# Toward Pathway Engineering: A New Database of Genetic and Molecular Pathways

Minoru Kanehisa
*Institute for Chemical Research, Kyoto University*

## From Genome Sequences to Functions

The Human Genome Project was initiated in the late 1980s as a natural consequence of the technology developments in molecular biology and with the expectation of new biomedical applications. The project will uncover the complete DNA sequence of the human genome consisting of 3 billion base pairs and 100 thousand genes. In 1977, a small virus genome, $\phi$x174, consisting of just 5,000 base pairs and 11 genes was determined by the emerging technology of DNA sequence determination. After two decades of technology developments the first complete genome of a free-living biological organism, *Haemophilus influenzae*, was determined in 1995. The bacterial genome consisting of 1.8 million nucleotides and 1,700 genes is already followed by the explosion of complete genomic sequences of a number of organisms from bacteria to eukaryotes. As of mid 1996, the genome sequencing projects have been completed for budding yeast (12 Mbp) and for several bacteria including cyanobacteria (3.6 Mbp) which was carried out by the Kazusa DNA Research Institute in Japan.

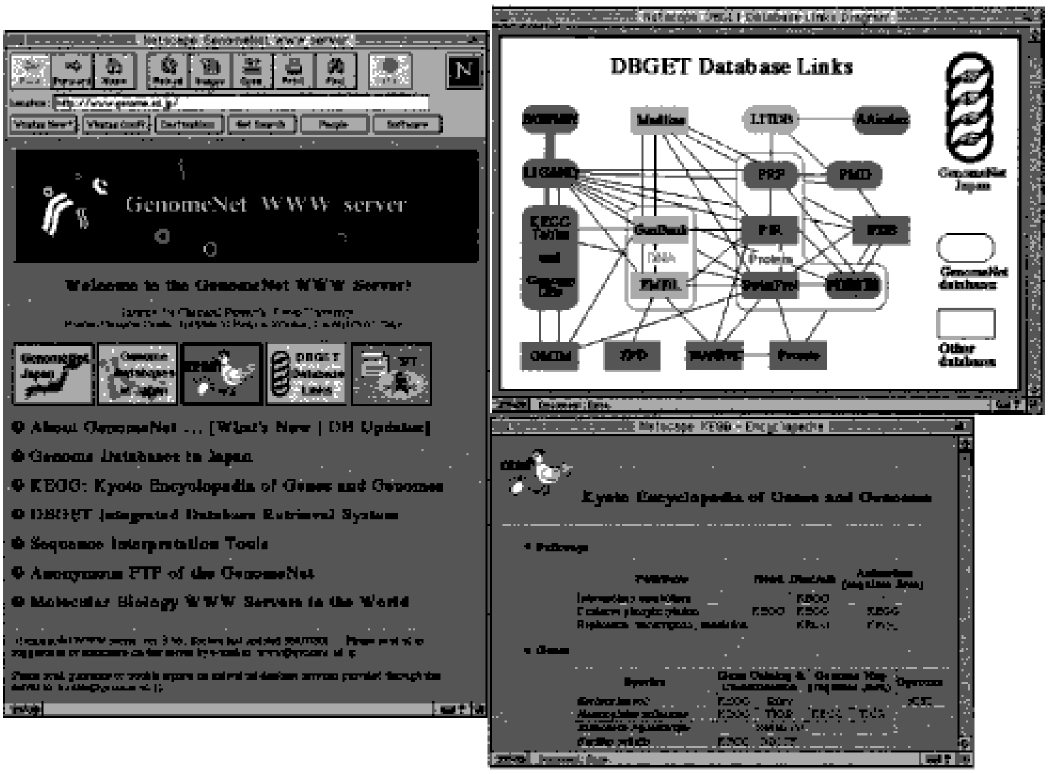
For the first time in human history we are beginning to have the data at hand which leads toward a basic understanding of the fundamental problems in life sciences including the origin and evolution of life and the conception and development of an individual. The data will also stimulate practical applications in medical, pharmaceutical, and agricultural sciences. However, somewhat contrary to public notion, the sequence data obtained by genome projects do not by themselves provide direct answers to such fundamental problems or practical applications. The sequencing of a genome is an easier part than the understanding of functional implications of when, where, and how genes and molecules function in living organisms. Fortunately, our knowledge of the functioning of genes and molecules is also rapidly expanding owing to the advancement of experimental technologies in the broad areas of molecular and cellular biology. In order to make full use of the information obtained by genome projects, it is essential that such functional data are properly computerized in databases and informatics technologies are developed for functional prediction.

## From Gene Gatalogs to Pathways

The functional data that relate to sequence information are currently stored as annotations to sequence data, for example, in the so-called features tables, in the sequence databases of DNAs and proteins. However, these basically represent the sequence-function relationships of single molecules, i.e., the individual components of a biological system, and they do not contain higher level information, i.e., wiring diagrams, of genetic interactions and molecular interactions. It is obvious that without such wiring-diagrams a biological system could never be described or understood.

We have thus initiated a project named KEGG, Kyoto Encyclopedia of Genes and Genomes, to computerize the current knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting genes or molecules. The basic data item in KEGG is a pairwise interaction of genes or molecules that is represented by what is termed a binary relation. An expert in the field would synthesize a pathway from a collection of binary relations obtained by experimental observations. In order to cope with the rapidly expanding body of information it is necessary to computerize the process of synthesizing pathways, in addition to computerizing known pathways derived by human experts.

Currently, we are focusing our attention on the metabolic pathways, although we intend to computerize other pathways, such as signal transduction and cell cycle pathways and the genetic pathways of the early stages in fruit fly development. We expect that once such data are properly computerized, it will become feasible to assist experiments, facilitate understanding, and even perform logical simulations of information pathways controlling all aspects of living organisms.



Welcome to the GenomeNet WWW Server!

- About GenomeNet ... [What's New | DB Updates]
- Genome Databases in Japan
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- DBGET Integrated Database Retrieval System
- Sequence Interpretation Tools
- Anonymous FTP of the GenomeNet
- Molecular Biology WWW Servers in the World

Figure 1: The home page of the GenomeNet WWW server at http://www.genome.ad.jp/ and the DBGET and KEGG search windows.

## GenomeNet Database Service

In 1991 we established a computer network named GenomeNet under the Genome Informatics Project, a part of the Human Genome Program of the Ministry of Education, Science, Sports and Culture (Monbusho). The aim of GenomeNet is not simply a network connection; it is to establish the informatics infrastructure for genome research and related research areas in molecular and cellular biology. In view of the subsequent government funding of Internet activities in Japan, we are only currently maintaining the connection between Tokyo, Kyoto and Fukuoka. Originally, we envisioned a network community where the informatics needs of individual researchers and individual projects would be realized on their local machines by integrating databases and computational resources distributed over the network. We believe that the wide-ranging database service in GenomeNet, which is jointly provided by the Supercomputer Laboratory of the Institute for Chemical Research, Kyoto University and the Human Genome Center of Institute of Medical Science, the University of Tokyo, has greatly contributed toward that end.

The most popular mode of access to the GenomeNet database service is to use our WWW server shown in Fig. 1, which provides among others, the DBGET integrated database retrieval system and the sequence interpretation tools including sequence similarity and motif search programs. The server receives tens of thousands of queries per day, one-third of which are from abroad. Although the GenomeNet database service is a result of technological developments in Japan, for example, DBGET was developed in my laboratory, most of the databases that we offer originated in the U.S. or Europe. Even the databases which claimed to be organized in Japan actually heavily depend on the systems and protocols developed in other countries.

KEGG is an attempt to advance our original concepts and technologies, and actual data collection efforts. Although we have not yet made a formal announcement of KEGG, a preliminary version has been publicly available through the GenomeNet WWW server since December 1995. The target date of the first release of KEGG is October 1996. We plan to distribute compact discs in addition to the service over the Internet.

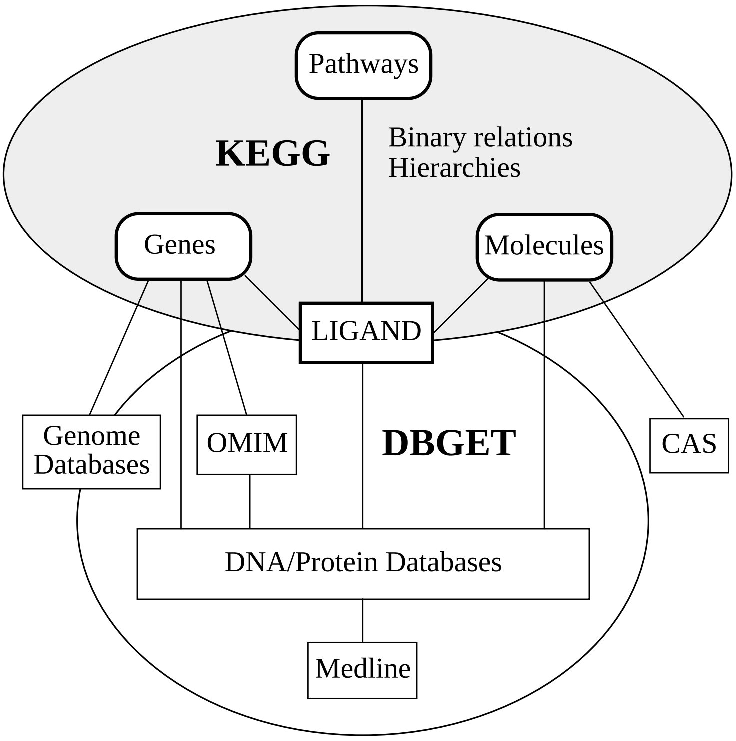


Figure 2: The concept of KEGG and its relation to DBGET. In KEGG functional aspects of genes and molecules are represented by binary relations, hierarchies, and pathways.

## The Concepts

Currently, KEGG is composed of three interconnected sections: pathways, genes, and molecules, which are also linked to a number of existing databases through DBGET (Fig. 2). Both conceptually and practically, KEGG and DBGET are tightly coupled systems. DBGET provides an integrated view of various databases in molecular biology, where the basis of integration is the link (binary relation) between related entries in different databases. In KEGG, an organism may be considered a database of genes and gene products, and the link between them is used for synthesizing a pathway. Thus, both KEGG and DBGET contain an aspect of the deductive database where new relations can be deduced from relations stored in the database.

Another important concept in KEGG is the hierarchy that represents functional, structural, and evolutionary relationships of genes and molecules. For example, the degree of similarity in sequences and 3D structures of proteins is used for classifying superfamilies and 3D folds. The taxonomy is the classification of organisms, which is important in extending sequence and 3D structural similarities to functional similarities. These and other classifications are taken from appropriate sources and implemented in KEGG.

While the binary relation represents flat, horizontal relationships, the hierarchy represents vertical relationships. Both are naturally integrated in the process of deduction. The concept of relation and deduction is thus the basis of our KEGG project. For our logic-based activities, we acknowledge the past collaborations with the Fifth Generation Computer Project team members in ICOT and the researchers in the Genome Informatics Research Projects 1991-1995 and 1996-2000.

## The Technologies

KEGG makes full use of the advancements in the database and networking technology, including deductive and object-oriented databases, the multimedia environment of WWW, and the mobile agent, Java. Although we maintain the logic-based formalism of relation and deduction, we take a practical, flexible approach in the actual implementation. For example, we use the CORAL deductive database system for experimenting the deduction process, but in the actual implementation of KEGG we have developed our own C++ library for manipulating binary relations and hierarchies.

With a similar philosophy, DBGET does not depend on any database management system. The entire system has been developed in house. Actually, DBGET has its roots in the IDEAS sequence analysis package that I developed in the early 1980s in the U.S. National Institutes of Health. DBGET aims at integrating different databases and different types of data in molecular biology. The integration, however, is at the level of data entries; for example, entries in different databases can be retrieved uniformly and links are made between related entries in different databases. In this loosely-coupled integration, the schema or the format of how an entry is organized by data items is left to each database. This should be contrasted with using the same relational database and enforcing a unified schema for entries coming from many different sources, which necessarily involves the process of data conversion.

The proliferation of WWW was a boon to our approach of loose integration; the link capabilities of DBGET fit the mechanism of WWW very nicely. The text-based DBGET system was easily extended to the multimedia environment, where 3D graphics, 2D graphics, and images are now retrievable in the WWW version of DBGET. In addition, because the update procedure does not involve data conversion, DBGET has been and will continue to be able to cope with the ever increasing number and volume of daily updated databases.

KEGG inherits all these DBGET capabilities. Furthermore, the graphics handling of pathway diagrams and chromosome maps has been implemented by Java. The new capabilities of logical inference and simulation are still under development, but we hope to make the first test version available shortly.

## Data Collection Efforts

However efficiently the system is developed as a database, the most critical thing is the quality of data that it contains. Especially, since the data we handle require biological knowledge in specific domains, quality control is far more difficult than, for example, DNA and protein sequence databases. Among the many different subjects of molecular and genetic pathways, the metabolic pathways are probably the easiest to computerize because of the well-established knowledge and existing compilations. We have been computerizing metabolic pathways, in collaboration with an expert in biochemistry, mostly from the Boehringer wall chart and the compilation by the Japanese Biochemical Society, and partly from other textbooks and on-line databases.

KEGG currently contains most of the known metabolic pathways represented by about 80 graphical diagrams. An enzyme is a clickable object in the diagram to retrieve the corresponding entry of the LIGAND database and then, through DBGET, a number of related entries in different databases. LIGAND is a database of enzyme reactions and metabolic compounds that we organize in a separate project. It provides links between the new PATHWAY database and the existing databases of nucleotide sequences, amino acid sequences, 3D structures, sequence motifs, amino acid mutations, genetic maps, genetic diseases, and literature.

In addition to computerizing known pathways, we are developing methods to compute pathways from binary relations. The metabolic pathway is best suited for this purpose as well, for there is a chemical basis of the binary relation between a substrate and a product. KEGG contains the substrate-product relationships and the relationships of two consecutive enzymes that appear in the known metabolic pathways.

One of the major objectives of the KEGG project is to link the structural data (gene catalogs) obtained by genome projects and the functional data obtained in specialized fields of molecular and cellular biology. Once the genome sequencing is complete, it is customary to attempt to classify all genes according to their functions, for example in the scheme developed by Monica Riley. We plan to make a more objective classification based on the pathway data being entered. At the moment, we only perform the classification of enzyme genes.

## Future Directions

Perhaps, the most challenging task of the KEGG system is the inference capabilities that will help human beings to make logical reasoning processes. These capabilities have not yet been developed or implemented, but here are some examples (Fig. 3).

Given a list of enzymes (EC numbers) that are found in the gene catalog of an organism, KEGG automatically generates the organism specific pathways by marking the enzymes found. Then, the connectivity and completeness of marked enzymes can be used to assess the correctness of functional assignments in the gene catalog. The existence of a missing element implies either the gene catalog is wrong or there is an unknown reaction pathway that utilizes different enzymes in the catalog. For the latter possibility, the deduction from binary relations of substrates and products is useful. In any case, the pathway information is critical in the finding and functional assignment of genes in the genome projects.

As in the sequence alignment and 3D structure alignment, the pathway alignment will become an important tool to identify global and local similarities between two pathways or a consensus among many pathways. For example, the comparison of organism-specific pathways will identify functional similarities and differences, as well as evolutionary relationships, between organisms. Because pathway data are linked to a diverse range of data in KEGG and DBGET, they can be analyzed in many different perspectives. For example, by examining where the enzymes in the same operon appear on the pathway will give insights into the regulation of gene expression, as well as the evolutionary implications of gene structures.

The pathway computations described above will have direct practical applications, which may be collectively called pathway engineering. For example, from pathway comparisons and analyses, an effective pesticide or a side-effect free drug may better be designed. Genetic engineering was based on the DNA sequence information, and protein engineering was based on the protein 3D structures. With the availability of new types of data on the wiring-diagrams of living systems, pathway engineering is bound to emerge as a new biotechnology in the 21st century.

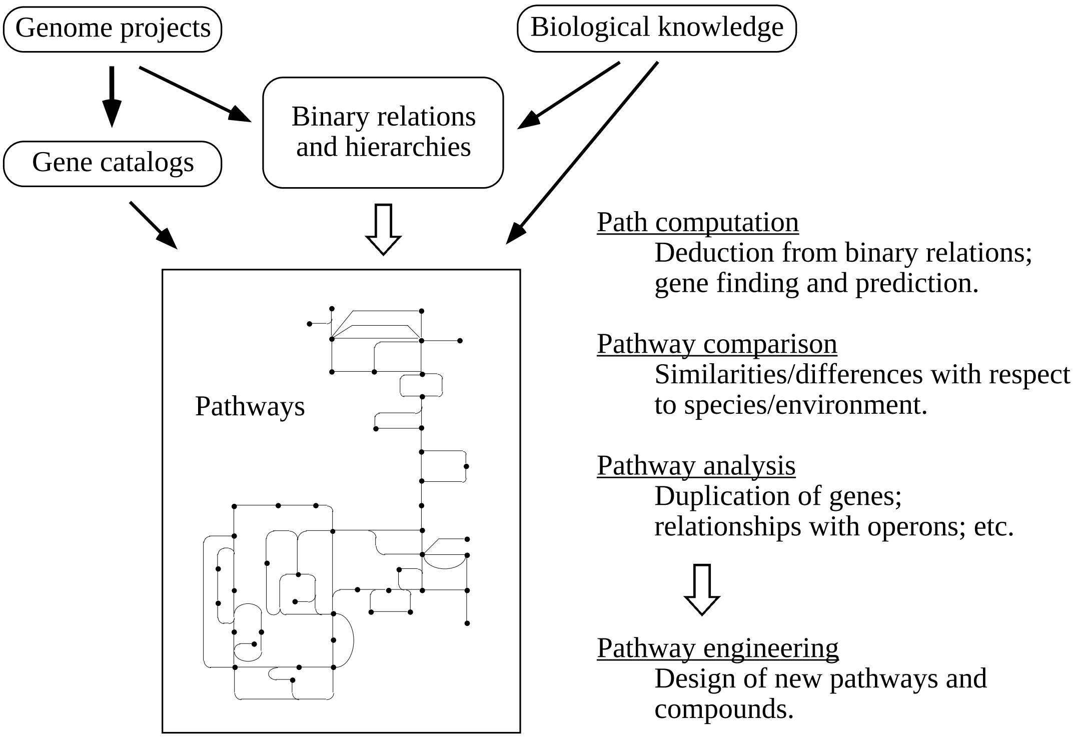

Genome projects
Biological knowledge
Binary relations and hierarchies
Gene catalogs
Pathways
Path computation
Deduction from binary relations; gene finding and prediction.
Pathway comparison
Similarities/differences with respect to species/environment.
Pathway analysis
Duplication of genes; relationships with operons; etc.
Pathway engineering
Design of new pathways and compounds.

Figure 3: Pathway engineering will become feasible once data and knowledge are properly computerized and new computational methods are developed.