DNA Data Bank of Japan dealing with large-scale data submission

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

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

ABSTRACT

The DNA Data Bank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp) has developed a software system for mass submissions to cope with a recent expansion of EST and genome data submissions. The system is composed of four parts, the WWW data submission, large-scale submission, submission management and storing. Using this system one can submit data on a large number of sequences or a very long sequence while checking the consistency between the annotation and sequence without much effort. DDBJ has received large scale data of Homo sapiens, Arabidopsis and Pyrococcus from Japanese researchers who made full use of the new submission system.

INTRODUCTION

When we began data bank activity at the DNA Data Bank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp) more than 10 years ago, we expected that the amount of submitted data would double over the following five years or so. However, the rate is presently less than two years beyond our expectations. There are two reasons for this which were unexpected at that time.

The first one is due to the commencement of the cDNA project which produced a large number of partially sequenced DNA fragments of a few hundred nucleotides in expressed genes called the expressed sequence tags (EST) (1). EST data have also been sent to DDBJ as mass submissions sometimes including more than 10 000 sequences in a single submission. We allocate an accession number to each of the submitted ESTs. The EST data continue to be produced by many projects worldwide and submitted to DDBJ, GenBank and the EMBL Nucleotide Sequence Database. Today, ESTs occupy more than 75% of the total data that the three data banks have collaboratively collected, processed and released.

Since EST data provide our users with information only about the source organism and the corresponding gene being just expressed in the cell with unknown function, one might consider that they have a narrower range of usage than ordinary sequence data. One can, however, take advantage of their richness in number when making use of them for biological study. For example, we recently carried out research in gene hunting for a HLA genome region, and found EST data quite useful (2). Namely, we made homology search of possible exons detected by a gene finding tool in this region against the dbEST, and picked up almost identical ESTs with the exons. Some of the ESTs were then found to form a gene which was similar to an extant functional gene.

The second reason lies in the beginning of genome projects, particularly for prokaryotes. The projects have produced the complete genome sequences for 10 or more prokaryotic species (3,4 and others) and that of Saccharomyces cerevisiae (5). Furthermore, the complete genome sequence of eukaryotic species such as Caenorhabditis elegans and Arabidopsis thaliana will soon be available, and then the completion of the human genome sequencing projects will follow. One of the usages of the complete genome sequence data is to investigate the evolution of genome structure. In S.cerevisiae it has been indicated that the whole genome of the yeast experienced genome duplication about 100 million years ago (6). Similar observation and discussion has been made for the human genome (7), though the complete genome sequence is not yet available.

In this article we report the modification of our data submission tool, Sakura (8), and database management system, Yamato II (9). The modification resulted in the ability of the tools to handle and process mass submissions of ESTs and long DNA stretches of a genome sequence more efficiently than the original ones. We also briefly discuss the extension of large scale data submission and processing on the basis of the object oriented database management system.

A LARGE-SCALE DATA SUBMISSION SYSTEM

DDBJ originally developed two World Wide Web (WWW) interface-oriented systems, Sakura and Yamato II; Sakura is used for data submission, whereas Yamato II is utilized in data annotation and management. Both tools have been successful over the past few years in terms of service and operation, however, the systems apparently lacked the ability to handle a large number of sequences in a set and/or very long sequences. In recent times, DDBJ has experienced a dramatic increase in massive data submission (Table 1). Therefore, DDBJ has introduced a new system (modified from Sakura and Yamato II) which has the capacity to accept and annotate this influx of data.

OVERVIEW OF DATA FLOW IN DDBJ

When submitting ordinary data, a submitter has several options (Fig. 1). Firstly, Sakura, which requires only a WWW browser,
It is clear that the number of EST and genome sequence submissions steeply increased, while that of ordinary data submissions has remained steady over the two years.

Table 1. Numbers of monthly submitted and updated entries in 1997 and 1998

<table>
<thead>
<tr>
<th>Year</th>
<th>Jan</th>
<th>Feb</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST</td>
<td>2</td>
<td>3,993</td>
<td>9,218</td>
<td>5,689</td>
<td>3,299</td>
<td>11,176</td>
</tr>
<tr>
<td>Genome</td>
<td>363</td>
<td>1</td>
<td>43</td>
<td>256</td>
<td>777</td>
<td>126</td>
</tr>
<tr>
<td>Patent</td>
<td>1,083</td>
<td>718</td>
<td>1,157</td>
<td>825</td>
<td>889</td>
<td>716</td>
</tr>
<tr>
<td>Others</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TOTAL</td>
<td>955</td>
<td>1,035</td>
<td>1,191</td>
<td>1,454</td>
<td>1,633</td>
<td>1,744</td>
</tr>
<tr>
<td></td>
<td>2,222</td>
<td>1,872</td>
<td>1,786</td>
<td>1,375</td>
<td>1,560</td>
<td>1,732</td>
</tr>
<tr>
<td></td>
<td>2,040</td>
<td>5,726</td>
<td>11,506</td>
<td>7,968</td>
<td>5,821</td>
<td>15,535</td>
</tr>
<tr>
<td></td>
<td>2,615</td>
<td>1,917</td>
<td>1,837</td>
<td>1,433</td>
<td>2,338</td>
<td>1,859</td>
</tr>
</tbody>
</table>

It is clear that the number of EST and genome sequence submissions steeply increased, while that of ordinary data submissions has remained steady over the two years.

The DDBJ’s new system for managing large-scale data submissions primarily consists of four separate parts which are (i) the WWW data submission system, (ii) large-scale data submission system/off-line (installed at the submitter side), (iii) data submission management system and (iv) data storing system (Fig. 2). The WWW data submission system incorporates the Internet environment so that a submitter is able to interactively send an inquiry to DDBJ, and to receive an ID, a password and a template ID from DDBJ. The system is, therefore, designed to properly handle all pieces of information necessary for the data submission. An interface of the system is available in both Japanese and English depending on the submitter’s preference.
The large-scale data submission system/off-line is publicly available by downloading the program from an ftp server at DDBJ. The system rigorously checks the file and annotations automatically excluding any invalid characters. It also allows a submitter to verify the file formats and consistency concerning annotation and sequence prior to actually sending the data files by Email. Two types of files are used for the submission; one is used for annotation and the other is employed for recording sequences. The file for annotation is in a tabular format which popular word processors, spreadsheet and database management systems can handle. Therefore, a submitter is not required to purchase an expensive platform in addition to their conventional system. Nucleotide sequences are submitted in the FASTA file format. The large-scale data submission system/off-line has another important function, which is to automatically verify the file format and consistency. This is important for submitting large-scale genome data, because it greatly alleviates the number of efforts, which exhaust the human resources of reviewers and annotators at DDBJ. The third system is called the data submission management system, which manages the submitted files and monitors the submission progress. After receiving the Excel and FASTA files sent from a submitter, a record is made for each submission by operating the system, and the message is sent to other staff members of DDBJ by Email. The fourth is called the data storing system. This system automatically performs more rigorous checking before completely loading the files to the master database in addition to checking the consistency regarding the annotation and sequence. The template information ranging from the locus, definition, accession number, the feature information and to the source organism is also loaded to the master database by the system. Finally, the system issues an accession number to the administrator of DDBJ who notifies the number to a submitter by Email.

By operating the large-scale data submission system, DDBJ has recently responded to some of the major genome data submitted from institutes and universities in Japan. For example, Kazusa DNA Research Institute has submitted Arabidopsis data with the size of 7 472 343 nucleotides which is the longest sequence data ever submitted to DDBJ on a single submission. The Institute has made another submission of Homo sapiens data of 67 914 nucleotides. Kitasato University has also submitted Homo sapiens data with the sizes of 4 842 948 and 5 561 026 nucleotides, separately. In addition, the Product Assessment Technology Center of the Ministry of Industry and Trade has submitted Pyrococcus horikosii data with the size of 1 738 505 nucleotides. All of those data are now available at DDBJ, GenBank and EMBL Nucleotide Database.

In addition to the new systems mentioned above, there are three more systems served for data management. First, is the master database system called the ddbj, which exclusively stores the files despite the size of the data per entry. The database is based upon Sybase with the UNIX operating system controlled by the SUN server. The second one is the so-called group manager, which updates the annotations after loading the files into the database. The last one is the distribution manager, which releases the new data into the public arena.

There was a special case for submitting GSS (Genome Survey Sequence) data (Fig. 3). Kitasato University submitted the data using a tool called the GSSin which was specially developed by DDBJ. In this case, the files, which are formatted in FASTA, are...
Figure 3. Submission system for GSS (Genome Survey Sequence).

sent to DDBJ through ftp. After receiving the files by a server named the supernig, they are then automatically checked and registered to the Sybase database by using another tool called the GSSsub, which was also developed by DDBJ on a UNIX server, Generous. During this process GSSsub also automatically monitors the submitted files to ensure that they are valid. In addition, the system automatically registers and releases the updated GSS data submitted from the university every night. In the event of a registration error the system sends an error message to the DDBJ administrator by Email. The system has been successfully operating since October 1997.

Although the large-scale data submission system is not well known, its ability to efficiently process data has led it to be regarded as the flagship of DDBJ. As shown previously, over the last few years there has been a substantial increase in sequence submissions to DDBJ. From our experience with data management, it has become obvious that we need to enhance our ability to develop more efficient and effective data submission tools. We aim to offer efficient, cost effective and user friendly services, thereby making DDBJ more competitive. DDBJ is now considering upgrading the data submission and processing systems by introducing a new type of architecture such as CORBA, which is an object oriented distributed platform. The new systems will be well suited for more efficiently managing the submitted data by reducing data handling labor and time.

REFERENCES


