveys in relation to drug discoveries. SNPs are found on average every one thousandth nucleotide, so that three million SNPs are supposed to characterize the haplotype of the patient genome, which would be related to the disease and drug response. Hence it would be of main interest for clinical application of genome to realize "Personalized medicine".

(2) System theoretic approach to diseases – disease modeling

Disease might be due to the defect of a single gene (mono-genic disease) or caused by defects of more than one gene (polygenic disease). Since mono-genic diseases seem to be mostly explored, polygenic diseases, which cover most "common diseases" such as hypertention, diabetes and ischemic heart diseases, are now attracting more attention. In the polygenic common diseases, diseases are thought to form themselves in the combined manner of the various gene defects and the environment. For example, more than 20 genes are related to the occurrence of diabetes. So like systems biology for normal biosystems, a systems-pathological approach or systematic disease modeling would be of great value to comprehensive understanding of polygenic diseases.

One of the promising approaches in disease modeling is the "Virtual

Patient”. Entelos Inc. has developed the virtual patient system that is used to model obesity, diabetes and asthma. The virtual patient model involves various levels of knowledge, such as related to the genetic, pathophysiologic and life-style factors and both top-down and bottom-up approach between genetic level to symptomatologic level are employed. The day will come soon when we use these virtual patient models for clinical decision making.

In the paper by Sreekumar et.al, they show the many examples wherein by using comparative genomics and computational sequence analysis, especially for domain analysis of functional protein, many human disease-related genes can be identified and the etiology is accessed. It could be considered as a preliminary trial of “systems pathology” or “disease modeling”.

4. Conclusions

In the bioinformatics field, new research topics to solve post-genomic challenges are emerging. In this synopsis, whole genome informatics (comparative and functional genomics), pathway analysis, systems biology and clinical bioinformatics were especially discussed. Collected papers have strong relation to these topics.

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome, Science 2001;291:1304-51.
3. Koonin.EV. The emerging paradigm and open problems in comparative genomics. Bioinfomatics 1999;15(4):265-6.
4. Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, et al. Global transposon mutagenesis and a minimal mycoplasma genome. Science 1999;286:2165-9.
5. Koonin EV, Mushegian AR. Complete genome sequences of cellular life forms. Curr opin genet dev 1996;6(6):757-62.
6. Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. Cell 2000;101(6):573-6.
7. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001 Jun;2(6):418-27.
8. Raychaudhuri S, Sutpin PD, Chang JT, Altman RB. Basic micro array analysis:grouping and feature reduction. Trends Biotechnol 2001;19(5):189-93.
9. Liang S, Fuhrman S, Somogyi R. REVEAL,a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput 1998:18-30.
10. D’haeseleer P, Liang S, Somogyi R. From co-expression clustering to reverse engineering. Bioinfomatics 2000;16(8): 707-26.
11. Kitano H. Systems biology: a brief overview. Science 2002; 295:1662-4.
12. Fell D, Snell K, editors. Understanding the Control of Metabolism. Portland Press;1997.
13. Tomita M. Whole cell simulation: a grand challenge of 21st century. Trends Biotechnol 2001;19(6):205-10.
14. Kauffman S. Origins of Life. Oxford;1994.
15. Babarasi A. Linked. Perseus; 2002.
16. Watts D. Small Worlds. Princeton; 1999.
17. Jeong H, Tombor B, Albert R, Oltavi ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature 2000; 407(6804):651-4.
18. Jeong H, Mason SP, Barabasi AL, Oltavi ZN. Centrality and lethality of protein networks. Nature 2001;411(6833):41-2.

Address of the author:
Hiroshi Tanaka
Department of Bioinformatics
Medical Research Institutes
Tokyo Medical and Dental University
1-5-45, Bunkyo
Tokyo 113-8510, Japan
Tel: 0081/3-5803-5839
Fax: 0081/3-5684-3618
E-mail: tanaka@cim.tmd.ac.jp