PERGAMON

Progress in Biophysics & Molecular Biology 80 (2002) 23–42

Progress in Biophysics & Molecular Biology

www.elsevier.com/locate/pbiomolbio

Review

Bioinformatics for the genomic sciences and towards systems biology. Japanese activities in the post-genome era

Toru Yao*

RIKEN Genomic Sciences Center, 1-7-22, Suehiro, Tsurumi, Yokohama 230-0045, Japan

Abstract

The knowledge gleaned from genome sequencing and post-genome analyses is having a very significant impact on a whole range of life sciences and their applications. 'Genome-wide analysis' is a good keyword to represent this tendency. Thanks to innovations in high-throughput measurement technologies and information technologies, genome-wide analysis is becoming available in a broad range of research fields from DNA sequences, gene and protein expressions, protein structures and interactions, to pathways or networks analysis. In fact, the number of research targets has increased by more than two orders in recent years and we should change drastically the attitude to research activities. The scope and speed of research activities are expanding and the field of bioinformatics is playing an important role.

In parallel with the data-driven research approach that focuses on speedy handling and analyzing of the huge amount of data, a new approach is gradually gaining power. This is a 'model-driven research' approach, that incorporates biological modeling in its research framework. Computational simulations of biological processes play a pivotal role. By modeling and simulating, this approach aims at predicting and even designing the dynamic behaviors of complex biological systems, which is expected to make rapid progress in life science researches and lead to meaningful applications to various fields such as health care, food supply and improvement of environment.

Genomic sciences are now advancing as great frontiers of research and applications in the 21st century.

This article starts with surveying the general progress of bioinformatics (Section 1), and describes Japanese activities in bioinformatics (Section 2). In Section 3, I will introduce recent developments in Systems Biology which I think will become more important in the future.

Keywords: Bioinformatics; Systems biology; Genome; Protein; RIKEN; Japan

*Tel.: +81-45-503-9295; fax: +81-45-503-9155.
E-mail address: yao@riken.go.jp (T. Yao).

0079-6107/02/$ - see front matter
PII: S0079-6107(02)00011-1

Contents

1. Recent developments in bioinformatics

1.1. Genome and gene analysis

It was a monumental event for mankind to obtain the "Human Genome Sequence" even in the draft form in February 2001 (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). In addition, genome sequences of more than 70 other species

(www.ncbi.nih.gov/Genomes) have already been obtained and the genome sequencing of hundreds of species is in progress. The GenBank (DNA Database) has more than 20 billion of ATGC letters and is expanding very rapidly (www.ncbi.nih.gov/Genbank). These huge data banks are now under examination using computers by various research groups. There is an unprecedented level of such research.

1.1.1. Identification and annotation of genes

In the human genome sequence, both groups, IHGSC and Celera Genomics, using various gene prediction tools, estimated the number of genes at around 30,000 (IHGSC, 2001; Venter et al., 2001). However, the estimations are not coincident (Hogenesch et al., 2001), because the performance of current prediction methods is not yet optimal (Stormo, 2000; Guigo et al., 2000). Gene finding programs need improvement especially for the mammalian genomes that contain many introns and repeats.

The analysis of the regulatory regions of genes is one of the important research topics of genome analysis (Pilpel et al., 2001; Davului et al., 2001). Suzuki et al. identified the potential promoter regions of 1031 kinds of Human Genes by using the cDNA libraries constructed by them (Suzuki et al., 2001a, b). Many programs have been recently developed for promoter finding analysis (Ohler and Niemann, 2001).

After the identification of genes, the annotation of those genes needs to be done. The Ensembl is one of the annotation systems for the human draft sequence (www.ensembl.org). The annotation of genome genes by using the cDNA information of mouse (The RIKEN Genome Exploration Research Group Phase II Team, 2001) or human (Yudate et al., 2001) is being carried out.

1.1.2. Comparative genomics

Comparative genomics is an emerging research field (Koonin, 1999) after the first genome was sequenced in 1995 (Fleischmann et al., 1995). Starting with microbial genomes (Makarova et al., 1999; Bansal, 1999; Perriere et al., 2000), comparative analysis came to be done among three kingdoms including archaebacteria and yeast (Tatusov et al., 1997), and led to the establishing of the Conserved Orthologous Group (COG) database (www.ncbi.nlm.nih/COG). Furthermore, comparative genomics moved on to multi-cellular organisms including c-elegans and fruit fly (Liu and Rost, 2001) and the research on mammals, including human, mouse and chimpanzee, is now in progress (O'Brien et al., 1999; Cyranoski, 2001; Fujiyama et al., 2002).

Many new interesting research results have been reported from these analyses. For example, Aravind found how the protein domains and repair systems have been conserved and evolved in DNA repair systems (Aravind et al., 1999). Kihara found that about half of the membrane proteins in each genome form tandem clusters by analyzing 16 complete genomes (Kihara and Kanehisa, 2000). Koonin expressed the new paradigm of evolutionary biology and highlighted 10 open problems in Comparative Genomics (Koonin, 1999). Another interesting research approach is to assess the minimum complement of genes (Fukuda et al., 1999). Experimental works should be incorporated to find these minimum genes (Hutchinson, 1999).

1.1.3. SNP (Single nucleotide polymorphism) analysis

SNP analysis is one of the big research areas in the post-genome era (The International SNP Map Working Group, 2001). The information obtained from SNP analysis is expected to bring to

us not only the understanding of individuality but also applications to personalized medicine or health care in the near future (see Section 1.4).

1.1.4. Information science and computer power

The above-mentioned analyses are mainly based on homology search methods which incorporate advanced information science, such as dynamic programming, neural networks and the hidden Markov model.

Computer power is indispensable for these analyses, especially for the comparative genomics for mammals because billions of letters are compared in a two-dimensional manner.

1.1.5. Full-length cDNA

Another big area is the cloning and sequencing of the full-length cDNAs from various tissues or organs of mouse (The RIKEN Genome Exploration Research Group, 2001; Yudate et al., 2001). This cDNA information is very useful for the analysis and annotation of genomes. Kondo applied the full length mouse cDNAs to the human draft sequence for the finding of new genes (Kondo et al., 2001). His group successfully aligned 35,141 non-redundant genomic regions with the mouse cDNAs, and found 1141 candidates for novel genes.

1.2. Computational analysis of protein structure and function

The conventional analyses of proteins using the computer have entered into a new stage with the arrival of the genome and the post genome-era.

1.2.1. Structure prediction

The Homology Modeling method (Greer, 1981; Blundell et al., 1987), which has been widely used since the beginning of the 1980s as the most useful technology to predict a tertiary structure from an amino acid sequence of a protein, is now applied to whole genes that are obtained from genome information (Sanchez and Sali, 1999). This method is producing more than 300,000 modeled structures (http://guitar.rockefeller.edu/modbase), compared with about 18,000 measured structures in the PDB (www.rcsb.edu/pdb). About 30% of genes from genomes can be made into modeled structures. It is anticipated that the number of modeled structures will increase rapidly as the genome information on various organisms are deciphered. In this case, the contribution to progress of basic research and applications to the drug design will be invaluable.

The 3D-1D method (Bowie et al., 1991) and succeeding methods, such as the fold recognition method and the threading method (Jones et al., 1992), which were developed in the early 1990s can now sometimes detect similarities of the fold with known structures even if sequence similarities are low. These methods are expected to play great roles once almost all fold types are determined by the structural genomics projects mentioned below.

On the other hand, the ab-initio prediction method, which many researchers have challenged for years, has come to be applied to small proteins and to show good performance (Simon et al., 1999; Baker, 2000; Kihara et al., 2001). This method seems to expand the range of practical use with the improvement of computer capabilities.

The CASP, worldwide competition of protein structure prediction, which has been held in every other year since 1994, has been making a great contribution to accelerate the improvement of the above-mentioned methods (http://predictioncenter.llnl.gov) (Moult et al., 1999; Murzin, 2001).

1.2.2. Structural genomics

The Structural Genomics Project (Stevens et al., 2001), which aims to systematically determine protein structures, is coming to the front as a big theme of the post genome era. This project is based on the premise that if almost all of the representative structures of basic folds and basic family members can be determined, then modeled structures can be made, across species, by the homology modeling method and the threading method, and eventually almost all of the protein structures deriving from genomic genes can be determined.

In this project, so-called "Protein Informatics" that includes the classification of protein structures (Murzin et al., 1995), target selections (Brenner, 2000; Vitkup et al., 2001) and most importantly structure predictions (Jones, 2000; Baker and Sali, 2001) play a key role.

1.3. The development of functional genomics and proteomics

With the progress of various technologies, such as the Microarray, DNA chip, Protein chip, 2D-PAGE, Yeast two hybrid, TOF/MAS (Dhand et al., 2000), systematic or exhaustive gene expression data and protein interaction data are coming to be obtained. Using these technologies, the dynamic change of gene expression profiles and the different expression patterns of normal states and abnormal states are being examined (Seki et al., 2001). Through the advancement of those experiments, databases, such as gene expression database (http://genome-www4.stanford.edu) and protein interaction database (http://dip.doe-mbi.ucla.edu), etc. are now available via the internet. Accordingly, new research approaches are emerging, such as how those data can be arranged and how useful information can be extracted from those data using a new type of bioinformatics (Bassett et al., 1999).

On the other hand, the prediction of protein functions from sequence information and from structure information is advancing. Functional motifs can be extracted by comparing sequence information on the same or similar functions of various species. The extracted functional motif database (http://www.expasy.ch/prosite) is available, and can be used for the prediction of functions. Some advances are seen in the methods of predicting binding sites, reaction sites etc. from protein structures (Orengo et al., 1999; Zhang et al., 1999; Weir et al., 2001; Aloy et al., 2001).

1.4. Impact on drug development

The above-mentioned genome-wide analysis is having a big impact on medicine and medical treatment. One example is the so-called "Genome-Based Drug Discovery". The new approach is expected to increase the number of drug targets dramatically. It is said that the receptors of the drugs developed so far form about 500 targets (Drews, 2000), many of which tertiary structures have not been determined yet. With genome analysis, however, several thousand genes related to diseases may be identified in the near future. Also hundreds of pathogenic microbial genomes each

of which has hundreds or thousands of genes will be identified. These achievements will help increase the number of drug targets (Sanseau, 2001).

Secondly, the Structure-Based Drug Design (SBDD) will become more active. As described above, advances in the Structural Genomics projects and the Structure Prediction methods will, no doubt, bring us a lot of structural information within a couple of years. In the progress of the SBDD, the increase in the number of chemical compounds cannot be ignored. The number of candidate compounds used for the SBDD has reached more than millions, and virtual compounds generated by the computer can exceed tens of millions. The computer screening for lead compounds by the docking study of a huge combination can be executed (Bajorath, 2001). The Chemo-informatics, including chemical structures, functions and properties of those compounds, become more and more important.

Thirdly, personalized medicines and medical treatments based on the genetic variations of individuals are expected to be realized in the near future (Ginsbug and McCarthy, 2001). The progress of the SNP project will explore relationships between geno-type and disease susceptibility or drug responsibility. The association studies using family data and the relational studies using disease/non-disease group data are highly dependent on statistical or mathematical methods. Not only the diseases dependent on a single gene mutation but also diseases dependent on multiple gene mutations will be analyzed using a lot of data.

2. Bioinformatics activities in Japan

In this section, recent developments in Japan regarding the above-mentioned genome and protein analyses are described in two parts—the RIKEN and other initiatives.

2.1. Establishment of the Genomic Sciences Center of RIKEN

The Genomic Sciences Center of RIKEN (Wada, 2001) (Director, Akiyoshi Wada) was established in October 1998 as a large scale integrated genome research body (www.gsc.riken.go.jp). As described below, the center was built so as to form a comprehensive research organization, and is conducting a broad range of research themes in genomic sciences including genome, gene, protein, animal, plant and bioinformatics research from geno-type to pheno-type.

2.1.1. History of the projects

Wada, at that time Professor at the University of Tokyo, took an initiative in the development of automated DNA sequencer in 1981, and announced completion of the development of the basic part of the sequencer in 1987 (Wada, 1987). He organized the first international meeting on the automation of the DNA sequencer in Japan in 1987 (Swinbanks, 1987). However, the Japanese government did not support the budget of the equipment development on a large scale. One of the researchers, Kambara continued the development of the Multiple Sheath-Flow Capillary (Takahashi et al., 1994) that was adopted in the sequence machine ABI 3700 in 1998.

The Japanese human genome project also suffered from a budgetary problem. Prof. Kenichi Matsubara, of Osaka University at that time, organized the Japanese part of the international human genome project in 1991, but the budget was not sufficient for a world-class contribution.

These two facts show the lack of recognition of the importance of genome analysis at that time in Japan.

Two young researchers of RIKEN made important research proposals around 1995. Yoshihide Hayasizaki proposed to collect and sequence exhaustively the mouse full-length cDNA clones, and annotate functions of them to establish the mouse encyclopedia for the better understanding of the human genome and medical phenomena. Shigeyuki Yokoyama advocated that the importance of structural biology should be more recognized and that the systematic analysis of protein structures would greatly contribute to the post genome analysis.

In 1996, Wada organized a team for the planning of consolidation of these two proposals and the Japanese human genome project that had just started in 1996 as the second phase under the leadership of Prof. Yoshiyuki Sakaki, University of Tokyo. The plan to establish a big research center incorporating these three projects was finally approved by the government in 1998. Fortunately, the Japanese government had just begun to promote life science research according to the basic plan of science and technology (1996–2000).

2.1.2. Current status of RIKEN/GSC

After the establishment of the basic three groups—genome research, gene research and protein research, the GSC added, in 1999, two more experimental groups—animal phenotype research and plant phenotype research and further added one more group, bioinformatics research in 2000. GSC has six interrelated research groups from genotype to phenotype. GSC also has two facilities for the protein synthesis/crystallization and the informatics infrastructures. The total organization chart is shown at the web-site (www.gsc.riken.go.jp). The following is the policy of GSC/RIKEN:

(A) Genomic sciences should be integrated as a big science involving various researchers from bioscience fields and non-bioscience fields. Thus, GSC hires various expertise and actively collaborates with various outside research groups.
(B) GSC should pursue both basic research (Newness) and applied research (Usefulness). In addition to the discovery of new phenomena, materials and methods, useful materials, technologies and applications should be developed.
(C) Collaborations between internal groups and outside groups should be strengthened. While keeping the independency of each of the six groups, collaborations among them are highly recommended. In order to grasp the needs of society, collaborating with private companies from the early stage of researches are very important.

Under the above policy, there are 12 collaborations among inside groups (Wada, 2001) and collaborations with private companies exceed 30 cases already. The research results by such collaborations will emerge in a few years.

2.1.3. Recent topics of RIKEN/GSC

The following section introduces three major topics that have been tackled.

(A) The establishment of the mouse encyclopedia by full-length cDNA of mouse (The Genome Exploration Research Group headed by Y. Hayashizaki).

This plan consists of three stages; the first stage is the exhaustive collection of mouse full-length cDNAs, the second one is the full sequencing, and the third one is the full annotation.

At present, the group has already finished the first stage, is in the middle of the second stage and has partly entered into the third stage.

At the first stage, they collected more than one million full-length cDNA clones from more than 50 tissues of mouse and selected non-redundant 128,000 clones by reading the terminal sequences. Those data were published in May 2000 (Cyranoski, 2000). The sequencing of full-length cDNA in the second stage is about half finished and the completion is scheduled for the spring of 2002. As part of the third stage results, the annotation of 21,000 sequences was published in Nature in February 2001 (The RIKEN Genome Exploration Research Group Phase II Team, 2001). For the annotation, the Functional Annotation of Mouse (FANTOM) meeting was held in Tsukuba, Japan for 2 weeks in August 2000 by gathering about 60 researchers from the world (Quackenbush, 2000; Nature News, 2000). In parallel, the group mapped the data to the mouse genome (Yamanaka et al., 2001).

The FANTOM results were used for the annotation of the human genome (Kondo et al., 2001; International Human Genome Consortium, 2001).

The feature of this group is that they have developed themselves the main parts of the technologies for their own researches. For example, to acquire the full-length cDNAs, they have developed the high temperature durable trehalose (Carninci and Hayashizaki, 1999). The RIKEN Integrated Sequence Analysis (RISA) system (Shibata, 2000) that functions 4 times faster than the ABI3700 is another development of theirs for the sequencing of the cDNAs. They have also developed technologies for the measurement of gene expressions (Miki et al., 2001) and protein–protein interactions (Suzuki et al., 2001a, b).

Having a strong team of bioinformatics within the group, they can effectively and efficiently gather data, determine sequences and annotate functions, and greatly contribute to the world's database publication (Bono et al., 2002a, b).

They will complete the mouse encyclopedia by the FANTOM-2 meeting that is scheduled in 2002.

(B) The Human genome and the Chimpanzee genome (Human Genome Research Group headed by Y. Sakaki).

This group has been engaged in the human genome sequencing as one of participating groups in the International Human Genome Project. One achievement is the 21st chromosome analysis the result of which was published in May 2000 (Hattori et al., 2000). This group also contributed to the publication of the first human draft sequence in Feb. 2001 (International Human Genome Sequencing Consortium, 2001). The bioinformatics team of this group published the HGREP, a system for the analysis and visualization of genome structures (www.hgrep.u-tokyo.ac.jp). The current main focus of this group is the finished sequence of 11th and 18th chromosomes, and they intend to complete the research by Spring 2003. It should be mentioned that the mapping team of this group has been greatly contributing to every aspect of the human genome project (International Human Genome Mapping Consortium, 2001).

This group has recently initiated the next project of comparative genomics on human genome and chimpanzee genome by organizing the international consortium (Cyranoski, 2001). One of the research results has just been published (Fujiyama et al., 2002).

(C) Structural Genomics (Protein Research Group headed by Yokoyama).

Wada and Yokoyama advocated the importance of structural genomics earlier (Nature, 1996). Half of their proposals were approved by the government as one of the research projects of GSC in October 1998. After the 2-year construction of the facility, the first part was set in operation in October 2000 (Yokoyama et al., 2000a) (http://protein.gsc.riken.go.jp). In the meantime, international collaborations on structural genomics were discussed among several countries, and the agreement was made at the meeting of April 2001 to solve about 10,000 structures in 5 years (Stevens et al., 2001). Japan is given a role to solve 30% of the structures, and 80% of them will be finished by the RIKEN group called the RIKEN Structural Genomics Initiative (www.rsgi.riken.go.jp).

The Protein Research Group of GSC has developed and improved the cell-free protein synthesis method (Kigawa et al., 1999), and is now proceeding to synthesize proteins from various sources including the mouse cDNAs, human cDNAs, Arabidopsis, Thermus thermophilus HB8, etc.

In this project, the Structural Bioinformatics (Yokoyama et al., 2000b) that concern the classification of structures, the selection of targets for structure determination, and the modeling of many structures are playing important roles. Following these activities, the SBDD are under development and many pharmaceutical companies have already entered into close collaborations with GSC in work on specific proteins.

2.2. *Other bioinformatics activities in Japan*

The promotion of bioinformatics activities by the Japanese government is being accelerated recently though the start was delayed for years compared with Europe and USA.

(1) The Ministry of Economy, Trade and Industry (METI) established in Tokyo in April 2001, two centers related to the bioinformatics within the National Institute of Advanced Industrial Science and Technology (AIST).

One is the Biological Information Research Center (JBIRC—Director: Y. Kyogoku) that consists of three groups (www.jbirc.aist.go.jp). The first group is the structural analysis group, especially aiming at membrane proteins. Prof. Y. Fujiyoshi who is known for the electron beam analysis of membrane proteins using the cryo-electromicroscopy is participating in the group as one of the leaders. The second one is the functional genomics group that conducts proteome analysis and microarray analysis using human cDNA clones. And the third one is the integrated database group headed by Prof. T. Gojyobori who is in charge of the DDBJ (DNA Database of Japan), and is expected to integrate various kinds of databases.

Another center is the Computational Biology Research Center (CBRC—Director Y. Akiyama) that is to conduct the pure bioinformatics research without any wet laboratory (www.cbrc.aist.go.jp). Their focus is on the development of the methodologies and their applications. They have already made notable achievements including a new multiple alignment program (Goto), a gene finding program (Asai), and the exhaustive analysis of the GPCR regions of the human genome (Suwa). This group has a uniquely large scale PC-cluster system for the genome-wide analyses.

(2) The Japan Biological Informatics Consortium (JBIC) (www.jbic.or.jp) was established in Tokyo in September 2000. Eighty-five companies have participated in the JBIC, and the

profile of participants varies from bioindustries, such as pharmaceutical and food companies, to non-bioindustries, such as information technology and measurement companies. Active interactions with academic groups and public sectors are taking place. The JBIC is expected to promote the development of bioinformatics tools and databases, and help grow bioinformatics businesses with the support of governmental funds.

(3) A new bioinformatics research grant was founded in April 2000 as one part of the genome research grants under the Ministry of Education, Science and Technology (MEXT). More than 200 academic researchers will receive support from this grant.

(4) The grant of the BIRD (Center for the Bioinformatics Research) (http://bird.jst.go.jp) was just set up in April 2001, and is providing fund to DDBJ, PDB/JAPAN, KEGG etc. to promote the biological database development.

(5) The 'Genome Informatics' project has been conducted since 1998 under the support from METI (MITI at that time). The project has been carried out on a contract basis with private sectors. As the main focus is the development of various measurement technologies for genome and post-genome analyses, the name of 'Informatics' does not fit properly, but some of the technologies will bring promising results to the progress of bioinformatics.

(6) Two academic societies were established in Japan in 2000. One is the Japan Bioinformatics Society (M. Kanehisa) (www.jsbi.org). The Society was organized as the successor to the GIW (Genome Informatics Workshop) that had been held for the past 10 years.

Another relevant society is the Chemistry, Biology and Informatics (CBI) Society (www.cbi.org). Prior to the Society, the organization called the CBI Forum was initiated by Dr. S. Kaminuma and had continued activities for more than 20 years. The CBI Forum had monthly meetings, involving researchers from industries and academy, and played a pioneering role for the progress of bioinformatics and chemoinformatics. The Society takes over the history.

(7) The Protein Science Society of Japan was organized in 2001 by integrating four interest groups of protein into one big group. The first annual meeting held in Osaka in June 2001 was very successful because researchers' general interests are shifting from genome to protein.

(8) The Japanese information scientists and IT engineers are paying close attention to bioscience. The Special Interest Group on Molecular Biology Informatics (SIGMBI) was organized in 1998 by the researchers of the Artificial Intelligent Society, and the Initiative of Parallel Architecture for Biology (IPAB) was established in 2000 by the engineers from IT industries.

While the movements in both the public and private sectors are active the establishment of new bioinformatics courses at universities is very much delayed because of the lack of flexibility of university systems in Japan. The attention to the bioinformatics is, however, generally very strong and special seminars or forums on bioinformatics are very well attended.

3. Systems biology—biosystems simulation

Following the genome sequence data, other huge data, such as gene expressions, protein interactions and protein structures, are being acquired at the genome-wide level. In order to extract valuable information from those data, bioinformatics plays an indispensable role.

In parallel with such data-driven research, however, a new approach called "Model-driven research" is being developed.

3.1. Systems biology—toward model-driven research

3.1.1. Systems biology

Model-driven research takes the approach that sets up a biological model by combining the knowledge of the system with related data and simulates the behavior of the system in order to understand the biological mechanism of the system. It is simply called, "Systems Biology".

The living body is composed of numerous subsystems. These include various subsystems, large and small, by which the flows of energy, material and information are controlled. This is a hierarchical system working consistently with many metabolic systems, transcriptional control systems, signal transduction systems, cell cycles, apoptosis systems, various physiological and pathological systems, organ systems, and other systems from the molecular level, cellular level, tissue level, organ level to the body level. Systems biology aims to model and simulate such various systems and visualize the results for the better understanding of life mechanisms.

3.1.2. Merits of systems biology

There are some important features and merits of this approach. One of the aims is to take important knowledge in the form of qualitative biological theories and try to express this as explicitly and quantitatively as possible. Thus implicit knowledge can be transferred to become explicit knowledge and disparate human knowledge can accumulate in an integrated way.

This approach also tries to model the dynamic behavior of the system. Life systems are inherently dynamic, but papers or books cannot fully express the dynamism. Computer models can handle and visualize such dynamic behavior. Thirdly, this approach will make us recognize the lack of knowledge through model building. There are many unknown pathways or mechanisms and also unknown parameters that govern the mechanisms. Conducting research or measurement of those unknown regions per se is one of the merits. Simulation can identify missing components. Furthermore, we can propose appropriate design of experiments with the prediction of results from such simulations.

3.1.3. Ultimate goal of systems biology

The ultimate goal of this approach is to develop a "Life Simulator", which will be attained, step by step, hierarchically from subsystem simulators of subcellular mechanisms, whole cell simulators, cell development simulators, organ simulators, physiological simulators, pathological simulators to body simulators.

3.2. Movements in systems biology in the United States

Through the acquisition of genome-wide information and the accumulation of such knowledge, the need for this type of approach is increasing rapidly so that systems biology is actually being put into practice. The following lists recent movements in systems biology in the United States.

3.2.1. NIH grants

NIH/NIGMS started the support of this field under the strong commitment of M. Cassman, Director of NIGMS. In September 2000, NIGMS decided to support the Alliance for Cellular Signaling (AFCS) project (www.afcs.org) that is led by W. Gilmann. This project aims to examine the mechanism of the signal transduction system inside cells, involving 21 research institutions and 52 principal researchers across the States. They are analyzing network pathways in the Mouse's G-Protein Coupled Signaling System in which more than 1000 proteins work highly cooperatively.

NIGMS also decided in September 2001 to support two large research groups. One is the research group of Cellular Communication that consists of more than 40 researchers (Consortium for Functional Glycomics—www.glycomics.scrips.edu), and another one is the Cell Migration Consortium—www.cellmigration.org.

NIGMS is also supporting many other small groups conducting quantitative analysis of complex biological systems using both computers and experiments (www.nigms.nih.gov/funding/complex-systems). DOE began to support the virtual cell research to examine the mechanism of radiation damage.

3.2.2. Institute for Systems Biology

In addition to the above publicly funded projects, Prof. L. Hood established the Institute for Systems Biology in Seattle in February 2000 (www.systemsbiology.org). His concept is that "The major challenge for biology and medicine in the 21st century is the understanding of biological systems, and the Institute's mission is to revolutionize biology and medicine with the realization that biology is an informational science".

According to his plan, systems biology research is to be done by recruiting 25 faculty members, with half of them being biologists focusing on systems approaches to the area of their interest, such as development, immunity, cancer biology, autoimmunity, yeast biology, and microbial, plant, and animal genomes, and with another half being cross-disciplinary—chemists, computer scientists, engineers, applied physicists and mathematicians in order to develop high-throughput tools and model the complexities of biological systems. The Institute will employ 400 staff in total, harnessing high-throughput equipment and powerful computational machines for genome, gene, protein and network analysis.

The Institute has recently attained a good example of the approach of the systems biology by integrating genomic and proteomics analyses for the understanding of metabolic pathway dynamics (Ideker et al., 2001).

3.3. The connection of the higher level simulation to the molecular level simulation (cell, organ, or body)

There are several levels for the analysis of systems of life:

1. Inside cell simulation.
2. Cell development simulation.
3. Organ simulation.
4. Body simulation.

Let me introduce some of the examples of these levels.

(A) Virtual Patient—4)

ENTELOS (www.entelos.com) has developed the virtual patient system "PhysioLab" and constructed the virtual obesity model, diabetes model and asthma model using this system. They adopt the top-down approach at first to list up various factors to affect a particular disease and next to break down those factors to various subsystems successively towards the metabolic pathway level for some of the subsystems. For example, in the case of the virtual obesity patient, various factors such as the metabolism of nutrition, energy consumption and accumulation, and nerve activities which include complex networks of neuro-peptides. This system consists of genetic, physiologic, and life-style factors and permits researchers to investigate the underlying pathophysiology that causes seemingly similar patients to respond differently to the same drug therapy.

(B) The Virtual Heart Project (Noble, 2002)—3)

The Virtual Heart Project that has been led by several researchers including Prof. Peter Hunter (Auckland) and Prof. Noble, of Oxford University, UK, constructed a huge model of the heart mechanism with the collaboration of more than 80 international researchers, with long-term data accumulations, and with the supercomputer of Oxford University. This model was demonstrated at the Millennium Dome in Greenwich in 2001.

This virtual heart model contains more than 1 million cells or elements each of which has the internal complex biochemical reactions, and is governed by more than 30 million equations in total.

The virtual heart has been constructed from the cellular level in the early stage, but recently the model is being connected to the genetic level, for example, by reconstructing the effects of particular mutations such as sodium-channel mutation that affect the voltage dependence.

Prof. Noble says that this virtual heart is just the beginning and similar approaches are underway not only on many organs, such as lungs and pancreas, but also on more complex immune systems. He also emphasizes that the full interpretation of genome will lie with this new technology formed from the bringing together of mathematics and medicine (www.balliol.ox.ac.uk). This work forms part of the International collaboration called the IUPS Human Physiome Project (www.physiome.org).

(C) Cell Development (Ko, 2001)—2)

As the recent progress of large-scale genomics approach, it has become possible to measure gene expressions at various developmental stages and at various mutational conditions on an unprecedented scale. Ko proposed the word "Embryogenomics" which expresses the idea that developmental biology meets genomics.

Recently, a CALTECH group published an exciting research result on the genomic regulatory network for development (Davidson et al., 2002). They investigated a gene regulatory network that controls the specification of endoderm and mesoderm in the sea urchin embryo, using a large-sale perturbation analyses, in combination with computational methodologies, genomic data, *cis*-regulatory analysis and molecular embryology. They could infer various system-level insights into the developmental process through this approach.

3.4. Several examples of systems biology research

Systems biology approaches are now emerging very rapidly besides the big research projects mentioned above, some examples of which are introduced as follows:

(1) Odell (Prof., University of Washington in Seattle) is simulating the movement of myosin and actin on the microtubule using about 9000 differential equation models (Foe et al., 2000). Moreover, he has obtained a good result of the analysis of the signal transmission system by the cell interaction model of fly (Dassow et al., 2000).
(2) Voit et al. made a model by combining genome-wide gene expression data of yeast with the enzymatic reaction process based on the Biochemical Systems Theory (BST), and examined the time response of ATP and Trehalose by a heat shock from 27°C to 37°C. They think that this approach will help infer new knowledge on the control mechanism of the metabolic pathway (Voit and Radivoyevitch, 2000).
(3) The group of B.O. Palsson made a metabolic model of the microorganism, and obtained results corresponding very closely to the experimental data (Covert et al., 2001; Edwards et al., 2001). A quantitative relation between the genotype and the expression type can be obtained by such approach.
(4) According to Staudt, research on the physiology and the pathology phenomenon of the immunity system using exhaustive gene expression data is now increasing in number (Staudt, 2001). He shows that the bottom-up approach that predicts a model of the response of the entire cell to environmental changes by the individual modeling of each signal pathway is effective.
(5) Endy and Brent wrote an overview of the current state and the future of the modeling of cell behavior (Endy and Brent, 2001). In the article, they say that although model building approaches have not always been successful so far, the situation is now changing rapidly. Pointing out that useful and predictive model building is being materialized by the increasing biological knowledge from the advanced measurement technologies and the gaining of computational power, they mention that qualitative modeling will be available first, followed by the quantitative modeling. Although there is the shortage of knowledge on genes, proteins and networks, the future challenges and guidelines are discussed.
(6) Peterson and Fraser have shown recent research results on the minimum number of genes required to maintain cell life (Peterson and Fraser, 2001). They identify a necessary protein group for basic biological functions, such as the amino acid synthesis, cell membrane formation, and the energy production and transportation, based on the comparison of two minimum microorganisms, *M. genetalium* (480 proteins) and *M. pneumonia* (685 proteins) of which genomes were fully sequenced. The minimum number of genes of *M. genetalium* is estimated to be 180–215, according to research on the evolution of minimum genes, the theoretical comparative analysis and the experiments of the transpozon insertion.
(7) Strogatz explains a general rule of the formation of the network structures including both artificial and natural networks (Strogatz, 2001). He points out the similarity of the networks between the biological systems, such as the cell cycle, the signal transduction, the metabolic pathway, and the neural network and the artificial systems, such as the power supply network,

the World Wide Web, and the telephone network. In any case, the structure, growth mechanism, dynamics, oscillation, non-linearity of networks are the common topics to be addressed. These topics will be useful for the understanding and the treatment of diseases.

3.5. *Movements in Japan in the field of systems biology*

In Japan, there are several unique activities going on and the development of this field seems to be promising.

(1) The basis of network analyses has been placed on network databases that cover biological processes (Karp, 2001). M. Kanehisa (Kyoto University), paying attention to the importance of the network database earlier, has been involved in the development of the metabolic pathway database KEGG (www.genome.ad.jp/kegg) since 1995. In this database, around 6550 pathways including metabolic and regulatory pathways are identified, and 311,758 genes of 76 genomes and 8868 compounds are interconnected in the networks. With Miyano and Akutsu joining in 2001, his laboratory has been enhanced to a new bioinformatics center for the network simulation (www.bic.kyoto-u.ac.jp).

(2) Tomita (2001) initiated the development of the E-cell simulation system (www.e-cell.org) in 1995, and has applied the system to the simulation of the red blood cell and the mitochondria etc. He founded in Yamagata in May 2001 a new center—Advanced Life Science Institute (www.bioinfo.sfc.keio.ac.jp/IAB) that consists of four experimental groups and one computer-modeling group. This formation is unique in the world. Aiming at the whole cell simulation of model organisms, such as *Escherichia coli* and *Bacillus subtilis*, exhaustive measurements of gene expressions and enzyme kinetics are being conducted.

(3) Kitano (2002) started up the Systems Biology Laboratory as a non-profit organization in June 2001. He has developed the Systems Biology Workbench (www.sbml.org) (Kyoda and Kitano, 2000) under the ERATO project (www.symbio.jst.go.jp) in cooperation with Caltech, USA. He has insisted on the importance of the systems biology and organized the 1st International Conference of Systems Biology in Tokyo in December 2000. The second was held at Caltech in November 2001 (www.icsb2001.org).

(4) The Ministry of Agriculture and Forestry started up the project called "Rice Genome Simulator" in March 2001. In this project, 17 groups including experts in the field of plant and rice, above-mentioned Tomita and Kitano, and other IT enterprises participate to develop the rice simulator, and the database construction and modeling of subsystems such as the chloroplast have already started.

(5) There are many other research groups that are moving in the direction of the systems biology. GSC/RIKEN has received the grant to have two teams related to the systems biology from fiscal year 2000, and A. Konagaya, one of the team leaders, is now focusing on the gene network analysis. Y. Kohara, Genetic Institute, is developing the cell division simulation of c-elegans. Otake, Hiroshima University, is developing a virtual *E coli* model. Kuhara, Kyushu University is analyzing the networks using pathway databases and gene expression data. H. Tanaka, Tokyo Medical and Dental University, is focusing on the minimum gene model. Ueda, University of Tokyo, has analyzed the Circadian Rhythm of Drosophila (Ueda et al., 2001).

As many researchers begin to pay attention to the importance of the systems biology, the Japanese government will increase the budget for this field.

3.6. The future of systems biology

Systems biology as a quantitative discipline has just started worldwide. However, the rapid growth of the genome sequence, gene expression and protein interaction data will expand the biological and pathological knowledge vastly and making models to connect such knowledge and data in various integrated systems will be inevitable. This tendency will even be accelerated, as seen in the case of the USA.

Natural life systems have features different from the artificial systems. Adaptability, evolution, redundancy, robustness and emergence are mysteriously and cleverly combined to form very complex systems. In order to analyze and understand these systems, not only the approach of reductionism but also the holistic approach is necessary. The collaboration between life scientists and information scientists will be crucially important to proceed in this direction. For the purpose, the formalization or modeling of the tacit knowledge that biologists have in their minds into explicit knowledge in the form of concrete models must be achieved. Accordingly, hypothesis-driven or model-driven experiments in large-scale will become new ways of research in biology.

4. Closing remarks

What is the principle that rules the phenomena of life? I wonder whether the first principle of life will be found, like the discovery of the principles of physics and chemistry. Can the phenomena of life be simulated by the above-mentioned analytical methods? How should we understand the evolution of life? Should we grasp life as a complexity system?

Even though we have yet to answer such essential questions (Wada, 2000), we are gaining a huge amount of practical data and extracted knowledge day by day. The progress of the life sciences will contribute to mankind in various application fields, such as medical and health care, food supply, energy and the environment.

Information science has come to play important roles in this context. Advanced information science and technologies are being used. The dynamic programming, neural network, hidden Markov model, other searching technologies, optimization methods and data mining methods are being incorporated in biology researches. And, Internet technologies, database management systems, supercomputer, parallel computers, pc-cluster systems, dedicated computer for biology are also crucial for the biology. Visualization technologies, Image processing technologies, Information theory, Systems theory and Control theory etc. will be needed more. In sum, advanced information sciences and technologies are now penetrating into the biology research. This tendency will be accelerated in the 21st century. We expect the age of "In Silico Biology" to come.

Acknowledgements

I would like to express my sincere admiration to Prof. Akiyoshi Wada, Director of RIKEN/GSC, for his continued efforts and the world-wide contribution to promoting the genomic sciences and structural genomics. My thanks are also extended not only to GSC colleagues but also to all academic, industrial and governmental people who have contributed to the promotion of the bioinformatics and the systems biology.

References

Aloy, P., Sternberg, M., et al., 2001. Automated structure-based prediction of functional sites in proteins. J. Mol. Biol. 311, 395–408.

Aravind, L., et al., 1999. Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res. 27, 1223–1242.

Bajorath, J., 2001. Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. Drug Discovery Today 6, 989–995.

Baker, D., 2000. A surprising simplicity of protein folding. Nature 405, 39042.

Baker, D., Sali, A., 2001. Protein structure prediction and structural genomics. Science 294, 93–96.

Bansal, A., 1999. An automated comparative analysis of 17 complete microbial genomes. Bioinformatics 15, 900–908.

Bassett, D., Eisen, M., Boguski, M., 1999. Gene expression informatics? It's all in your mine. Nature Genet. Suppl. 21, 51–55.

Blundell, T., et al., 1987. Nature 326, 347.

Bono, H., et al., 2002a. FANTOM DB: database of functional annotation of RIKEN Mouse cDNA Clones. Nucleic Acids Res. 30, 116–118.

Bono, H., et al., 2002b. READ: RIKEN expression array database. Nucleic Acids Res. 30, 211–213.

Bowie, J., Luthy, R., Eisenberg, D., 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164–170.

Brenner, S., 2000. Target selection for structural genomics. Nature Struct. Biol. 7 (Suppl.), 967–969.

Carninci, P., Hayashizaki, Y., 1999. High-Efficiency of Full-Length cDNA Cloning. Methods in Enzymology, Vol. 303. Academic Press, Inc. San Diego, pp. 19–44.

Covert, M., Palsson, B., et al., 2001. Metabolic modeling of microbial strains in Silico. Trends Biosci. 26, 179–186.

Cyranoski, D., 2000. Japan opens access to mouse cDNA data. Nature 407, 279.

Cyranoski, D., 2001. Japan's ape sequencing effort set to unravel the brain's secrets. Nature 409, 651–652.

Dassow, G., Odell, G., et al., 2000. The segment polarity network is a robust developmental module. Nature 406, 613–615.

Davidson, E., et al., 2002. A genomic regulatory network for development. Science 295, 1669–1678.

Davului, R., Grosse, I., Zhang, M., 2001. Computational identification of promoters and first exons in the human genome. Nature Genet. 29, 412–417.

Dhand, R., et al., 2000. Nature insight: functional genomics. Nature 405, 819–865.

Drews, J., 2000. Drug discovery: a historical perspective. Science 287, 1960–1964.

Edwards, J., Palsson, B., et al., 2001. In Silico predictions of *E. coli* metabolic capabilities are consistent with experimental data. Nature Biotechnol. 19, 125–130.

Endy, D., Brent, R., 2001. Modelling cellular behavior. Nature 409, 391–395.

Fleischmann, H., Venter, C., et al., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

Foe, V., Field, C., Odell, G., 2000. Microtubules and mitotic cycle phase modulate spatio temporal distributions of F-Actin and Myosin 2 in Drosophila syncytial blastderm embryos. Development 127, 1767.

Fujiyama, A., et al., 2002. Construction and analysis of a human–chimpanzee comparative clone map. Science 295, 131–134.

Fukuda, Y., Washio, T., Tomita, M., 1999. Comparative study of overlapping genes in the genomes of *Mycoplasma genetalium* and *Mycoplasma pneumoniae*. Nucleic Acids Res. 27, 1847–1853.
Ginsbug, G., McCarthy, J., 2001. Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol. 19, 491–496.
Greer, J., 1981. Comparative model-building of the mammalian serine proteases. J. Mol. Biol. 153, 1027–1042.
Guigo, R., et al., 2000. An assessment of gene prediction accuracy in large DNA sequences. Genome Research 10, 1631–1642.
Hattori, Y., Sakaki, Y., et al., 2000. The DNA sequence of human chromosome 21. Nature 405, 311–319.
Hogenesch, J., et al., 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell 106, 413–415.
Hutchinson, C., Venter, C., et al., 1999. Global transposon mutagenesis and a minimal mycoplasma genome. Science 286, 2165–2169.
Ideker, T., Hood, L., et al., 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292, 929–933.
International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
International Human Genome Mapping Consortium, 2001. A physical map of the human genome. Nature 409, 934–941.
Jones, D., Taylor, W., Thornton, J., 1992. A new approach to protein fold recognition. Nature 358, 86–89.
Jones, D., 2000. Protein structure prediction in the post-genomic era. Curr. Opin. Struct. Biol. 10, 371–379.
Karp, P., 2001. Pathway databases: a case study in computational symbolic theories. Science 293, 2040–2044.
Kigawa, T., et al., 1999. Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett. 442, 15–19.
Kihara, D., Kanehisa, M., 2000. Tandem clusters of membrane proteins in complete genome sequences. Genome Res. 10, 731–743.
Kihara, D., Skolnick, J., et al., 2001. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. PNAS 98, 10125–10130.
Kitano, H., 2002. Systems biology: a brief overview. Science 295, 1662–1664.
Ko, M.S.H., 2001. Embryogenomics: developmental biology meets genomics. Trends Biotechnol. 19, 511–518.
Kondo, S., et al., 2001. Computational analysis of full-length mouse cDNA compared with human genome sequences. Mamm. Genome 12, 673–677.
Koonin, E., 1999. Editorial; the emerging paradigm and open problems in comparative genomics. Bioinformatics 15, 265–266.
Kyoda, K., Kitano, H., 2000. Construction of a Generalized Simulator for Multi-Cellular Organisms and it's Application to SMAD Signal Transduction. PSB-2000, 317-328.
Liu, J., Rost, B., 2001. Comparing function and structure between entire proteomes. Protein Sci. 10, 1970–1979.
Makarova, K., et al., 1999. Comparative genomics of the archea: evolution of conserved protein families, the stable core and the variable shell. Genome Res. 9, 608–628.
Miki, R., et al., 2001. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. PNAS 98, 2199–2204.
Moult, J., Hubbard, T., et al., 1999. Critical assessment of methods of protein structure prediction (CASP): round 3. Proteins, 3 (Suppl.) 2–6.
Murzin, A., 2001. Progress in protein structure prediction. Nature Struct. Biol. 8, 110–112.
Murzin, A., Brenner, S., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.
Nature News, 1996. Superconductivity Spurs Japanese Plan for NMR Research. Nature 381, 105.
Nature News, 2000. Autumn annotation. Nature Genetics 25, 371.
Noble, D., 2002. Modeling the heart-from genes to cells to the whole organ. Science 295, 1678–1682.
O'Brien, S., et al., 1999. The premise of comparative genomics in mammals. Science 286, 458–480.
Ohler, U., Niemann, H., 2001. Identification and analysis of eukaryotic promoters: recent computational approaches. TRENDS Genet. 17, 56–60.

Orengo, C., Todd, A., Thornton, J., 1999. From protein structure to function. Curr. Opin. Struct. Biol. 9, 374–382.
Perriere, G., Duret, L., Gouy, M., 2000. HOBACGEN: database system for comparative genomics in bacteria. Genome Res. 10, 379–385.
Peterson, S., Fraser, C., 2001. The complexity of simplicity. Genome Biol. 2, 1–8.
Pilpel, Y., et al., 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. Nature Genet. 29, 153–159.
Quackenbush, J., 2000. Viva la revolution! a report from the FANTOM meeting. Nature Genet. 26, 255–256.
Sanchez, R., Sali, A., 1999. ModBase: a database of comparative protein structure models. Bioinformatics 15, 1060–1061.
Sanseau, P., 2001. Impact of human genome sequencing for in silico target discovery. Drug Discovery Today 6, 316–322.
Seki, M., Shinozaku, K., et al., 2001. Monitoring the Expression pattern of 1300 Arabidopsis genes under drought and cold stresses using full-length cDNA microarray. Plant Cell 13, 61–72.
Shibata, K., Hayashizaki, Y., et al., 2000. RIKEN integrated sequence analysis (RISA) system—384-format sequencing pipeline with 384 multicapillary sequencer. Genome Res. 10, 1757–1771.
Simon, A., Baker, D., 1999. Ab initio protein structure prediction of CASP-3 targets using Rosetta. Proteins 3 (Suppl.) 171–176.
Staudt, L., 2001. Gene expression physiology and pathophysiology of the immune system. Trends Immunol. 22, 35–40.
Stevens, R., Yokoyama, S., Wilson, I., 2001. Global efforts in structural genomics. Science 294, 89–92.
Stormo, G., et al., 2000. Gene-finding approaches for eukaryotes. Genome Res. 10, 394–397.
Strogatz, S., 2001. Exploring complex networks. Nature 410, 268–276.
Suzuki, H., et al., 2001. Protein–protein interaction panel using mouse full-length cDNAs. Genome Res. 11, 1758–1765.
Suzuki, Y., Sugano, S., et al., 2001a. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. 11, 677–684.
Swinbanks, D., 1987b. Japanese plans to sequence human genome. Nature 326, 323.
Takahashi, S., Kambara, H., et al., 1994. Multiple sheath-flow gel capillary-array electrophoresis for multicolor fluorescent DNA detection. Anal. Chem. 66, 1021–1026.
Tatusov, R., Koonin, E., Lipman, D., 1997. A genomic perspective on protein families. Science 278, 631–637.
The International SNP Map Working Group, 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Science, 409, 928–933.
The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, 2001. Functional annotation of a full-length mouse cDNA collection. Nature 409, 685–690.
Tomita, M., 2001. Whole-cell simulation: a grand challenge of the 21st century. Trends Biotechnol. 19, 205–210.
Ueda, H., et al., 2001. Robust oscillations within the interlocked feedback model of Drosophila circadian rhythm. J. Theoret. Biol. 210, 401–406.
Venter, C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.
Vitkup, D., Melamud, E., Moult, J., Sander, C., 2001. Completeness in structural genomics. Nature Struct. Biol. 8, 559–566.
Voit, E., Radivoyevitch, T., 2000. Biochemical systems analysis of genome-wide expression data. Bioinformatics 16, 1023–1037.
Wada, A., 1987. Automated high-speed DNA sequencing. Nature 325, 771–772.
Wada, A., 2000. Editorial; bioinformatics—the necessity of the quest for 'first principle in life'. Bioinformatics 16, 663–664.
Wada, A., 2001. Genomic sciences center (RIKEN). Science Technol. Japan 77, 19–23.
Weir, M., Swindells, M., Overington, J., 2001. Insights into protein function through large-scale computational analysis of sequence and structure. TIBT 19, S61–S66.
Yamanaka, I., et al., 2001. Mapping of 19032 mouse cDNAs on mouse chromosomes. J. Struct. Funct. Genomics 2, L72–L86.

Yokoyama, S., et al., 2000a. Structural genomics projects in Japan. Nature Struct. Biol. 7 (Suppl.), 943–945.

Yokoyama, S., Matsuo, Y., et al., 2000b. Structural genomics projects in Japan. Prog. Biophys. Mol. Biol. 73, 363–376.

Yudate, H., et al., 2001. Hunt: launch of a full-length cDNA database from Helix Research Institute. Nucleic Acids Res. 29, 185–188.

Zhang, B., Godzik, A., et al., 1999. From fold prediction to function predictions: automation of functional site conservation analysis for functional genome predictions. Protein Sci. 8, 1104–1115.