Deposition
The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between DDBJ/NIG, EMBL/EBI, and GenBank/NCBI for 20 years (http://www.insdc.org/). In 2006, DDBJ assigned accession numbers to 2.13 million entries of 1178 Mbps in total, resulting in the contribution of 17% in entries and 11% in nucleotides to INSD. EMBL processed as many as DDBJ, and GenBank did the rest. The data deposited to DDBJ were mainly from Japan, including patent sequences from Japan Patent Office. DDBJ estimates the growth of the INSD as shown in Fig. 1(A) and Fig. 1(B). They are not radical estimations, considering the development of new sequencing technologies and the expanding interests in metagenome and gene variation analyses.

So far, nearly 90% data from the Asia Pacific region have been deposited by groups in Japan. DDBJ predicts that the deposition from other countries in the region will increase, and has prepared to accept it either through a web browser or by file transfer. The web deposition system named SAKURA has been open to the public and continuously refined since 1995. Using this system, one can interactively enter and submit nucleotide and translated amino acid sequences, functions and features of the sequences, and references as well as one’s name, affiliation and address. However, SAKURA is not suitable for data deposition in such cases as the following:
- the bandwidth of the internet connection is narrow;
- the deposition consists of large number of entries; or
- the submission involves long nucleotides and/or rich biological features (namely, annotation) of the sequence.
In these cases, DDBJ recommends contacting the help desk before preparing files for the deposition. Most of the depositors to DDBJ have actually used EXCEL files to send a set of fasta files and annotation files.

Analysis
DDBJ has made an effort to improve the usability of the primary database, which is an archive of data deposited by diverse wide communities from academia to industry. The information in the INSD ranges from very short sequences to full chromosome sequences with deep annotation. The data quality is also diverse.

An option of DDBJ is the development of derived databases (secondary databases) from the primary database: Genome Information Broker (GIB) databases, Gene Trek in the Prokaryote Space (GTPS) database and Genes to Protein (GTOP) database, for example.

The GIB databases at http://www.genome.nig.ac.jp/ are a family of complete genomes of microorganisms and viruses (GIB-M and GIB-V), sequences from environmental samples (GIB-ENV) and sequences of the IS region (GIB-IS). GTPS is also accessible from the above URL and provides high-quality ORFs in prokaryote genomes. In GTPS, DDBJ has repeated the prediction and evaluation of ORFs of all the microbial genomes in the primary database every year. GTPS discarded quite a few ORFs in INSD and added some new high quality ORFs that were not predicted by the original data depositor. GTOP is the result of a comprehensive analysis of ORFs to predict the secondary structure of corresponding proteins. GTOP has recently included the disordered region predicted in amino acids sequences as shown in Fig. 3.
Another option of DDBJ is the development of analytical tools. G-InforBIO is a stand-alone software package to compare multiple genomes for laptop computers. One can download genomes from DDBJ into G-InforBIO to compare them with one’s own genome. Moreover, web services (http://xml.nig.ac.jp/) at DDBJ provide network computing. One can use ARSA, blast, ClustalW, getentery, GIB, GTOP, GTPS and other DDBJ services from one’s Perl or Java program to construct workflows for solving bioinformatics problems. DDBJ will enrich these web services and workflows to make the life of bioinformaticians easier.

Training Courses

DDBJ organizes bioinformatics training courses both in Japan and abroad, especially in collaboration with China and Korea. In March 2007, a course was organized in Shanghai. It is worthwhile to visit the DDBJ website from time to time to find details of training courses in addition to new services.

The course covers not only basic sequence retrieval and analysis, but also population genetics, genomics, proteomics, and much more.

References


Contact Details:

Hideaki Sugawara, PhD
Address: Center for Information Biology and DDBJ
National Institute of Genetics
1111 Yata, Mishima
Shizuoka 411-8540, Japan
Email: hsugawar@genes.nig.ac.jp
URL: www.ddbj.nig.ac.jp