Constructing and Annotating GENES Database in KEGG

Susumu Goto [1]
goto@kuicr.kyoto-u.ac.jp
Kotaro Shiraishi [2]
kshirais@fqs.fujitsu.co.jp
Kayo Okamoto [1]
kayo@scl.kyoto-u.ac.jp
Hiroko Ishida [1]
hiroko@scl.kyoto-u.ac.jp
Sanae Asanuma [1]
sanae@scl.kyoto-u.ac.jp
Hidemasa Bono [1]
bono@kuicr.kyoto-u.ac.jp
Hiroyuki Ogata [1]
ogata@kuicr.kyoto-u.ac.jp
Wataru Fujibuchi [1]
wataru@kuicr.kyoto-u.ac.jp
Minoru Kanehisa [1]
kanehisa@kuicr.kyoto-u.ac.jp

[1] Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan
[2] Fujitsu Kyushu System Engineering Limited, 2-2-1 Momochihama, Sawara-ku, Fukuoka, 814-8589, Japan

1 Introduction

The KEGG (Kyoto Encyclopedia of Genes and Genomes) project is accumulating and computerizing vast knowledge of biochemistry and molecular biology. In addition to the PATHWAY database, the GENES database is a basic element of the KEGG system. It describes gene IDs, gene names, product names, chromosomal positions, gene classifications, EC numbers, codon frequencies, amino acid sequences and nucleotide sequences of over 20 species mainly with complete genomes.

Since the GENES database provides links between genomes and pathways, it can be an important tool for functional genomics, where information of both genomes and pathways should be utilized. It can also be used for comparative genomics, where whole genomes from several species are analyzed at a time. A major problem in such analyses is that there is no standard nomenclature for representing genes and gene products. The names of the genes and gene products are often different between species even if the products have the same function. The discrepancies are conspicuous when the genomes are sequenced by different organizations. The construction of the GENES database is an effort to fill the gap and to provide a standardized resource for functional and comparative genomics.

Last year, we reported main concepts of the GENES database [4]. This paper focuses on how the GENES database is constructed and how gene annotations are made to maintain consistent information among species.

2 Constructing GENES Database

The complete genome sequences are deposited in the GenBank database and distributed in a flat file format. We developed a parser program to extract information of genes from the GenBank flat files. It extracts gene IDs, chromosomal positions, names, products, EC numbers, amino acid sequences and nucleotide sequences. Because the format of the field representation often differs from species to species (i.e., from authors to authors) or even within species (i.e., containing variations and exceptions), the parser allows the user to specify, without changing the program, where and how the information should be extracted. It also computes codon frequencies of each gene and the whole genome.

At this stage, the annotation is not good enough. Many genes remain hypothetical even if they have similarity to known genes, and the majority of enzymes are not assigned EC numbers. This is partly because different authors adopt different criteria for interpretation of significant homologies and because some organizations do not invest much on informatics efforts of functional annotation. We

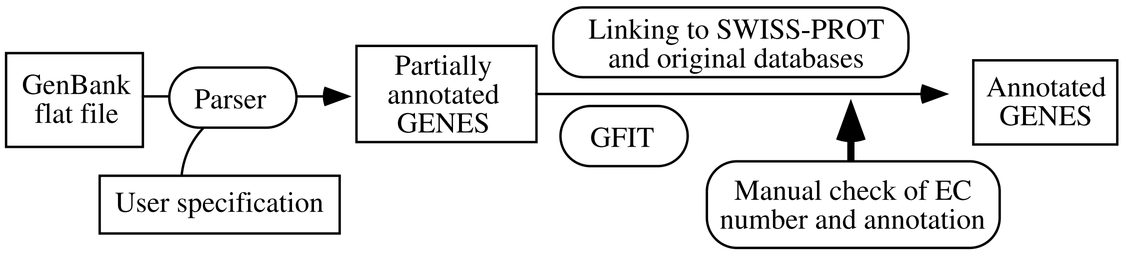

Figure 1: Construction flow of the GENES database.

find numerous cases where GenBank annotations are insufficient. In order to facilitate our annotation process, the GENES database is then integrated as part of the DBGET/LinkDB system [2] and it is also linked to the genome map browser of the KEGG system [3].

3 Annotating Genes in the GENES Database

The EC number assignment of GENES is performed through amino acid sequence similarity search against the whole genome sequences and SWISS-PROT by GFIT program [1]. GFIT reports possible functions by trying to determine orthologous genes and automatically assigns EC numbers from the orthologous genes with known functions. Another program is used to make links from each entry of GENES to SWISS-PROT and original genome databases, such as TIGR microbial database, by checking species and gene names described in both databases. The outputs of the latter program are checked manually and annotations are refined. Then the links to the PATHWAY database [3] and other databases are created automatically. The flow of the GENES database construction is summarized in Fig. 1. Currently we are developing tools for storing and manipulating the information of GENES in a relational database, and for providing interactive annotation of genes via a WWW interface.

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Science', from the Ministry of Education, Science, Sports and Culture of Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Bono, H., Ogata, H., Goto, S. and Kanehisa, M., Prediction of enzyme genes in complete genomes by reconstruction of metabolic pathways, *Genome Res.*, 8:203–210, 1998.

[2] Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y. and Kanehisa,M., DBGET/LinkDB: an Integrated Database Retrieval System, *Pac. Symp. Biocomput. '98*, 683–694, 1998.

[3] Kanehisa, M., A database for post-genome analysis, *Trends Genet.*, 13:375–376, 1997.

[4] Sato, K., Katsurada, T., Kimura, Y. and Kanehisa, M., Integrated GENES Database in KEGG, *Genome Informatics 1997*, 334–335, 1997.