Visualization of Sequence and Biological Data in DNA Data Bank of Japan - Genome Information Broker and the Enhancement of SAKURA

Kousuke Goto [1] kougoto@genes.nig.ac.jp
Hikaru Yamamoto [1] hikyamam@genes.nig.ac.jp
Takuro Tamura [2] tatamura@genes.nig.ac.jp
Toshitsugu Okayama [1] tokayama@genes.nig.ac.jp
Tomohiro Koike [1] tkoike@genes.nig.ac.jp
Satoru Miyazaki [1] smiyazak@genes.nig.ac.jp
Hirotada Mori [3] hmori@gtc.aist-nara.ac.jp
Takashi Gojobori [1] tgojobor@genes.nig.ac.jp
Hideaki Sugawara [1] hsugawar@genes.nig.ac.jp

[1] DNA Data Bank of Japan, National Institute of Genetics
1111 Yata, Mishima, Shizuoka 411, Japan

[2] Hitachi Software Engineering America, Ltd.

[3] NARA Institute of Science and Technology,
Research & Education Center for Genetic Information
8916-5 Takayama, Ikoma, Nara 630-01, Japan

Abstract

Data submitters, reviewers and users of DNA Data Bank of Japan (DDBJ) processes sequence data longer than 1M base pairs thanks to genome projects. In order to realize smooth and reliable submission, annotation and dissemination of the large scale genetic information, DDBJ developed systems which visualize sequences and relevant biological information. A newly developed data dissemination system named Genome Information Broker and the enhancement of Web data submission system SAKURA are introduced here from the viewpoint of visualization.

1 Introduction

The International Sequence Database, DDBJ/EMBL/GenBank, includes mega sequences, although each entry of the database was designed to store sequences of 20~350,000 base pairs for the sake of the rapid homology and other search. Thus mega sequences are separated into many entries in a large DDBJ/EMBL/GenBank database which is composed of more than 1.5 million entries. Therefore, it is obvious that the data bank and the user require a tool which provides a comprehensive view of a mega sequence like a genome sequence.

In 1996, DDBJ collaborated with the Japan *Escherichia coli* genome project team in the development of a system to disseminate the genome data to the public in a comprehensive way as soon as the data was submitted to DDBJ. In the development, we both recognized that the visualization of sequence, genes, clones and ORFs is indispensable to the retrieval of the large scale sequence. The development was successful and DDBJ applied the system named Genome Information Broker (GIB) to other microbial genomes whenever the sequence was completed.

In the meantime, DDBJ also recognized that graphical interface is also required for the data submission and annotation, even if the sequence is as short as 1,000 base pairs. It is often difficult to describe and evaluate the sites for exon, intron and other biological features along the string of A, T, G and C. Visualization of features along the sequence will greatly help data submission and annotation. DDBJ developed such visualization tool in the Web data submission system, SAKURA [1].

2 Methods

2.1 Genome Information Broker (GIB)

The core of GIB developed for *E. coli* occupies about 200MB disk in an NT4.0 machine with 130MB memory. The core of GIB for other microbial genomes utilizes also about 200MB disk in a UNIX workstation with 64MB memory. The blast search and the acquisition of sequence are done in a separate UNIX workstation.

The GIB is composed of two components:

- Java applet for the display of genome information
- CGI program which constructs Web pages with commands embedded in HTML files

2.2 Enhancement of SAKURA

The server of SAKURA is running on a UNIX workstation of 64MB memory and utilizes 3.1GB. In SAKURA, the data submitted from Web browser is converted into the flat file format, based on which Java applet visualize the features.

3 Results and Discussion

We compiled the International Nucleotide Sequence Database (DDBJ/EMBL/GenBank) entries that had been submitted by microbial genome project teams worldwide into databases specific to the species such as:

- *Escherichia coli* (The Japan Escherichia coli genome project team)
- *Escherichia coli* (Laboratory of Genetics, University of Wisconsin)
- *Haemophilus influenzae*
- *Mycoplasma genitalium*
- *Methanococcus jannaschii*
- *Synechocystis PCC6803*
- *Mycoplasma pneumoniae*
- *Helicobacter pylori*

It took 2 ˜ 3 hours to implement a new microbial genome data into GIB, once the data was disclosed from DDBJ/EMBL/GenBank. In GIB, you are able to use the numbers and names of clones, ORFs and genes, and sequences to retrieve the data in the database. Results are represented in interactive graphics and tables. We plan to apply GIB to the genome information of higher organisms than microbes.

In the case of the enhancement of SAKURA, the visualization is quite powerful to notify data submitters of the controversy in the features of their sequences.

More than just the data retrieval and the error checking, the visualization will stimulate our insight into the structure of genomes.

References

[1] Yamamoto, H., Tamura, T., Isono, K., Gojobori, T., Sugawara, H., Nishikawa, K., Saitou, N., Imahishi, T., Fukami-Kobayashi, K., Ikeo, K. and Tateno, Y.,*Proceedings of the Seventh Workshop on Genome Informatics*, Akutu, T. et al, 204-205, 1996.