Systematic Prediction of Orthologous Units of Genes in the Complete Genomes

Hidemasa Bono bono@kuicr.kyoto-u.ac.jp
Susumu Goto goto@kuicr.kyoto-u.ac.jp
Wataru Fujibuchi wataru@kuicr.kyoto-u.ac.jp
Hiroyuki Ogata ogata@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

Abstract

In order to fully make use of the vast amount of information in the complete genome sequences, we are developing a genome-scale system for predicting gene functions and cellular functions. The system makes use of the information of sequence similarity, the information of positional correlations in the genome, and the reference knowledge stored as the ortholog group tables in KEGG (Kyoto Encyclopedia of Genes and Genomes). The ortholog group table summarizes orthologous and paralogous relations among different organisms for a set of genes that are considered to form a functional unit, such as a conserved portion of the metabolic pathway or a molecular machinery for the membrane transport. At the moment, the ortholog group table is constructed for the cases where the genes are clustered in physically close positions in the genome for at least one organism. In this paper, we describe the system and the actual analysis of the complete genome of *Pyrococcus horikoshii* to identify ABC transporters.

1 Introduction

While an increasing number of complete genome sequences has become publically available, the biological function of roughly a half of the genes in each genome remains unknown Thus, efficient methods still need be developed to annotate functional properties for the entire set of predicted open reading frames (ORFs). The popular method that is widely used for functional annotation relies on searching for sequence similarity or motifs in the database. When a new genome is sequenced, the amino acid sequences of translated ORFs are usually searched independently against the non-redundant protein sequence database after the ORFs are determined by some gene finding methods. There are several problems in this approach. First, a proper threshold value cannot be predetermined to extend sequence similarity to functional similarity. Second, because some genes, such as for ABC transporters, have many homologous genes in the genome, it is difficult to assign orthologous relations that can be used to specify functions. Third, the so-called non-redundant database actually contains many duplicate entries and the similarity search against it often produces a long list of similar sequences that is not easy to process.

Thus, it is, first of all, desirable that additional information is incorporated to make it easier to interpret the result of similarity searches. Especially, the positional correlation in the genome, e.g., the operon structure, has turned out to be extremely useful information in the functional annotation of bacterial and archaeol genomes. Furthermore, it is necessary to develop a clean data set that can be used as reference for the functional annotation process.

We started the KEGG (Kyoto Encyclopedia of Genes and Genomes) project in 1995. It aims to make links from the gene catalogs generated by the genome sequencing projects to the biochemical pathways that may be considered wiring-diagrams of genes and molecules [5]. Under the project we

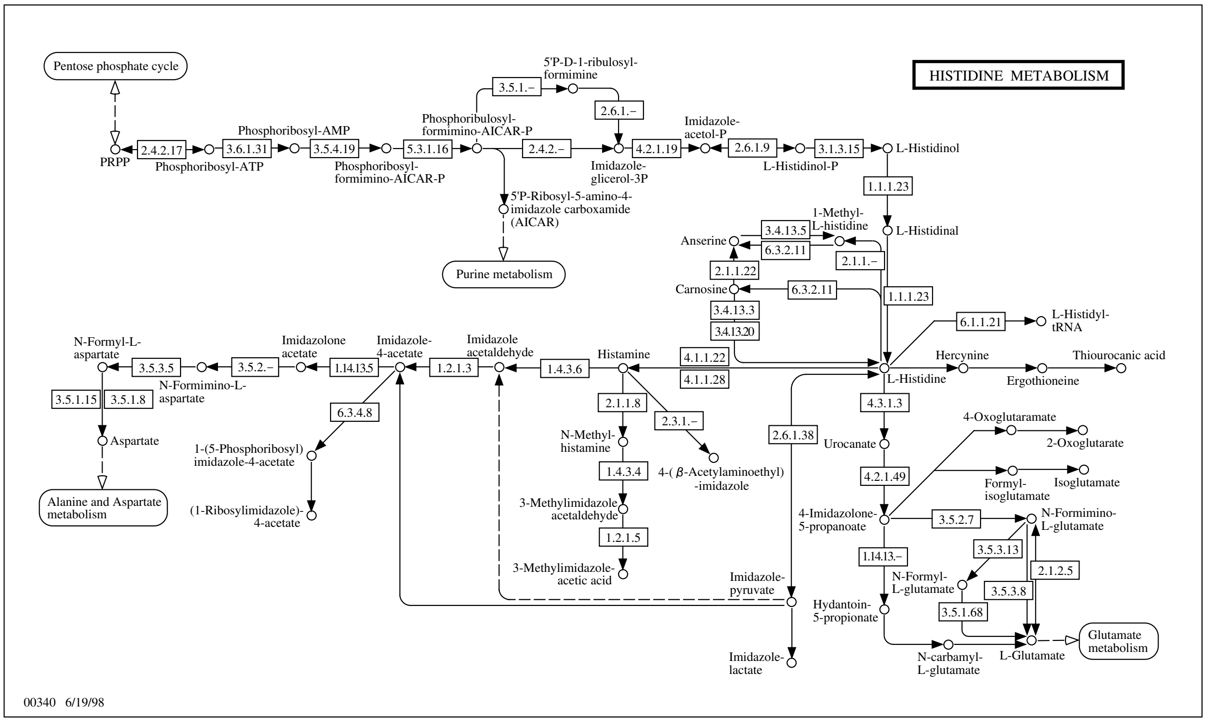

Figure 1: The KEGG standard pathway map for histidine metabolism.

have been developing a system for identifying gene functions, called GFIT (Gene Function Identification Tools), utilizing the orthologous relations among genes of the query organism and those of reference organisms [2]. By analyzing the actual genomes with GFIT, curated orthologous relations are stored in, what we call, the ortholog group tables in KEGG for units of genes that are functionally correlated.

We present here a new component of GFIT, a web-based tool that searches sequence similarity against the ortholog group tables and that considers the positional information of the query genome. It accepts the whole genome, i.e., the complete set of ORF sequences, and the result is presented by integrating with the KEGG resources. Thus, the tool can be used as a workbench for functional annotation of the whole genome.

2 Data and Methods

2.1 KEGG

Details of the KEGG systems are described elsewhere [8]. We mention here brief overview of KEGG system. KEGG is an attempt to computerize the knowledge of the information pathways of interacting biomolecules. Typical pathways include metabolic pathways, and an example of how KEGG represents the metabolic pathway information is shown in Fig. 1 for histidine metabolism. KEGG also attempts to computerize various regulatory pathways, for example signal transduction circuits, from the biological knowledge currently available.

The KEGG pathway maps are implemented as clickable GIF image maps to be used in the WWW. Thus the information about enzymes and compounds in the LIGAND chemical database [4], or genes in the KEGG gene catalogs can be retrieved through the in-house developed database management system, DBGET/LinkDB [3].

2.2 Identification of orthologs and paralogs

The so-called homologs of genes that share sequence similarity could be due to the following two mechanisms. Orthologs are the genes that are derived by a common ancestry; hence they are responsible for the identical function in different organisms. In contrast, paralogs are generated by gene duplications and in general have similar but not necessarily the same function.

With the availability of complete genomic sequences, practical procedures to distinguish orthologs and paralogs were proposed [7, 10]. Given two complete lists of genes, the amino acid sequence similarity is examined for each gene in one organism against all genes in the other organism. Only the gene pairs that show the similarity of statistical significance are to be considered. If the two genes, gene A in organism 1 and gene B in organism 2 are more closely related to each other than to any other genes it can be paired with, we define that gene A and gene B are orthologous. Of course, this is an operational definition of orthologs, and there may be complications resulting from the existence of high scoring paralogs within each organism, from the existence of multidomain proteins, and also from the inconsistencies of pairwise comparisons when multiple organisms are considered.

In KEGG the functional annotation of each gene in each organism is maintained in the GENES database [8]. For a newly sequenced organism, the EC number assignment for enzyme genes is made manually according to the orthologous relations identified by comparing against all organisms in the GENES database. The gene function annotations are continuously re-evaluated in KEGG by comparing with the KEGG/PATHWAY database, SWISS-PROT, and other databases. Consequently, the EC number assignment is also continuously updated.

2.3 Ortholog and paralog group tables

The sequence similarity search against the existing sequence databases, even against non-redundant database, often generates a long list of hits, which requires human efforts to find orthologous relations that can be used for gene function assignments. The ortholog group tables in KEGG are curated reference data set of orthologous relations that is intended to make this process easier. These tables also contain the information of the group of genes that is supposed to form a functional unit, such as a regulatory unit in the metabolic pathway or a molecular unit of assembly. The data representation of the ortholog group table is a simple HTML table, in which additional information, such as hyperlinks, can easily be added.

We are working to maintain and expand the ortholog group tables. As of August 1998, there are 53 tables, which are manually edited from biological viewpoint. The ortholog group tables listed in Table 1 are largely categorized into two groups. One is for the tables from metabolic pathways, and the other is for those from regulatory pathways. Fig. 2 shows the ortholog group table of histidine metabolism, which contains orthologous genes extracted from the pathway map in Fig. 1 for different organisms. In Fig. 2, the shaded cells in the same row represent genes that are closely located in the chromosome. It is supposed that they form an operon. In this figure, we can easily see that *E.coli, H.influenzae,* and *B.subtilis* have operon structures in the histidine metabolism pathway, but the other species may not.

The tables for the metabolic pathways contain well conserved sections of the pathway, which may be called pathway motifs, that are generated by the SIMIC (Simultaneous Linkage Clustering) program (Ogata, H et al., manuscript in preparation) for identifying correlated clusters of genes in the genome and the gene products in the pathway. The region is named functionally related enzyme clusters (FRECs), and it contains an operon-like structure of genes that codes for a unit of related enzymes in the pathway.

In contrast, the tables for the regulatory pathways are mostly collected by human efforts. The best organized ones at the moment are for the ABC transporters [9] and the two-component signal transducers [1] that often form large paralogous gene clusters. Other genes concerning cell processes and cell organization are also catalogized in the tables.

Table 1: List of ortholog group tables.

- Metabolism
  - Carbohydrate Metabolism
    - Glycolysis / Gluconeogenesis
    - Citrate cycle (TCA cycle)
    - Pentose Phosphate Cycle
    - Pentose and Glucuronate Interconversions
    - Fructose and Mannose Metabolism
    - Galactose Metabolism
    - Ascorbate and Aldarate Metabolism
    - Pyruvate Metabolism
    - Glyoxylate and Dicarboxylate Metabolism
    - Propanoate Metabolism
    - Butanoate Metabolism
  - Energy Metabolism
    - Methane Metabolism
    - Nitrogen Metabolism
    - Sulfur Metabolism
  - Lipid Metabolism
    - Fatty Acid Biosynthesis (Path 1)
  - Nucleotide Metabolism
    - Purine Metabolism
    - Pyrimidine Metabolism
    - Nucleotide Sugars Metabolism
    - Aminosugars Metabolism
  - Amino Acid Metabolism
    - Glutamate Metabolism
    - Alanine and Aspartate Metabolism
    - Glycine, Serine and Threonine Metabolism
    - Methionine Metabolism
    - Cysteine Metabolism
    - Valine, Leucine and Isoleucine Degradation
    - Valine, Leucine and Isoleucine Biosynthesis
    - Lysine Biosynthesis
    - Arginine and Proline Metabolism
    - Histidine Metabolism
    - Phenylalanine Metabolism
    - Phenylalanine, Tyrosine and Tryptophan Biosynthesis
    - Urea Cycle and Metabolism of Amino Groups
  - Metabolism of Other Amino Acids
    - beta-Alanine Metabolism
  - Metabolism of Complex Carbohydrates
    - Starch and Sucrose Metabolism
    - Peptideglycan Biosynthesis
  - Metabolism of Complex Lipids
    - Glycerolipid Metabolism
    - Metabolism of Cofactors, Vitamins, and Other Substances
    - Thiamine Metabolism
    - Nicotinate and Nicotinamide Metabolism
    - Biotin Metabolism
    - Folate Biosynthesis
    - One Carbon Pool by Folate
    - Porphyrin and Chlorophyll Metabolism
    - Ubiquinone Biosynthesis
  - Metabolism of Macromolecules
    - Aminoacyl-tRNA Synthetase
- Cell Process
  - Membrane Transport
    - ABC Transporters
    - PTS System
  - Signal Transduction
    - Two-Component System
  - Ligand-Receptor Interaction
    - G-protein coupled receptors
- Cell Organization
  - Molecular Assembly
    - Ribosome assembly
    - F1F0-ATPase
  - Molecular Components
    - Translation Factors

Ortholog/Paralog Groups in Histidine Metabolism

Probable operons are represented by color.

[P..Pathway map | G..Genome map | T..Title list]

Organism	2.4.2.17	3.6.1.31	3.5.4.19	5.3.1.16	2.4.2.-	4.2.1.19	3.1.3.15	2.6.1.9	1.1.1.23
	ATP phospho-ribosyl-transferase	phospho-ribosyl-ATP pyrophospho-hydrolase	phospho-ribosyl-AMP cyclo-hydrolase	phosphoribosyl-formimino-5-aminoimidazole carboxamide ribotide isomerase	amidotransferase	imidazole-glycerol-phosphate dehydratase	histidinol-phosphatase	histidinol-phosphate aminotransferase	histidinol dehydrogenase
eco [P | G | T]	b2019(hisG)	b2026(hisI)		b2024(hisA)	b2023(hisH)	b2022(hisB)		b2021(hisC)	b2020(hisD)
hin [P | G | T]	HI0468	HI0475		HI0473	HI0472	HI0471		HI0470	HI0469
bsu [P | G | T]	hisG	hisI		hisA	hisH	hisB		hisC	hisD
aae [P | G | T]	aq_1613		aq_1968	aq_1303	aq_181 aq_732	aq_039		aq_2084	aq_782
syn [P | G | T]	sll0900	slr0608		slr0652		sll0084 slr0500		sll1713 sll1958	slr0682 slr1848
mja [P | G | T]	MJ1204		MJ0302 MJ1430	MJ0703 MJ1532	MJ0411 MJ0506	MJ0698		MJ0955	MJ1456
mth [P | G | T]	MTH119 MTH1506		MTH245	MTH669 MTH843	MTH1343 MTH1524	MTH1467		MTH1587	MTH225
afu [P | G | T]	AF0590		AF1950				AF0985		AF0212
sce [P | G | T]	YER055C	YCL030C		YIL020C	YBR248C YM8021.09C	YOR202W	YFR025C	YIL116W	YCL030C

Organism	4.3.1.3	4.2.1.49	3.5.2.7	3.5.3.8
	histidine ammonia-lyase	urocanate hydratase	imidazolone-propionase	formimino-glutamase
bsu [P | G | T]	hutH	hutU	hutI	hutG

Last updated: July 16, 1998
Compiled by KEGG

Figure 2: The ortholog group table for histidine metabolism.

3 Genome-Scale Prediction of Biological Functions

3.1 New computational tool in GFIT

The genome-scale prediction of biological functions require a new generation of tools that examine a complete set of genes in the genome and to return functional prediction results after considering all dependencies. The initial version of the GFIT program provides one solution, where the program receives the entire set of ORFs in the genome as a query, compares against each of the completely sequenced organisms, and returns brief but informative results of similarities. As of August 1998, the complete genome sequences of 13 micro-organisms are available[1]. GFIT tentatively assigns orthologs of each ORF by the operation described in the Data and Methods section. Unfortunately, the automatic operation based on the bidirectional best hits is too strict and often misses real orthologous relations. This becomes obvious when assigning EC numbers by GFIT. The correctness of EC number assignment can be checked by whether the complete routes of metabolic pathways are properly reconstructed, i.e., whether any missing enzymes are present to make the pathway continuous [2].

With the availability of the clean data set of ortholog group tables, it is now possible to query the entire genome sequence for, at least, a selected aspects of biological functions. In the traditional similarity search of individual genes (or proteins) against repositories of non-redundant databases, it has always been problematic to determine an appropriate level of sequence similarity that can be extended to functional similarity. The program to search ortholog and paralog tables benefits from an additional feature that is used for interpretation of sequence similarity; namely, the requirement for reconstructing a complete functional unit from a set of genes or proteins. Utilizing this feature the functional inference can be better performed.

The program actually searches sequences in the ortholog group tables and reports the genes above a specified threshold. They can then be superimposed on the reference ortholog table with additional coloring showing the location and the degree of similarity.

3.2 Identification of ABC transporters

In this section, we show the result of using the new GFIT program. We performed the analysis of ABC transporters in newly sequenced bacterium, *Pyrococcus horikoshii* [6]. All ORFs of *Pyrococcus horikoshii* were searched against the ortholog (and paralog) group tables of ABC transporters [9].

Fig. 3 shows the top part of the whole result, in which the columns correspond to the three components of ABC transporters (binding protein, membrane protein, and ATP-binding protein) and the annotation in the original database (last column). The rows correspond to the ORFs of *Pyrococcus horikoshii* that have homology to at least one of these components. The numeral in each cell is the highest FASTA opt score between the ORF sequence and the database sequence, also showing to which components the similarity was found. The background color shows the percentage range of the database hits among paralogs. This representation of the result also contains the information about clustering of genes in the chromosome. The rows separated by thin lines are the genes that are located next to each other in the genome. The rows separated by thick lines are the genes that are apart. Therefore, a cluster of genes not separated by thick lines contain hits to all necessary components, then it is considered to be the functional unit of, in this case, the ABC transporter. One of the results we obtained is the cluster of ORFs from `pho:PHBC018` to `pho:PHBC015` (check boxes in Fig. 3).

A detailed picture of matches can be examined for these genes and Table 2 shows the summary of best hits (only the top three hits are indicated here) according to the FASTA opt scores. Because the database hits exist in the reference ortholog group table, they can be displayed by superimposing on the reference table. A portion of the superimposed table is shown in Fig. 4. Except for `b1123` all database hits are in the subgroup of 'Maltose / sn-Glycerol-3-phosphate' although the table of ABC transporter contains more than 250 gene clusters.

[1]`http://www.genome.ad.jp/kegg/java/org_list.html`

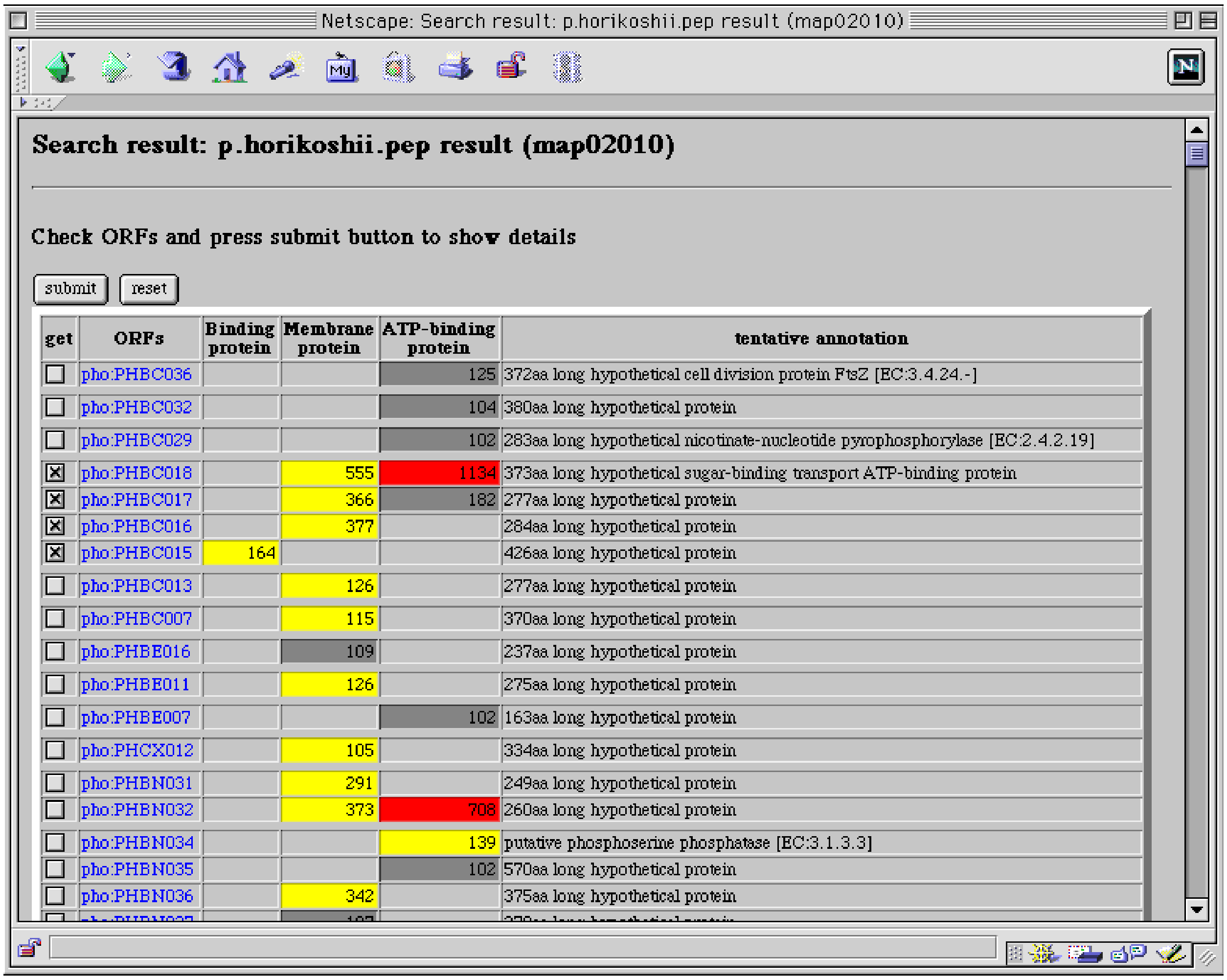

Figure 3: Genome scale identification of ABC transporters.

Table 2: A detailed content of the database hits.

ORF	database hits (opt scores) in ABC transporter ortholog group table			
pho:PHBC018	yurJ(1134)	slr0747(1081)	b3450(1049)	...
pho:PHBC017	yurM(366)	b1312(359)	slr0531(346)	...
pho:PHBC016	slr1202(377)	b1311(343)	yurN(330)	...
pho:PHBC015	b1123(164)	yurO(162)	b1310(140)	...

(Maltose / sn-Glycerol-3-phosphate)

Organism	Binding protein	Membrane protein	ATP-binding protein	Substrate
eco [P | G | T]	b4034(malE) b4037(malM)	b4033(malF) b4032(malG)	b4035(malK)	Maltose
eco [P | G | T]	b1310	b1311 b1312	b1318	
eco [P | G | T]	b3453(ugpB)	b3452(ugpA) b3451(ugpE)	b3450(ugpC)	sn-Glycerol-3-phosphate
bsu [P | G | T]	yurO	yurN yurM	yurJ	
mge [P | G | T]	MG186	MG188 MG189	MG187	
mpn [P | G | T]	E07_orf301	E07_orf329 E07_orf319	E07_orf586	
syn [P | G | T]		slr1202 slr1723	slr1224	
syn [P | G | T]		sll0530 sll0531	sll0747	

Figure 4: The portion of the ortholog group table used for functional prediction.

In the annotation by the original authors of *Pyrococcus horikoshii*, `pho:PHBC018` was tentatively assigned to be 'sugar-binding transport ATP-binding protein', but others were all left as hypothetical. We predict `pho:PHBC016` and `pho:PHBC017` are the membrane proteins and `pho:PHBC015` is the substrate binding protein. We also predict that the transporter is not for simple sugar (such as ribose and galactose) but for multiple sugar (maltose) or sn-glycerol-3-phosphate.

4 Summary and Perspective

KEGG organizes the knowledge of metabolic and regulatory pathways efficiently and usefully. The tool presented here is a first attempt to incorporate the information of well curated ortholog (and paralog) group tables and the information of chromosomal neighbors, as well as the information of sequence similarity, for functional prediction of ORFs. The KEGG ortholog group table representation is more informative than the KEGG pathway representation because it contains the positional information in the genome, it represents a multiple alignment of organisms, and it is far better curated in contrast to the automatically reconstructed pathway maps that contain many missing enzymes. However, the major drawback of the ortholog group table is that it covers only a small fraction of the pathway information that is present in KEGG. By comparative genomics, especially for identifying conserved gene gene clusters, we hope to identify more functional units that can be represented by the ortholog group tables.

Acknowledgments

This work was supported in part by the Grant-in-Aid for Scientific Research on the 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. Hidemasa Bono was supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

References

[1] Bono, H., Goto, S., Ogata, H., and Kanehisa, M., Genome scale prediction of two-component signal transducers from the knowledge of regulatory interactions, In *Genome Informatics 1997*, 260–261, Yebisu Tokyo Japan, Dec 1997, Universal Academy Press.

[2] Bono, H., Ogata, H., Goto, S., and Kanehisa, M., Reconstruction of amino acid biosynthesis pathways from the complete genome sequence, *Genome Res*, 8:203–210, 1998.

[3] Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., and Kanehisa, M., DBGET/LinkDB: an integrated database retrieval system, 3:683–694, 1997.

[4] Goto, S., Nishioka, T., and Kanehisa, M., LIGAND: Chemical database for enzyme reactions, *Bioinformatics*, 14:591–599, 1998.

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

[6] Kawarabayashi, Y. et al. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3, *DNA Res.*, 5:55–76, 1998.

[7] Mushegian, A.R. and Koonin, E.V., A minimal gene set for cellular life derived by comparison of complete bacterial genomes, *Proc. Natl. Acad. Sci. USA*, 93:10268–10273, 1996.

[8] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 1999. in press.

[9] Tomii, K. and Kanehisa, M., A comparative analysis of ABC transporters in the complete microbial genomes, *Genome Res*, 1998. in press.

[10] Watanabe, H., Mori, H., Itoh, T., and Gojobori, T., Genome plasticity as a paradigm of eubacteria evolution, *J. Mol. Evol.*, 44:S57–S64, 1997.