Biological Systems Database and Genome Information Science

Project Leader:

Minoru KANEHISA Professor, Institute of Chemical Research, Kyoto University

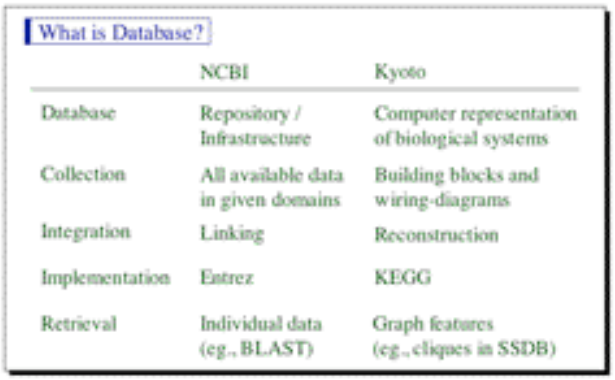
1. Objective:

The increasing amount of genomic information is the basis for understanding principles of how higher-order biological systems, such as the cell, the organism, and the biosphere, are formed, as well as for medical, industrial, and other practical applications. However, current informatics technologies cannot readily uncover higher-level complexity of such biological systems, although they are quite effective to find and characterize building blocks of genes and proteins. Here we develop knowledge-based methods for uncovering higher-order systemic behaviors of the cell and the organism from genomic information. The reference knowledge is stored in KEGG, Kyoto Encyclopedia of Genes and Genomes, and associated bioinformatics technologies are developed both for basic research and practical applications.

2. Summary

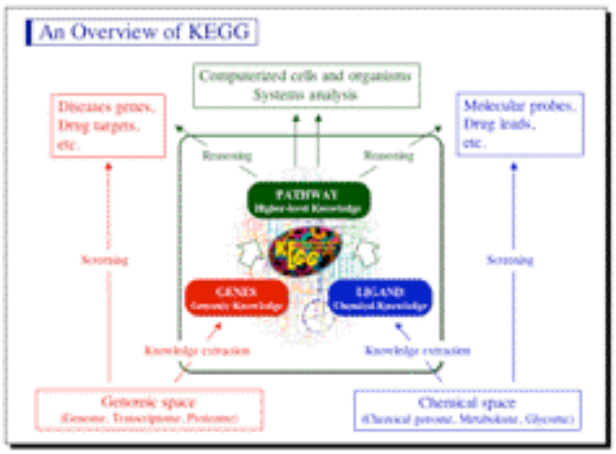
2.1 New concept of the database

An ultimate goal of bioinformatics is a complete computer representation of the cell and the organism, which will enable computational prediction of higher-level complexity, such as molecular interaction networks involving various cellular processes and phenotypes of entire organisms, from genomic information [1]. From this perspective, a new concept of the biological database has been developed. As shown below, the database is a computer representation of the biological system, and it has been successfully implemented as KEGG.



2.2 The KEGG database

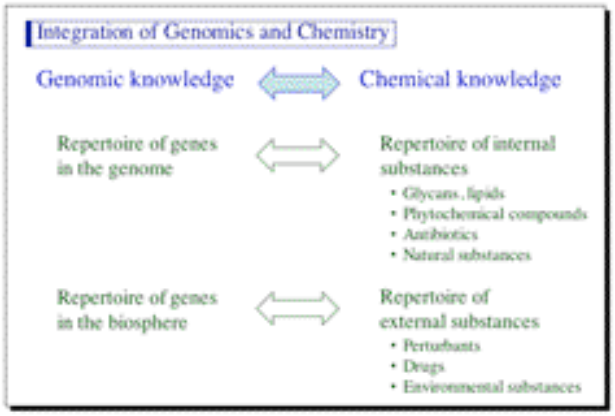
The KEGG database project was initiated in our laboratory in 1995, the last year of the first five-year phase of the Japanese Human Genome Program, and then continued in the second five-year phase, with supports from the Ministry of Education as grants-in-aid for scientific research on priority areas. With a new funding under the Millennium Project the current research-for-the-future program was launched in 2000, and KEGG was significantly expanded. As shown below, KEGG is a biological systems database, integrating genomic information (GENES database) and chemical information (LIGAND database) in terms of network information (PATHWAY database) [2, 3].



Traditional bioinformatics technologies have focused on finding useful genes and molecules, such as disease genes and drug targets, by screening of large-scale data. In contrast, our approach is first to understand wiring diagrams (molecular interaction networks) of building blocks and then to find functions and utilities of biological systems as a whole. KEGG is a reference knowledge base containing current knowledge on such wiring diagrams, and it is used worldwide as a unique resource for reconstructing metabolism and other cellular processes from genomic information and for understanding systemic functional meanings and utilities.

During the five years of this project we have developed various new features in KEGG, which are summarized below.

(1) For network information, we expanded the PATHWAY database from a collection of metabolic pathways to a more comprehensive collection containing signaling and various other regulatory pathways as well as human disease pathways. In addition, an XML version of pathway maps was made available to facilitate computational analysis of KEGG pathways. KEGG has become the international standard of pathway information.
(2) For genomic information in the GENES database, we introduced the KO (KEGG Orthology) system, and developed an automatic method of KO assignment and KEGG pathway mapping, enabling rapid analysis of genomic sequences and EST sequences. By improving the KO system and the automatic assignment program, we hope to make KEGG as the international standard for genome annotations.
(3) For chemical information in the LIGAND database, we introduced GLYCAN [4] in addition to COMPOUND and REACTION. We also developed graph-based algorithms for chemical compound structure comparison and glycan structure comparison, as well as the RC (Reaction Classification) system, which can be used for automatic assignment of EC numbers [5, 6]. These advanced activities resulted in the international collaborations with the NCBI, the EBI, and the Consortium for Functional Glycomics.
(4) The entire KEGG resource is made available at the enhanced KEGG website for general use, as well as through the newly developed KEGG API (application programming interface) for custom use to meet specific needs.



2.2 Integration of Genomics and Chemistry

The representation of the biological system (ontology) in KEGG is based on the concept of the graph, especially the nested graph and the line graph. The nested graph is a graph whose nodes can themselves be graphs. It is used for representing KEGG network hierarchy and for pathway reconstruction and functional inference.

The line graph is a graph derived by interchanging nodes and edges. The metabolic pathway can be viewed either as a network of genes (enzymes) or as a network of compounds, meaning that one can be generated from the other by the line graph transformation. With this concept, we have undertaken new research on integrated analysis of genomic and chemical information [4-6]. The gene repertoire in the genome would tell us about all substances that should be produced by an organism, and conversely chemical structures of natural substances would tell us about the genes that should be present in the genome. The integration of genomics and chemistry is becoming more important now that chemical genomics initiatives generate large amounts of experimental data.

3. Concluding Remarks

The number of accesses to the KEGG/GenomeNet website has increased four fold during the five-year period of this project. It is by far the best-used database site in Japan, despite the fact that it has been developed and maintained by a single laboratory. In fact, it is one of the major sites in the world as is apparent by the Google links search showing the number of links made from other sites. We acknowledge the Millennium funding for this achievement.

Number of links by Google

Database	Address	Links
NCBI	www.ncbi.nlm.nih.gov	29,800
ExPASy (SwissProt)	www.expasy.org	18,300
EBI	www.ebi.ac.uk	13,200
GenomeNet (KEGG)	www.genome.jp	9,430
DDBJ	www.ddbj.nig.ac.jp	620
JSNP	snp.ims.u-tokyo.ac.jp	55
PDBj	www.pdbj.org	23
H-invitational	www.h-invitational.jp	19

As of 16 July 2005.

Primary Publications

1. Kanehisa, M. and Bork, P. (2003) Bioinformatics in the post-sequence era. *Nat. Genet.* **33**, 305-310.
2. Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42-46.
3. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277-D280.
4. Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F., Ueda, N., Hamajima, M., Kawasaki, T., and Kanehisa, M. (2005) KEGG as a glycome informatics resource. *Glycobiology*, in press.
5. Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **125**, 11853-11865.
6. Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* **126**, 16487-16498.