Construction of Ortholog/Paralog Group Table for PTS

Takeaki Taniguchi takeaki@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp
Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611-0011, Japan

1 Introduction

The identification of the functions for all proteins in the genome is important to understand the features of the organism. With the increasing amount of complete genome sequences, the approaches based on comparative genomics have become particularly useful. There are attempts to automatically construct the ortholog table containing orthologous relations of genes in different organisms, but the complexity of organisms often requires manual efforts to add or remove specific cases and to improve the quality of data. In order to facilitate the process of constructing ortholog tables we have been focusing on specific aspects of protein functions [1, 2] rather than trying to cover the entire spectrum. Here we present a new addition of the ortholog/paralog group table for PTS.

The phosphoenolpyruvate:carbohydorete phosphotransferase systes (PTSs) are both transport and sensing systems in gram-negative and gram-positive bacteria. They take in and phosphorylate a large number of carbohydrates, and play as signal transducers to move to these carbon sources. In general they have unique gene structures in the genome. The PTS comprises 5 or 6 components (EI, HPr, EIIA, EIIB, EIIC, EIID) where EI and HPr are common components and there are multiple components (paralogs) of EIIs depending on substrates. The EII components often form gene clusters in the genome and some of them (EIIA, EIIB, EIIC) sometimes fuse into one protein. There are many variations in the order of the components and the fusion protein in the gene cluster.

In contrast to the bacterial ABC transport system [1], the PTS contains a large number of "fusion of components" that makes it difficult to cluster genes for functional grouping and to perform multiple alignment for identifying functional residue. In order to cope with this difficulty, the fusion proteins were divided into the components by the homology search against our collection of known components.

2 Method

(1) The sequences of PTS enzymes were extracted from SWISSPROT.

(2) Then the database of the sequences of each component (EI, HPr, EIIA, EIIB, EIIC, EIIAB, EIIABC, EIIBC) was constructed.

(3) The fusion sequences of EIIAB,EIIABC,EIIBC were aligned as queries with the sequences of EIIA,EIIB,EIIC by ssearch3 and the locally aligned regions were taken as candidates of the components.

(4) Each component region in (3) was confirmed whether correct or not with a tool that visualizes the alignment (Fig. 1). The criterion to determine a component region is based on the overlap lengths, the alignment scores, and the existence of PTS motifs. After manual inspection, the component sequences were in the database of EIIA, EIIB, EIIC.

(5) All the protein sequences of the complete genome in KEGG [3] were used as queries to search the sequences of EIIA,EIIB,EIIC by ssearch3.
To the candidate sequences obtained by this query the steps (3) and (4) were applied again.

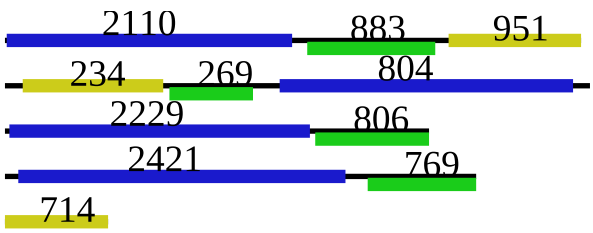

Figure 1: The visualizing of the results of homology search for *Escherichia coli.* The bold lines are aligned regions and the numbers over them are the scores of the homology search. The darkness indicates the kinds of the components.

(6) The clustering of the sequences in the EIIA,EIIB,EIIC database was done for each component.

(7) According to the grouping from the result of the clustering and by examining the gene clusters and fusion proteins, the ortholog/paralog group table was constructed.

3 Result

The number of PTS proteins extracted from SWISSPROT was: 17 for EI, 15 for HPr, 27 for EIIA, 13 for EIIB, 12 for EIIC, 4 for EIID, 1 for EIIAB, 24 for EIIBC, and 21 for EIIABC. In the complete genomes in KEGG the number of PTS proteins found were: 53 in *Escherichia coli*, 6 in *Haemophilus influenzae*, 30 in *Bacillus subtilis*, 5 in *Mycoplasma genitalium*, 8 in *Mycoplasma pneumoniae*, 13 in *Borrelia burgdorferi*, and 1 in *Methanobacterium jannaschii.* As the result of the clustering EI and HPr fall in one group, EIIA in 3 groups, EIIB in 5 groups, and EIIC in 6 groups.

Acknowledgements

This work was supported in part by the Grant-in-Aid for Scientific Research on the Priority Area 'Genome Science' form 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] Tomii, K. and Kanehisa, M.,A comparative analysis of ABC transporters in the complete microbial genome, *Genome Research*, in press, 1998.

[2] Bono, H., Goto, S., Fujibuchi, W., Ogata, H. and Kanehisa, M., Systematic prediction of orthologous units of genes in the complete genomes, *Genome Informatics 1998*, Universal Academy Press, 1998.

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