A New File Format and Tools for the Large-Scale Data Submission to DNA Data Bank of Japan (DDBJ)

Satoru Miyazaki[1] smiyazak@genes.nig.ac.jp
Hiroyuki Hashimoto[2] hirohash@genes.nig.ac.jp
Akemi Shimada[1] akshimad@genes.nig.ac.jp
Yoshio Tateno[1] ytateno@genes.nig.ac.jp
Hideaki Sugawara[1] hsugawar@genes.nig.ac.jp

[1] Center for Information Biology, National Institute of Genetics,
1111 Yata, Mishima, Shizuoka 411-8540, Japan
[2] Hitachi Software Engineering Co., Ltd. Yokohama 231-0015, Japan

Keywords: DNA sequence, genome, large-scale submission, JAVA, object-oriented

1 Introduction

Thanks to the genome projects and the rapid development of high-throughput technique to determine DNA sequences, two types of very large-scale data have been submitted to DDBJ/EMBL/Genbank International Nucleotide Sequence Databases (INSD). One is a whole genome sequence of an organism. Another is a set of sequences for the phylogenetic study based on a ubiquitous gene among hundreds of organisms. It is very important for researchers to get accession numbers of their sequences in a timely fasion. INSD has to meet their needs even if the data is massive.

To make submission of data easier, INSD has developed and provided such tools for stand-alone usage and Web submission as Authorin, Sequin, SAKURA, Webin, BankIt. These submission tools, however, target relatively small scale data and are not suitable for massive data.

Therefore we at DDBJ developed a new data file format and off-line tools to support the submission of the large-scale sequence data.

2 Method and Results

2.1 A new file format

Authorin and Sequin use the transaction file format and ASN.1 format respectively. EMBL flat file format is required to send your data to EMBL database. These formats, however, have some points we have to consider. For instance, files described by ASN.1 format are difficult to edit directly, but easy to find some syntax error by computers. EMBL format is easy to edit by yourself, but difficult to describe many kinds of combination of feature keys and qualifiers, that are two level data descirptors, without syntax errors. To realize readability and easiness to parse, we designed a new file format for the description of biological annotation to sequences. It is a tab-delimited text format with five columns (Entry, Feature, Location, Qualifier and Value) (See Fig. 1).

2.2 Mass Submission Tool (MST)

MST is a stand-alone application program and is able to handle the annotation data consisting of over 100 features with various qualifiers for more than 10,000 sequences. This application is consist of three modules: annotation editor, sequence editor and parser tool. MST is developed based on the object-oriented concept by use of a library implemented by JAVA language. The features from the viewpoints of programming are:

1. executable on all-most all operating system
2. easy to modify user interface for the addition and the deletion of feature keys and qualifiers
3. each module is also executable as an application by minor modification of source code

Fig. 2 is sample windows of annotation editor on MST.

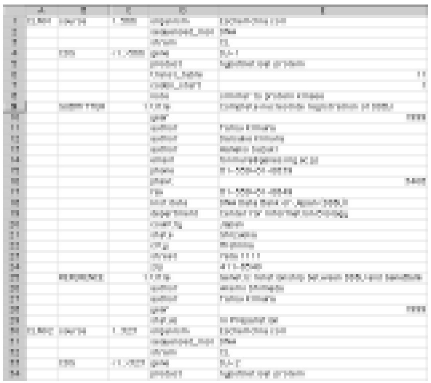

Figure 1: The format of Annotation file

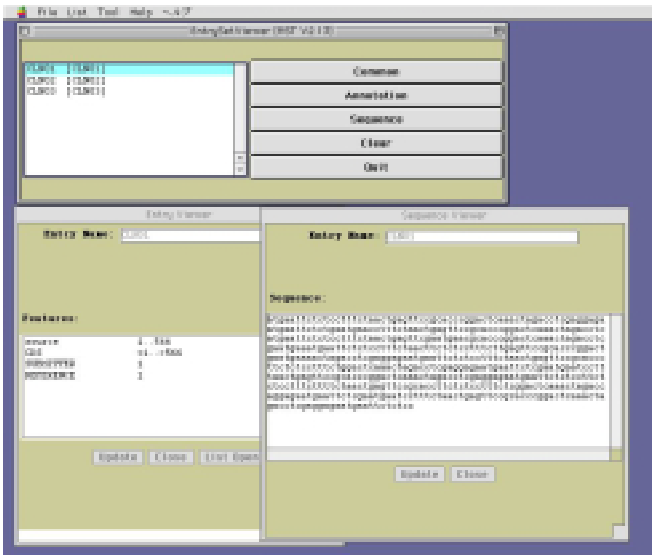

Figure 2: Sample windows on MST

References

[1] Sugawara, H., Miyazaki, S., Gojobori, T. and Tateno, Y., DNA Data Bank of Japan dealing with large-scale data submission, *Nucleic Acids Res.*, 27(1):25–28, 1999.

[2] Tateno, Y., Fukami-Kobayashi, K., Miyazaki, S., Sugawara, H. and Gojobori, T., DNA Data Bank of Japan at work on genome sequence data, *Nucleic Acids Res.*, 26(1):16–20, 1998.

[3] Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H. and Gojobori, T., DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams [In Process Citation], *Nucleic Acids Res.*, 28(1):24–26, 2000.