Chemical Information

THE KEGG (Kyoto Encyclopedia of Genes and Genomes).

This is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. KEGG is utilized for bioinformatics research and education, including data analysis in genomics, metagenomics, metabolomics and other omics studies, modelling and simulation in systems biology, and translational research in drug development.

Introduction

The KEGG database project was initiated in 1995 by Minoru Kanehisa, Professor at the Institute for Chemical Research, Kyoto University, under the then ongoing Japanese Human Genome Program. Foreseeing the need for a computerized resource that can be used for biological interpretation of genome sequence data, he started developing the KEGG PATHWAY database. It is a collection of manually drawn KEGG pathway maps representing experimental knowledge on metabolism and various other functions of the cell and the organism. Each pathway map contains a network of molecular interactions and reactions and is designed to link genes in the genome to gene products (mostly proteins) in the pathway. This has enabled the analysis called KEGG pathway mapping, whereby the gene content in the genome is compared with the KEGG PATHWAY database to examine which pathways and associated functions are likely to be encoded in the genome.

According to the developers, KEGG is a "computer representation" of the biological system. It integrates building blocks and wiring diagrams of the system — more specifically, genetic building blocks of genes and proteins, chemical building blocks of small molecules and reactions, and wiring diagrams of molecular interaction and reaction networks. This concept is realized in the following databases of KEGG, which are categorized into systems, genomic, chemical, and health information.

- Systems information
  - PATHWAY — pathway maps for cellular and organismal functions
  - MODULE — modules or functional units of genes
  - BRITE — hierarchical classifications of biological entities
- Genomic information
  - GENOME — complete genomes
  - GENES — genes and proteins in the complete genomes
  - ORTHOLOGY — ortholog groups of genes in the complete genomes
- Chemical information
  - COMPOUND, GLYCAN — chemical compounds and glycans
  - REACTION, RPAIR, RCLASS — chemical reactions
  - ENZYME — enzyme nomenclature
- Health information
  - DISEASE — human diseases
  - DRUG — approved drugs
  - ENVIRON — crude drugs and health-related substances

Databases

Systems information

The KEGG PATHWAY database, the wiring diagram database, is the core of the KEGG resource. It is a collection of pathway maps integrating many entities including genes, proteins, RNAs, chemical compounds, glycans, and chemical reactions, as well as disease genes and drug targets, which are stored as individual entries in the other databases of KEGG. The pathway maps are classified into the following sections:

- Metabolism
- Genetic information processing (transcription, translation, replication and repair, etc.)
- Environmental information processing (membrane transport, signal transduction, etc.)
- Cellular processes (cell growth, cell death, cell membrane functions, etc.)
- Organismal systems (immune system, endocrine system, nervous system, etc.)
- Human diseases
- Drug development

The metabolism section contains aesthetically drawn global maps showing an overall picture of metabolism, in addition to regular metabolic pathway maps. The low-resolution global maps can be used, for example, to compare metabolic capacities of different organisms in genomics studies and different environmental samples in metagenomics studies. In contrast, KEGG modules in the KEGG MODULE database are higher-resolution, localized wiring diagrams, representing tighter functional units within a pathway map, such as subpathways conserved among specific organism groups and molecular complexes. KEGG modules are defined as characteristic gene sets that can be linked to specific metabolic capacities and other phenotypic features, so that they can be used for automatic interpretation of genome and metagenome data.

Another database that supplements KEGG PATHWAY is the KEGG BRITE database. It is an ontology database containing hierarchical classifications of various entities including genes, proteins, organisms, diseases, drugs, and chemical compounds. While KEGG PATHWAY is limited to molecular interactions and reactions of these entities, KEGG BRITE incorporates many different types of relationships.

Genomic information

Several months after the KEGG project was initiated in 1995, the first report of the completely sequenced bacterial genome was published. Since then all published complete genomes are accumulated in KEGG for both eukaryotes and prokaryotes. The KEGG GENES database contains gene/protein-level information and the KEGG GENOME database contains organism-level information for these genomes. The KEGG GENES database consists of gene sets for the complete genomes, and genes in each set are given annotations in the form of establishing correspondences to the wiring diagrams of KEGG pathway maps, KEGG modules, and BRITE hierarchies. These correspondences are made using the concept of orthologs. The KEGG pathway maps are drawn based on experimental evidence in specific organisms but they are designed to be applicable to other organisms as well, because different organisms, such as human and mouse, often share identical pathways consisting of functionally identical genes, called orthologous genes or orthologs. All the genes in the KEGG GENES database are being grouped into such orthologs in the KEGG ORTHOLOGY (KO) database. Because the nodes (gene products) of KEGG pathway maps, as well as KEGG modules and BRITE hierarchies, are given KO identifiers, the correspondences are

established once genes in the genome are annotated with KO identifiers by the genome annotation procedure in KEGG.

Chemical information

The KEGG metabolic pathway maps are drawn to represent the dual aspects of the metabolic network: the genomic network of how genome-encoded enzymes are connected to catalyse consecutive reactions and the chemical network of how chemical structures of substrates and products are transformed by these reactions. A set of enzyme genes in the genome will identify enzyme relation networks when superimposed on the KEGG pathway maps, which in turn characterize chemical structure transformation networks allowing interpretation of biosynthetic and biodegradation potentials of the organism. Alternatively, a set of metabolites identified in the metabolome will lead to the understanding of enzymatic pathways and enzyme genes involved. The databases in the chemical information category, which are collectively called KEGG LIGAND, are organized by capturing knowledge of the chemical network. In the beginning of the KEGG project, KEGG LIGAND consisted of three databases: KEGG COMPOUND for chemical compounds, KEGG REACTION for chemical reactions, and KEGG ENZYME for reactions in the enzyme nomenclature. Currently, there are additional databases: KEGG GLYCAN for glycans and two auxiliary reaction databases called RPAIR (reactant pair alignments) and RCLASS (reaction class). KEGG COMPOUND has also been expanded to contain various compounds such as xenobiotics, in addition to metabolites.

Health information

In KEGG, diseases are viewed as perturbed states of the biological system caused by perturbants of genetic factors and environmental factors, and drugs are viewed as different types of perturbants. The KEGG PATHWAY database includes not only the normal states but also the perturbed states of the biological systems. However, disease pathway maps cannot be drawn for most diseases because molecular mechanisms are not well understood. An alternative approach is taken in the KEGG DISEASE database, which simply catalogs known genetic factors and environmental factors of diseases. These catalogs may eventually lead to more complete wiring diagrams of diseases. The KEGG DRUG database contains active ingredients of approved drugs in Japan, the US, and Europe. They are distinguished by chemical structures and/or chemical components and associated with target molecules, metabolizing enzymes, and other molecular interaction network information in the KEGG pathway maps and the BRITE hierarchies. This enables an integrated analysis of drug interactions with genomic information. Crude drugs and other health-related substances, which are outside the category of approved drugs, are stored in the KEGG ENVIRON database. The databases in the health information category are collectively called KEGG MEDICUS, which also includes package inserts of all marketed drugs in Japan.

Subscription model

In July 2011 KEGG introduced a subscription model for FTP download due to a significant cutback of government funding. KEGG continues to be freely available through its website, but the subscription model has raised discussions about sustainability of bioinformatics databases.

BioCyc database collection

The BioCyc database collection is an assortment of organism specific Pathway/ Genome Databases (PGDBs). They provide reference to genome and metabolic pathway information for thousands of organisms. As of December 2016, there are 9300 databases within BioCyc. SRI International, based in Menlo Park, California, maintains the BioCyc database family.

Categories of Databases within BioCyc:

Based on the manual curation done, BioCyc database family is divided into 3 tiers:

Tier 1: Databases which have received at least one year of literature based manual curation. Currently there are seven databases in Tier 1. Out of the seven, MetaCyc is a major database that contains almost 2500 metabolic pathways from many organisms. The other important Tier 1 database is HumanCyc which contains around 300 metabolic pathways found in humans. The remaining five databases include, EcoCyc (*E. coli*), AraCyc (*Arabidopsis thaliana*), YeastCyc (*Saccharomyces cerevisiae*), LeishCyc (*Leishmania major Friedlin*) and TrypanoCyc (*Trypanosoma brucei*).

Tier 2: Databases that were computationally predicted but have received moderate manual curation (most with 1–4 months curation). Tier 2 Databases are available for manual curation by scientists who are interested in any particular organism. Tier 2 databases currently contain 43 different organism databases.

Tier 3: Databases that were computationally predicted by PathoLogic and received no manual curation. As with Tier 2, Tier 3 databases are also available for curation for interested scientists.

Software Tools within BioCyc:

The BioCyc website contains a variety of software tools for searching, visualizing, comparing, and analyzing genome and pathway information. It includes a genome browser, and browsers for metabolic and regulatory networks. The website also includes tools for painting large-scale ("omics") datasets onto metabolic and regulatory networks, and onto the genome.

Use of BioCyc Database Collection in Research:

Since BioCyc Database family comprises a long list of organism specific databases and also data at different systems level in a living system, the usage in research has been in a wide variety of context. Here, two studies are highlighted which show two different varieties of uses, one on a genome scale and other on identifying specific SNPs (Single Nucleotide Polymorphisms) within a genome.

AlgaGEM

AlgaGEM is a genome scale metabolic network model for a compartmentalized algae cell developed by Gomes de Oliveira Dal'Molin et al. based on the *Chlamydomonas reinhardtii* genome. It has 866 unique ORFs, 1862 metabolites, 2499 gene-enzyme-reaction-association

entries, and 1725 unique reactions. One of the Pathway databases used for reconstruction is MetaCyc.

SNPs

The study by Shimul Chowdhury et al. showed association differed between maternal SNPs and metabolites involved in homocysteine, folate, and transsulfuration pathways in cases with Congenital Heart Defects (CHDs) as opposed to controls. The study used HumanCyc to select candidate genes and SNPs.

EcoCyc

In bioinformatics **EcoCyc** is a biological database for the bacterium *Escherichia coli* K-12. The EcoCyc project performs literature-based curation of the *E. coli* genome, and of *E. coli* transcriptional regulation, transporters, and metabolic pathways. EcoCyc contains written summaries of *E. coli* genes, distilled from over 36,000 scientific articles. EcoCyc is also a description of the genome and cellular networks of *E. coli* that supports scientists to carry out computational analyses. Data objects in the EcoCyc database describe each *E. coli* gene and gene product. Database objects also describe molecular interactions, including metabolic pathways, transport events, and the regulation of gene expression. EcoCyc provides several genome-scale visualization tools to aid in the analysis of omics data, such as by painting gene expression or metabolomics data onto the full regulatory network of *E. coli*. EcoCyc can be accessed through the EcoCyc web site, as a set of downloadable files, and in conjunction with the Pathway Tools software that can be installed locally on Macintosh, PC/Windows, and PC/Linux computers. The downloadable software provides capabilities that go well beyond the web version of EcoCyc.

MetaCyc

The **MetaCyc** database is one of the largest metabolic pathways and enzymes databases currently available. The data in the database is manually curated from the scientific literature, and covers all domains of life. MetaCyc has extensive information about chemical compounds, reactions, metabolic pathways and enzymes. MetaCyc has been designed for multiple types of uses. It is often used as an extensive online encyclopedia of metabolism. In addition, MetaCyc is used as a reference data set for computationally predicting the metabolic network of organisms from their sequenced genomes; it has been used to perform pathway predictions for thousands of organisms, including those in the BioCyc Database Collection. MetaCyc is also used in metabolic engineering and metabolomics research. MetaCyc includes mini reviews for pathways and enzymes that provide background information as well as relevant literature references. It also provides extensive data on individual enzymes, describing their subunit structure, cofactors, activators and inhibitors, substrate specificity, and, when available, kinetic constants. MetaCyc data on metabolites includes chemical structures, predicted Standard energy of formation, and links to external databases. Reactions in MetaCyc are presented in a visual display that includes the structures of all components. The reactions are balanced and include EC numbers, reaction direction, predicted atom mappings that describe the correspondence between atoms in the reactant compounds and the product compounds, and computed Gibbs free energy. All objects in MetaCyc are clickable and provide easy access to related objects. For example, the page for L-lysine lists all of the

reactions in which L-lysine participates, as well as the enzymes that catalyze them and pathways in which these reactions take place.

Plant MetabolicNetwork (PMN)Databases

PlantCyc is a metabolic pathway reference database containing more than 800 pathways and their catalytic enzymes and genes, as well as compounds from over 350 plant species (See Database Statistics).

- The majority of the pathways have been curated from experimental literature by curators at the PMN and collaborators' sites. In addition, PlantCyc includes hypothetical pathways that are published in peer-reviewed journals based on the educated conjectures of experts, and computationally predicted pathways that have been manually validated by PMN curators.
- Similarly, enzymes in PlantCyc may have experimental support or may be based solely on computational predictions.
- For both pathways and enzymes, evidence codes are assigned to clearly indicate the type of support associated with these database items.

- PlantCyc Pathways
- PlantCyc Content Statistics
- Taxonomic range:primarily Viridiplantae
- Protein sequence source: Varies according to source: All enzymes have been imported from PMN databases or MetaCyc.
- Enzyme functional annotation method: Varies according to source: All enzymes have been imported from PMN databases or MetaCyc.
- Enzyme evidence:
    - Substantial manual curation of enzymes
    - In addition, large-scale computational predictions of enzyme function not subject to curator review

- Pathway prediction program: All pathways have been manually imported from PMN databases or MetaCyc
- SAVI pathway validation lists: UPP 6.0, NPP 6.0, MCP 1.0, AIPP 3.0, and CAPP 3.0
- Pathway evidence:
    - Substantial manual curation of pathways
    - Some computational prediction of pathways followed by SAVI refinement and curator review
    - Additional pathways imported from MetaCyc based on curator inference

REFERENCES AND FURTHER READING

Kanehisa M, Goto S (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes". Nucleic Acids Res. 28 (1): 27–30.

Kanehisa M (1997). "A database for post-genome analysis". Trends Genet. 13 (9): 375–6.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006). "From genomics to chemical genomics: new developments in KEGG". Nucleic Acids Res. 34 (Database issue): D354–7.

Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. by R. Fleischmann, M. Adams and J. Merrick (1995)

Chemical and genomic evolution of enzyme-catalyzed reaction networks by M. Kanehisa (2013)

Goto S, Nishioka T, Kanehisa M. LIGAND database for enzymes, compounds and reactions. Nucleic Acids Res. 1999 Jan 1;27(1):377-9.

Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, Kawasaki T, Kanehisa M. KEGG as a glycome informatics resource. Glycobiology. 2006 May.

Muto A, Kotera M, Tokimatsu T, Nakagawa Z, Goto S, Kanehisa M. Modular architecture of metabolic pathways revealed by conserved sequences of reactions. J Chem Inf Model. 2013 Mar 25.

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010 Jan.

SRI International home page - https://www.sri.com/

Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD. The MetaCyc database of metabolic

pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2012 Jan

Karp PD, Caspi R. A survey of metabolic databases emphasizing the MetaCyc family. Arch Toxicol. 2011 Sep

Chowdhury S, Hobbs CA, MacLeod SL, Cleves MA, Melnyk S, James SJ, Hu P, Erickson SW. Associations between maternal genotypes and metabolites implicated in congenital heart defects. Mol Genet Metab. 2012 Nov

Karp PD, Riley M, Paley SM, Pelligrini-Toole A. EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. Nucleic Acids Res. 1996 Jan.