2

Trend Report on Bioinformatics

MARIKO SHOJI AND SHIN-ICHI MOGI
Life Science and Medical Research Unit

2.1 Introduction

In February 2001, the outline of the human genome was individually reported by the International Analysis Team and Celera Genomics. Now that the genome sequencing of 60 or more living species have been determined, the so-called post-genome sequencing studies are getting into full swing including gene expression analysis, protein structure analysis, proteome analysis, and intermolecular interaction analysis. These studies are intended to efficiently organize and analyze a vast amount of diverse biological information for elucidating the biological and medical implications. To fulfill such a mission successfully, bioinformatics is essential technology.

In this article, we give an overview of bioinformatics, focusing on the human genome studies, and list the challenges in this region.

2.2 Definition of Bioinformatics

Originally, the term "bioinformatics" has been used to refer to a specific region of study, in which the viewpoints and philosophy of informatics were incorporate into bioscience, although the meaning of it has been steadily extended.

Prof. Toshihisa Takagi of the Institute of Medical Science, Tokyo University defines bioinformatics as follows.

> "Information technology allows a boundless searching space (for example, the number of genes or proteins, or combinations of interactions between them), which has required manual examination or experiments for verification so far, to be narrowed down and its basic principle"

In the United States, the Biomedical Information Science Technology Initiative (BISTI) consortium of the National Institutes of Health (NIH) differentiates the term "bioinformatics" from the term "computational biology."

> **Bioinformatics**
> Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data
> **Computational biology**
> The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Both of the two terms have almost the same underlying concept and NIH, perhaps, differentiates between them to represent immediate challenges specifically.

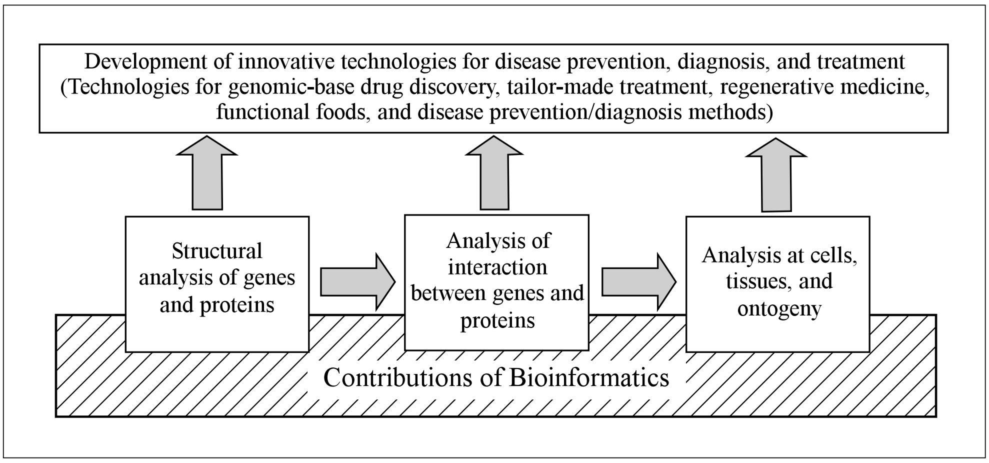
This section describes "What is bioinformatics?," closely focusing on the concept of bioinformatics defined by NIH.

2.3 Positioning and Categorization of Bioinformatics

2.3.1 Positioning of Bioinformatics

Bioinformatics plays an important role as one of basic technologies supporting researches in the life science area, mainly using genome analysis (Figure 1).

Figure 1: Positioning of bioinformatics in the life science area


Development of innovative technologies for disease prevention, diagnosis, and treatment
(Technologies for genomic-base drug discovery, tailor-made treatment, regenerative medicine, functional foods, and disease prevention/diagnosis methods)
Structural analysis of genes and proteins
Analysis of interaction between genes and proteins
Analysis at cells, tissues, and ontogeny
Contributions of Bioinformatics

2.3.2 Categorization of Bioinformatics in the Post-genome Research Area

Table 1 shows the direction of the post-genome researches, the database of bioinformatics supporting it, and the data analysis methods. The post-genome researches have been shifted from those on the "structure field" into the "relationality field" and oriented toward researches on cellular and ontogenetic functions (the "function field") for the systematic elucidation of life.

In the "structure field", various types of analyses including the DNA sequences and protein structure analyses are categorized. Homology search is one of data analysis methods commonly used in this field. This method compares DNA sequences, and extract knowledge of gene structures, functions from those homologies. Furthermore, another methods have been developed, which allows us to estimate the specific sites of genes based on the statistical characteristics observed in the gene sequences.

On the other hand, in the "relationality field," the gene expression analysis, which determines whether genes are switched on or off, and the intermolecular interaction analysis, which detects any interactions among proteins and others, are categorized. These analyses treat gene classification based on gene expression information under various conditions and intracellular localized site prediction, which predicts how proteins behave in the cells based on the physiochemical characteristics of proteins estimated from amino-acid sequences.

In the "function field", the analysis methods are grouped and include those for elucidating signal transmission among the cells and the ontogenetic mechanism.

Table 1: Categorization of bioinformatics in the post-genome research region

	Subject	Database	Data analysis
Structure	Sequence, Structure	DNA sequences, Gene polymorphism, Protein-amino acid sequences, Protein structure, etc.	Homology search, Gene discovery, Motif extraction, Protein structure Prediction, etc.
Relationality	Expression, Localization, Interaction	Gene expression information, Intermolecular interaction, Proteome, etc.	Intracellular localized site prediction, Intermolecular interaction prediction, Gene expression clustering, etc.
Function	Cell function, Ontogenetic function	Signal transmission, Ontogenetic/ Physiological function, Immune function, Brain function, etc.	Pathway comparison, Computer simulation, etc.

R & D is under progress.

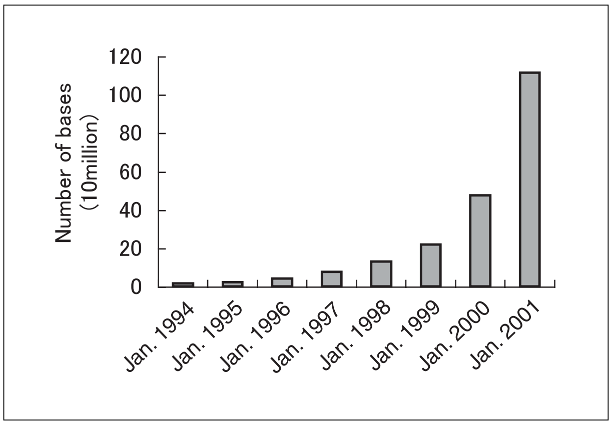
Some type of future strategic motivation is critical.

Source: Authors' compilation based on materials from Prof. Toshihisa Takagi, the Institute of Medical Science, Tokyo University

In this field, however, few useful data analysis methods have been developed.

At present, public funds are invested almost exclusively in researches and developments in the "structure field", and its infrastructure is being consolidated. As opposed to this field, almost no large-scale attempts have been made toward database arrangement and development of data analysis methods with respect to the "relationality" and "function field". This is true, especially, of the latter field. To facilitate the evolution of these out-of-stream fields, some type of strategic policy for motivation must be developed in the near future.

Figure 2: Transition in the number of bases registered in DDBJ



Source: Authors' compilation based on DDBJ data

2.4 Current Status of Bioinformatics

2.4.1 Database

(1) Rapid increase in data amount

The parties concerned are obligated to register any obtained information on decoded DNA sequences and others into any of the public databases of GenBank (U.S.), EMBL (Europe), and DDBJ (DNA Databank of Japan). These three databases maintain consistency in data generation with each other, because they exchange data among themselves. In recent years, the data amount has drastically increased as can be seen from the fact that a vast amount of DNA sequences were registered, for example, 11.1 billion in January 2001 and 14.1 billion in October 2001, all of which reflect the recent, rapid progress in this area, which include registration of a large number of partial mRNA sequences (EST) into the databases and un-intermitted elucidation of the DNA sequences of various species of organisms (Figure 2).

In data organization, not only base strings are enumerated but they are also annotated with various data clarified by analyses, such as the locations of gene information coding regions as well as the structures and functions of genes, and with the literatures related to these genes, all of which are registered together into the databases. In the future, it is important to make annotation attached to sequence data both qualitatively and quantitatively complete, as sequence data described here increases.

Additionally, the data structures come into the complex phase as data on gene expression under various combinations of conditions and on intermolecular interaction are added. For this reason, various requirements constantly would arise such as further speed-up and increased disk capacity of information system devices. Besides, another important challenge to be solved is to store high quality data in the databases through researcher's updating of annotation.

(2) Examples of databases

Table 2 shows typical databases commonly used in the genome research region and others. Generally, DNAs are easier to be purified than proteins and their sequences can be determined with less difficulty using a DNA sequencer, which makes the database consisting of DNA sequences the largest of all the databases. In addition to those described in section (1), the public databases include the databases containing DNA sequences of human genes and Single Nucleotide Polymorphisms (SNPs).

Additionally, there exist various types of databases containing gene information such as amino-acid sequences of proteins, motif sequences useful in predicting gene functions, and protein structures. It is said that about 400 types of databases can be found throughout the world.

Note that the journal "Nucleic Acids Research," published on January 1 every year, features these databases.

In Japan, for most of the databases, frequent access from abroad is not authorized except for some, as example, the pathway database. For us, Japanese researchers, to lead the post-genome researches in

Table 2: Examples of databases

Content in database	Database name (country)	Content in database	Database name (country)
DNA sequences	GenBank (U.S.) EMBL (Europe) DDBJ (Japan)	Amino-acid sequences	SWISS-PROT (Europe) PIR (U.S.)
DNA sequences of human genes	UniGene (U.S.)	Amino-acid sequence domains	Pfam (Europe)
Single Nucleotide Polymorphisms	dbSNP (U.S.) JSNP (Japan)	Amino-acid motifs	PROSITE (Europe) BLOCKS (U.S.)
Genetic diseases	OMIM (U.S.) Mutation Database (Europe)	Protein structures	PDB (U.S.), SCOP (Europe) CATH (Europe)
General human sequence informations	HGREP (Japan) Ensembl (Europe)	Pathways	KEGG/PATHWAY (Japan)
General human informations	LocusLink/Refseq (U.S.) GDB (Canada)	Literatures	MEDLINE (U.S.)

the world, it may be required to build original databases on a larger scale to deliver gene information to the countries throughout the world.

(3) Integration of databases

To successfully perform data analysis using bioinformatics, combinations of various databases and searching systems are often used instead of only one kind of database. For this reason, systems, which link various databases and searching systems into one integrated system, have been developed. Table 3 shows examples of typical integrated database searching systems.

One problem about database integration to be solved lies in that no standardization and coordination have been made for the methods for displaying database data and for describing searching conditions. In other words, it is important that a systematic theory (ontology) be built using the refined lexis system and describing method rather than using the concept and terminology specific to each subject of search.

For example, in the United States, Interoperable Informatics Infrastructure Consortium (I3C), which consists of more than 40 biochemical businesses and information system and/or service providers such as INCOGEN and Oracle, was founded in January 2001 and has commenced activities on propelling standardization efforts for data exchange, management, and others in the life science area. At present, I3C is working on the standardization task of XML data description and the communications protocol.

Table 3: Examples of typical database searching systems

Integrated system	Service provided from
DBGET (Japan)	Institute for Chemical Research, Kyoto Univ., Institute of Medical Science, The Univ. of Tokyo
Entrez (U.S.)	National Center for Biotechnology Information (NCBI)
SRS (Europe)	European Molecular Biology Laboratory (EMBL)

2.4.2 Hardware

As described in 2.4.1 (1), in the bioinformatics, it has been increasingly required that a high level information system, based on a series of high performance computers (super computers), be built to address the rapid increase in data amount. Table 4 shows the estimation of computer performance needed for genome analysis. Assuming some possible evolution in the technologies for describing and simulating the complex systems such as the life system, a parallel computer system with a performance ranging from several-ten to several-hundred Tera flops,

Table 4: Computer performance needed for major genome analyses

Genome analysis	Performance (flops)
Protein family classification	1 Tera
Phylogenetic diagram	10 Tera
Sequence assembly	10^2 Tera
Sequence comparison	$>10^2$ Tera
Gene modeling	10^5 Tera

Source: Authors' compilation by making reference to the materials provided by Professor Nishikawa based on the data of the 2000 Report "Advanced Computational Structural Genomics", U.S. DOE Scientific Simulation Initiative (SSI) which consists of several thousands or even several-ten thousands of processors, may be required ("Strategy for Genome Information Science in Japan"—the Genome Science Committee, Life Science Working Group, Council for Science and Technology Policy, November 2000).

For example, ASCI white, which was ranked first out of the "TOP 500 worldwide high-performance computers" announced in November 2001, has 12.3 Tera flops of peak performance.

In the bioinformatics, parallel processing systems and parallel computers are under development. On the other hand, software development lags behind, for example, delays due to difficulties in integrating applications running on high-performance computers into the parallel system. (Note: flops is a measure of calculation performance indicating the speed a computer can calculate floating-point calculations. Tera=10^{12})

2.4.3 Data Analysis Methods and Software

Table 5 lists the data analysis methods for identifying the desired knowledge from databases and software executing these methods.

2.4.4 Industrialization of Bioinformatics

In recent years, bio-related venture businesses have proactively developed their activities centering around developed countries in the world; as of 2000, about 160 in Japan, about 1,300 in U.S., and about 700 in Europe.

In the bioinformatics, such bioinformatic businesses actively participate in this industry

Table 5: Main data analysis methods and software

Method	Descriptions of analyses and example software
Homology search	Compares among sequences to extract knowledge based on their homology and is most commonly used. BLAST, FASTA, and Smith-Waterman, as well as high-sensitive programs, which can extract even weak homology, such as PSI-BLAST, SAMT99.
Gene discovery	Estimates genes based on statistical characteristics observed in gene sequences and is used to extract unknown genes, which cannot be discovered by homology search. GENSCAN and DIGIT.
Motif extraction	Identifies characteristically short sequences (motif), which are found in DNA binding sites and functional sites such as an oxygen active center. Compares sequences against the databases containing motifs of amino-acid sequences, such as PROSITE.
Intracellular localized site prediction	Based on hydrophobicity indexes of amino acids, physicochemical properties of electric charges, and the sequences of localized signals, etc., predicts where biosynthesized proteins go within cells. Signal P, which predicts the positions of signal sequences, PSORT, which predicts localized positions, and SOSUI and TMHMM, which predict membrane-penetrating regions.
Protein structure prediction	Swiss-Model and MODELLER homology modeling software, which predict protein structures based on homology among sequences. DALI and MODBASE, which determine similarities based on the results of comparison among protein structures.

Source: Authors' compilation by referencing the "Genomic Medical Science and Introduction to Bioinformatics", Experimental Medicine Vol. 19, No. 11 (extra number) P. 61-66, P. 73-81

Table 6: Main bioinformatics-related venture businesses in the United States and Europe

Company	Business
Incyte Genomics (U.S.)	cDNA database, gene expression (DNA chips), etc.
Human Genome Sciences (U.S.)	Secretory protein database, membrane protein database, gene drug development, etc.
Celera Genomics (U.S.)	Genome database (human, mouse, and Drosophila), proteome, SNP, etc.
Gene Logic (U.S.)	Gene expression database, etc.
CuraGen (U.S.)	Gene expression database, SNP database, etc.
Genset (France)	Disease analysis using SNP database, etc.
deCODE genetics (Iceland)	Clinical database, genealogy database, polymorphism database, etc.

Source: Authors' compilation by making reference to the materials provided by Professor Nishikawa based on the "Genome Information Ventures (P. 96-102), Leading Edge of Genome Medical Science and Worldwide Bio-Ventures, Yodosha, 2001 such as those providing the databases and software, which are customized for user requirements. To successfully perform research and development using genome information, any data management system and search assistance system, which can satisfy the requirements including higher security, faster analysis, and adaptation to individual researches, are essential in addition to public databases and software commonly available. Table 6 shows main bioinformatics-related venture businesses in the United States and Europe.

2.5 Activities for Propelling Bioinformatics

2.5.1 Political Principle

Recently, political emphasis has been laid on bioinformatics in many countries throughout the world, assuming that it is critical to propel researches on genome. Table 7 shows the main propelling bases for bioinformatics in the United States and Japan. Besides, active measures are conducted mainly in the laboratories of universities.

In Japan, bioinformatics was listed as one of the challenges to be exclusively and strategically tackled in the life science area in the 2nd-stage Science and Technology Basic Plan (decided by the Government of Japan, March 2001). In addition, it was listed as one of the emphasized issues in the "2001 Emphasized Guideline on Advancement of Science and Technology" (Policy Committee, Council for Science and Technology Policy, June 2000), with 10.4 billion yen budgeted for 2001. In 2002, the Ministry of Education, Culture, Sports, Science and Technology founded the Institute for Bioinformatics Research and Development (Chiyoda-ku, Tokyo) and the Ministry of Economy, Trade and Industry formed the Japan Biological Information Research Centre (Kohto-ku, Tokyo) and the Computational Biology Research Center (Kohto-ku, Tokyo), respectively. The Japan Biological Informatics Consortium (JBiC), consisting of 87 businesses in the private sector, is making great efforts in bioinformatics research and development.

In the United States, from an early stage, bioinformatics actively advanced under the leadership of NIH. NLM, which has jurisdiction over NCBI, playing a core role in bioinformatics in the United States, was budgeted about 246 million dollars (about 29 billion yen) for 2001.

In addition, in 2001, the Center for Bioinformatics and Computational Biology (CBCB) was founded as an organization, which fosters bioinformatics

Table 7: Main bases for propelling bioinformatics in Japan and the U.S.

Japan	Ministry of Education, Culture, Sports, Science and Technology	Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics Human Genome Center, Institute of Medical Science, Tokyo Univ. Bioinformatics Center, Institute for Chemical Research, Kyoto Univ. Genomic Sciences Center, Institute of Physical and Chemical Research (RIKEN) Institute for Bioinformatics Research and Development, Japan Science and Technology Corporation(JST)
	Ministry of Economy, Trade and Industry	Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology(AIST) Japan Biological Information Research Centre (JBIRC), National Institute of Advanced Industrial Science and Technology (AIST)
		Japan Biological Informatics (JBiC), etc.
U.S.	Department of Health and Human Services (DHHS)	National Institutes of Health (NIH) ├ National Library of Medicine (NLM) │ └ National Center for Biotechnology Information (NCBI) ├ National Institute of Biomedical Imaging and Bioengineering(NIBIB) ├ National Institute of General Medical Sciences (NIGMS) │ └ Center for Bioinformatics and Computational Biology (CBCB) ├ National Human Genome Research Institute (NHGRI) └ Biomedical Information Science and Technology (BISTI) Consortium
	Department of Energy (DOE)	Office of Biological and Environmental Research (BER)
	National Science Foundation (NSF), etc.	

researches, and will subsidize about 10 million dollars (about 1.2 billion yen) to universities and other institutes during its the initial year. Furthermore, DOE, NSF, DARPA(Defense Advanced Research Projects Agency) and others propel researches in the bioinformatics area.

In Europe, EMBL(European Molecular Biology Laboratory) and EBI(European Bioinformatics Institute ; one of the divisions of EMBL) as well as Sanger Center (U.K.) are playing core roles in developing bioinformatics. The public bioinformatics budget of Europe for 2000 is 100 million Euro dollars (about 10 billion yen), among which 10 million Euro dollars is allocated to EBI.

2.5.2 Activities for Training

With respect to recent researches in the life science area, the quality and performance of bioinformatics tools to be used and techniques for making full use of them are the main factors contributing to the progress of research and development. For this reason, the need for talents who can use these sophisticatedly tools and people who can develop superior algorithms and software, has rapidly increased. However, since this bioinformatics is a new study, a sufficient amount of personnel has not been trained and training has become a major problem to be solved in many countries.

"Strategy of Genome Information Science in Japan" (November 2000, Genome Science Committee, Life Science Working Group, Council for Science and Technology Policy) defines the tentative requirements for training as follows.

1) Training of adaptable talents: Development of education and training programs to make full use of existing talents, as well as provision of training opportunities and incentives.
2) Training near-future and future talents: Formation of bases for giving places where communication of research information and experiment of test and fault is achieved by reorganizing graduate school courses and major subjects of universities.

With respect to requirement 1), for example, Institute for Bioinformatics Research and Development, Japan Science and Technology Corporation(JST), commenced the "Genome Literacy Course," a practical training program, in cooperation with the Human Genome Center, Institute of Medical Science, Tokyo University, in 2001. In this course, a training program is provided for researchers to learn the methods of using databases and analytical software.

With respect to requirement 2), a new business for training was started in 2001 ,by allocating the 2001 Special Coordination Funds for Promoting Science and Technology (SCF), and four themes were adopted. For example, Keio University's "System biologist education and training program" was initiated, in which a new life information course will be established in the Department of Science and Engineering.This program is intended to help students learn about experiments and computer science as one method for understanding biology based on chemistry, physics/information, and mathematics. This program has as its target that about 40 graduated students and about 25 masters will be provided to pharmaceutical and computer companies and consulting firms.

In the United States, the subject of training is actively attacked as well. The divisions of NIH include the Fogarty International Center (FIC), the National Cancer Institute (NCI), the National Institute of Aging (NIA), the National Institute of General Medical Sciences (NIGMS), the National library of Medicine (NLM), and the National Human Genome Research Institute (NHGRI). Out of them, for example, NLM invested funds in 12 training programs such as those in Yale University and Columbia University.

2.6 Conclusion

Bioinformatics can be viewed as not only a basic technology supporting the researches in the life science area but also as a new region bearing the life science aera, a fused interdisciplinary area.

At present, to analyze DNA sequences and protein structures, databases and data analysis methods are being reorganized. In the future, research and development of databases and data analysis methods needed for researches on cell functions and ontogenetic functions must be strategically advanced.

Since databases are the foundation for data analysis, it is necessary to always attain high quality data. It is important that appropriate data management, such as update of annotation data, is maintained.

With respect to hardware, even the current highest-performance computers are insufficient for various types of genome analyses, and further improvement of performance is required. Similarly, software must be improved so that applications running on high-performance computers can be integrated into one parallel system.

In Japan, efforts in forming the bases for propelling bioinformatics and in developing and implementing programs for training have just started. Considering the importance of bioinformatics in the future life science area, it is desirable that further continuous policies are implemented.

Acknowledgement

We would like to express our deep thanks to Prof. Toshihisa Takagi of the Institute of Medical Science, The University of Tokyo for his valuable advice and the provision of the related materials.

(Original Japanese version: published in December 2001)