1

Trends in Human Genome Analysis by the International Human Genome Sequencing Consortium and by Celera Genomics, and Future Prospects of Post Genome Research in Japan

SHIN-ICHI MOGI, MARIKO SHOJI, HIROKO EBIHARA, AND AKIHIRO HASEGAWA
Life Science and Medical Research Unit

1.1 Introduction

Human genome is the genetic blueprint of the human being, which conveys all genetics information necessary for a human. In general, the term "human genome" refers to the DNAs on 22 pairs of autosomal chromosomes and 2 sex chromosomes. DNAs contain 4 kinds of bases (adenine, guanine, cytosine and thymine), and human genome is composed of about 3 billion base pairs. It has been said that, among all the base pairs making up the human genome, only about 3% corresponds to genes (regions encoding genetic information).

The concept of the sequencing of the whole human genome was already created in the mid 1980s. However, for reasons such as the fact that an extremely long time was required to sequence the whole human genome with the use of techniques for sequencing at that time, sequencing of the whole human genome was thought to have only limited feasibility. However, research on human genome has gradually become active, and, in 1990, the International Human Genome Sequencing Project was launched on a full scale under the initiative of the United States. While the project was originally planned to be completed in 15 years, the date for the project's completion has been brought forward because of the increase in performance of genome sequencers as well as the entry of venture capital companies into the project. As a result, in February 2001, the Human Genome Sequencing Consortium and Celera Genomics, a venture capital company in the United States, separately published that they have almost finished the sequencing of the whole human genome.

The research conducted by utilizing the determined base sequence of human genome is called "post genome research." Since post genome research includes many factors that may lead directly to future commercial applications, cutthroat world-wide competition involving private companies has been carried out in this field.

This article summarizes the international trends in human genome sequencing, trends in research performed by Celera Genomics, a noteworthy venture capital company in the United States, as well as trends in post genome research in Japan.

1.2 International trends in human genome sequencing

In February 2001, the International Human Genome Sequencing Consortium and Celera Genomics separately published articles on human genome sequencing in the journals "Nature" and "Science," respectively (Chart 1).

1.2.1 Sequencing by the International Human Genome Sequencing Consortium

Chart 2 shows the number of bases in finished human sequences among the human genome sequences entered into GenBank, a public base sequence database.

With regard to the complete sequencing of a specific pair of chromosome, a joint research team composed of researchers from Japan, the United Kingdom and the United States (Keio University, Sanger Centre in the United Kingdom, as well as the University of Oklahoma and Washington University in the United States) finished the

complete sequencing of chromosome 22. In addition, Keio University, the Institute of Physical and Chemical Research, etc., collaboratively finished the complete sequencing of chromosome 21 in May 2000. The accuracy of the sequencing of those chromosomes (chromosomes 21 and 22) is as high as 99.99%, earning an excellent reputation worldwide. Chromosome 22 has the gene responsible for Parkinson's disease, genes related to autoimmune diseases, etc., and chromosome 21 has relations with Down's syndrome (an extra chromosome 21 is detected in Down's syndrome), a gene related to Alzheimer's disease, etc. The sequencing of human genome including such disease-related genes as those mentioned above is expected to greatly expedite the progress in research on the mechanism of expression of those genes.

However, there still remain draft sequences, i.e., regions that are difficult to be completely sequenced on chromosomes other than chromosomes 21 and 22. The International Human Genome Sequencing Consortium is still pursuing the sequencing project, aiming to complete the sequencing of the whole human genome with high accuracy by the end of 2003.

1.2.2 Human Genome Sequencing by Celera Genomics

Celera Genomics did not sequence each chromosome separately but adopted the whole genome shot-gun method where all the 24 chromosomes (22 pairs of autosomal chromosomes and 2 sex chromosomes) were

Chart 1: Publishers of the human genome sequence

Journal Title	Nature February 15, 2001 issue "Initial sequencing and analysis of the human genome"	Science February 16, 2001 issue "The Sequence of the Human Genome"
Publisher	International Human Genome Sequencing Consortium ● United States (Washington University; Joint Genome Institute, Department of Energy; Baylor College of Medicine; Whitehead Institute, Massachusetts Institute of Technology, etc.) ● United Kingdom (Sanger Centre, etc.) ● France (Genoscope, etc.) ● Japan (Institute of Physical and Chemical Research [Riken], Keio University, etc.) ● Germany, etc.	Celera Genomics (American venture capital company)

Chart 2: Finished human sequences entered into GenBank by institutions participating in the International Human Genome Sequencing Consortium

Order of entry	Research Organization	Number of bases sequenced (kb)
1	The Sanger Centre (United Kingdom)	284,353
2	Washington University Genome Sequencing Center (United States)	175,279
3	US DOE Joint Genome Institute (United States)	78,486
4	Baylor College of Medicine Human Genome Sequencing Center (United States)	53,418
5	Genoscope (France)	48,808
6	Whitehead Institute, Center for Genome Research (United States)	46,560
7	Department of Genome Analysis, Institute of Molecular Biotechnology (Germany)	17,788
8	Institute of Physical and Chemical Research (Japan)	16,971
9	University of Washington Genome Center (United States)	14,692
10	Keio University (Japan)	13,058
	Others	92,614
	Total	842,027
	(Total number of bases on the whole human genome.)	about 3,200,000

Prepared by the Science and Technology Foresight Center based on Table 3 on page 868 in the February 15, 2001 issue of the journal Nature.

simultaneously analyzed. At the first step of Celera's whole genome shot-gun protocol, sampled human genome was randomly physically sheared into small fragments of the same size (2 kb or 10 kb) by using ultrasound, and a library of the fragments was constructed. At the next step, the regions composed of about several hundred base pairs at both ends of those fragments were analyzed in DNA sequencers. Scientists then used powerful computers to assemble the fragments back into place to determine the sequence of the whole human genome. According to Celera Genomics, with the whole genome shot-gun method, 95% of the whole human genome could be sequenced with an accuracy of as high as 99.96%.

1.2.3 Access to the data on base sequences of human genome

While the data on human genome sequences entered into GenBank by the International Human Genome Sequencing Consortium can be accessed at no charge, you need to make a contract with Celera Genomics in order to gain access to the data on human genome sequences determined by the company.

(1) Human genome sequences determined by the International Human Genome Sequencing Consortium

In the sequencing activities by the International Human Genome Sequencing Consortium, respective participating institutions are assigned specific regions to be sequenced, and, if the sequencing of a region is finished, the data on the sequence of the region are to be immediately entered into public databases including GenBank and then made available for public view at no charge. Those data entered on such public databases include data on the base sequences of the finished human sequences such as those of chromosomes 21 and 22 as well as data on the regions on draft sequences. As mentioned in section 1.2.1, the sequencing of the whole human genome must be completed by the International Human Genome Sequencing Consortium before free access and utilization of data on the sequence of the whole genome with high accuracy become possible.

(2) Human genome sequences determined by Celera Genomics

Researchers, institutions, companies, etc., cannot freely access or utilize the data on human genome sequences obtained by Celera Genomics unless they make a contract with the company. In Australia, for example, a contract with Celera Genomics has been made at a nation level, and national institutions have access to the data on human genome sequences determined by the company. In Japan, some universities and private companies individually contracted with Celera Genomics to access the data on sequences determined by the company.

1.3 Accomplishments in genome sequence analysis achieved by Celera Genomics and strategies for pursuing post genome research

In March 2001, Dr. J. Craig Venter, president of Celera Genomics at the time, visited Japan, and he commented as follows at his lecture, etc., about genome sequence analysis and strategies for pursuing post genome research:

(1) Accomplishments in genome sequence analysis

Celera Genomics, which was founded in 1998, has determined the sequences of genomes of the Drosophila (fruit fly), human beings, and mice (3 strains of mice including 129SvJ, DBA/2 and A/J). Currently, the company is pursuing the sequence analysis of canine and rat genomes. The company has been able to accomplish a succession of achievements in a short period of time against the backdrop of; i) the introduction of 300 units of a new automated sequencer (ABI3700), ii) development of software for the whole genome shotgun sequencing technique, and iii) the introduction of the latest computers. In other words, Celera Genomics has succeeded in climbing on the bandwagon in terms of both hardware and software.

(2) Current state of human genome sequence analysis

The evolutionary process can be traced in human genome. When viewing the sequences of the human genome, many regions with similar sequences can be seen on different chromosomes, indicating that they have evolved from a common ancestral chromosome.

In addition, while individual difference is noted in the sequences of some regions on the human genome, roughly 3 million single nucleotide polymorphisms (SNPs) have been found which are DNA sequence variations that occur when a single nucleotide (A, T, C or G) in the genome sequence is altered. Ultimately, 4 million SNPs will be found. Among them, 1% or less exists on coding regions of the human genome (genes), and only a fraction of those SNPs cause amino acid substitution. On the other hand, it is known that SNPs that occur in non-coding regions could predispose people to disease or influence their response to a drug, so such SNPs (SNPs in the regions not encoding proteins) are also of consequence.

(3) Strategies for pursuing post genome research

At the facility for protein analysis (Proteomics Factory), which has been established by Celera Genomics, one million samples of proteins can be structurally analyzed in a day with the use of mass spectrometers. Celera Genomics thinks that about 250 thousand kinds of proteins are expressed from roughly 30 thousand human genes by way of genetic transcription, translation and post-translational processing, and expects that causal relationships can be found between abnormalities of proteins and diseases. They have targeted cancer and aim to develop cancer diagnostic methods and cancer-specific vaccines.

According to the media, Celera Genomics has made a capital investment in a Japanese venture capital biotechnology company, which has SNPs analysis technology and aims to reveal, for example, the association between SNPs in Japanese and diseases.

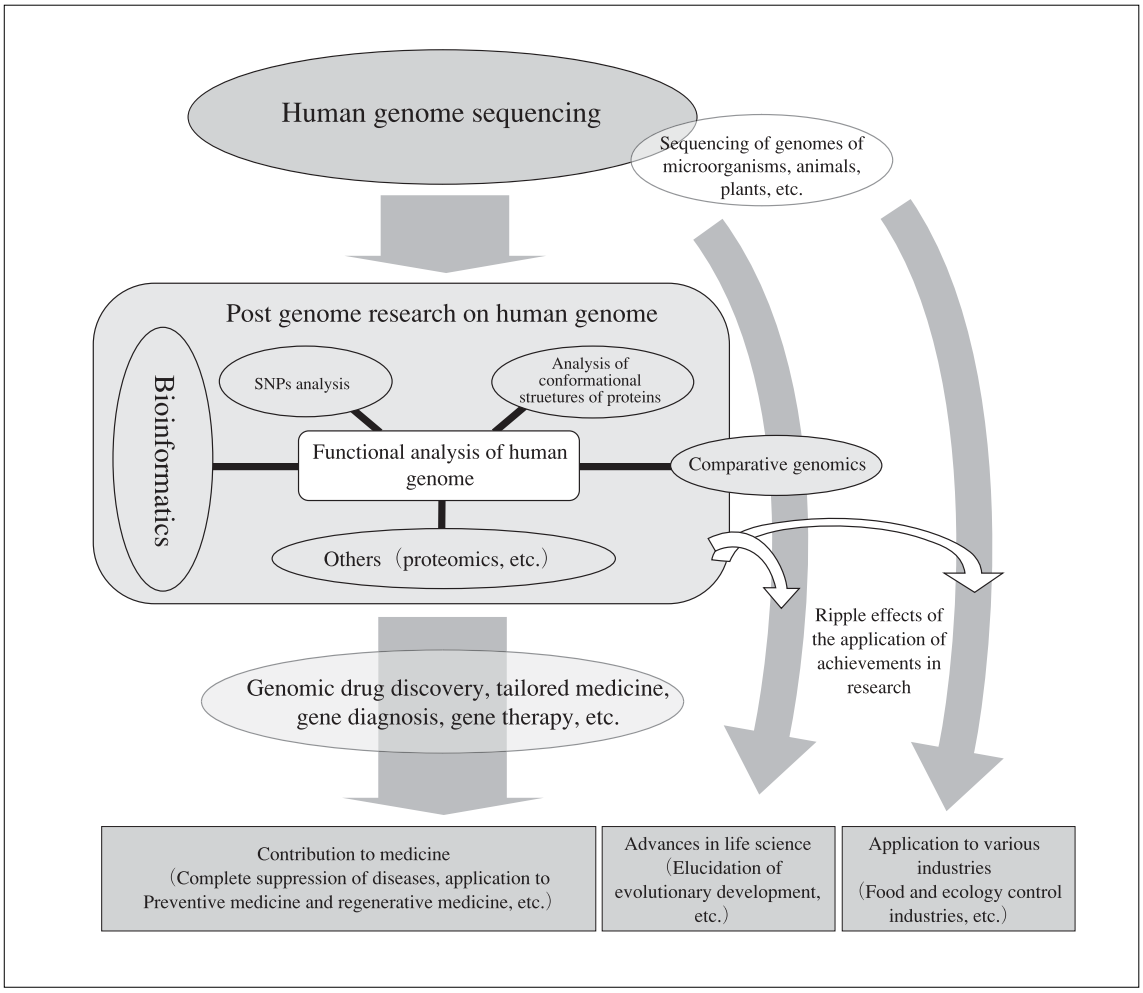
1.4 Trends in post genome research in Japan

Post genome research on human genome may have ripple effects on various fields in the form of, for example, contributions to medical care, facilitation of progress in advances in life science, contributions to various industries, etc. shown in Chart 3.

Such post genome research has come to be accepted as important in the Phase Two Technology Master Plan, which was adopted at the Cabinet meeting held in March 2001. The plan states that, among the areas of life science given priority to in research and development in Japan, post genome research should be given high priority and pursued strategically among others.

Future directions of post genome research in Japan are concretely charted in the report entitled "Promotion of strategies for pursuing post genome research" (report on a round-table conference concerning the strategic promotion of post genome research held by the Policy-making committee of the Technology Council) published in December 2000. The report states that the areas of post genome research that must be urgently pursued in Japan include the analysis of human genome variability and genes related to diseases, structural and functional analysis of proteins, bioinformatics, as well as functional analysis of genome. In addition, the report said that Japan will be able to move ahead of other countries in post genome research through urgent pursuit of such research by utilizing technologies in Japan that are superior to those in other countries such as the techniques for creating full-length cDNAs containing the complete genetic information of the mRNA encoding amino acids, which are used to form proteins (technique developed by Dr. Sugano et al. at the Institute of Medical Science, the University of Tokyo, and that developed by Dr. Hayashizaki et al. at the Institute of Physical and Chemical Research), and techniques for inducing the expression of proteins in the cell-free system (technique developed by Dr. Yokoyama et al. at the Institute of Physical and Chemical Research for efficiently inducing the expression of, and preparing the sample of many

Chart 3: Post genome research on human genome



Prepared by the Science and Technology Foresight Center

kinds of proteins), as well as by utilizing "facilities" in Japan that are better equipped than those in other countries including the large-scale nuclear magnetic resonance (NMR) park (Genomic Science Center, Institute of Physical and Chemical Research) and "SPring-8," a third-generation synchrotron radiation facility, at the Japan Synchrotron Radiation Research Institute (JASRI).

In the budget plan for fiscal year 2001, 98.3 billion yen was earmarked for the "strategies of promoting research and development in the areas of biotechnology with an eye toward life science of the 21st century" by the Ministry of Education, Culture, Sports, Science and Technology, the Ministry of Agriculture, Forestry and Fisheries, and the Ministry of Health, Labor and Welfare. Of the 98.3 billion yen, 60.6 billion yen (compared to 44.4 billion yen in fiscal year 2000) was allocated to; i) analysis of human genome structure and variability (structural and functional analysis of human full-length cDNAs, research on SNPs, etc.) and, ii) functional analysis of the human genome (structural and functional analysis of proteins, bioinformatics, etc.) to further facilitate such research.

Concerning post genome research pursued in Japan, this article describes trends in analysis of conformational structures of proteins, studies on SNPs found in Japanese as well as bioinformatics.

(1) Analysis of conformational structures of proteins

No methods, which are based on the sequences of the genome and are universally applicable, have been found for predicting the functions of proteins that are to be expressed from genes via transcription and translation. However, projects for analyzing conformational structures of proteins, which may be closely associated with their functions, are underway at institutions including the Institute of Physical and Chemical Research.

So far, conformational structures of 2,000 to 3,000 kinds of proteins have been elucidated, most of which have been clarified in the United States.

The international research team composed of researchers from Japan, the United States and European countries is planning to elucidate the conformational structures of 10,000 or more kinds of proteins within the next 5-10 years. Japan is aiming to complete the analysis of conformational structures of about 3,000 kinds of proteins under the initiative of the Institute of Physical and Chemical Research within the next five years. Moreover, the United States is targeting for the completion of conformational analysis of roughly 5,000 kinds of proteins within the next 5 years. While only a few percent of the whole sequences determined by the International Human Genome Sequencing Consortium was elucidated by Japanese researchers, it can be expected that a larger percentage will be represented by conformational structures of proteins determined in Japan among whole structures identified in the world.

In the First International Structural Genomics Meeting jointly held under the auspices of Japan, the United States and the United Kingdom in April 2001, discussion was held by the representatives of 10 countries on the conceptual framework of international cooperation in protein analysis. In the conference, it was agreed that situation reports on research should be made through common procedural steps based on the rule that every piece of information obtained in such analysis must be made public. Moreover, with regard to the timing of the publication of data on the conformational structures of proteins, it was agreed that such data must be made public within 6 months at the maximum after obtaining the relevant data, by giving consideration to the possible unfairness that may be caused due to the difference in patent application systems among countries and in view of the fact that such data may directly lead to future commercial applications such as the development of diagnostic methods and pharmaceutical products.

(2) Study on SNPs found in Japanese

Individual differences are seen in the sequences of many regions of the human genome. Such differences in base sequences are called genetic polymorphisms which include; i) SNPs, ii) insertion or deletion polymorphisms (polymorphisms due to the insertion or deletion in some regions of DNA), as well as iii) VNTR (variable number of tandem repeat) polymorphisms and microsatellite polymorphisms (polymorphisms caused by the difference in the number of repeat of a specific sequence made up of two to several tens of bases; those due to the differences of repeat number of sequences made up of a few to several tens of bases and 2-4 bases are called VNTR polymorphisms and microsatellite polymorphisms, respectively).

While VNTR and microsatellite polymorphisms can be found in several thousands and several tens of thousand regions on the genome, respectively, it has been said that SNPs are found in about as much as 4 million regions (according to Dr. J. Craig Venter, former president of Celera Genomics) or in as much as 3-10 million regions (Prof. Yusuke Nakamura, "Frontiers of Genetic Medicine," Yodosha Co., Ltd., 2000). In addition, since high-speed and high throughput SNPs analyzers are nearing practical use, SNPs may be more useful in exploring disease-related genes. Therefore, researchers assign a higher importance to SNPs as compared with other polymorphisms.

Under these circumstances, analysis of some of the SNPs found in Japanese was conducted at the Institute of Medical Science, University of Tokyo, and yielded results that suggest Japanese have descended from a relatively genetically isolated population, indicating the possibility that there might be disease-related SNPs specific to Japanese. The Institute of Medical Science, University of Tokyo, has compiled data on SNPs into a computer database and put it up for public view on the Internet. In addition, the Pharma SNP Consortium (organized by the Institute of Physical and Chemical Research, Tokyo Women's Medical University, as well as 43 pharmaceutical companies) is pursuing studies that may lead to the promotion of tailored medicine including the collection of DNA samples from a general population (about 1,000 healthy volunteers) for use in the analysis of polymorphisms of the 165 genes related to drug metabolizing enzymes.

It is expected that data on regions on the human

genome in which individual differences are seen in base sequences will be the key to future individualized medical care, i.e., medical care designed for each individual to offer maximum therapeutic benefits in light of accurately analyzing his/her genetic predisposition. On the other hand, ethical problems have loomed about how to handle individual genetic information. In March 2001, the Ministry of Education, Culture, Sports, Science and Technology, the Ministry of Health, Labor and Welfare, and the Ministry of Economy, Trade and Industry jointly drew up the "Ethical Guidelines concerning the Analysis of Human Genome and Genes," which are the concrete guidelines for ethical aspects of research on human genome and genes in general. In the guidelines, responsibilities to be assumed by researchers and institutions included the pursuit of genetic studies after gaining written informed consent from the candidate suppliers of DNA samples following detailed explanation of, for example, the purposes of the relevant studies as well as the guaranteed protection of privacy in terms of genetic information. Moreover, 8 genetic medicine-related societies in Japan including the Japan Society of Human Genetics have jointly drawn up draft guidelines as to the way human genetic tests should be carried on.

(3) Bioinformatics

Rapid progress in human genome analysis has created a need for processing enormous amounts of data on base sequences. In order to accelerate the progress in post genome research from now, it may be essential to develop software that enables us to not only manipulate data on base sequences but also to predict the structures, functions, metabolic pathways, etc., of proteins based on amino acid sequences. Therefore, further facilitation of studies in the field of bioinformatics, in which databases of information that may form the foundation of post genome research are constructed to associate the extensive data to one another, is thought to be an urgent necessity. In particular, development of human resources, such as specialists who are experts in both biology and information science, has long been considered to be a matter of high priority.

One example of specific efforts made by the Japanese government is that bioinformatics was selected as one of the fields (two fields in total were selected) from among the research fields proposed by the general public as those for which monetary support for the development of human resources must be provided from the Funds for the Coordination of Advancement of Technology for fiscal year 2001.

In addition, one example of a recent noteworthy movement is that a seminar, "Information Biology Teki-juku (training course for people with Aptitude)" (Director: Dr. Kenichi Matsubara, professor at Nara Institute of Science and Technology), was held during the period from late March to early April 2001 at the International Institute for Advanced Studies, under the sponsorship of the Ministry of Education, Culture, Sports, Science and Technology, with the aim of training professionals on information technology necessary for studies in the field of life science who can then be trailblazers and explore new areas. These realities suggest that researchers also recognize the significance of bioinformatics.

In the United States, biology is compulsory for all students regardless of their specialties at some universities, which would give a clue in devising strategies for promoting the progress in bioinformatics.

1.5 Conclusion —Future prospects of the movement surrounding human genome sequences

(1) Human genome sequencing by the International Human Genome Sequencing Consortium is currently in progress

Currently, human genome sequencing is being pursued by the Finishing Group (composed of researchers from the United States, the United Kingdom, Japan and France) of the International Human Genome Sequencing Consortium, with the aim to complete the sequencing of the whole human genome with high accuracy. In the Finishing Group, there are coordinating centers that are responsible for the sequencing of specific chromosomes as well as participating centers that are to cooperate with sequencing chromosomes where possible.

In Japan, the Institute of Physical and Chemical Research, a coordinating center, is pursuing the sequencing of chromosomes 11 and 18 in collaboration with the Whitehead Institute, Massachusetts Institute of Technology. On the other hand, Keio University, a participating center, is engaged in the sequencing of chromosome 8.

(2) Applicability of genome research to disease suppression

In order to apply the achievements in genome research to disease suppression, it is necessary to analyze such data on genome sequences as those on SNPs in association with data on clinical characteristics. The association method is one example of such analytical methods with which candidate disease-related genes will be identified through comparing gene sequences of each population of several hundred to several thousand people affected by a specific disease with those of each population of several hundred to several thousand people free of the disease. For such analysis, blood samples must be collected from a relatively great many volunteers.

Japan lags behind some other countries in the establishment of foundations on which to conduct studies using such analytical method mentioned above. It is difficult to collect blood samples from a great many people at the level of each individual institution and, therefore, some researchers obtain such blood samples from abroad.

From here forward, in order to build up a solid foundation on which to do post genome research in Japan, it is urgently required to establish resource centers where samples and clinical data collected from many people, while following specified procedures such as gaining informed consent, are stored and managed at a national level.

(3) Promotion of post genome research

In recent post genome research, the United States has taken the international initiative, against the backdrop of the aggressive establishment of venture capital biotechnology companies, to develop strategies for obtaining and protecting U.S. patents. For example, about 1,500 venture capital biotechnology companies were reported to be established in the United States in 2000, in contrast to about 150 companies in Japan (Masamichi Ooishi, "Workings of Human Genome," Nippon Jitsugyo Publishing Co., Ltd., 2001).

Nevertheless, in Japan too, such various projects as those mentioned under Section 1.4 of this article have been prepared or launched under the initiative of government. Private companies including pharmaceutical companies are also hurriedly promoting larger investments in research targeting at genomic drug discovery, and forming alliances for the sake of improving their competitiveness in the field of bioinformatics as well as promoting technical cooperation. Furthermore, joint research projects between such public institutions as the Institute of Physical and Chemical Research and private companies have been pursued, raising expectations that the achievements in publicly funded studies will lead to future commercial applications.

Post genome research will have an enormous impact on Japan's future life and economy, and will be deeply involved in various aspects of society, primarily concerning the ethical aspects. Therefore, it may be required that research funds and manpower be provided, while flexibly responding to changes in the social climate surrounding post genome research as well as in the environment for such research in Japan and abroad, for its acceleration.

Commentary

Percentage of base sequences determined by Japanese among whole sequences analyzed by the International Human Genome Sequencing Consortium

It has been frequently reported by the media, etc., that the rate of Japan's contribution to human genome sequencing pursued by the International Human Genome Sequencing Consortium is about 6%. This value was calculated based on the percentage (5.59%) of the total of sequences (3,934,884 kb) determined by the Institute of Physical and Chemical Research (188,056 kb), Keio University (20,105 kb), and Tokai University together with two societies for the study of cancer (11,783 kb in total) in the whole draft sequences determined by the Consortium.

As of February 2001, when the article contributed by the International Human Genome Sequencing Consortium appeared in the journal Nature, as much as 4,338,224 kb were contained in the whole draft sequences, far exceeding (about 1.4 times) the estimated number of bases making up the human genome. Such increase in bases in draft sequences is due in part to the fact that institutions in countries other than Japan entered the data on the sequences of the regions not assigned to them, which they determined in order to explore SNPs entered into databases.

The values in Chart 2 are not the numbers of bases in the draft sequences, but those in finished human sequences.