Genome Scale Prediction of Enzyme Genes Utilizing the Knowledge of Metabolic Interactions

Hidemasa Bono
bono@kuicr.kyoto-u.ac.jp
Hiroyuki Ogata
ogata@kuicr.kyoto-u.ac.jp
Susumu Goto
goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611, Japan

Abstract

Thanks to various genome projects, genome scale protein sequence data have become available. Predicting gene locations and gene functions by computational methods is a very important stage in the genome project. We are developing a systematic method of predicting enzyme genes for an organism with its entire genomic sequence known, utilizing the knowledge organized in KEGG (Kyoto Encyclopedia of Genes and Genomes) [1][2]. The system can also help reconsider the functional assignments of the genomes previously determined and the gene catalog can be fixed accordingly.

1 Introduction

Genome sequencing has become a reality. The first complete genome of a free-living organism, *Haemophilus influenzae*[3], was determined in 1995, which would be followed by the explosion of complete genomic sequences and complete gene catalogs of many organisms, from Bacteria and Archaea[4] to Eucarya. The prediction of gene locations and gene functions is a crucial problem when sequencing is completed. The functional assignment of predicted genes, or open reading frames (ORFs), is usually done by searching sequence similarities and/or sequence motifs against the databases[5]. This assignment is done separately for each gene. As the result, many ORFs are left without any functional assignments and multiple ORFs can be assigned to the same and similar functions. We are developing a system of assigning enzyme genes by comparing all the ORFs of a genome against all known enzymes in different organisms stored in KEGG. The system examines the completeness of the organism-specific pathways formed as well as the completeness of the enzyme catalog derived, which is the feature not incorporated in the other gene prediction systems.

2 System and Methods

The data sets of protein sequences were taken from the original sources for the organisms of *Escherichia coli, Haemophilus influenzae*[3], *Bacillus subtilis, Mycoplasma genitalium*[6], *Methanococcus jannaschii*[4], *Synechocystis* sp.[7] and *Saccharomyces cerevisiae*. Enzymes were identified either by the definition given or by the similarity search against the sequence databases, and the EC numbers were assigned accordingly. This resulted in a table of EC numbers with known sequences for the seven organisms. When the gene catalog of a new organism becomes available, this table can be used for identifying enzymes and associated EC numbers. Furthermore, the EC number assignment in the table itself can be checked by taking out one organism and by comparing against the rest of the table.

Next, utilizing the knowledge stored in KEGG, we examined whether the enzyme genes identified constituted metabolic pathways correctly. When a pathway is interrupted because of missing enzymes, it is necessary to re-examine the assignment of enzyme genes. We performed the re-examination of the EC number assignment table and the search of additional enzymes for each organism.

After mapping of enzyme genes to the functional map of metabolic pathways, there still remained a number of enzymes in the EC number assignment table. We examined if any new pathways could be formed from the enzymes that had not been mapped.

Acknowledgment

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 in Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Ogata, H., Bono, H., Fujibuchi, W., Goto, S., Kanehisa, M., *Proc. of the Seventh Workshop on Genome Informatics*, this issue.

[2] Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K., Kanehisa, M., *Pacific Symposium on Biocomputing '97*, (1997) in press.

[3] Fleischmann, R.D. *et.al. Science* **269**, 496-512, (1995)

[4] Bult, C.J. *et.al. Science* **273**, 1058-1073, (1996)

[5] Koonin, E.V., Tatusov, R.L, Rudd, K.E. *Proc. Natl. Acad. Sci. USA*, **92**, 11921-11925, (1995)

[6] Fraser, C.M. *et.al Science* **270**, 397-403, (1995)

[7] Kaneko, T. *et.al. DNA Research* **3**, 109-136, (1996)