Ortholog Identifiers for Integration of Genomic and Pathway Information 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

Toshi Nakatani [1]
toshi@scl.kyoto-u.ac.jp

Tomoko Deno [1]
tomoko@scl.kyoto-u.ac.jp

Minoru Kanehisa [1]
kanehisa@kuicr.kyoto-u.ac.jp

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

1 Introduction

KEGG (Kyoto Encyclopedia of Genes and Genomes) [1] is an integrated database system for reconstruction of systemic functional behaviors of the cell, which are represented by networks of interacting molecules, from the complete set of genes in the genome. The KEGG/PATHWAY database contains reference knowledge on the actual networks of interacting molecules in various cellular processes, such as metabolism, membrane transport, signal transduction, cell cycle, and apoptosis. The KEGG/GENES database is a collection of gene catalogs and associated sequence data for all the completely sequenced genomes and some partial genomes including human and mouse. The reconstruction is considered as a process of matching nodes in the two graphs: the reference pathway and the genome. For example, genes in the genome are matched against proteins (enzymes) in the reference metabolic pathway to reconstruct an organism-specific pathway. The automatic reconstruction of metabolic pathways has been successfully implemented in KEGG because the common identifiers, namely the EC numbers, can be utilized for the matching process. We report here the introduction of the ortholog identifiers as an extension of the EC numbers for automatic reconstruction of both metabolic and regulatory pathways.

2 Limitations of the EC numbers

An obvious limitation of the EC numbers is that they can be assigned only to a specific class of genes or proteins involving metabolism. In addition, when relating EC numbers to gene functions, there are complications due to the fact that the EC number is a representation of a chemical reaction rather than a molecular entity. For example, a single reaction may be catalyzed by an enzyme complex, in which case the same EC number is assigned to different subunits without distinguishing different functional roles. Conversely, a single gene product may catalyze multiple reactions, in which case different EC numbers have to be assigned to multiple domains. The EC numbering system is hierarchical, but the hierarchy is not unequivocal. Some numbers are assigned to a broad class of reactions, while others are very specificically defined. When the same reaction is viewed differently, a different EC number is sometimes more appropriate. Due to these complications, the EC numbering requires some modifications when it is used to link with genomic information of enzyme genes.

3 Ortholog identifiers

An ortholog is defined in KEGG as the functional equivalence, namely the same node, in the network of interacting molecules. In practice it is identified by sequence similarity with consideration of additional information, such as the positional coupling of genes on the chromosome. It is tempting to pre-define a functional hierarchy of genes as in Riley's *E. coli* gene classification (http://genprotec.mbl.edu/start) or in Gene Ontology by FlyBase/MGD/SGD consortium (http://www.ebi.ac.uk/~ashburn/GO/). However, we do not start from a pre-defined hierarchy of genes. Because the hierarchy is inherent in the pathway information, it is automatically introduced to the genomic information as the result of mapping each ortholog to a pathway node.

For the metabolic pathways, the ortholog identifier is a natural extension of the EC number, distinguishing different subunits in an enzyme complex or isozymes that function under different conditions. For example, the ortholog identifier for the two subunits of succinyl-CoA synthetase are: E6.2.1.5A and E6.2.1.5B. For various regulatory pathways and assemblies, we try to use standard mnemonics as the ortholog identifiers. For example, a ribosomal protein subunit is represented by: RP-L1. It must be noted that some orthologs are conserved among all species while others are valid only within a limited number of species. Such a distinction is introduced in the reference pathway diagram to be matched, which is associated with the information on the applicable group of species.

4 KEGG annotation tool

The information about gene annotations is stored in a Sybase relational database in KEGG. It contains descriptions of gene functions according to the GenBank database (original annotation by the authors), the on-line genome databases which may be more up-to-date, and the SWISS-PROT database, as well as additional annotations by KEGG based on the ortholog identification and pathway reconstruction. We have been developing a Web based annotation tool for the Sybase database [2]. A most unique feature of this annotation tool is its capability to simultaneously annotate a group of genes across species. Once a unique ortholog identifier is given to the group, the standard description of the ortholog may be entered in the DEFINITION field of each entry in the group. The use of the annotation tool, hence the update of the relational database, is limited to in-house KEGG annotators, but the updated information is reflected on the following day in the publicly available version of the KEGG/GENES database.

Acknowledgements

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

References

[1] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res.*, 27:29–34, 1999.

[2] Goto, S., Shiraishi, K., Okamoto, K., Ishida, H., Asanuma, S., Bono, H., Ogata, H., Fujibuchi, W., and Kanehisa, M., Constructing and annotating GENES database in KEGG, *Genome Informatics*, 9:226–227, 1998.