Nucleic Acids Research, 2025, **53**, D672–D677
https://doi.org/10.1093/nar/gkae909
Advance access publication date: 17 October 2024
Database issue

KEGG: biological systems database as a model of the real world

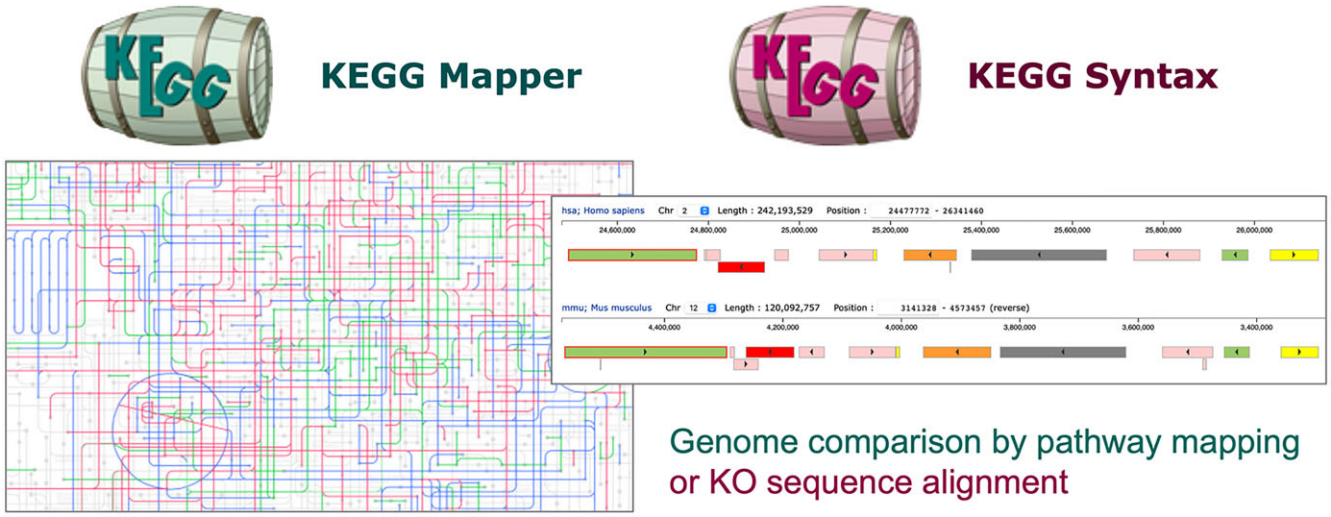
Minoru Kanehisa [1,*], Miho Furumichi[1], Yoko Sato[2], Yuriko Matsuura[2] and Mari Ishiguro-Watanabe[3]

[1]Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
[2]Pathway Solutions Inc., 2-16-3 Higashi-Shinbashi, Minato-ku, Tokyo 105-0021, Japan
[3]Human Genome Center, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108-8639, Japan
[*]To whom correspondence should be addressed. Tel: +81 774 38 4521; Fax: +81 774 38 4523; Email: kanehisa@kuicr.kyoto-u.ac.jp

Abstract

KEGG (https://www.kegg.jp/) is a database resource for representation and analysis of biological systems. Pathway maps are the primary dataset in KEGG representing systemic functions of the cell and the organism in terms of molecular interaction and reaction networks. The KEGG Orthology (KO) system is a mechanism for linking genes and proteins to pathway maps and other molecular networks. Each KO is a generic gene identifier and each pathway map is created as a network of KO nodes. This architecture enables KEGG pathway mapping to uncover systemic features from KO assigned genomes and metagenomes. Additional roles of KOs include characterization of conserved genes and conserved units of genes in organism groups, which can be done by taxonomy mapping. A new tool has been developed for identifying conserved gene orders in chromosomes, in which gene orders are treated as sequences of KOs. Furthermore, a new dataset called VOG (virus ortholog group) is computationally generated from virus proteins and expanded to proteins of cellular organisms, allowing gene orders to be compared as VOG sequences as well. Together with these datasets and analysis tools, new types of pathway maps are being developed to present a global view of biological processes involving multiple organism groups.

Graphical abstract



Introduction

Since 1995 the KEGG database (1,2) has been developed as a computer model of biological systems, such as the cell and the organism, by capturing and organizing knowledge reported in literature. The KEGG model consists of molecular building blocks of genes and molecules, molecular networks of interactions and reactions, and a mechanism to link from building blocks to networks. This is implemented as a collection of databases. Most notably, genes in the genome (GENES database) are linked to KEGG pathway maps (PATHWAY database) through the KEGG Orthology (KO) system (KO database). Pathway maps and other KEGG molecular

Received: September 14, 2024. **Revised:** September 30, 2024. **Editorial Decision:** October 1, 2024. **Accepted:** October 2, 2024

Category	Database	Data object	kid prefix or combined form
Systems Information	PATHWAY	KEGG pathway maps	map | ko, ec, rn, <org>
	BRITE	BRITE functional hierarchies and tables	br, jp, ko | <org>
	MODULE	KEGG modules	M | <org>_M
		Reaction modules	RM
Genomic Information	KO	KO groups for functional orthologs	K
	GENES	KEGG organism genes and proteins	<org>:<gene>
		Virus genes and proteins	vg:<gene>
		Virus mature peptides	vp:<gene-no>
		Functionally characterized proteins from literature	ag:<protein>
	GENOME	KEGG organisms	T, gn:<org>
		KEGG viruses	gn:<vtax>
Chemical Information	COMPOUND	Metabolites and other small molecules	C
	GLYCAN	Glycans	G
	REACTION	Biochemical reactions	R
	RCLASS	Reaction class	RC
	ENZYME	Enzyme nomenclature	ec:<ecnum>
Health Information	NETWORK	Network elements	N
		Network variation maps	nt
	VARIANT	Human gene variants	hsa_var:<gene_vno>
	DISEASE	Human diseases	H
	DRUG	Drugs	D
	DGROUP	Drug groups	DG

<org>	three- or four-letter KEGG organism code	<vtax>	NCBI virus taxonomy ID
<gene>	NCBI gene ID or locus_tag	<ecnum>	EC number
<gene-no>	Gene ID followed by peptide number	<gene-vno>	Gene ID followed by variant number
<protein>	NCBI protein ID or PubMed ID and number	|	separates reference and expanded objects

Figure 1. KEGG consists of various data objects stored in sixteen databases in four categories. Each object (database entry) is identified by the KEGG identifier (kid) as defined here. It takes one of two forms: a simple form consisting of a dataset-dependent prefix followed by a five-digit number (such as map01310) or a combined form consisting of a dataset name and an entry name separated by a colon (such as hsa:116337).

networks are developed in a generic way using functional orthologs of KOs as network nodes, in order to generalize experimental knowledge in specific organisms to other organisms. Each KO is manually defined from experimental evidence and the grouping of each KO is expanded both manually and computationally to cover a set of complete genomes called KEGG organisms. Thus, once genes in any genome are assigned KO identifiers (K numbers), organism-specific versions of molecular networks can be reconstructed, enabling to uncover cellular and organism-level features hidden in the genome.

With the increasing repertoire of KOs and the increasing number of KEGG organisms, the KEGG model may have potential for helping to analyze the biosphere as an open system with the Earth environment. We have recently released a simple pathway map for the nitrogen cycle (map01310), which indicates the roles of different organism groups in different chemical transformation processes of a biogeochemical cycle. Another aspect of biosphere analysis is co-evolution of cellular organisms and viruses. We have developed a computationally generated dataset of virus ortholog groups (VOGs) among viral proteins (3), which are then expanded to proteins of cellular organisms in order to characterize conserved genes and conserved gene clusters in virus-organism relationships. This paper reports these and other developments in the last two years.

Overview of KEGG

Database

KEGG (https://www.kegg.jp) is a database resource for representation and analysis of biological systems. As shown in Figure 1, it consists of sixteen manually curated databases for various data objects representing (i) molecular network systems in the systems information category, (ii) genetic building blocks in the genomic information category, (iii) chemical building blocks in the chemical information category and (iv) disease-related perturbed systems in the health information category. Each data object is identified by the KEGG identifier (kid), which takes one of two forms. A simple form is used for KEGG-original datasets, consisting of a dataset-dependent prefix followed by a five-digit number. A combined form is used for datasets introduced from outside sources, consisting of a dataset name and an entry name separated by a colon. For the three databases in the systems information category, PATHWAY, BRITE and MODULE, the molecular network objects of pathway maps, Brite hierarchies and KEGG modules are expanded from manually created reference objects to computationally generated organism-specific objects, such as from map01100 (reference metabolic pathways) to hsa01100 (human metabolic pathways).

The KEGG database is stored internally as an Oracle relational database. For outside services, flat files are generated for handling by the DBGET system (4), which has been used as

Table 1. KEGG data viewers

Data object	Viewer	URL form[a]
All KEGG data objects	DBGET viewer	https://www.kegg.jp/entry/<kid>
Pathway maps	Pathway viewer[b]	https://www.kegg.jp/pathway/<kid>
Brite hierarchies	Brite viewer[b]	https://www.kegg.jp/brite/<kid>
KEGG modules	Module viewer	https://www.kegg.jp/module/<kid>
Genomes	Genome browser[b)]	https://www.kegg.jp/genome/<kid>
Network variation maps	Network viewer	https://www.kegg.jp/network/<kid>

[a] <kid> represents KEGG identifier shown in Figure 1.
[b] Part of KEGG Web Apps shown in Table 2.

Table 2. KEGG analysis tools

Tool	Feature	URL
KEGG Mapper	KEGG mapping tools including Reconstruct, Search, Color, Join and MWsearch	https://www.kegg.jp/kegg/mapper/
KEGG Web Apps	Pathway viewer, Brite viewer and Genome browser with mapping capabilities	https://www.kegg.jp/kegg/webapp/
KEGG Syntax	Analysis of conserved genes, gene sets and gene orders in organism/virus groups	https://www.kegg.jp/kegg/syntax/
BlastKOALA	Automatic KO assignment	https://www.kegg.jp/blastkoala/
GhostKOALA		https://www.kegg.jp/ghostkoala/

the backbone retrieval system in KEGG. Currently, however, its search capabilities are being replaced by SQLite interfaces and DBGET is used mostly for retrieving and viewing data specified by the KEGG identifier. The DBGET viewer presents flat file views of all data objects in KEGG, and can be invoked by simply appending /entry/kid to the base URL as shown in Table 1. Specialized viewers are available for the five types of molecular network objects (Table 1), including genomes that are considered one-dimensional networks of genes. Among them, Pathway viewer, Brite viewer and Genome browser are JavaScript based applications called KEGG Web Apps (Table 2) with many operations performed on the client side.

Analysis tools

KEGG analysis tools have been expanded and reorganized as shown in Table 2. KEGG Mapper is a collection of KEGG mapping tools, which started as a simple tool for searching and coloring pathway maps at the beginning of the KEGG project and was significantly expanded over the years (5,6). Recently, a special-purpose search tool called MWsearch was added for analyzing mass spectrometry data. With the availability of Pathway viewer and Brite viewer (KEGG Web Apps), which are capable to perform mapping operations on the client side (6), KEGG Mapper is integrated with these viewers allowing, whenever possible, to split server-side database search operations and client-side coloring and other mapping operations.

KEGG Syntax (Table 2) is a new name given to the collection of existing tools, including ortholog table and taxonomy mapping tools, enhanced with a new tool for gene order analysis. As of September 2024 the GENOME database contains over 10 thousand complete genomes of cellular organisms (KEGG organisms), covering a wide range of taxonomic distributions. The corresponding GENES database contains over 50 million genes with the KO assignment rate of about 53%. In contrast, the KO assignment rate for viruses is very low, only about 8%. In order to supplement KOs, virus ortholog groups (VOGs) are computationally generated from 670 thousand viral proteins as described below. Thus, KEGG Syntax allows analysis of conserved genes (KOs), conserved gene sets forming functional units (KEGG modules) and conserved gene orders (conserved synteny) in the context of taxonomic grouping, which may help better understand the genetic building blocks of the biosphere.

New developments in KEGG

VOG (virus ortholog group)

VOG (Virus Ortholog Group) is a computationally generated dataset using the same resource already established for KO annotation. All genome pairs in KEGG are subject to SSDB (Sequence Similarity DataBase) computation using the SSEARCH program for both amino acid sequences (protein coding genes) and nucleotide sequences (RNA genes). For each gene an organism-based list of top-hit similarity neighbors is generated and displayed in a tabular form, called the GFIT table, which is the most basic dataset for KO annotation. In the SSDB computation the vg (virus gene) category is treated as a single organism, and the similarity relation among virus genes is displayed in the paralog GFIT table. The measure of similarity is defined by a modified identity score with weighting of min(1, overlap*2/(aalen1 + aalen2)) for the identity score of the overlap (aligned) region given by SSEARCH. Paralog GFIT tables for all viral proteins are processed in the order of decreasing table sizes, and VOGs are generated by a heuristic method effectively performing single linkage clustering (3). In practice, three VOG datasets are generated with the modified identity threshold of 30%, 50% and 70%, and each VOG is given a six-digit number identifier starting with 3, 5 and 7, respectively. This is not a stable identifier and may change when the GENES vg category taken from RefSeq (7) is updated. Furthermore, all proteins of cellular organisms (KEGG organisms) are compared against the three datasets to see if any of them can be considered to belong to a VOG.

The current statistics of the VOG dataset is available in the KEGG Virus page (https://www.kegg.jp/kegg/genome/virus.html). About 90% of viral proteins belonged to VOGs of size 2 or larger when the threshold of 30% was used. The

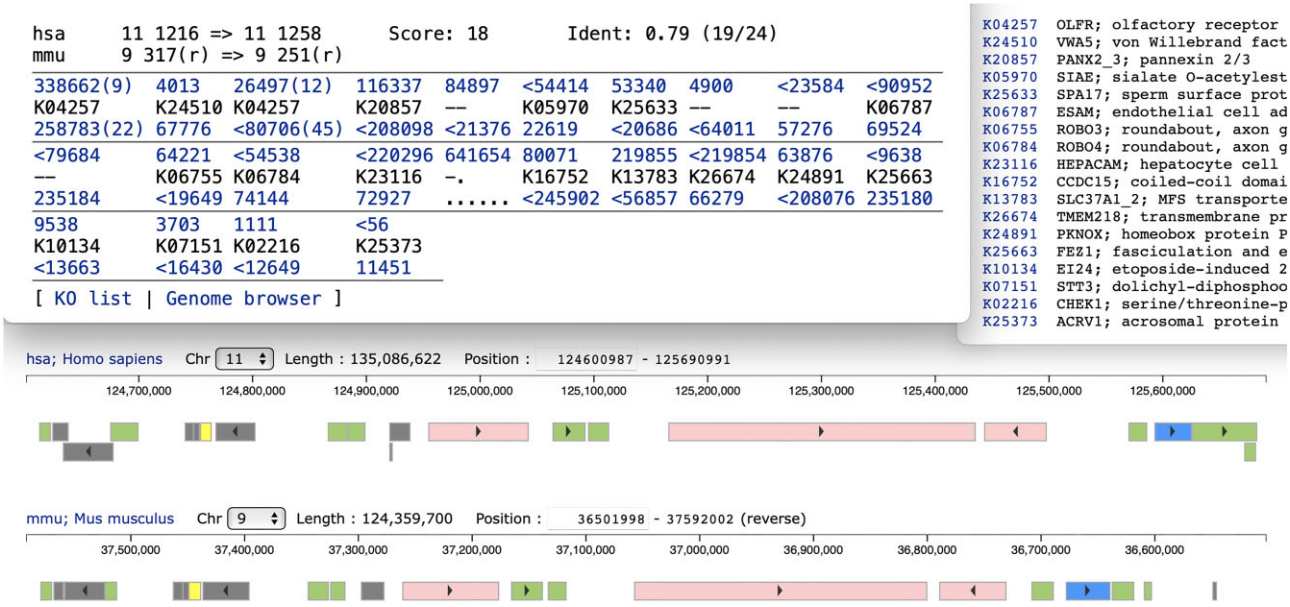


Figure 2. One of the local gene order alignments obtained by comparing KO sequences converted from 21 325 human genes and 22 435 mouse genes. Gene identifiers are aligned with matching K numbers in the middle (upper left) and the list of KOs can be viewed from the link (upper right). This particular alignment in human chromosome 11 and mouse chromosome 9 (reversed) contains olfactory receptor repeats at first and the two genome maps shown here starts with hsa:116 337 and mmu:208098. The coloring of genes indicates functional categories of KOs. To reproduce this result, access https://www.kegg.jp/kegg/syntax/gnalign.html, enter the organism codes hsa and mmu as Genome1 and Genome2, respectively, and click on 'Align by KOs' button.

largest VOG contained 8% of viral proteins, while all the other groups were much smaller, each containing 0.7% or less. The majority of viral proteins taken from RefSeq were phage proteins (80%) and the largest VOG was dominated by phage proteins as well (93%). According to RefSeq annotation, one third of proteins belonging to the largest VOG were hypothetical proteins, and the most frequently annotated term was HNH endonuclease. When cellular organisms were included, about 5% of 50 million proteins shared similarity with viral proteins.

Gene order alignment

Genome alignment is usually done by aligning nucleotide sequences of two genomes. Here, the genome is considered as a sequence of genes identified by KOs (K numbers) or VOGs (VOG numbers) and the genome alignment is done by aligning sequences of matching K numbers or VOG numbers. As mentioned, KOs are assigned to 53% of cellular organism genes and VOGs are assigned to 90% of virus genes. Thus, this approach significantly simplifies the problem of gene order alignment. We have developed a new tool for finding all instances of locally similar gene orders in two genomes above a given threshold using the dynamic programming algorithm developed by Goad and Kanehisa (8) at Los Alamos in the early 1980s. The essence of this algorithm is to perform pruning of paths by taking a logical product of forward and reverse path matrices, in addition to the pruning associated with the weighting scheme of not allowing negative score values, which was also used in the Smith-Waterman algorithm (9) implemented as the SSEARCH program.

The new tool is made available for comparison of two genomes as part of the KEGG Syntax suite (Table 2). As an example, Figure 2 shows one of about 1 000 local alignments with at least 3 matching K numbers when the gene orders of human and mouse are compared. The alignment displays gene identifiers of two genomes and matching K numbers in the middle, with < sign indicating the complementary strand. The list of K numbers may be examined to see if there are any

Table 3. Ortholog table and related tools

Tool	Feature
Ortholog table	• For a given set of KOs, currently assigned genes are shown in a tabular form with coloring of cells for adjacent gene sets • KO assigned genes in both cellular organisms and viruses • Linked to taxonomy mapping summary view • Interface available in KEGG Annotation and KO database pages
Gene cluster	• For a given gene and up to ten adjacent genes on both sides, similar genes in GFIT tables are displayed in a tabular form with coloring of cells for adjacent gene sets • Cellular organisms only • Linked from 'Gene cluster' button in each GENES entry page
VOG cluster	• For a given virus gene and up to five adjacent genes on both sides, genes belonging to the same VOGs are displayed in a tabular form with coloring of cells for adjacent gene sets • VOG assigned genes in both viruses and cellular organisms • Linked to taxonomy mapping summary view • Linked from 'VOG cluster' button in each virus gene entry page

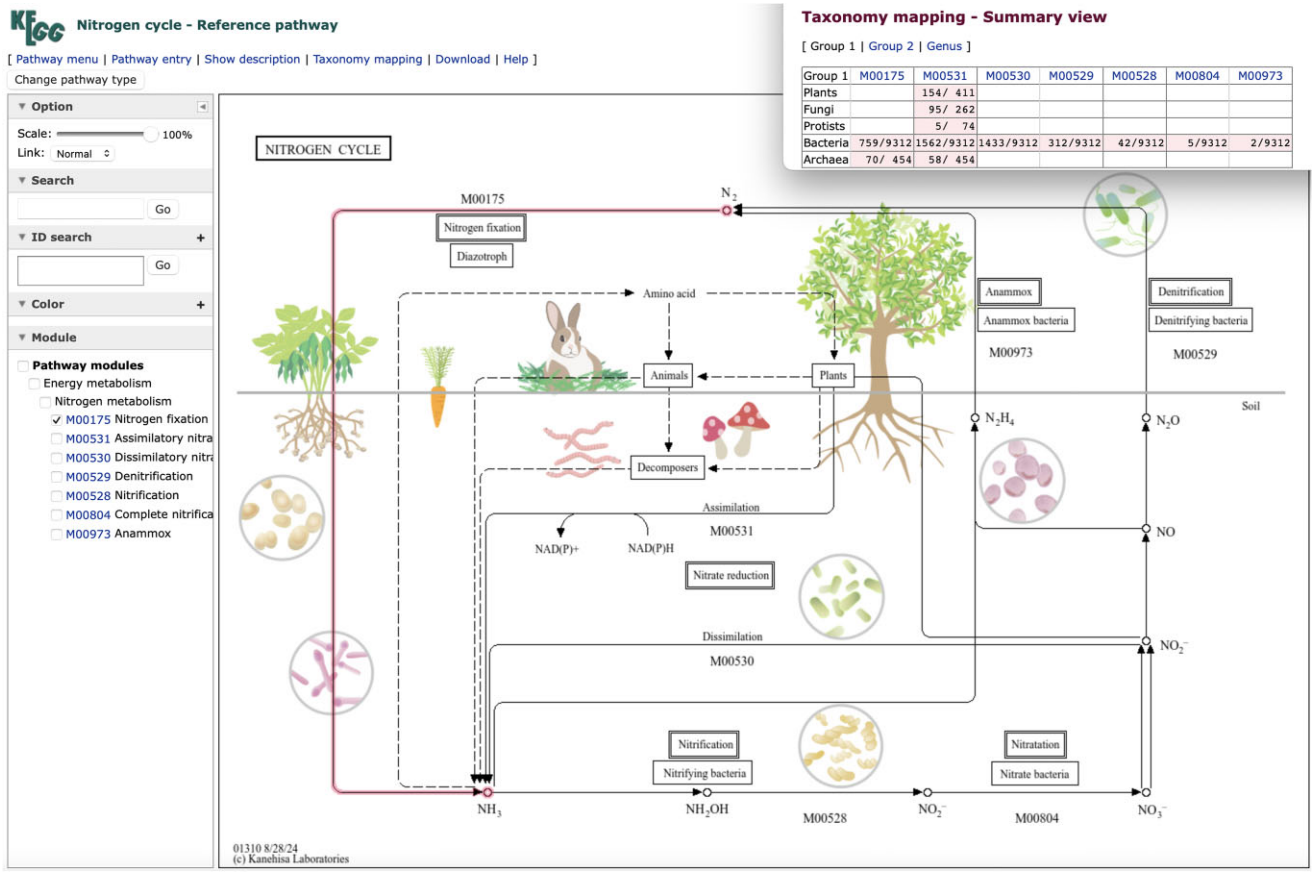


Figure 3. KEGG pathway map for nitrogen cycle (https://www.kegg.jp/pathway/map01310), a new biogeochemical cycle map. KEGG modules in the left panel can be used to display a specific chemical transformation process as a red colored segment, such as M00175 for nitrogen fixation, and also to examine enzyme genes involved. The map contains a link to Taxonomy mapping, which displays in a separate window (top right corner) taxonomic categories of organism groups involved according to the seven constituent modules.

functional correlations. The tool compares two genomes twice in two directions, forward-forward and forward-reverse, the latter indicated by (r) in the second genome. When genes with the same K numbers are repeated, they are combined into a single unit with the number of repeats in parentheses in the output. In this example, varying numbers of olfactory receptor (K04257) repeats are matched. The gene orders after these repeats are shown as a part of the genome map for human and a part of the reversed genome map for mouse.

In order to enable a more comprehensive analysis, precomputed datasets are being created for conserved gene orders using both KO sequences and VOG sequences. In addition, other tools summarized in Table 3 may be used to obtain multiple alignment views (without gaps) of conserved genes and conserved gene orders. The ortholog table tool existed from the beginning of the KEGG project. For a given set of K numbers it displays the current assignment of genes in KEGG organisms and now viruses as well. The same coloring of cells in the same row means that genes are adjacent to each other. A more direct way to examine conserved gene orders is to use 'Gene cluster' button and 'VOG cluster' button in the GENES entry page for cellular organisms and viruses, respectively. The former is based on sequence similarity in GFIT tables, while the latter is based on VOG assignments. Both display alignments in a tabular form with coloring of cells in a similar way as the ortholog table.

Taxonomy mapping of pathway maps

The KEGG database uses the NCBI taxonomy (10) for classification of cellular organisms and viruses, where different versions of taxonomy trees are implemented as multiple Brite hierarchy files. The default file for cellular organisms (br08611) is a classification of three- or four-letter KEGG organism codes according to the fixed levels of taxonomic ranks: phylum, class, order, family, genus and species. The default file for viruses (br08621) is a classification of vtax identifiers (Figure 1) according to the fixed levels of taxonomic ranks: realm, kingdom, phylum, class, order, family, genus and species (11).

Taxonomy mapping is the process to map genomic contents of KOs (K numbers), modules (M numbers) and VOGs to a KEGG taxonomy file. The result is displayed by the KEGG taxonomy browser, which is a special-purpose Brite hierarchy viewer (6). The browser has a zooming capability to change the bottom level of the taxonomic rank. Another display of taxonomy mapping is recently introduced as a summary view of broad taxonomic categories with the number of mapped organisms or viruses in each category.

The original concept of KEGG pathway maps was to manually create generic maps with nodes represented by KOs, which can then be adapted for each organism by converting KOs to specific gene IDs, resulting in organism-specific pathway maps. The generic (reference) pathway map may be applied to an organism group, rather than a single organism, and even to a collection of organism groups. Figure 3 is a new

pathway map for nitrogen cycle (map01310), which is essentially the same as the existing map for nitrogen metabolism (map00910), but it emphasizes how chemical compounds are transformed as a biogeochemical cycle and how different organism groups are involved in specific transformation processes. The involvement of organism groups is represented by taxonomy mapping of seven KEGG modules, each linked to a specific transformation process. This can be displayed as a red segment in Pathway viewer by selecting a module, such as M00175 for nitrogen fixation (Figure 3).

Other improvements of KEGG

KO annotation

The number of KEGG organisms is increasing, at the moment, by about 80 per month. The KEGG annotation procedure of assigning KOs has been streamlined to cope with this accelerated increase. First, protein coding genes in a new genome are compared by BLAST with a small reference sequence dataset, which is the same as the one provided for the BlastKOALA server (Table 3). Second, both protein and RNA coding genes are compared by SSEARCH with the entire GENES dataset and an automatic annotation is performed using the new KOALA program. In addition to computational genome-based annotations, manual KO-based annotations are performed by creating groups of sequences whenever new KOs are defined or existing KOs are modified. The consistency of the entire GENES annotation is checked every day, and additional candidates and possible misannotations are presented for human intervention.

Network-disease association

KEGG MEDICUS is a practical resource integrating the health information category of KEGG (Figure 1) with drug labels of marketed drugs in Japan and the USA by assigning D number identifiers. Japanese drug labels are obtained from JAPIC (Japan Pharmaceutical Information Center) and incorporated in the KEGG Oracle database. FDA's National Drug Code (NDC) directory is used to create links to the DailyMed database for drug labels in the USA. In contrast to the other three categories of KEGG, which model molecular systems at the cellular, organism and biosphere levels in a generic way, the health information category models human molecular systems, especially perturbed systems associated with human diseases. Network variation maps in the NETWORK database present an integrated view of how reference molecular networks are perturbed by human gene variants, viruses, etc., how perturbations are associated with specific diseases, and what drugs and which targets are available. Network variation maps have been developed for a number of metabolic and signaling networks, most of which are linked to KEGG pathway maps. As a result, an increasing number of entries in the DISEASE database are linked to network variation maps, currently about 30%, showing network-disease associations.

Data availability

KEGG is a self-sustaining database. Without any substantial public funding, it is based mainly on the 'community funding' model, whereby the KEGG user community contributes financially to the development and maintenance of the database. KEGG is updated daily and made available at the KEGG website (https://www.kegg.jp/). The content is mirrored to the GenomeNet website (https://www.genome.jp/kegg/) one day later. Major updates of database contents and web services are announced every three months with the release number.

Acknowledgements

Computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Funding

NBDC Database Integration Coordination Program JPMJND2203 of the Japan Science and Technology Agency (partial funding for KEGG MEDICUS). Funding for open access charge: NBDC Program of the Japan Science and Technology Agency.

Conflict of interest statement

None declared.

References

1. Kanehisa,M. (2019) Toward understanding the origin and evolution of cellular organisms. *Protein Sci.*, **28**, 1947–1951.
2. Kanehisa,M., Furumichi,M., Sato,Y., Kawashima,M. and Ishiguro-Watanabe,M. (2023) KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.*, **51**, D587–D592.
3. Jin,Z., Sato,Y., Kawashima,M. and Kanehisa,M. (2023) KEGG tools for classification and analysis of viral proteins. *Protein Sci.*, **32**, e4840.
4. Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa,M. (1998) DBGET/LinkDB: an integrated database retrieval system. *Pac. Symp. Biocomput.*, 683–694.
5. Kanehisa,M. and Sato,Y. (2020) KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci.*, **29**, 28–35.
6. Kanehisa,M., Sato,Y. and Kawashima,M. (2022) KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci.*, **31**, 47–53.
7. Haft,D.H., Badretdin,A., Coulouris,G., DiCuccio,M., Durkin,A.S., Jovenitti,E., Li,W., Mersha,M., O'Neill,K.R., Virothaisakun,J., *et al.* (2024) RefSeq and the prokaryotic genome annotation pipeline in the age of metagenomes. *Nucleic Acids Res.*, **52**, D762–D769.
8. Goad,W.B. and Kanehisa,M.I. (1982) Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucleic Acids Res.*, **10**, 247–263.
9. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
10. Schoch,C.L., Ciufo,S., Domrachev,M., Hotton,C.L., Kannan,S., Khovanskaya,R., Leipe,D., Mcveigh,R., O'Neill,K., Robbertse,B., *et al.* (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**, baaa062.
11. Siddell,S.G., Smith,D.B., Adriaenssens,E., Alfenas-Zerbini,P., Dutilh,B.E., Garcia,M.L., Junglen,S., Krupovic,M., Kuhn,J.H., Lambert,A.J., *et al.* (2023) Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV). *J. Gen. Virol.*, **104**, 001840.

Received: September 14, 2024. **Revised:** September 30, 2024. **Editorial Decision:** October 1, 2024. **Accepted:** October 2, 2024