Genetwork

A database for post-genome analysis

When the Human Genome Project was initiated in the late 1980s, it was promoted as the ultimate project to uncover the blueprint of life. Although the goal of sequencing the entire 50 billion base pairs of the human genome by 2005 is likely to be achieved, whether we will have the blueprint of life at that time is quite questionable. First of all, as we have learned from the complete genomes of yeast and several bacteria, the biological function of a large fraction of the genes (a third to over a half depending on the organism) is still uncharacterized. Secondly and more importantly, because the genes and gene products are only the individual components that make up a biological system, the understanding of how each component works is not sufficient to understand the entire system. The post-genome analysis, as we define it here, includes both experimental and informatics approaches to uncover systematically the interactions and pathways of genes and molecules, which can be considered as the wiring diagrams of the biological system. The complete catalog of components and the complete catalog of wiring diagrams together can be called the blueprint of life.

KEGG (Kyoto Encyclopedia of Genes and Genomes) is an informatics project for the post-genome analysis, which we initiated in 1995 under the Human Genome Program of the Ministry of Education, Science, Sports and Culture in Japan. Its objectives are threefold. (1) To computerize the current knowledge of molecular pathways and genetic pathways from the experimental observations of genetics, biochemistry, and molecular and cellular biology. In the past two years, KEGG contained only the metabolic pathways, but starting in July 1997 a number of regulatory pathways, such as signal transduction, cell cycle and developmental pathways, are being placed online. (2) KEGG maintains the gene catalog of every organism that has been sequenced, and each component in the catalog is to be mapped on to the KEGG pathways. (3) In addition to these database efforts, KEGG aims at developing new informatics technologies that are associated with interactions and pathways[1].

KEGG is a part of the Japanese GenomeNet WWW server[2] and is linked to all the major molecular biology databases by the DBGET LinkDB system[3]. Figure 1 shows a portion of the KEGG metabolic pathway diagram for phenylalanine, tyrosine and tryptophan biosynthesis, where each box represents an enzyme with the EC number inside. The box is clickable to retrieve the corresponding enzyme entry of the LIGAND database[4], which is the starting point of retrieving related entries of chemical compounds, molecular sequences,

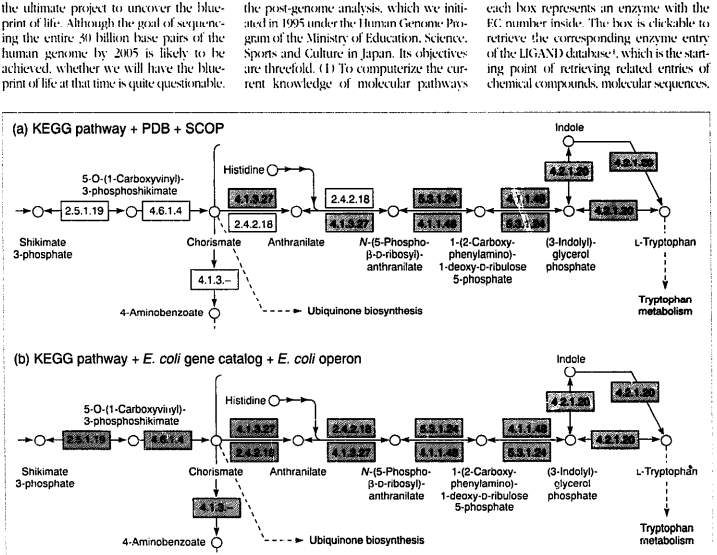

FIGURE 1. Examples of using KEGG at http://www.genome.ad.jp/kegg. (a) From the KEGG table of contents choose the molecular catalog 'Enzymes by SCOP 3D-folds', select alpha and beta (α β), copy all the description in the category of alpha beta (TIM)-barrel, select the 'Pathway' option, paste in the search box, and search against '3D structures in PDB'. Several pathways, each of which contains at least one enzyme with a TIM barrel according to SCOP, will be indicated in the results screen. The view shown is part of 'pdb00400 phenylalanine, tyrosine and tryptophan biosynthesis'. The blue boxes are the enzymes whose three-dimensional structures are known and those marked with red are the ones found to correspond to the SCOP classification. (b) From the KEGG table of contents choose the catalog of *Escherichia coli* operons, copy all the description in the *trp* operon, select the 'Pathway' option, paste in the search box, and search against '*Escherichia coli*'. Select the 'eco00400 phenylalanine, tyrosine and tryptophan biosynthesis' pathway. The green boxes are the enzymes whose genes exist and those marked with red are the ones found to correspond to the *trp* operon.

0168-9525/97/$17.00
PII: S0168-9525(97)01227-7

three-dimensional structures, and genetic diseases among others. KEGG maintains structural and functional classifications of molecules and genes in the form of, what we call, hierarchical texts, in which the headings and subheadings are clickable to unfold or fold branches. Figure 1(a) is a result (marked in red) of matching the βα (TIM) barrel proteins in the hierarchical table derived from the SCOP database[5] with the KEGG metabolic pathway diagrams where the enzymes with known PDB (Protein Data Bank) structures are shown in blue boxes. This indicates possible gene duplications in the formation of the tryptophan biosynthetic pathway[6].

One of the most unique aspects of KEGG is the automatic generation of organism-specific pathways by matching the gene catalogs being produced by the genome sequencing projects and the reference pathway diagrams manually drawn and updated. In Figure 1(b) the enzymes colored in green indicate that the corresponding genes are found in the *Escherichia coli* gene catalog. Those marked in red belong to the tryptophan operon, and the genome map section of KEGG can also be utilized with a Java-compatible browser, for example, to see any correlation between the physical proximity of genes in the genome and the functional proximity of gene products in the pathway. An important consequence of mapping gene products on the pathway diagrams is the validation of the initial gene assignments. In case the pathway is not continuous because of missing gene products, KEGG provides computational tools to assist re-examination of gene function assignment[7] and further analysis of possible existence of alternative paths[8].

While KEGG tries to cover a diverse range of pathways at a high level of abstraction there are complementary resources that contain more detailed data and knowledge in specific pathways. We have started collaboration with WIT (Ref. 9) for metabolic pathways and are open to any other collaborations. The mirror sites of KEGG are being established in the USA and UK (Ref. 10). In addition to the Internet version, KEGG is available in CD-ROM for Macintosh and Windows where pathway diagrams, hierarchical texts, and genome maps are all to be handled by a Java-compatible browser. The content of CD-ROM can be downloaded by anonymous FTP (Ref. 11) and used in UNIX as well. We also plan to start distributing the KEGG server to be mirrored in a local environment.

Minoru Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp
http://www.kuicr.kyoto-u.ac.jp/kanehisa/

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

References

1 Goto, S. *et al.* (1996) Pacific Symposium on Biocomputing 1997, pp. 175–186
2 http://www.genome.ad.jp
3 http://www.genome.ad.jp/dbget/dbget.links.html
4 http://www.genome.ad.jp/htbin/show_man?ligand
5 Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) *J. Mol. Biol.* 247, 536–540
6 Wilmanns, M. and Eisenberg, D. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 1379–1383
7 http://www.genome.ad.jp/kegg/comp/GFIT.html
8 http://www.genome.ad.jp/kegg/comp/pathcomp.html
9 http://www.cme.msu.edu/wit/
10 http://kegg.com
11 ftp://kegg.genome.ad.jp/CD/