DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams

Yoshio Tateno*, Satoru Miyazaki, Motonori Ota, Hideaki Sugawara and Takashi Gojobori

Center for Information Biology, National Institute of Genetics, Yata, Mishima 411-8540, Japan

Received September 8, 1999; Revised September 16, 1999; Accepted October 4, 1999

ABSTRACT

We at DDBJ (http://www.ddbj.nig.ac.jp) process and publicise the massive amounts of data submitted mainly by Japanese genome projects and sequencing teams. It is emphasised that the collaboration between data producing teams and the data bank is crucial in carrying out these processes smoothly. The amount of data submitted in 1999 is so large that it alone exceeds the total amount submitted in the preceding 10 years. To cope with this situation, we have developed tools not only for processing such massive amounts of data but also for efficiently retrieving data on demand.

MASS DATA PROCESSING TOOLS

Enormous amounts of scientific data have been produced and processed in many fields including biology. This raises the serious problem of how to deal with such massive amounts of data. We foresee that it will be of paramount importance not only to collect, process and store data, but also to efficiently extract the necessary information from the jungle of data. Otherwise one could easily be lost in the jungle, which would devalue the data itself.

At DDBJ we have been flooded with mass submissions of DNA sequence data. At present we are trying to cope with the massive data flow for processing and publicising. In particular, various sequencing projects in Japan are now actively producing large amounts of data that are in turn submitted to DDBJ. The recent growth of submitted data to DDBJ is shown in Figure 1. The growth in 1999 is particularly steep and noteworthy. The number of entries to be processed for this year alone will exceed the total number processed in the preceding 10 years. The main driving forces behind this explosive growth are the Japanese human genome project, mouse cDNA project and *Caenorhabditis elegans* project, of which further details will be mentioned below.

As we have reported previously (1), we developed a large-scale data submission system. We have improved this system and now call it MSS (Mass Submission System). MSS includes an off-line tool, MST (Mass Submission Tool), which functions by arranging data into a form ready for submission and also acts as a parser at a data producing site. The parser, however, can only detect trivial errors, and promotes the submitter to make their submission as error-free as possible. At DDBJ we use MSS to monitor the processing of submitted data and to install the processed data into the database.

With MSS and other data processing tools, we can currently process submitted data at a rate of 20 000 entries per day. This is more than four times the rate processed in 1998. However, errors in submitted data found at DDBJ drastically reduce the rate at which the data can be processed as they necessitate extra work and communication with the submitter. We therefore believe that collaboration with submitters is key to faithful rendering of their data.

DATA RETRIEVAL TOOLS

When one performs retrieval, one is often concerned with sequences derived from a particular species. To address such a demand, we have developed a device by which to reorganise our entire database into a species-oriented database in which the data are divided into a species as a unit. At present (DDBJ release 38, July 1999), there are 50 700 species in total. It is noted that this DDBJ release includes data processed not only by DDBJ but also by GenBank and the EMBL Data Library. The three data banks organise the International Nucleotide Sequence Database and exchange data collected and processed at each bank on a daily basis. Therefore, the quality and quantity of the data are maintained to be equivalent among the three data banks. We have developed a tool (http://ftp2.ddbj.nig.ac.jp:8000/orgstart-e.html) by which one can first specify a species of interest by its scientific name among the 50 700 species, and then carry out a keyword or homology search against the data for that species alone. This tool is expected to be useful particularly for examining whether a particular gene sequence is available for the species in question. As data accumulate in our database at an ever-increasing rate, the tool will provide a means of reducing retrieval time.

The same tool as above is applied to ESTs. If one is interested in ESTs of a particular species, one might carry out a homology search against the data of that species only by giving a probe sequence and specifying the scientific name of the species. As the amount of ESTs grows tremendously, this tool will help reduce retrieval time when one is concerned with a particular species. This is better understood when one realises that >70% of the rapidly increasing data are ESTs (see Fig. 1).

Another device we recently developed is a tool (mblast@watson.genes.nig.ac.jp) that allows one to give multiple probes at once and individually retrieve homologous or similar sequences to those probes. If one uses those two tools together, one would easily examine whether a set of sequences for a

*To whom correspondence should be addressed. Tel: +81 559 81 6857; Fax: +81 559 81 6858; Email: ytateno@genes.nig.ac.jp

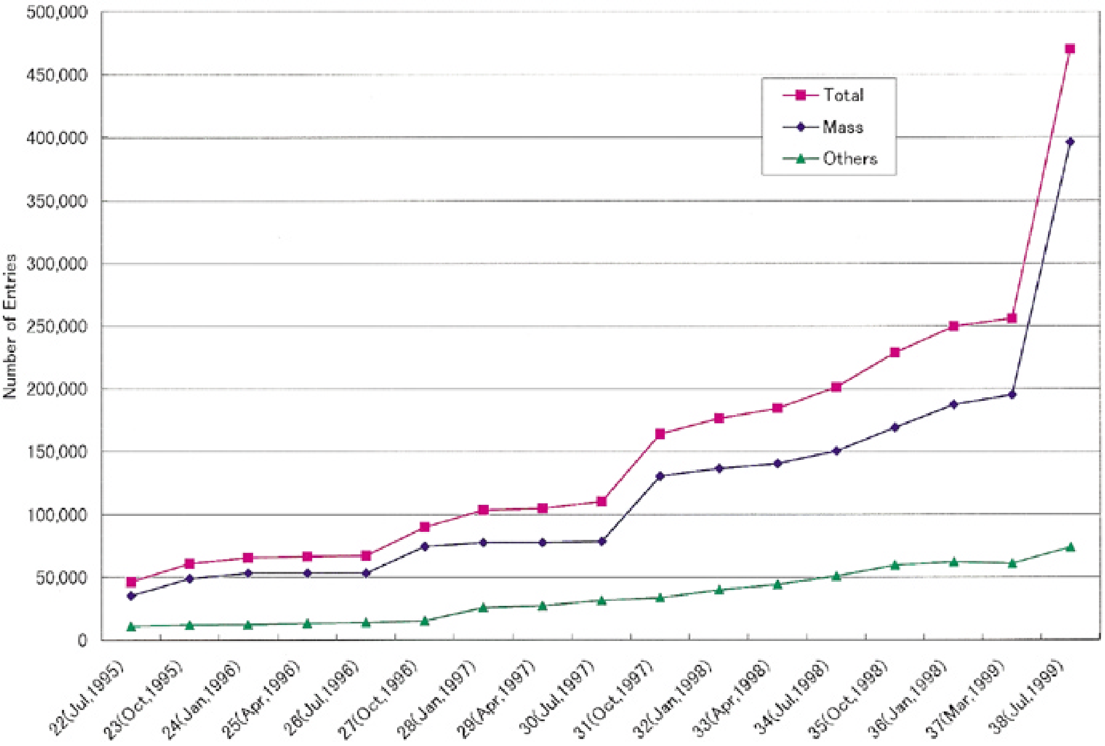

Figure 1. Recent growth of data submissions to DDBJ. Mass represents massively submitted data such as ESTs, HTGSs, STSs and GSSs, and Others mostly represents data of complete sequences.

particular biological function is available for a species of interest.

For retrieval of the complete genome data, the Genome Information Broker (GIB, http://mol.genes.nig.ac.jp/gib/) (2) has been actively used worldwide. Since the first implementation of GIB, we have repeatedly revised it and installed new complete genome sequence data into it whenever such data becomes available. Currently, GIB includes the genome data of *Saccharomyces cerevisiae* and 22 prokaryote species including those of *Escherichia coli*, *Synechocystis sp*. and *Pyrococcus horikoshii* that were sequenced by Japanese teams. By use of GIB one can now search for a particular gene not only for one species but also across the 23 species. In this way, one can study, for example, the genomic organisation of the gene and its neighbour for different species.

HUMAN DATA

There are four major teams in the Japanese human genome project: the Sakaki and Hottori (the Genome Science Comprehensive Research Center of RIKEN), Shimizu (Keio University), Inoko (Tokai University) and Nakamura (the Japanese Foundation of Cancer Research) teams. The data produced by the four teams have been submitted to DDBJ. In particular, the Sakaki and Hattori team has sequenced >24 Mb of chromosomes 11 and 21, and submitted them to DDBJ as HTGSs (High Throughput Genome Sequences). The data are accessible at http://www.ddbj.nig.ac.jp/ddbjnew/990513-e.html . The team continues to sequence chromosome 21, and will finish the whole chromosome soon. The Shimizu team has submitted >1 Mb of chromosome 8 as HTGSs. The Japanese human genome project is expecting to contribute ~10% of the entire human genome to the International Human Genome Consortium.

In collaboration with the Sakaki and Hattori team, we have established an automation process of data handling of GSSs (Genome Survey Sequences). This process is implemented by a tool, which, similarly to MSS, operates both at the Sakaki and Hattori team and DDBJ. At the Sakaki and Hattori team this tool checks and arranges produced data into a form ready for submission and then sends the data to DDBJ. At DDBJ the tool functions as a monitor of the processing of submitted data and as an installer of the processed data into the database. The tool has greatly facilitated data processing of GSSs at both sites.

MOUSE DATA

At the newly established Genome Science Comprehensive Research Center of RIKEN the Hayashizaki team has produced a large number of mouse sequences which are expressed in tissues (3,4). The team recently obtained 175 734 complete mouse cDNAs and sequenced them from the 3′ terminus for a few hundred bases upward. The sequence data may be distinct from ESTs in that all the sequences subject to sequencing are the complete cDNAs, while ESTs do not necessarily come from complete cDNA. At any rate, the team submitted the data on the 175 734 sequences at once to DDBJ. The total number of bases is 46 032 374, which implies that the average length over the total sequences is 262 bases.

We had never experienced such a massive submission before but could finish processing the mouse data in a week or so and made them public because the Hayashizaki team worked in collaboration with DDBJ. As can be seen, there are many duplicated sequences among the data. Therefore, the total data reflect the expression profile of the tissue to some extent. If one is interested in a set of non-redundant data, one may refer to http://genome.trc.riken.go.jp at RIKEN. For elucidating the function of a human sequence the mouse data will play a significant role, because the mouse counterpart is easily retrieved against these mouse data. The retrieved data are used as a probe to single out the corresponding mouse complete cDNA which can in turn be fully sequenced. One can then perform appropriate experiments on the sequence in mice, which one cannot in humans. The accession numbers of the mouse data use the continuous range from AV000001 to AV175734. The Hayashizaki team has recently informed us of their plan to submit another set of 170 000 mouse sequences soon.

Caenorhabditis elegans DATA

As reported by the *C.elegans* Sequencing Consortium (5), the genome of the nematode has been sequenced, and the data are now available worldwide. However, the expression profiles and functions of the genes on the genome mostly remain to be elucidated. The Kohara team of our institute has carried out research in the expression stage and profile of the nematode genes in order to understand their functions and relationships. The team has accordingly produced a mass of ESTs from the nematode which have been submitted to DDBJ. Since the team and DDBJ are on the same campus, we have established a good collaboration with respect to data submission and processing. Recently, the team submitted 28 278 ESTs to DDBJ where they were processed and made public in a few days. Though they were classified as ESTs, the data on the expression stage of a sequence and the sex of the organism are also given, as one can see by use of our getenry retrieval tool (http://ftp2.ddbj.nig.ac.jp:/8000/getstart-e.html) against their accession numbers, AV175735–AV204012. The team will continuously produce data and submit them to us.

CONCLUSION

DNA sequence data are being produced worldwide at an enormous rate, while the International Nucleotide Sequence Database is responsible for collecting, processing and publicising them as soon as possible. Therefore, the collaboration between the three data banks will be more and more critical, though it has been kept in excellent condition. It is also important that the DNA data banks work in collaboration with mass sequencing teams to facilitate data processing and release.

We also have to think of reducing retrieval time in order to cope with mass production of data. One of the ways to realise this is to develop software tools by which to guide database users to what they really obtain as economically as possible. We have made efforts not only to collect and process massive amounts of data but also to develop tools for efficient retrieval.

ACKNOWLEDGEMENTS

We thank K. Goto, S. Misu, A. Shimada, H. Tsutsui and H. Hashimoto of DDBJ for their contributions to the present report.

REFERENCES

1. Sugawara,H., Miyazaki,S., Gojobori,T. and Tateno,Y. (1999) *Nucleic Acids Res.*, **27**, 25–28.
2. Tateno,Y., Kobayashi-Fukami,K., Miyazaki,S., Sugawara,H. and Gojobori,T. (1998) *Nucleic Acids Res.*, **26**, 16–20.
3. Carninci,P., Nishiyama,Y., Westover,A., Itoh,M., Nagaoka,S., Sasaki,N., Okazaki,Y., Muramatsu,M. and Hayashizaki,Y. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 520–524.
4. Sasaki,N., Izawa,M., Watahiki,M., Ozawa,K., Tanaka,T., Yoneda,Y., Matsuura,S., Carninci,P., Muramatsu,M., Okazaki,Y. and Hayashizaki,Y. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 3455–3460.
5. *C.elegans* Sequencing Consortium (1998) *Science*, **282**, 2012–2018.