A Survey on *E. coli* Enzymes: Correlation between Metabolic Pathway and Gene Location

Hiroyuki Ogata and Minoru Kanehisa

Deciphering the meaning of the gene location on the chromosome is one of the basic demands of molecular biology. Here we report a correlation between the position of enzymes on the metabolic pathways and the locality of the genes on the chromosome. We performed a window base search to identify functionally related enzymes coding segments (FRECS) for the *Escherichia coli* chromosome. Among 52 FRECS identified, 32 (~60%) were related to the known operons. It is also suggested that gene duplication has no connection to the observed correlation.

Keywords: Function and structure / Metabolism / Genome / Chromosome / Operon / Gene duplication

What function is coded in the gene location along the chromosome? Many examples in which the gene location has critical roles for the expression of gene functions have emphasized the importance of this basic question on genome structure and evolution. However, the recent completion of sequencing the bacterial gemomes and the extensive comparisons of gene order revealed that gene location is highly shuffled in the course of bacterial evolution and that the gene order is not a conservative entity except for the short-range colinearity (1). Although this striking feature of the gene disposition could be interpreted as the absence of functional role of the long-range gene order, direct comparisons of the gene locations with their functions must be necessary before deriving the answer to the question.

When this kind of questions are addressed, a major difficulty is in the definition of gene function. A recently developed database by Kanehisa *et al*. dealing with biological pathways provides excellent opportunities for the investigation of gene functions (2,3). The database named KEGG (Kyoto Encyclopedia of Genes and Genomes) stores wide coverage of the known metabolic pathways in a computable form called "binary relations", which enabled us to introduce a new measure to capture the degree of functional link between two enzymes (4).

In this study, we performed a statistical analysis on the relation between the distance of *Escherichia coli* enzyme genes along the chromosome and the functional link measured by the shortest path length between these enzymes on the metabolic pathways.

The information of *E. coli* metabolic pathways, operons

MOLECULAR BIOLOGY AND INFORMATION —Biological Information Science—

Scope of research

This laboratory aims at developing theoretical frameworks for understanding the information flow in biological systems in terms of genes, gene products, other biomolecules, and their interactions. Toward that end a new deductive database is being organized for known molecular and genetic pathways in living organisms, and computational technologies are being developed for retrieval, inference, and analysis. Other studies include: functional and structural prediction of proteins from sequence information and development of sequence analysis tools.

Prof
KANEHISA,Minoru
(D Sc)

Instr
GOTO, Susumu
(D Eng)

Instr
FUJIBUCHI,
Wataru

Instr
OGATA, Hiroyuki

Research Fellows
SATO, Kazushige
SUGIYAMA, Yukiteru (D Agr)

Students
TOMII, Kentarou (DC);
SUZUKI, Kenji (DC);
KIHARA, Daisuke (DC);
BONO, Hidemasa (MC);
HATTORI, Masahiro (MC);
KAWASHIMA, Shuichi (MC);
PARK, Keun-joon (MC);
IGARASHI, Yoshinobu (MC);
KATAYAMA Toshiaki (MC)

and the protein sequences of the enzymes were taken from KEGG. The genomic map positions of the enzymes were from EcoCyc (5). When an enzyme represented by an EC number is related to multiple genes as in the case of a multi-component enzyme, we omitted it from the results for the simplicity of the assignment on the genomic map.

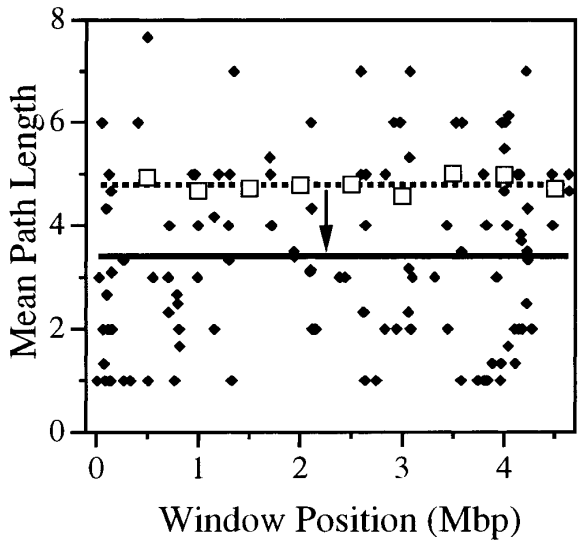

Figure 1. The mean path length of enzyme pairs is plotted for different sizes of the search window. White boxes and black boxes are those for 1 Mbp and 10 Kbp size window, respectively. Dotted and solid lines show the avarage levels of the path length.

Investigation here focused on whether the functional link of the enzymes has any relation to the physical distance of the genes on the *E. coli* chromosome. To this purpose, we performed a window search for the identification of the sets of enzymes coded on close positions along the chromosome, and calculated the mean path length of enzyme pairs appearing in the window. In Figure 1, the mean path length was plotted for two different window sizes. Interestingly, when we changed the window size from 1 Mbp to 10 Kbp, the path length was decreased, on the average, from 4.8 (10 samples) to 3.4 (139 samples) as indicated by the arrow in the Figure 1. If the enzymes appearing in close positions on the metabolic pathways are coded randomly on the chromosome, the average values of path length are expected to be the same for different window sizes. We considered that this correlation would be due to the existence of relatively short (~10 Kbp) DNA segments that encode enzymes playing their rolls at close position on the metabolic pathways. Here we call these segments functionally related enzymes coding segments (FRECS). After merging overlapped 10 Kbp segments, we obtained 52 FRECS with the mean path length less than 4. We examined these segments in terms of two possibilities that might explain this observation.

First, operon enzymes were examined. If we omitted operon enzymes from the above analysis, the average value of the path length raised from 3.4 to 4.2, and the number of the FRECS mentioned above decreased from 52 to 20. The enzymes coded in these twenty FRECS were 45 in all.

Second, we examined the chromosomal distance between 64 paralogous enzyme pairs identified by the BLASTP search (6). In the previous study, we revealed that the paralogous enzymes often play their rolls at close positions in the metabolic pathways (4). Thus we expected that this heterogeneity of the *E. coli* metabolic pathways might be reflected on the disposition of these duplicated enzymes. However, there were no paralog pairs in any of the 52 FRECS.

In conclusion, we identified 52 DNA segments that code enzymes appearing in close position on the metabolic pathways. Most of these segments (60%) were related to the operon enzymes. These segments contributed to the apparent correlation between the metabolic pathway and the gene location (Fig. 1). Further investigation on the remaining 20 DNA segments will elucidate the meaning of the feature that connects the *E. coli* chromosome and metabolic pathways .

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

References

1. Mushegian AR and Koonin EV, *Trends Genet.*, **12**, 289-290 (1996).
2. Kanehisa M, *Science & Technology Japan*, No. 59, pp. 34-38 (1996).
3. Goto S, Bono H, Ogata H, Fujibuchi W, Nishioka T, Sato K and Kanehisa M, *Proc. Pacific Symposium on Biocomputing '97*, pp. 175-186 (1996).
4. Ogata H, Bono H, Fujibuchi W, Goto S and Kanehisa M, *Proc. 7th Workshop on Genome Informatics*, pp. 128-136 (1996).
5. Karp PD, Riley M, Paley SM and Pelligrini-Toole A, *Nucleic Acids Res.*, **24**, 32-39 (1996).
6. Altschul SF, Gich W, Miller W, Myers EW and Lipman DJ, *J. Mol. Biol.*, **215**, 403-410 (1990).