Reconstruction of Amino Acid Biosynthesis Pathways from the Complete Genome Sequence

Hidemasa Bono, Hiroyuki Ogata, Susumu Goto, et al.

Genome Res. 1998 8: 203-210
Access the most recent version at doi:10.1101/gr.8.3.203

RESEARCH

Reconstruction of Amino Acid Biosynthesis Pathways from the Complete Genome Sequence

Hidemasa Bono, Hiroyuki Ogata, Susumu Goto, and Minoru Kanehisa[1]

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

The complete genome sequence of an organism contains information that has not been fully utilized in the current prediction methods of gene functions, which are based on piece-by-piece similarity searches of individual genes. We present here a method that utilizes a higher level information of molecular pathways to reconstruct a complete functional unit from a set of genes. Specifically, a genome-by-genome comparison is first made for identifying enzyme genes and assigning EC numbers, which is followed by the reconstruction of selected portions of the metabolic pathways by use of the reference biochemical knowledge. The completeness of the reconstructed pathway is an indicator of the correctness of the initial gene function assignment. This feature has become possible because of our efforts to computerize the current knowledge of metabolic pathways under the KEGG project. We found that the biosynthesis pathways of all 20 amino acids were completely reconstructed in *Escherichia coli*, *Haemophilus influenzae*, and *Bacillus subtilis*, and probably in *Synechocystis* and *Saccharomyces cerevisiae* as well, although it was necessary to assume wider substrate specificity for aspartate aminotransferases.

Whereas the complete genome sequences of an increasing number of organisms have become publicly available (Fleischmann et al. 1995; Fraser et al. 1995; Bult et al. 1996; Himmelreich et al. 1996; Kaneko et al. 1996; Blattner et al. 1997; Goffeau et al. 1997; Kunst et al. 1997; Tomb et al. 1997), the functional identification of predicted genes still remains largely incomplete. The only effective method that is widely used for functional identification relies on searching for sequence similarities in the databases. If no sequence similarities are found, or only similarities to unidentified sequences are found, this informatics approach cannot give any clues to functions. Consequently, one-third to two-thirds of the genes are left unidentified in every genome thus far sequenced. This ratio may become smaller if one tries to overinterpret the degree of sequence similarity or the existence of a sequence motif. The ratio may also vary depending on how the function is defined. For example, is a gene functionally identified if it is revealed to code for a protein kinase? Perhaps it is not identified until the target protein that is phosphorylated is identified, or even until the role on the biochemical pathway is identified.

Obviously, the functional prediction from a complete genome imposes an additional constraint of the completeness; for example, the total number of aminoacyl tRNA synthetases, the group of proteins that form an ABC transport system, and the group of enzymes that form a specific biosynthetic pathway have to be complete. In the standard sequence similarity search, suppose that a query sequence is found to be similar to a database sequence and the database sequence is known to have a certain biological function. These two observations are used to deduce the conclusion that the query sequence would also have the same biological function. This type of syllogism is the basis of functional prediction. However, the reasoning here is done on a piece-by-piece (gene-by-gene) basis without regard to the completeness of the gene catalog or the completeness of the biochemical pathway that should be formed by the gene products.

In view of the expanding body of information being generated by the genome sequencing projects, the gene function prediction must be improved by incorporating such global features of the biological system. However, when the similarity found is only marginal, it is critical to combine additional information to arrive at a final functional

[1]Corresponding author.
E-MAIL kanehisa@kuicr.kyoto-u.ac.jp; FAX 81-774-38-3269.

assignment, because an appropriate level of sequence similarity that can be extended to functional similarity cannot be predetermined. In bacterial genomes in which the operon structure is relatively well conserved, the clustering of predicted genes in the genome is a good indicator to reconstruct a functional unit of, for example, the ABC transport system. Furthermore, experts' knowledge on biochemical pathways, molecular assemblies, and other functional units must often be utilized in human interpretation of sequence similarity search results. If we can computerize such biological knowledge in a proper way, the functional prediction can be improved and better automated.

From this perspective there are efforts, including ours, to develop new databases and computational technologies, to perform metabolic reconstruction for a number of organisms, and to make the resources publicly available, such as in the WIT (Overbeek et al. 1997; Selkov et al. 1997), Ecocyc (Karp et al. 1997), and Hincyc (Karp et al. 1996) systems. In 1995, we initiated the KEGG (Kyoto Encyclopedia of Genes and Genomes) Project (http://www.genome.ad.jp/kegg/) to computerize the current knowledge of metabolic and regulatory pathways, which we consider wiring diagrams of genes and gene products, and to link them with the gene catalogs generated by the genome projects (Goto et al. 1996; Kanehisa 1997a; Ogata et al. 1998). Here we report the use of the KEGG metabolic pathway database and the orthologous gene grouping for function identification of enzyme genes.

RESULTS

Reference Pathways

The metabolic pathway section of KEGG has been primarily organized from the compilations of the Japanese Biochemical Society (Nishizuka 1980, 1997) and the wall chart of Boehringer Mannheim (Gerhard 1992) and has also been verified with a number of printed and on-line sources. The current knowledge of the intermediary metabolism and the secondary metabolisms is represented by ~100 pathway diagrams that are manually drawn and continuously updated. An example of the pathway diagram is shown in Figure 1, in which an enzyme is represented by a box with the EC number inside, and a chemical compound by a circle. Thus, the KEGG pathway diagram represents a network of interacting molecules, which is at a higher level of abstraction than the existing molecular biology databases that describe various aspects of individual molecules. KEGG is linked to such detailed molecular information through the DBGET/LinkDB system (Fujibuchi et al. 1997; Kanehisa 1997b); in Figure 1 a box or a circle is clickable under the World Wide

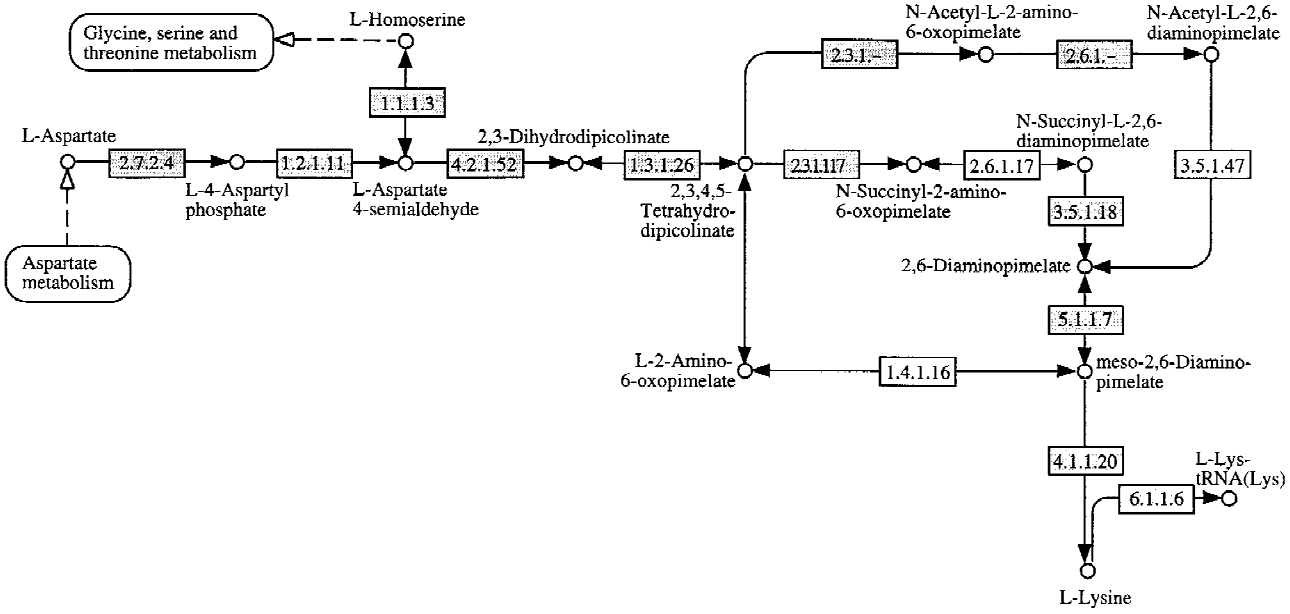

Figure 1 The lysine biosynthesis pathway taken from KEGG release 4.1 (December 1997) at http://www.genome.ad.jp/kegg/. The enzymes are shown in boxes with the EC numbers inside. The shaded boxes represent those enzymes whose genes are identified, in this case, in *E. coli.* Thus, the sequence of shaded boxes corresponds to the *E. coli*-specific pathway of lysine biosynthesis.

Web address to retrieve an enzyme or a compound entry of the LIGAND database (Suyama et al. 1993; Goto et al. 1998) as well as related entries in the existing databases.

There is also a genome section in KEGG that contains the gene catalogs of a number of organisms, including the complete genomes of *Escherichia coli, Haemophilus influenzae, Helicobacter pylori, Bacillus subtilis, Mycoplasma genitalium, Mycoplasma pneumoniae, Methanococcus jannaschii, Synechocystis,* and *Saccharomyces cerevisiae.* Once all the enzyme genes are identified for an organism and the EC numbers are properly assigned, an organism-specific pathway is automatically generated in KEGG by matching the EC numbers in the gene catalog and the EC numbers on the metabolic pathway diagrams. In fact, the chain of shaded boxes in Figure 1 represents a specific pathway, in this case, for *E. coli.* Therefore, the manually drawn diagrams are used as references of all known enzymatic reactions and pathways, and organism-specific pathways are reconstructed by comparing the gene catalog against the reference pathways.

EC Number Assignment

The quality of this reconstruction largely depends on the quality of the initial EC number assignment. Because this information is not usually provided by the original authors of the genome sequencing projects, we have been doing the assignment both manually and computationally. As described in Methods, we have developed a computational tool, Gene Function Identification Tool (GFIT), to make use of the orthologous relations of known genes for identifying new gene functions (http://www.genome.ad.jp/kegg/comp/GFIT.html). In particular, this tool has been utilized for automatically identifying enzyme genes and assigning EC numbers. Before developing the GFIT program, we relied on manual assignment of EC numbers in which the investigators' description of ORFs was used to search other database entries, mostly SWISS–PROT entries, that contained EC numbers. At the moment, we use GFIT for the initial screening, which is followed by human inspection for the final assignment.

Table 1 summarizes the EC number assignment in KEGG as of November 20, 1997 for the nine complete genomes. The most updated version of this table can be viewed in the WWW (http://www.genome.ad.jp/kegg/java/org__list.html). The fraction of enzyme genes is ~20% in bacteria, whereas it is ~14% in yeast. According to our EC number assignment, ~75%–80% of the enzyme genes in each organism have been mapped on the KEGG reference pathway diagrams.

Missing Enzymes

Whether the gene function assignment is biologically meaningful can be checked by examining the reconstructed pathways in KEGG. Namely, when a specific metabolic pathway formed is complete, it is likely that the enzyme genes have been correctly assigned. When the metabolic pathway is incompletely reconstructed from the enzyme genes predicted to exist in the genome, there are basically two possibilities. One is that the gene function identification has not been done extensively enough, resulting in an incorrect assignment of the EC num-

Table 1. Complete Genomes Mapped on the KEGG Metabolic Pathways

Species	Genome size (nucleotides)	Genes		Enzymes		GenBank accession no.
		protein	RNA	total	mapped	
M. jannaschii	1,664,987	1681	43	356	295	L77117
E. coli	4,639,221	4289	108	952	703	U00096
H. influenzae	1,830,135	1717	74	446	340	L42023
H. pylori	1,667,867	1566	43	317	253	AE000511
B. subtilis	4,214,807	4021	117	654	503	AL009126
M. genitalium	580,073	467	36	127	101	L43967
M. pneumoniae	816,394	678	33	150	118	U00089
Synechocystis	3,573,470	3166	49	636	489	synecho
S. cerevisiae	12,069,313	6046	262	861	654	yst_chr_[1–16]

As of November 20, 1997: http://www.genome.ad.jp/kegg/java/org_list.html.

ber. We use GFIT for an extensive search of homologs and alternative assignments to resolve this problem. At the moment, however, we do not reanalyze the genomic nucleotide sequence for possibly identifying additional ORFs, but take the authors' list of ORFs as the starting point. It will eventually become necessary to reanalyze the genomic sequence as well.

The other possibility is that our knowledge of the metabolic pathway is not sufficient. There may be an alternative enzyme or a set of alternative enzymes that can fill the missing reactions. For this possibility, the other tool, PATHCOMP, (http://www.genome.ad.jp/kegg-bin/mk__pathcomp__html) is useful for searching alternative reaction paths (Goto et al. 1996). The lysine biosynthesis pathway in *E. coli* shown in Figure 1 contains missing enzymes, and we believe that this is because of the second possibility as described below.

Reconstruction of Amino Acid Biosynthetic Pathways

Table 2 summarizes the current status of the reconstruction of amino acid biosynthetic pathways in KEGG for the seven organisms, *E. coli*, *H. influenzae*, *H. pylori*, *B. subtilis*, *M. jannaschii*, *Synechocystis*, and *S. cerevisiae*. The other two in Table 1, *M. genitalium* and *M. pneumoniae*, do not appear to contain any pathway for amino acid synthesis. The result of reconstruction is classified into five categories in Table 2. (1) The reconstruction is complete and is reflected in the current version of KEGG. (2) The reconstruction would probably become complete because there is a candidate gene identified by GFIT. (3) The reconstruction can be complete because there is evidence that the reference pathway diagram needs to be updated (see below). (4) Most of the pathway is reconstructed, but the missing enzymes cannot be filled yet either by additional genes or by alternative reactions. (5) The pathway does not seem to exist at all.

In Table 2 the enzymes with the EC number hierarchy of 2.6.1 are aminotransferases (transaminases). Aminotransferase is an enzyme that adds the amino group to the precursor of a specific amino acid; for example, EC 2.6.1.6 is aminotransferase for leucine. Although aspartate aminotransferase (EC 2.6.1.1) is generally considered to have ligand speci-

Table 2. Results of Reconstructing Amino Acid Biosynthesis Pathways from Seven Complete Genomes

KEGG map ID	Amino acid	*Eco*	*Hin*	*Hpy*	*Bsu*	*Mja*	*Syn*	*Sce*
00251	Gln	++	++	++	++	++	++	++
	Gln	++	++	++	++	++	++	++
00252	Asn	++	++	?	++	++	++	++
	Asp	++	++	?	++	++	++	++
	Ala	2.6.1	2.6.1	++	++	++	++	++
00260	Ser	++	++	++	++	++	++	++
	Gly	++	++	++	++	++	++	++
	Thr	++	++	++	++	++	++	++
00271	Met	++	++	–	++	–	+	++
00272	Cys	++	++	++	++	–	++	++
00290	Val	++	++	–	++	++	++	++
	Leu	++	++	–	++	+	++	++
	Ile	++	++	–	++	++	++	++
00300	Lys	2.6.1	2.6.1	2.6.1	2.6.1	?	?	?
00330	Arg	++	+	–	++	++	+	++
	Pro	++	++	++	+	–	++	++
00340	His	++	++	–	++	++	+	+
00400	Phe	++	++	?	++	?	++	++
	Tyr	++	++	?	++	?	++	++
	Trp	++	++	+	++	?	++	++

(++) Reconstruction is complete in the current version of KEGG; (+) reconstruction becomes complete by assigning a homologous gene; (2.6.1) reconstruction becomes complete by assuming wider specificity for an enzyme; (?) the pathway apparently exists but all missing enzymes are not yet resolved; (–) the pathway does not appear to exit.

ficity for aspartate and glutamate, certain experiments suggest the enzyme activity toward other amino acids as well. Mavrides et al. (1978) showed that chicken aspartate aminotransferase had an activity toward aromatic amino acids. Birolo et al. (1991) found by the active site analysis that aspartate aminotransferase in *Sulfolobus solfataricus* was fairly active when alanine was used as a substrate. The sequence analysis by Mehta et al. (1989, 1993) indicated that aminotransferases for L-aspartate, L-alanine, L-tyrosine, and L-phenylalanine could be classified to the same subgroup.

Taking these observations into account, aspartate aminotransferase (EC 2.6.1.1) that does exist in the *E. coli*, *H. influenzae*, *H. pylori*, and *B. subtilis* genomes can be assumed to act on alanine and lysine as well. This makes the biosynthetic pathways for all 20 amino acids complete in *E. coli*, *H. influenzae*, and *B. subtilis* (Table 2). The lysine pathway in *Synechocystis* still contains a few more missing enzymes, but we expect they can be filled as we improve our EC number assignment system. *S. cerevisiae* appears to use an entirely different pathway for lysine biosynthesis from the bacterial pathway, and there are still several enzymes that we have not been able to assign.

As indicated in Table 2, the archea *M. jannaschii* apparently lacks methionine, cysteine, and proline pathways, and the gastric pathogen *H. pylori* probably lacks methionine, valine, leucine, isoleucine, arginine, and histidine pathways. These observations are likely to reflect the life cycle and the evolution of the organisms.

DISCUSSION

The knowledge of enzymatic reactions and metabolic pathways was accumulated from numerous biochemical experiments in the past. In general, these experiments were performed with enzymes isolated from specific organisms and tissues. It is possible that a different organism has a different set of functionally related enzymes, such as aminotransferases, that perform, as a group, its biological function. Figure 2 shows the current knowledge in KEGG for the roles of aminotransferases in the amino acid biosynthesis pathways. Aspartate aminotransferase (EC 2.6.1.1) is used in four places with different substrates. This enzyme in the tyrosine and phenylalanine pathways has just recently been added to KEGG according to the new compilation of the Japanese Biochemical Society (Nishizuka 1997). We hope our prediction of its role on alanine and lysine will be tested by biochemical experiments in respective organisms.

In the current version of KEGG, the number of genes assigned to the aminotransferase class (2.6.1) is 16 for *E. coli*, whereas it is 7 for *H. influenzae*,

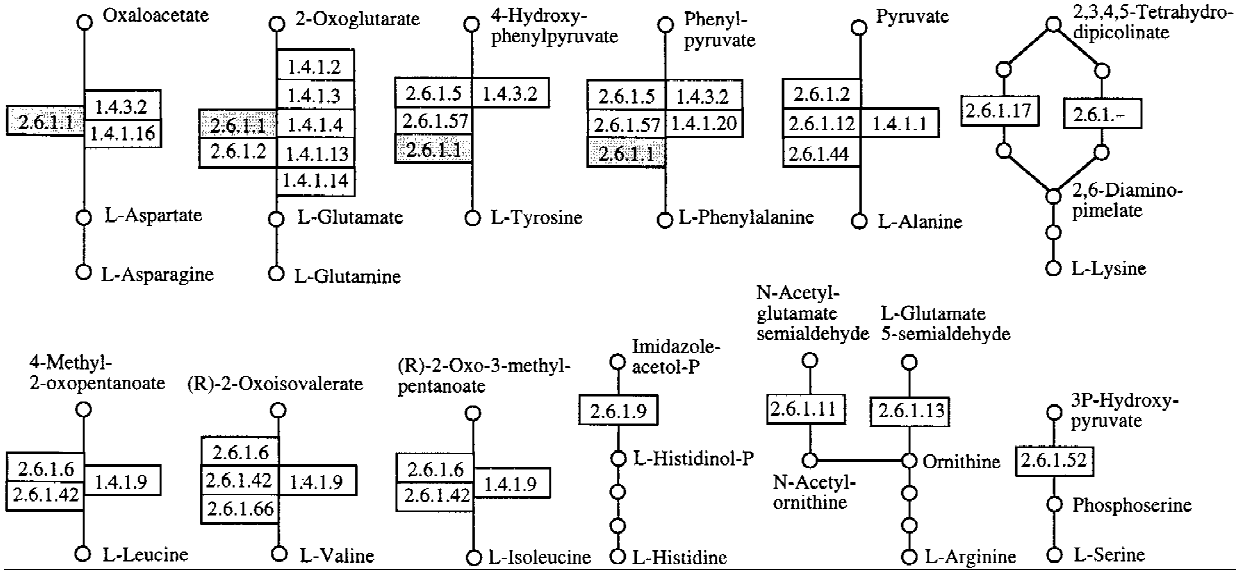

Figure 2 The missing enzymes in the amino acid biosynthesis pathways in *E. coli, H. influenzae,* and *B. subtilis* often involved the EC number hierarchy 2.6.1 of aminotransferases. This is a summary of reaction steps catalyzed by aminotransferases in the amino acid biosynthesis pathways according to KEGG release 4.1 (December 1997). The shaded boxes represent aspartate aminotransferase (EC 2.6.1.1), which is assumed to have wide specificity for ligands (see text for details).

although this organism also seems to have all the amino acid biosynthesis pathways as shown by Tatusov et al. (1996) as well. This may suggest the possibility that an enzyme in *H. influenzae* plays multiple functional roles that are taken by different enzymes in *E. coli*. For example, *E. coli* has tyrosine aminotransferase EC 2.6.1.5 (gene accession no. b4054) that has a high sequence similarity to aspartate aminotransferase EC 2.6.1.1 (gene accession no. b0928), whereas there is one aspartate aminotransferase in *H. influenzae* (gene accession no. HI1617) that apparently catalyzes the reactions both for aspartate and tyrosine (and phenylalanine, as well). The number of aminotransferases in *H. pylori* is only three, which is in agreement with the observation that this organism lacks many of the amino acid biosynthesis pathways.

During the pathway reconstruction process, we have noticed wide variations in the degree of annotation in different complete genome sequences. *E. coli* (Blattner et al. 1997) is the best annotated genome reflecting the fact that it is the best studied organism by biochemical, genetic, and other experiments. In contrast, a number of hypothetical proteins are left unassigned in *Synechocystis* (Kaneko et al. 1996) partly because the authors had made relatively conservative interpretation of sequence similarities found, but largely because the additional information, especially the conserved operon structure, was scarce in this organism. The reference pathways organized in KEGG can be used as a functional catalog that will help make a uniform and systematic assignment of gene functions in all organisms.

We also note that there is a danger of relying too much on the sequence similarities and believing the description given in the similar sequence entries in the database. It is possible that an erroneous description is propagated to a number of entries without knowing where the error actually originated. The biochemical knowledge of metabolic pathways is a different level of information that can be utilized to check the validity of assignments made by sequence similarities.

The enzymes represent ~20% of the genes in bacterial genomes, and the proportion becomes smaller in higher organisms. Under the KEGG project, we started to computerize, in addition to the metabolic pathways, a number of regulatory pathways, such as membrane transport, signal transduction, cell cycle, and developmental pathways. Because the regulatory pathways seem more divergent in different organisms and because the amount of biochemical knowledge is limited to selected organisms, we represent the knowledge of regulatory pathways in an organism-specific way. We consider, among others, *E. coli*, *B. subtilis*, *S. cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, and *Drosophila melanogaster* as reference organisms and provide ways to compare against these references for reconstruction of regulatory pathways.

METHODS

Orthologous Gene Table

The availability of complete genomic sequences has enabled us to identify orthologous relations of genes in the following manner (Mushegian and Koonin 1996). 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. If, for example, there are n genes in organism 1 and m genes in organism 2, this requires $n \times m$ sequence comparisons. When gene *A* in organism 1 is most similar to gene *B* among all genes in organism 2, and when gene *B* in organism 2 is most similar to gene *A* among all genes in organism 1, we conclude 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 (duplications) within each organism and also from the inconsistencies of pairwise comparisons when multiple organisms are considered. We are working to maintain an orthologous gene table in KEGG, which, from a biological viewpoint, is manually edited for a number of organisms.

GFIT Program

The sequence similarity search used for functional prediction is actually a process of assigning a sequence to a group of similar sequences. The reasoning from ''A is similar to B'' and ''B has function C'' into ''A has function C'' is equivalent to the reasoning from ''A is similar to B'' and ''B belongs to group C'' into ''A belongs to group C.'' From this perspective, it should be useful to identify all candidates of orthologous gene groups from a set of complete gene catalogs even though the functions of certain groups are not yet known.

We have been developing the computational tool GFIT to make use of the orthologous relations for identifying gene functions (http://www.genome.ad.jp/kegg/comp/GFIT.html). When a complete genome is newly determined for an organ-

ism, the list of candidate genes is given to the GFIT program, which performs the organism-by-organism comparison with FASTA (Pearson and Lipman 1988) to identify orthologs; namely, all the amino acid sequences in the gene catalog of the query organism are compared against all the amino acid sequences in each of the gene catalogs in the KEGG database, and the orthologs are identified as mentioned above. At the moment, GFIT tentatively assigns the gene function according to the top-scoring ortholog. However, it is up to the user to make a final assignment by examining the members in the ortholog group from different organisms. As the quality of the KEGG orthologous gene table improves, we plan to better automate the final assignment process.

ACKNOWLEDGMENTS

This work was supported in part by the Grant-in-Aid for Scientific Research on the Priority Area 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.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Birolo, L., M.I. Arnone, M.V. Cubellis, G. Andreotti, G. Nitti, G. Marino, and G. Sannia. 1991. The active site of Sulfolobus solfataricus aspartate aminotransferase. *Biochem. Biophys. Acta* 1080: 198–204.

Blattner, F.R., G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453–1462.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. Fitzgerald, R.A. Clayton, J.D. Gocayne et al. 1996. Complete genome sequence of methanogenic archaeon, *Methanococcus jannaschii. Science* 273: 1058–1073.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496–512.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397–403.

Fujibuchi, W., S. Goto, H. Migimatsu, I. Uchiyama, A. Ogiwara, Y. Akiyama, and M. Kanehisa. 1997. DBGET/LinkDB: An integrated database retrieval system. In *Pacific Symposium on Biocomputing '98* (ed. R.B. Altman, K. Dunker, L. Hunter, and T.E. Klein), pp. 683–694. World Scientific, Singapore.

Gerhard, M. (ed.). 1992. *Biological pathways*, 3rd ed., Boehringer Mannheim, Mannheim, Germany.

Goffeau, A., R. Aert, M.L. Agostini-Carbone, A. Ahmed, M. Aigle, L. Alberghina, K. Albermann, M. Albers, M. Aldea, D. Alexandraki et al. 1997. The yeast genome directory. *Nature* (Suppl.) 387.

Goto, S., H. Bono, H. Ogata, W. Fujibuchi, T. Nishioka, K. Sato, and M. Kanehisa. 1996. Organizing and computing metabolic pathway data in terms of binary relations. In *Pacific Symposium on Biocomputing '97* (ed. R.B. Altman, K. Dunker, L. Hunter, and T.E. Klein), pp. 175–186. World Scientific, Singapore.

Goto, S., T. Nishioka, and M. Kanehisa. 1998. LIGAND: Chemical database for enzyme reactions. *Bioinformatics* (in press).

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B.C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae. Nucleic Acids Res.* 24: 4420–4449.

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

———. 1997b. Linking databases and organisms: GenomeNet resources in Japan. *Trends Biochem Sci.* 22: 442–444.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3: 109–136.

Karp, P.D., C. Ouzounis, and S. Paley. 1996. HinCyc: A knowledge base of the complete genome and metabolic pathways of *H. influenzae. Intell. Systems Mol. Biol.* 4: 116–124.

Karp, P.D., M. Riley, S.M. Paley, A. Pellegrini-Toole, and M. Krummenacker. 1997. EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* 25: 43–50.

Kunst, F., N. Ogasawara, I. Moszer, A.M. Albertini, G. Alloni, V. Azevedo, M.G. Bertero, P. Bessieres, A. Bolotin, S. Borchert et al. 1977. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature* 390: 249–256.

Mavrides, C. and P. Christen. 1978. Mitochondrial and cytosolic aspartate aminotransferase from chicken: Activity

toward aromatic amino acids. *Biochem. Biophys. Res. Comm.* 85: 769–773.

Mehta, P.K., T.I. Hale, and P. Christen. 1989. Evolutionary relationships among aminotransferases. Tyrosine aminotransferase, histidinol-phosphate aminotransferase, and aspartate aminotransferase are homologous proteins. *Eur. J. Biochem.* 186: 249–253.

———. 1993. Aminotransferases: Demonstration of homology and division into evolutionary subgroups. *Eur. J. Biochem.* 214: 549–561.

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

Nishizuka, T. (ed.). 1980. *Metabolic maps.* Biochemical Society of Japan, Tokyo, Japan.

———. 1997. *Cell functions and metabolic maps.* Biochemical Society of Japan, Tokyo, Japan.

Ogata, H., S. Goto, W. Fujibuchi, and M. Kanehisa. 1998. Computation with the KEGG pathway database. *BioSystems* (in press).

Overbeek, R., N. Larsen, W. Smith, N. Maltsev, and E. Selkov. 1997. Representation of function: The next step. *Gene* 191: GC1–9.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* 85: 2444–2448.

Selkov, E., N. Maltsev, G.J. Olsen, R. Overbeek, and W.B. Whitman. 1997. A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene* 197: GC11–26.

Suyama, M., A. Ogiwara, T. Nishioka, and J. Oda. 1993. Searching for amino acid sequence motifs among enzymes: The enzyme-reaction database. *Comput. Appl. Biosci.* 9: 9–15.

Tatusov, R.L., A.R. Mushegian, P. Bork, N.P. Brown, W.S. Hayes, M. Borodovsky, K.E. Rudd, and E.V. Koonin. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli. Curr. Biol.* 6: 279–291.

Tomb, J.F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty et al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori. Nature* 388: 539–547.

Received December 1, 1997; accepted in revised form February 2, 1998.