9 Research and Study on Desired Search Methods in the Biotechnology Field

The drastic advancements of gene analysis technology in the field of biotechnology in recent years have brought about an increase in the number of related patent applications as well as an increase in the amount of DNA sequence data contained in each patent application. These increases are having an effect on the sequence search system of the Japan Patent Office (JPO). Under such circumstances, it has become necessary to review the situation of the sequence search system and the manner in which the sequence data is accumulated, while also attempting to improve the search response of the sequence search system. With this in mind, this research and study involved consideration of a next-generation sequence search system while examining such matters as the optimum form of searching in the field of biotechnology that is suitable for patent examinations from various perspectives, including (1) study of the present situation and problems with the sequence search system along with tentative countermeasures, (2) study of data accumulation, (3) study of the effectiveness of compression of sequence data and actual demonstration testing of the data compressions, and (4) the opinions and requests to the JPO.

I Present Situation and Problems of DNA Sequence Databases

1 Database in the Japan Patent Office

The current DNA search system used in the Japan Patent Office(JPO) was constructed in 1998. However, in consideration of the current situation in which the number of patent applications has increased and the amount of DNA sequence data accumulated within the JPO has also increased due to the following development of DNA sequence analysis technology and the diversification of the subjects of analysis, it has become necessary to examine how proper accommodations should be made from various perspectives.

(1) Present Situation of Examination of Patent Applications Relating to DNA Sequences

Examination of patent applications relating to DNA sequences are different from that of patent application for typical technologies in that the sequence data in the patent application is compared with sequence data described in the prior art by means of a homology search. In examinations of patent applications relating to DNA sequences that include unpublished sequence data, all of the search format, search results and other findings have to be handled with confidentiality, since sequence data itself contained in the search format of the homology search consists of the substantial part of invention.

(2) Present Situation of the DNA Search System in the JPO

The DNA search system in the JPO is broadly divided into search software, hardware including the CPU and hard disk, and sequence data serving as the target of the search.

Typical search software such as FASTA, BLAST and Smith & Waterman is used for the search software. This is because in the prior art search, it is important to determine whether or not a sequence can be found in the homology search using ordinary search software, rather than special search software. Although hardware is periodically enhanced based on predictions of increases in the number of applications, amount of data, usage frequency and other factors, hardware enhancement is no longer able to keep up with increases that are exceeding the predictions for the amount of sequence data in recent years. The sequence data include sequences described in patent applications, and public sequences being registered with registration organizations.

(3) Problems with the DNA Search System in the JPO and Tentative Countermeasures

(i) Negative Impact on the System and Examinations

Due to the sudden increase in data relating to DNA sequences filed as patent applications in recent years, there are concerns over a shortage of database capacity in the JPO. In addition, negative impact on examinations is also beginning to appear, including the potential for longer search times.

(ii) Tentative Countermeasures

① NR (Non-Redundant) processing by public databases (GenBank, DDBJ, EMBL)

It has become possible to secure available capacity and shorten search times of the database in the JPO by eliminating duplication of public databases based on GenBank since the end of 2000.

② Simplification of keyword search functions

(deletion of index data)

③ Extension of system operating hours

Even though the above tentative countermeasures have been implemented, it is clear that there will be shortage of database capacity again in the near future. It is therefore necessary to proceed with considerations from more radical viewpoints such as data compression or database outsourcing.

2 Public Databases

(1) National Institute of Genetics (NIG)

(i) DNA Data Bank of Japan (DDBJ)

Data directly registered in DDBJ is evaluated (evaluation of data format, data notation and fundamental biological significance) and accumulated, and is then made available to public after being integrated with data sent daily from EMBL and GenBank and with data from the JPO.

(ii) Task for the future

① Ideal situation of public databases

Ideal situation of public databases can be characterized by the following:

* Long-term, stable operation
* Function as a system for public release of private databases
* Standardization of data format and notation
* Efforts to verify compatibility between data from different sources
* Presentation of a guideline for development of a medical-biological database, and promoting technology transfer

② Tasks of DDBJ

* Establishment of an information environment

It is necessary to enhance the computer system corresponding to the growing size of the nucleotide sequence database. In addition, since there is expected to be an increase in the demand for integrated databases with a backbone database such as nucleotide sequence database and databases distributed on networks, and the distribution of image data originating in new experimental technologies, it is necessary for leading public databases, including DDBJ, to connect to broadband networks.

* Securing of personnel

Personnel having capabilities equivalent to Ph.D. in the fields of medicine and biology are essential for the course of a key project such as the construction of database, and it will be necessary to give ample considerations to compensation and evaluation in order to secure such personnel.

③ Specific tasks of DDBJ

Examples of specific task of DDBJs include the following:

* Long-term stable operation consisting of data reception, evaluation and public disclosure
* Prompt data updating
* Deployment from current basic evaluation to genuine annotation
* Efforts to develop secondary databases through processing the primary nucleotide sequence database
* Study of archiving of new types of data other than nucleotide sequence data
* Integrated utilization of medical-biological data and deployment of analytical tools

④ Accommodation of outsourcing

Since DDBJ is widely accessible to the public, in case users require anonymity with respect to their use, it will be possible to make accommodations through the use of private ISDN lines, SSL and user authorization.

(2) The Institute of Medical Science, the University of Tokyo HGC (Human Genome Center)

(i) Present situation and problems

The database of the Institute of Medical Science, the University of Tokyo contains data from both internal and worldwide public databases, and is available to the public. The problems, such as a shortage of database capacity and delays in service programs occurred due to the growing amount of data so that recording disks are being added, network services are being enhanced and reliability is being improved. In the future, a large-scale center and a powerful network accessing that center will be required.

(ii) Accommodation of outsourcing

The greatest problem with respect to outsourcing is protection of security. This is being accommodated through the introduction of network monitoring systems and access control.

(iii) Database topics

① Easy-to-use and reliable database

In order to create a database that is both easy to use and reliable, it must provide a fast response to searches, the latest information, the establishment of a relationship between sequences and three-dimensional structural data for new types of data, interaction between organic molecules and expression information, and in terms of data quality, a database that guarantees the quality and uniformity of data from database service side.

② Development status of DNA search system

HGC supports development of a system that combines WWW search engine technology and associative information search technology, as known as BACE by the Hitachi Central Research Laboratory Group, Hitachi, Ltd.

③ Future image of DNA databases

In the near future, such a database will be required that is able to comprehensively handle various databases, an example of which is a database that can efficiently carry out comparative

genome research. It is expected that research will be actively conducted on simulation of biological phenomena using a database that gathers knowledge relating to intermolecular interactions and other related subject matter.

(3) Institute for Protein Research, Osaka University

(i) Present situation and problems of the protein three-dimensional structure database PDB(*1) (Protein Data Bank)

The PDB database is accessible free of charge throughout the world over the Internet.

① Administrative problems

The personnel, computer, software and other costs required for PDB administration are not adequate. Activities are required to improve the database and create added value in peripheral areas of the PDB.

② Functional problems

Instead of simply accumulating primary data, it is necessary to enhance database services such as by structuring the database to enable linkage with other databases, and by providing algorithms for predicting the similarity or active sites of other proteins besides key word searches. In addition, in the field of education, a simpler search system is desired. In the near future, it will be necessary to improve the efficiency and reliability of entry editing software in preparation for increases in the rate of growth of PDB entries accompanying the advancement of structural genome projects.

(ii) Accommodation of outsourcing

The current database maintenance system does not allow them to undertake outsourcing and will require a considerable increase in personnel and specialists in the field of data processing. In addition, since the implementation of a risk management system and the establishment of a private line are difficult to realize at universities, significant changes will be necessary in order to undertake outsourcing, and this problem is considered to be extremely challenging.

3 Private Databases

(1) Fujisawa Pharmaceutical Co., Ltd.

(i) Reason for requiring an in-house gene database system

The reason is that there are cases in which analysis of nonpublic sequence information and sequence information from joint research partners or contract partners is prohibited on the Internet to avoid leakage of that information.

(ii) Problems and countermeasures

Since there are limits in terms of costs and personnel on the installation of giant databases and analytical systems within a private company, they are beginning to consider the use of data centers providing outsourcing services.

(2) Other Typical Private Databases

(i) Reason for constructing in-house databases

The reason for private companies to construct in-house databases is that it was judged that when private companies use public databases for searching, there is risk caused by leakage of research information when external public database servers is accessed through the Internet and it is greater than the cost of constructing an in-house database. However, this does not mean that all high-tech related corporations have constructed their own in-house databases.

(ii) Problems and countermeasures

Under the present circumstances, there is no public site provided with security technology over the Internet. Thus, there is no guarantee for ensuring a situation in which there is no risk of leakage of research information. Several private companies are engaged in outsourcing businesses for biotechnology database services. There is a high probability that such services will proliferate if the costs associated with providing outsourcing services are reduced.

4 Situation of DNA Sequence Databases at Overseas Patent Offices

(1) USPTO (United States Patent and Trademark Office)

(i) Database

The USPTO has constructed three types of databases consisting of the CRF ISSUED, CRF PENDING and CRF PUBLISHED, in which only information relating to applications is recorded. The USPTO also has an amino acid sequence database containing PIR, GENSEQ, SWISS-PROT and TREMBL data, and a nucleotide sequence database containing GENBANK, GENBANK -NEW, GENSEQ and EMBL data. In addition, commercial databases such as CAS, Questel-Orbit, LEXIS and NEXIS for chemical structure and literature searches can be connected with the USTPO by telephone line.

(ii) Sequence searching and analysis

The USPTO is currently using Smith & Waterman and considering the use of BLAST2. In addition to Smith & Waterman, it is also using GCG Word Search and the multiple alignment

(*1) The database of the Institute for Protein Research, Osaka University is an official Asia-Oceania region archive of the database being managed by a consortium called RCSB(Research Collaboratory for Structural Bioinformatics) consists of the group of Rutgers University, U.S.A., and of the University of California, San Diego, U.S.A.; URL (http://pdb.protein.osaka-u.ac.jp/pdb/).

program, Gen Align (included in GCG). It is also currently examining automation of manual processing required for long or large numbers of sequences.

(iii) Hardware and personnel

A dedicated board, Compugen Bio XL/P4 is used for homology searches.

12 full-time searchers are requested to conduct searches. There are also five examiners who conduct searches on their own (there are a total of about 1,600 USTPO examiners and 500 of them are in charge of biotechnology fields).

(iv) Future outlook

The USTPO is currently testing three companies (CELERA, COMPUGEN and Bisant (an MIT joint venture business)) in order to examine the feasibility of outsourcing. It is using SSL(*2) or VPN(*3) for security, and employs user authorization and a firewall. In addition, it is also currently examining the use of a private T1 line.

However, a database of patent applications is constructed and maintained within the USTPO. Backup system required for managing sequence searches is also intended to maintain within the USTPO. Consequently, on a short-term basis, the USTPO is scheduled to enhance its hardware. On a long-term basis, it is expecting MPSRCH software, including cutting back on costs.

(2) EPO (European Patent Office)

(i) System requirements

The EPO recognizes the need for a search system that covers all public databases, patent databases and "membership" databases available for third party use (e.g., CELERA and Incyte).

(ii) Examination system

Sequence data is contained in a dedicated database. The EPO is switching from the STRAND system constructed within the EPO to the EBI-EPO system that uses the nucleotide sequence database of EMBL (European Molecular Biology Laboratory). GenSeq is installed at EMBL and the EPO is commissioning EMBL with the development of an interface exclusive for the EPO. In case of referring to PubMed and other references based on the results of a homology search, seamless operation is attained by switching from private line connection to the Internet connection.

Actual use of the EBI-EPO system involves roughly 100 examiners and is accessed roughly 20,000 times annually.

Functional improvements include the goal of storing the results of homology searches (for a minimum of 2 years and ideally for 25 years), and the EPO recognizes the need for evaluation of the homology search program and knowledge regarding setting of parameters.

The EPO is also engaged in standardization of examination methods (for example, setting parameters) and implementation of training for accommodation of new technology. It is also considering entrusting with EBI.

With respect to the relationship with industry and academia, meetings are held once a year with major patent applicants along with workshops. The EPO examiners and corporate applicants are able to access the same databases.

(iv) Future prospects

The EPO is aiming to create a system that not only will provide homology search functions, but that will also allow comprehensive referral of all types of related literatures and other information. In addition, it is also planning the development of a validation system that can be used by applicants filing with the EPO.

II Outsourcing of the DNA Sequence Databases in the JPO

When considering next-generation DNA search systems, three countermeasures have been proposed regarding the urgent issue of dealing with the accumulation of the increasing amount of sequence data.

Proposal 1: Enhance hardware (hard disks) and continue to accumulate data within the JPO.

Proposal 2: Utilize hardware at maximum efficiency through data compression and continue to accumulate data within the JPO.

Proposal 3: Minimize accumulation of data within the JPO, and instead outsource the majority of that data.

Here, although a preliminary study was conducted on the feasibility of the third proposal, this study was not conducted for the purpose of determining which of proposals 1 through 3 should be employed for constructing a next-generation DNA search system. This matter should be examined carefully in the future.

1 Need for Outsourcing

In terms of both accumulation and search capabilities, the burden of enhancing the DNA search system in the JPO is considerable. On the other hand, DNA sequence data in patent applicationsis accumulated in duplicate in public databases and the DNA search system in the JPO after patent applications are published. Database outsourcing is one way of dealing with this.

(*2) Secure Socket Layer.
(*3) Virtual Private Network.

2 Prospects with Outsourcing

(1) Problems in Publication of Patent Applications

Generally, patent applications are not opened to the public for 18 months after filing. Information before the publication is also subject to the prior art search, and in particular, since DNA sequence data itself represents the substantial part of an invention, it requires to be handled with caution in the form of confidential information.

(2) Problems with Data Accumulation of Information before Publication by Outsourcing

Examples of problems associated with data accumulation of information before publication by outsourcing include: ① physical and network maintenance of the server accumulating data, ② obligation of outsourcing personnel to keep confidentiality, ③ penalties for leaking confidential information and impairing data, ④ work for transition from information before publication to information after publication at the time of publication, ⑤ guarantee of search capabilities, ⑥ maintaining confidentiality when searching, and ⑦ organization having a high degree of public visibility that is not affected by specific individuals.

(3) Communication Security

Since homology searches use a search formula that includes unpublished DNA sequence information thereby resulting in the substantial part of invention itself to be carried along communication lines, in the case of examination before the application is published, it is necessary to ensure that strict security measures are implemented for that sequence data.

(4) Search Security

In the case of DNA sequence searches, since the search formula itself represents the substantial part of invention, the search formula must not be retained by the outsourcing service. There is no other way of accomplishing this other than by guaranteeing that logs are not stored at the outsourcing service in accordance with the terms of a contract and so forth.

(5) Others

Data compression technology offers an alternative to outsourcing as countermeasures of handling the growing amount of DNA sequence data. Combining data compression technology and outsourcing not only enables data accumulation to be performed more efficiently, but also it is expected to offer effects such as improved search efficiency in the form of communication by compressed data and searching compressed data.

3 Specific Outsourcing Methods

(1) Study Elements

The following indicates examples of elements to be considered in feasibility study on outsourcing: ① large amount of data as much as possible, ② convenience during searching, ③ absence of obstructions in search response, ④ reliable data management, ⑤ ability to shift data from a database before publication to a database after publication at the time of publication, ⑥ ability to send a search formula to two locations as well as correlate search results in case databases are found at two locations, ⑦ ability to standardize the data format, and ⑧ easy and reliable security measures.

(2) Proposals for Outsourcing Methods

(i) Formulated basic policy

① Maintaining present functions

② Ensuring the present level of security

③ Reduction of amount of data retained in the JPO as much as possible

(ii) Database Classification (Table 1)

Table 1

Database classification		Database contents
Database before publication		Patent information before publication
Database after publication	A	Sequence information
	B	Keyword information
	C	Flat files (containing all data)

Note: "Database" includes GenBank, EMBL, DDBJ, SWISS-PROT and PIR.
All information before publication is premised on being retained within the JPO from the viewpoint of maintaining security.

(iii) Security measures

① Network security: Use of a private line and encrypted communications

② Server security

(a) Deletion of task query information, search logs and other history information

(b) Server protection: Restricted access to server installation site, authorization for room access, and prohibition of access to the server from the outside

(c) Unauthorized access monitoring function: Function required for monitoring unauthorized access to the server in case of using the Internet

③ User validation: Required in case of accessing the server of an outsourcing service involving client validation by SSL (a system that only allows access to clients who have been installed with a specific disclosure key

validation statement)

(iv) Proposals for database configuration

Data with potentiality to be configured at outsourcing services consists of sequence information a, keyword information b and flat files c contained in a database after publication.

① Database configuration proposal 1

Databases after publication a, b and c are all outsourced.

② Database configuration proposal 2

Only keyword information b and flat files c of databases after publication are outsourced.

Since sequences searching itself is conducted within the JPO, the query formulae of sequence searches are not disclosed to the outside. Although there is also the characteristic of not leaking search logs to the outside, keyword search queries are disclosed to the outsourcing service.

③ Database configuration proposal 3

Only flat files of databases after publication are outsourced.

In this proposal, since sequence and keyword searches themselves are performed within the JPO, search queries are not disclosed to the outside. Search logs are also not disclosed to the outside.

(3) Study of Problems with Outsourcing

The results of conducting a study on the various problems associated with outsourcing are indicated below.

① Problems with publication of patent application and data accumulation

There is a considerable burden to introduce high level of security due to the obligation of secrecy of handlers of information before publication. The task of shifting information before publication to information after publication at the time of publication is not a significant problem by accommodations such that information before publication is set to be undisclosed until a specified date for each of the databases. In case that information before publication is contained within the JPO while information after publication is contained outside the JPO, a system will be required that enables their respective searches to be performed smoothly.

② Communication security

A private line is used in combination with encrypted communications.

③ Search security

No storage of search formula logs is guaranteed by the outsourcing service according to the terms of a contract and so forth.

④ Others

The application of data compression technology will be examined in consideration of the amount and contents of data remaining in the system within the JPO.

⑤ Security levels of individual data types

* Flat data

Accumulation of data before publication requires sophisticated security. Flat data before publication is preferably accumulated within the JPO, while flat data after publication can be outsourced for a considerable reduction in the amount of data.

* Index data

Although index data is the key to invention and data before publication requires security, data after publication only requires a security level similar to that of flat data, and since the amount of data is large, data after publication is preferably outsourced.

* Sequence data

Since sequence data before publication require the highest level of security, it will be accumulated within the JPO. It is not desirable to send the search formulae of sequence data before publication to outsourcing services for a search before publication. The accumulation of sequence data by dividing into the data before and after publication results in a considerable burden such as the construction of a new system.

On the basis of the above, reliability and performance of searching can be improved by accumulating all sequence data within the JPO, and performing sequence searches within the JPO only. Accordingly, database configuration proposal 2, calling for only outsourcing keyword information b and flat files c of databases after publication, is thought to allow a considerable reduction in the amount of data, is adequately effective for handling the growing amount of data, and provides adequate security measures.

4 Opinions and Requests Relating to Outsourcing

(1) Security

(i) Because of the possibility of leaking information through internal personnel engaged in system operation, a strict information management is desired by the outsourcing service. In order to ensure reliability, the exclusion of data before publication from outsourcing should also be considered.

(ii) Outsourcing offers the advantage of considerable system cost reductions within the JPO. In addition, providing examiners with the latest user environments and customization will lead to increased accuracy and faster examinations, thereby also resulting in benefits for patent applicants. If outsourcing is provided with adequate security measures, it is evaluated to be worth deploying positively.

(iii) Ensuring security

There are two types of security: ① security

during data transfer between the outsourcing service and the JPO, and ② data security in the outsourcing service. Current technology is adequate for ensuring a certain level of security under certain conditions. The only system that is applying these technologies in the field of biotechnology is the JBIC (Japan Bio Information Consortium) system. The JBIC system adopts the following approach to security:

(a) protection of data on the Internet is provided by SSL encryption;

(b) all actual sequence data in the server is processed in internal process memory (internal process memory refers to memory that is dynamically allocated when required and cannot be accessed from other locations); and,

(c) temporary files containing sequences in the server are encrypted as necessary.

(2) Outsourcing Services

Public agencies are preferable as outsourcing services. In case of using a competing corporation or corporation affiliated with a competing corporation as an outsourcing service, there is increased concern over the possible information leakage as compared with a public agency.

(3) Others

① Although the entity that receives outsourcing data guarantees a certain degree of security, since it cannot be considered to be completely safe, it is thought that a considerable number of exclusions will be incorporated in contracts. The party requesting outsourcing service will be required to provide an adequately convincing explanation that these exclusions will not cause problems for patent applicants.

② When considering outsourcing, it is necessary to make a comparative study from the aspects of personnel, hardware, budget and so forth between the burden in case of maintaining and enhancing the system within the JPO and the burden in case of performing outsourcing.

Ⅲ Compression of Sequence Data

1 Compression and Data Types

A study was made of data compression premised on increasing searching speed using technology available at present. The result is shown below with regard to the respective types of data together with the effectiveness of data compression.

(1) Sequence Data (Ineffective)

If the number of character types to be used are limited to four of a, t, c and g, a single character can be represented with 2 bits. It is difficult to significantly increase the compression rate beyond this level. In addition, although "compression methods based on the finite property of the genome" are effective techniques in case all sequence data is available, effectiveness at the present time is uncertain. In case it is necessary to use a plurality of existing search programs that employ their own file formats, there is hardly any advantage of compressing sequence data independently under the present circumstances.

(2) Index Data (Ineffective)

Index data refers to an index structure used for searching for the location of a pattern character sequence P in flat data T, and is able to achieve a considerable increase in searching speed. It is necessary to effectively install this data by selecting a search structure according to the target search method. In case the index data is constructed by a program selected and installed according to the search objective, attempts to compress it from the outside with a different method are meaningless. Instead, the required search specifications should be decided, and a suitable index structure should be selected and installed.

(3) Flat Data (Effective)

The method of saving files following compression and uncompressing those files in response to a request such as indication of results and so forth is considered as effective from the viewpoint of saving on disk capacity. If a compression method is used that allows rapid uncompressing, practical performance can be adequately obtained. However, in case of requiring only partial uncompressing, it is necessary to employ a suitable compression method. If file size is reduced by using suitable compression, searching speed can be expected to increase.

(4) Summary

Research trends adopting the approach of perceiving compression and searching as an integrated entity have recently appeared on the scene, and the situation is changing rapidly. Thus, it will be necessary to continue to carefully examine compression of sequence data and so forth in the future even it is judged to be ineffective at the present time.

2 Demonstration Testing

Since most of the nucleotide sequence data added after decoding genome nucleotide sequences is duplicate data, it is expected that this portion of the data can be compressed to a high degree. In this demonstration, when the change in compression efficiency was investigated for nucleotide sequence data beyond the size of the genome, the trend was observed in which additional data following genome decoding was

able to be compressed to 1/20 (0.4 bits per base) or less of its original size.

(1) Compression of Nucleotide Sequence Data

(i) Conventional General-Purpose Compression Methods

① compress: A LZW(*4) dictionary-based compression method is faithfully installed.

② gzip: This method is based on LZ775(*5) and generally able to realize a higher compression rate than ①.

③ gzip.9: It refers to the addition of the -9 option to gzip that provides a mode offering optimum compression.

(ii) Conventional compression methods dedicated to nucleotide sequence

① Use of references

In case of later finding a nucleotide sequence portion identical to a nucleotide sequence that has been previously registered, this method represents the latter sequence with data that represents a reference relationship with the former. Here, all of the sequence data that can be used as a reference destination is referred to as a "dictionary".

(a) Reuse of matching portions

There are a method of referring only to portions that completely match (complete match), and a method of referring by adding difference information to portions that approximately match (approximate match).

(b) Reuse of complementary portion (Reverse Complement)

For example, "AATTGCGC" and "GCGCAATT" are in the relationship of mutually complementary sequences. In case that one of the sequences is already in the dictionary, that sequence is referred to as a complementary sequence to represent the other sequence.

② Use of encoding

In this method, A, T, G and C having high incidences are represented with short bit strings, while symbols such as N that are rarely used are represented with long bit strings. A well-known example of this method is Huffman coding(*6). In addition, Context Tree Weighting, in which the encoding method is not fixed, but rather predicts the next character to appear according to the context of the nucleotide sequence so as to optimize the code, is also effective for compressing nucleotide sequences.

③ formatdb is a program attached to BLAST, and generates a compressed database that is directly searchable.

④ Even the best compression methods currently available have a compression rate of about 1.74 bits per base on average.

It is reported that a method combining "encoding based on Context Tree Weighting" and the "use of references" allows compression of nucleotide sequences at the highest compression rate(*7). If uncompressed data consists of 1 byte per base, data size after compression is reduced to about 21-22% of the uncompressed.

(iii) Conditions required for nucleotide sequence data compression methods in this demonstration work

① The data structure must be similar to that which allows homology searches even after compression.

② Compression capability must increase accompanying increases in sequence data.

③ Compression must also be effective for factors causing increases in the number of sequences due to addition of annotation.

It should also be taken into consideration that it may be necessary to extract partial sequences contained in ORIGIN (the item describing the entire sequence contained in a single record in the data records of disclosed nucleotide sequences) and include them in the database as independent sequences.

④ There must be no data overflow even if all known nucleotide sequences are maintained.

(iv) Improvements added in this demonstration work

Improvements are added so that the finite nature of the genome, which is one of its biological characteristics, is used as the principle of compression.

① Genome finite nature (based on the unique properties of DNA sequences)

With the exception of the case of artificially creating random nucleotide sequences, the nucleotide sequences that are handled on a routine basis are fragments that have been copied from a portion of a nucleotide sequence that exists in the genome. Although the total amount of nucleotide sequence data increases as compared with performing sequencing, there are numerous mutually duplicate portions between

(*4) Welch, T. "A Technique for High-Performance Data Compression", IEEE Computer, Vol. 17, No. 6, pp.8-19 Jun.1984.

(*5) Ziv, J. and Lempel, A. "A universal algorithm for sequential data compression", IEEE Transactions on Information Theory, Vol. 23, No. 3, pp.337-343, May 1977.

(*6) Huffman, D. A., "A method for the construction of minimm-redundancy codes", Proc. of the IRE, Vol. 40, No. 9, pp. 1098-1101, Sep. 1952

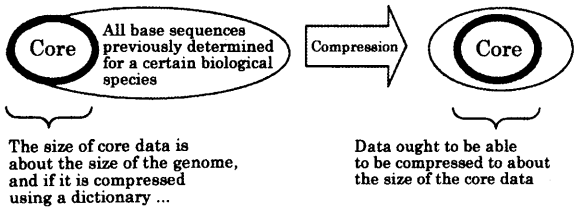
(*7) Matsumoto, T., Sadakane, K. and Imai, H., "Bioloical Sequence Compression Algorithms", Genome Inforatics, 11: 43-52, Dec. 2000.

the resulting sequence data, thus resulting in a limit on the total amount of unique fragments. This is because the length of the genome is finite and even if additional data is obtained, only partially duplicate data increases. The use of the shotgun method that applies this principle has been used in recent years to successfully decode the entire human genome. This principle has also been applied to data compression methods.

② Expanding dictionary size to the entire width of the genome

According to the finite nature of the genome, a nucleotide sequence obtained from an organism represents the information of a fragment that has copied a portion of the nucleotide sequence present in the genome, and the minimum amount of required information is roughly equal to the size of the genome. If this is used as a dictionary, since nearly all portions of nucleotide sequences can be represented by referring to that dictionary, a high compression rate can be expected to be achieved (see Fig. 1). This compression method is characterized by expanding the size of the dictionary to that genome size or more.

Fig. 1



③ Compression method focusing only on "dedicated search files created with formatdb"

The compression program developed in this demonstration further compressed nucleotide sequence data for BLAST searching generated by formatdb.

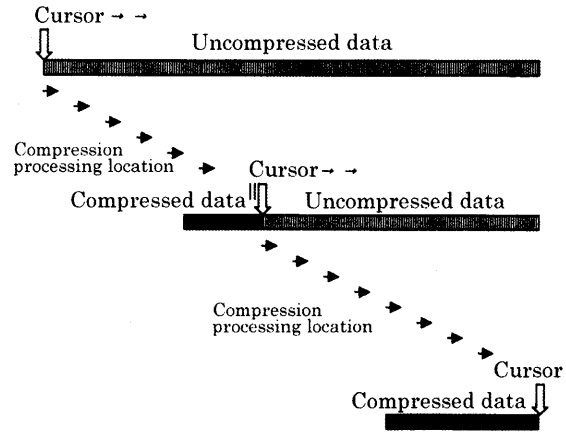
Nucleotide sequence data (1 base = 1 byte)
↓ formatdb
Data for BLAST (1 base = 2 bits)
↓ Compression method by this research
Highly compressed data
↓ Simultaneous execution of uncompressing and BLAST searching
Data for BLAST

④ Verification of compression effects by referring to a genome-wide dictionary

In order to accurately verify the compression effects obtained by using the finite nature of the genome for the compression principle, a program was created only with the method "reference to a genome-wide dictionary" followed by measurement of its compression effects. A summary of the program is explained below.

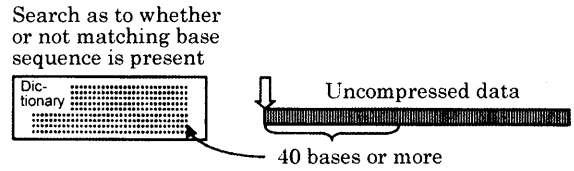
Step (a) As shown in Fig. 2, the processing location during compression is referred to as a cursor. In the compression method of this research, since compression is processed every 2 bit (1 base) from the start of the file, the cursor advances two bits at a time.

Fig. 2



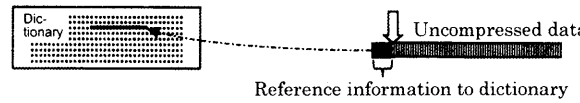
Step (b) As shown in Fig. 3, a search is performed as to whether or not a sequence exists in the dictionary that matches the nucleotide sequence fragment of at least 40 bases from the cursor (while also considering complementary sequences).

Fig. 3



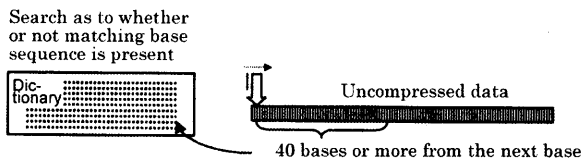
Step (c) As shown in Fig. 4, in case a matching sequence is found, the reference information is replaced with the nucleotide sequence data. The cursor then moves to the next base and this process is repeated from the step (b).

Fig. 4



Step (d) As shown in Fig. 5, in case a matching sequence is not found, the cursor skips to the next base by one base only and repeats the process from step (b). The skipped base is then added to compressed data.

Fig. 5

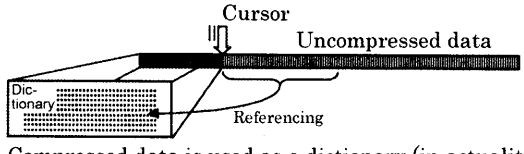


(e) Two methods were attempted to determine whether or not to fix the dictionary.

Method 1: Method in which all previously compressed base data is used as a dynamic dictionary

As shown in Fig. 6, when consecutively compressing multiple nucleotide sequences, base data on which compression processing has already been performed is used as a dynamic dictionary. The advantage of this is that the optimum compression effects can be obtained. On the other hand, the disadvantage is that, since dictionary size gradually increases, compression rate decreases correspondingly.

Fig. 6

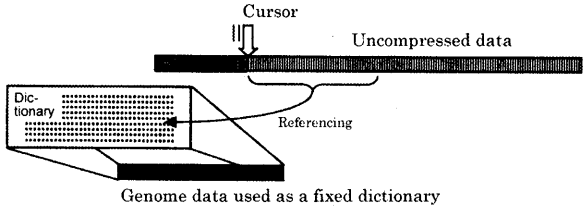


Compressed data is used as a dictionary (in actuality, this involves referring to data following decoding)

Method 2: Method using genome data as a fixed dictionary

As shown in Fig. 7, this method involves using genome data as a fixed dictionary and is likely to be used most commonly in the future. Although the compression rate remains constant since the size of the dictionary is fixed, it is not possible to take maximum advantage of compression effects as in Method 1 using a dynamic dictionary.

Fig. 7



Genome data used as a fixed dictionary

(v) Verification method using actual data

The compression method using a fixed dictionary (genome nucleotide sequences) and the compression method using a dynamic dictionary (compressed nucleotide sequences) were applied to the entire disclosed nucleotide sequence of E. coli and S. cerevisiae followed by measurement of compression capability and compression rate (omitted here) under respective conditions.

(vi) Speeding up compression processing

Dividing a nucleotide sequence to be compressed into several sections, assigning a CPU to each section and performing compression processing in parallel has also been tested. After compressing each section, the sections are coupled to complete compression processing. The problem with parallel compression processing is that sections of low matching rate with the dictionary have lowered compression rates and this becomes controlling rate.

(vii) Verification results of compression method

① Compression of public nucleotide sequence data of E. coli

All the nucleotide sequences originating in E. coli were extracted from the file 39 "ddbjbct.seq" released by DDBJ (October 1999) and compressed. The data sizes after compression by various general-purpose compression methods and the compression methods developed in this demonstration are shown in Table 2.

Table 2 Compression Method Verification Results

Compression Method	Size after compression	Size ratio	Size per base
Uncompressed	15,668kb	100.0%	8.00 bits
compress	4,234kb	27.0%	2.16 bits
gzip	4,000kb	25.5%	2.04 bits
gzip-9	3,865kb	24.7%	1.97 bits
formatdb	3,921kb	25.0%	2.00 bits
gzip-9 following formatdb	3,647kb	23.3%	1.86 bits
Method 1 of this research	1,501kb	9.6%	0.77 bits
Method 2 of this research	527kb	3.4%	0.27 bits

The size of the public nucleotide sequence data of E. coli before compression was 15,668 kb, and this included 4,663 kb of complete data of the E. coli genome (Escherichia coli K-12MG1655).

(a) Compression method 1 using dynamic dictionary

The size of public nucleotide sequence data of E. coli after compression was 1,501 kb, indicating that the data was able to be reduced to 9.6% of the size before compression. The size per base was 0.77 bits.

A breakdown of the compressed data was such that the size rate as calculated for the genome complete data only was 24.5%, and the size per base was 1.96 bits. This value was similar to that of formatdb, and compression effects were not very high. The portion other than the genome

complete data was compressed to 360 kb. The size ratio of this portion was 2.3%, and the size per base was 0.18 bits, indicating that extremely high compression effects were obtained. Moreover, even if nucleotide sequences were added, a similar degree of compression effects can be expected for this additional data as well.

(b) Compression method 2 using the genome for a fixed dictionary

The size of public nucleotide sequence data of E. coli after compression was 527 kb, indicating that data was able to be reduced to 3.4% of the size prior to compression. The size per base was 0.27 bits. It is preferable to classify the nucleotide sequences desired to be compressed in advance based on similarities with genome sequences.

② Compression of public nucleotide sequence data of S. cerevisiae

The entire nucleotide sequence of yeast (Saccharomyces cerevisiae) disclosed in the SGD (Saccharomyces Genome Database) of Stanford University (*8) was compressed to assess compression effects using yeastGenBank (as of February 14, 2001). The data sizes following compression by various general - purpose compression methods and the compression methods developed in this demonstration are shown in Table 3

Table 3 Compression Method Verification Results

Compression Method	Size after compression	Size ratio	Size per base
Uncompressed	27,602kb	100.0%	8.00 bits
compress	7,369kb	26.7%	2.14 bits
gzip	7,286kb	26.4%	2.11 bits
gzip-9	6,983kb	25.3%	2.02 bits
formatdb	7,011kb	25.4%	2.03 bits
gzip-9 following formatdb	6,481kb	23.5%	1.88 bits
Method 1 of this research	3,537kb	12.8%	1.02 bits
Method 2 of this research	1,043kb	3.8%	0.30 bits

(a) Compression method 1 using dynamic dictionary

The size of public nucleotide sequence data of yeast following compression was 3,537 kb, indicating that data was compressed to 12.8% of the size prior to compression. The size per base was 1.02 bits.

A breakdown of compressed data was such that the size rate as calculated for the genome complete data only was 23.9%, and the size per base was 1.92 bits. In other words, the yeast genome itself can be said to contain few duplications and is therefore difficult to compress. The portion other than the genome complete data was compressed to 679 kb. The size ratio of this portion was 4.3%, and the size per base was 0.35 bits, indicating that extremely high compression effects were obtained. Moreover, even if further nucleotide sequences were added, a similar degree of compression effects can be expected for this additional data as well.

(b) Compression method 2 using the genome for a fixed dictionary

The data size after compression in the case of using the complete data of the S. cerevisiae genome as a fixed dictionary was 1,043 kb, indicating that data was compressed to 3.8% of the size prior to compression. The size per base was 0.30 bits. The data size of the genome serving as the dictionary is not included in this compression rate.

In case further nucleotide sequences are coded and S. cerevisiae nucleotide sequence data is added, it is presumed that the additional data can be compressed at the above compression rate by using the genome as the dictionary.

(2) Evaluation of Verification Results

This evaluation was only based on the actual verification results obtained in this demonstration and further studies are required for practical application.

(i) Compression effects increase dramatically when the dictionary size exceeds the genome size.

(ii) Since partial sequences (sequences extracted from the complete sequence) are duplicate sequences, the resulting effects are considerable when compressed together with the complete sequence.

(iii) Although the dictionary covers the entire genome and the processing required for compression becomes extremely slow, the compression speed can be increased by separating the compressing regions and processing in parallel computers.

(iv) The trend was observed in which data added after genome decoding can be compressed to 5% or less when compressed using the genome as a dictionary.

(v) The dictionary covers the entire genome and requires memory (just under 1 GB for human genome).

(vi) It is difficult to obtain compression effects when using the genome as a dictionary if not classified for individual organisms.

(*8) SGD: URL (http://genome-www.stanford.edu/Sacharomyces)

Ⅳ Next-Generation DNA Search Systems

1 System Requirements

An example of functions desired in next-generation DNA search systems is indicated below.

① Display of search results

* Display of differences between sequences when comparing sequences
* Display of matching score for each region
* The ability to mark portions essential to an invention contained within a sequence in a search formula, and the use of a display form in which that marker is reflected at all times.

② Ability to sort search result sets according to not only order of homology but also various keys

③ Avoiding duplicate searches by making it possible to store and reproduce history of examination procedure such as the status of a previous search in the case of, for example, examinations of an amendment following notice of grounds for rejection of the same patent application.

④ Efficient searching for outsourcing data

In case of accumulating data before and after publication by dividing into data accumulated within the JPO and data accumulated outside the JPO, when search formulae are simultaneously transmitted to databases both inside and outside the JPO, considerable differences in the response time are expected due to difference in the amount of data. It is therefore necessary to implement countermeasures to resolve this problem. In addition, in case of designating either database earlier for conducting a search, it is necessary to consider the desirable order for conducting searches.

Moreover, searches are conducted by designating only the database outside the JPO in case that searching data before publication is clearly unnecessary. However, in case searching for only a part of data before publication (for example, only that immediately prior to publication) is required, searching for all data before publication leads to decreased search efficiency.

In the current DNA search systems, since restrictions with date cannot be set, all data ends up being searched which contributes to longer response times. It is desired to provide, for example, a function limiting such condition that only data prior to a specific date is searched.

2 Requests to the JPO

(1) Access to Outsourcing Gene Sequence Analysis Data by General Users

If it is possible to analyze gene sequences under the same conditions as patent examinations and guidelines are indicated for evaluation methods employed in actual examinations (including database to be used, analysis tools and parameters), since this would enable applicants to assess the propriety of a patent application in advance, the number of invalid applications and their examinations can be hopefully decreased. Access to outsourcing gene sequence analysis systems by general users is therefore desired.

(2) Database Development and Disclosure

It is also desired that links be established with cited patents and literatures, citating patents, patents of US, European and other foreign countries, and gene and protein databases, that search functions be provided that enable searches to be conducted based on the results of sequence analyses of BLAST or PFAM (Protein families database of alignments and HMMs) or according to various categories, applicants and other bibliographic items, and that highly useful patent database and analysis systems be developed and constructed that contain such information as examination status, patent family status in the US, Europe and other countries and patent term in an easily understandable manner, and that those databases be made available for the general public.

Moreover, it is desired that sequence data described in patent applications be compiled and made downloadable so that private sectors can easily create databases with high added value allowing further sophisticated key word searches and so forth.

(3) Disclosure of Sequence Information in Patent Applications to Public Databases

It is desired that sequence information in published patent applications be promptly included in public databases. In addition, it is desirable to construct a system that the time of publication of a patent application coincides with the time of disclosure of sequence data in public databases, and in case a patent application is withdrawn prior to publication, the sequence data itself is prevented from being disclosed alone.

3 Others

① It is also preferable to introduce "deposition system" for sequences into the system. More specifically, this involves deposition to a specific institute of sequences to be identified with the application number (or other ID if prior to filing) and the sequence number without submission of sequences list. The depositor is able to refer to the deposited sequences in a patent application and, a third party can cite the deposited sequences after publication of the patent application.

If this system is employed throughout the world, it is certain that the administrative burden of not only the JPO, but also the PCT receiving office and patent offices of other countries will be considerably reduced, thereby making this an appealing system not only for patent offices, but also from the viewpoint of reduction of applicants' workload.

② It is also desirable to construct a system that allows the submission of sequence data to be omitted when submitting translations of PCT applications.

V Summary

1 Present Situation and Problems of DNA Sequence Databases

Firstly, in this research and study, a discussion has been made regarding the present situation and problems associated with DNA sequence databases along with the DNA sequence search system used in the JPO based on the present situation of searches of patent applications for DNA sequence using the JPO databases.

In order to construct a database that satisfies the needs of examiners engaged in examination of patent applications relating to DNA sequences, it is clearly necessary to conduct a comprehensive study regarding not only the problem of the amount of data of DNA sequence databases, but also the ideal form of databases and qualitative issues including reliability. It is then considered to be essential to discuss the advantages and disadvantages of outsourcing databases, following that study. However, since there has been very little discussion of qualitative problems associated with databases in the committee of this research and study due to the objective of the committee and time constraints, that issue will be a subject for future discussion.

With respect to public databases, detailed introductions have been provided regarding the present situation of databases that have actually been constructed and are in use at the National Institute of Genetics, the Institute of Medical Science, the University of Tokyo and Institute for Protein Research, Osaka University which the members of the committee belong to. Specific and valuable proposals were also made regarding future problems based on numerous issues including technical, administrative and qualitative problems. Public databases are different not only in the types of data to be handled but also in the objectives of database construction. Consequently, present situation and problems of the databases are not always shared. However, numerous problems have been pointed out, including technical problems such as increases in the amount of data, updating and transmitting of data and problems in outsourcingas well as organizational problems such as database management problems. With respect to qualitative problems confronting databases, public databases are required to operate stably for a long time, to provide a function allowing disclosure of data including private data, to standardize data formats and notation, and to provide compatibility between the data of different databases. It has also been pointed out that it is important to construct databases easy to use and reliable with the emphasis on networking of those databases. Since all of these factors are considered to be common to DNA sequence search systems in the JPO, it is expected that further concrete and detailed studies will be conducted in the future.

Moreover, although the present situation and many of the problems of private databases are shared by public databases from the standpoint of users, and there are also problems with respect to leakage of data and system security, it has been reported that outsourcing will be unavoidable from the viewpoint of increasing data volume and management costs.

In addition, reports and discussions have been presented regarding the situation surrounding DNA databases at the USPTO and EPO. As the data retained by the trilateral patent offices is considered to be duplicated, both of these patent offices along with the JPO are promoting projects such as common databases and exchange of data through Trilateral Patent Offices Conference. When the JPO constructs a DNA database, discussions should be adequately conducted towards handling of duplicate data and the unification of the format and notation of the database, while paying close attention to the situations at both of these patent offices, after which studies should be made regarding the feasibility of database outsourcing.

2 Outsourcing of DNA Sequence Databases in the JPO

There is an opinion to doubt the necessity and effectiveness of outsourcing from the viewpoint of countermeasures for accumulating data and improving the search response of databases. Since this involved raising questions regarding lack of the basic policy in terms of implementing outsourcing, it is first necessary to clarify the basic policy regarding outsourcing.

In fact, if outsourcing is to be implemented, after conducting a detailed and specific discussion concerning the evaluation of outsourcing itself, specific issues in the implementation of outsourcing should be discussed.

3 Compression of Sequence Data

According to an evaluation of data compression and the results of demonstration testing, it was confirmed that compression of flat data only is effective from the viewpoint of speeding up sequence database searching. In addition, it was also confirmed that compression of duplicate data in demonstration testing provides a certain degree of effectiveness under specific conditions. Although it is possible that these evaluations may change depending on the progress of future research in the field of data compression, this most likely serves as a reference when a DNA sequence database is constructed in the JPO.

4 Next-Generation Search System

With respect to a next-generation search system, discussions should be conducted after clearly identifying the positioning within the entire next-generation search system of the JPO, namely the positioning of a DNA sequence database within the entire information policy of the JPO, including system requirements such as display of search results and efficient searching for outsourcing data, along with the requirements to the JPO such as access to the system by general users, database development and disclosure, and disclosure of sequence information contained in patent applications in public databases.

On the basis of the above, although it was not possible to reach a final conclusion due to time constraints and the remarkable technical advances in the field of databases, it is believed that a definite direction was able to be obtained. Furthermore, the opinions presented in the report on this research and study and this bulletin are merely based on personal opinions, and do not necessarily represent the consensus of the committee.

(Researcher: Takashi Ikegami)