京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

京都大学
KYOTO UNIVERSITY

Title	The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals.
Author(s)	Kotera, Masaaki; Hirakawa, Mika; Tokimatsu, Toshiaki; Goto, Susumu; Kanehisa, Minoru
Citation	Methods in molecular biology (2012), 802: 19-39
Issue Date	2012
URL	http://hdl.handle.net/2433/152398
Right	The final publication is available at www.springerlink.com
Type	Journal Article
Textversion	author

Kyoto University

The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals

Masaaki Kotera*, Mika Hirakawa, Toshiaki Tokimatsu, Susumu Goto and Minoru Kanehisa

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan.

*To whom correspondence should be addressed. Tel: +81 774 38 3270; Fax: +81 774 38 3269, Email: kot@kuicr.kyoto-u.ac.jp

Abstract

In this chapter, we demonstrate the usability of the KEGG (Kyoto Encyclopedia of Genes and Genomes) databases and tools, especially focusing on the visualization of the omics data. The desktop application KegArray and many web-based tools are tightly integrated with the KEGG knowledgebase, which helps visualize and interpret large amount of data derived from high-throughput measurement techniques including microarray, metagenome and metabolome analyses. Recently developed resources for human disease, drug and plant research are also mentioned.

Keywords: Pathway map, KEGG Orthology (KO), BRITE hierarchy, KEGG API, KegArray.

1. Introduction

"Omics" is a general term for a research field of life science analyzing massive amounts of interactions of biological information objects, including genome, transcriptome, proteome, metabolome, and many other derivatives. As omics data has been rapidly accumulating as the result of recent development of high-throughput measurement techniques, the needs for omics-data integration have been becoming more important. In general, bioinformatics techniques have been developed and utilized to computationally process a vast amount of biological data. However, only the collection and computation of these data is not sufficient to understand the complete and dynamic system of life programmed in the genome sequence. These data must be described as the knowledge on life science, *i.e*., network diagram of various interactions such as cellular functions, signaling/metabolic pathways and enzyme reactions. Thus, we have been focusing on generating the integrated knowledge database named KEGG (Kyoto Encyclopedia of Genes and Genomes) ***(1)*** by the high-quality manual curation.

KEGG can be seen as an efficient viewer of living systems. The main page is located at http://www.kegg.jp/ (**Fig. 1**), and it can also be reached from GenomeNet http://www.genome.jp/. KEGG and GenomeNet have a search option named "*dbget*" ***(2)***, by which the user can use any term without knowing the database structure, just like to "*google*" without knowing how web pages are linked to each other in the Internet. The user can find many similar search boxes in many different pages in KEGG, which can generally be used in the same way, with the mere differences in the selection of databases being searched and the display style. The user needs not know which database contains the data of interest, since the dbget searches all relevant data throughout all databases. This integrity is a big advantage with which the user cannot only look up the data of interest, but can also trace the links to collect and understand the relevant information.

At the first sight, the KEGG data structure seems quite complicated, because there are many web pages (which we refer to as "entry points") focusing on different objects and different purposes, even though they occasionally reach the same data. However, this becomes actually advantageous when the user learns the basics about the KEGG data structure. **Figure 2** describes the grid-shaped relationships of the KEGG data. KEGG can be divided into the four main databases, PATHWAY, BRITE, GENES and LIGAND, from one perspective. GENES consists of genes and genomes (see **Note 1** for the detail), while LIGAND contains the other objects, *e.g*., metabolites and reactions ***(3)***. PATHWAY describes inter-molecular networks such as regulatory or metabolic pathways, and BRITE is a collection of hierarchical classifications (ontology) of biological or pharmaceutical vocabularies. In other words, GENES and LIGAND are the databases of "components", while PATHWAY and BRITE are those of "circuits" of living systems. On the other hand, the recently developed resources, *e.g*., DISEASE, DRUG and PLANT, view the data in different ways. They focus on human diseases, pharmaceutical compounds and plants, respectively, with the same usability of GENES, LIGAND, PATHWAY and BRITE. Thus, the user can use the same data and tools with the most efficient way depending on the situation and purpose.

2. Experience the structure of PATHWAY / BRITE

KEGG PATHWAY (http://www.kegg.jp/kegg/pathway.html) had started as a computational description of metabolic pathways, and still keeps growing and expanding to represent the phenomenon (such as metabolism, cellular processes and human diseases) manually compiled from published literatures. KEGG has about 400 maps where the genes from genome-sequenced organisms are assigned, and the number of the organisms and pathway maps keeps increasing. In other words, the user is able to compare the genomes in the viewpoint of about 400 phenomenon just by viewing this database.

Browsing the pathway map using KEGG PATHWAY is similar to searching a restaurant using the Internet. The user might want to view and understand the content (the collection of the genes, proteins and small molecules) and context (their interaction) in the organism of interest. The user might input the name of the restaurant into the search box, or narrow down the search area from the map. The KEGG PATHWAY can be used just in the same way, *i.e.*, the user can search the gene or any substances in whichever pathway, or browse many pathways in a specified organism, or compare the specified pathway in many species, just by choosing options or clicking links.

KEGG PATHWAY entries generally do not focus on a specific organism. Reference pathways are defined as the combined pathways that are present in a number of organisms and are consensus among many published papers. Only the reference pathway map is manually drawn; all other organism-specific maps are computationally generated. The KEGG pathway map is manually drawn with in-house software called KegSketch, which generates the KGML (KEGG Markup Language; see http://www.genome.jp/kegg/xml/) file. This xml files contain graphics information and also KEGG entry, relation, and reaction information.

GENES and PATHWAY can be viewed in two different ways (**Fig. 2**): the limited search in an organism of interest, and the comprehensive search throughout the all genome-sequenced organisms. The former method is explained in **Note 2**. Here, we explain the latter method. **Figure 3a** is a screenshot of the inositol phosphate metabolism pathway, which can be seen by clicking one of the links on the PATHWAY main page. In this graphic, rectangles and circles represent gene products (mostly proteins) and other molecules (mostly metabolites), respectively. This black-and-white graphic is one of the reference pathways for which no organism has been specified.

The user can view the organism-specific pathways by using the pull-down menu. **Figure 3b** is taken as an example PATHWAY page of a specified organism. The colored rectangles in this page indicate that there are links to the corresponding GENE pages, which means the specified organism possesses the corresponding genes or proteins in the genome. White rectangles indicate that there are no genes annotated to the corresponding function. Note that this does not necessarily mean the organism does not really have the corresponding genes. It is possible that the corresponding genes have not been identified yet.

Coloring the rectangles in the organism-specific pathways is based on the KEGG Orthology (KO). KO is a collection of the classes of orthologous genes having a common function and the same evolutional origin. An orthology (KO entry) in principle corresponds to more than one genes derived from more than one organisms. Genes assigned to the same orthology correspond to the same rectangle in a PATHWAY map (**Fig. 3a**). The corresponding genes in the PATHWAY maps are assigned for the individual organisms through the KO, so that the user can view the corresponding pathway for the specific organism. When the user specifies an organism, then the genes in the organism corresponding to the KO are linked to the rectangles. The rectangle becomes colored and clickable when the corresponding KO contains genes in the specified organism (**Fig. 3b**). KO entries for GENES (complete genomes) are manually defined and annotated by the KEGG expert curators based on the phylogenetic profiles and functional annotations of the genes. On the other hand, KO for DGENES (draft genomes) and EGENES (EST sequences) are automatically annotated by KAAS (see **Note 3**). DGENES and EGENES have relatively less number of colored rectangles (and less links) due to the less number of genes annotated to KO.

Changing organisms by using the pull-down menu enables the comparison of pathways among organisms. The menu is very long because it contains the entire set of organisms registered in KEGG. Therefore we provide a useful option to customize the menu (**Note 4**). The user can emphasize any genes or chemical compounds using any color to customize the pathway map for presentation (see **Section 3**). KEGG PATHWAY is also useful for understanding the relationships of the genes identified in experiments such as microarray analysis. The user can quickly obtain the graphics representing the functions to which the genes up- (down-) regulated in microarray experiments are related (see **Section 4**).

KEGG PATHWAY recently incorporated new types of pathway maps, named "Global Maps" (**Fig. 3c**), which are also reachable from the PATHWAY top page. The user can map any set of genes to grasp the overview by using the Global Maps. We expect this will become more valuable for the interpretation of metagenome and pangenome studies. We also developed a new graphical interface, KEGG Atlas ***(4)***, to map smaller functional units (such as pathway maps and pathway modules) in the Global Maps with zooming and navigation capabilities (**Fig. 3d**).

KEGG BRITE (http://www.genome.jp/kegg/brite.html) represents the hierarchy of vocabularies used in papers, references and academic communities. It contains the widely accepted classifications derived from other databases or references, and hierarchical classifications that we originally compiled (See **Section 5** and **Fig. 8c** for a DISEASE example), as well as the hierarchy of the substances defined in KEGG (such as KO). The BRITE functional hierarchies contain tab-delimited fields, which can be handled by the desktop application KegHier (downloadable from the KEGG homepage; see **Fig. 1**).

3. Customize the PATHWAY / BRITE as you like

The user can color KEGG PATHWAY/BRITE as necessary. As explained above, when the user specifies an organism, the gene products are colored in pathway maps (**Fig. 3b)**. There is also an option to specify multiple organisms at a time (See **Note 5**). Additionally, when the user inputs the term of interest into the search box, the corresponding objects are colored (as explained in **Fig. 3**). Here, we provide more flexible options to color PATHWAY or BRITE (http://www.genome.jp/kegg/tool/color_pathway.html and http://www.genome.jp/kegg/tool/color_brite.html, respectively). **Figure 4a** is reachable from the KEGG sitemap (see **Fig. 1b)**. The user can easily find any objects of interest (genes, metabolites, *etc*.) in the KEGG PATHWAY or BRITE by coloring them (**Figs. 4c and 4d**). The objects have to be specified by the KEGG IDs. Therefore, if the objects of interest are represented by the identifiers of other databases, they have to be converted into the KEGG IDs (described in **Note 6**).

Another flexible option is available through the KEGG API (http://www.genome.jp/kegg/soap/). KEGG API is a web service to use the KEGG system from the user's program via SOAP/WSDL. The service enables the user to develop software that accesses and manipulates a massive amount of online KEGG contents that are constantly refreshed. KEGG API provides many useful functions, including those for coloring pathways that colors the given objects on the pathway map with the specified colors and returns the URL of the colored image.

For the users who would like to deal with the pathways that are not still present in KEGG PATHWAY, we provide a number of options. See **Note 7** for the detail.

4. Use the KegArray application

KegArray is a Java application that provides an environment to analyze either transcriptome/proteome and metabolome data. Closely integrated with the KEGG database, KegArray enables the user to easily map those data to KEGG resources including PATHWAY, BRITE and genome maps. It can be downloaded from the KegTools page (http://www.genome.jp/kegg/download/kegtools.html) linked from the KEGG homepage (**Fig. 1a**).

KegArray can read the transcriptome data format of the KEGG EXPRESSION database (http://www.genome.jp/kegg/expression/) or tab-deliminated text similar to the EXPRESSION format. Each entry of EXPRESSION consists of brief descriptions about experiment, reference information, and a set of intensity values or ratios of two-channels derived from a DNA microarray. Examples for intensity values and for expression ratios between two-channels are given in **Figures 5a** and **5b**, respectively. KegArray also deals with the metabolome data, although only ratio values can be available as shown in **Figure 5c**. To convert data in Microsoft Excel format for KegArray, the user needs to order the columns as in the KegArray format in advance and save them as a tab-delimited text.

Once KegArray is launched, the user can see the KegArray control panel (**Fig. 6**), where there are two tabs to select "Gene/Compound" or "Clustering" on the top. In the "Gene/Compound" pane, the user can load a data file of transcriptome and/or metabolome experiments from the local computer or the KEGG EXPRESSION database, by clicking the "Local" or "GenomeNet" buttons, respectively. The user can obtain the list of up- or down-regulated genes (or compounds) by choosing the option from the menu. The number of listed genes can be modified by changing the value in the box at upper-right of the pop-up table. The up- or down-regulated genes (or compounds) can be mapped onto PATHWAY, Genome map and BRITE for the user to understand the result (as the examples shown in **Figure 7**).

In the "Clustering" pane, the user can load several data files of transcriptome experiments and set an intensity threshold. Once the user selects more than one data files, the "Clustering" button

becomes active. Clicking this button performs hierarchical clustering of the gene expression profiles constructed from the files listed. A tree-view window is shown when the calculation is completed. The user can change the number of clusters (1 - 6) by specifying the number in the input box at the top of the tree-view window. Different clusters are shown in different colors. Clicking the "Set results" button saves the color-coding for further analysis using the Tools section.

5. Overview the DISEASE/DRUG resources

Before closing this chapter, we briefly explain recently released three resources for specific requirements: DISEASE, DRUG and PLANT. DISEASE database contains information of human molecular system perturbed by gene mutation, infection of pathogens, *etc*. DRUG database contains information of pharmaceutical compounds, identified with the chemical structures and classified hierarchically based on various perspectives: the Anatomical Therapeutic Chemical (ATC) Classification System, US pharmacopeia (USP) classification, Therapeutic category of drugs in Japan, *etc*. Plant species produce those with medical, nutritional and environmental values, which is one of the motivations for us to produce the PLANT resources and the EDRUG database.

KEGG DISEASE (http://www.genome.jp/kegg/disease/) is a new collection of disease entries capturing knowledge on genetic and environmental perturbations. There are a number of disease databases available, but they are mostly descriptive databases for humans to read and understand. Disease information in KEGG is in more computable forms, pathway maps and gene/molecule lists. The Human Diseases category of the KEGG PATHWAY database contains multifactorial diseases such as cancers, immune disorders, neurodegenerative diseases, and circulatory diseases, where known disease genes (genetic perturbants) are marked in red (**Fig. 8a**). Each disease entry contains a list of known genetic factors (disease genes), environmental factors, diagnostic markers, and therapeutic drugs (**Fig. 8b**), which may reflect the underlying molecular network. For single-gene diseases, perturbed pathway maps are not drawn, but causative genes are mapped to normal pathway maps through disease entries. It also contains some infectious diseases where molecular interaction networks of both pathogens and humans are depicted. Diseases with known genetic factors and infectious diseases with known pathogen genomes are being organized in KEGG DISEASE and classified in the BRITE hierarchy (**Fig. 8c**).

KEGG DRUG (http://www.genome.jp/kegg/drug/) is a unified drug information resource that contains chemical structures and/or chemical components of all prescription and over-the-counter (OTC) drugs in Japan, most prescription drugs in the USA, and many prescription drugs in Europe. All the marketed drugs in Japan are fully represented in KEGG DRUG and linked to the package insert information (labels information). These include crude drugs and TCM (Traditional Chinese Medicine) drugs, which are popular in Japan and some of which are specified in the Japanese Pharmacopoeia. Each KEGG DRUG entry distinguishes the chemical structure of chemicals or the chemical component of mixtures and crude drugs. It is associated with generic names, trade names, efficacy and target information, as well as information about the history of drug development. KEGG DRUG contains information about three types of molecular networks. The first is the drug degradation pathways by drug-metabolizing enzymes. The second is the molecular interaction network involving target and other molecules. The drug-target relationship is not simply a molecule-molecule relationship. The target is given in the context of KEGG pathways, enabling the analysis of drugs as perturbants to molecular systems. The last molecular network is the one representing drug development history ***(5)***. Many marketed drugs have been developed from lead compounds or existing drugs by introducing chemical structure transformations retaining the core chemical structures. KEGG DRUG structure maps graphically illustrate knowledge on such drug development in a manner similar to the KEGG pathway maps.

KEGG PLANT is a new resource for plant research, especially for understanding relationships between genomic and chemical information of natural products from plants. This is part of the EDRUG database (http://www.genome.jp/kegg/drug/edrug.html), a collection of natural products such as crude drugs and essential oils. Plants are known to produce diverse chemical compounds including those with medicinal and nutritional properties. The available complete genomes for plants are very limited in comparison to other organism groups such as animals and bacteria. Thus, massive EST datasets have been established for a number of plant species to generate the EGENES database ***(6)*** where EST contigs are treated as genes and automatically annotated with KAAS (see **Note 3**). We have been expanding the repertoire of KEGG pathway maps for plant secondary metabolism, as well as developing the Global Maps and several category maps. The category maps are used to classify plant secondary metabolites as part of the BRITE hierarchy.

Conclusion

In this chapter, we introduced main KEGG resources and their usability. Emphasis was put on the usage for omics studies; however, the KEGG resources are applicable for a variety of studies on life sciences. These useful characteristics of KEGG enable the user to find new idea or to determine future direction for omics analysis. For further reading, we recommend two publications of Wheelock *et al*. ***(7, 8)*** explaining other KEGG contents that are not mentioned in this chapter.

Notes

Note 1. KEGG GENES

KEGG GENES (http://www.genome.jp/kegg/genes.html) is a database of the genes derived from all organisms with the sequenced genomes publicly available. GENES contains nucleic and amino acid sequences, identifiers in KEGG and other databases and the functional KEGG annotation. For eukaryotes, there are DGENES and EGENES databases containing draft genomes and EST sequences, respectively. We also started to collect and annotate metagenome information that is stored as MGENES. Gene and genome sequences have been retrieved from Refseq in NCBI, and other public databases of the genome-sequencing organizations.

Note 2. KEGG Organisms and GENOME

KEGG Organism page (http://www.genome.jp/kegg/catalog/org_list.html) contains a list of organisms with complete genomes (**Figure 9a**). A KEGG Organism code of a complete genome consists of three alphabets, while the code of a draft genome and EST sequences consists of four alphabets beginning with "d" and "e", respectively. KEGG Organism codes are used for specifying organisms, and also used as the headers of the pathway map IDs (*e.g*., hsa00010). We recently started incorporating metagenome and pangenome sequences as well, in order to meet the future needs of environmental and health problems. KEGG Organism page (**Fig. 9a**) contains the links to the metagenome and pangenome data. In addition to the three or four letter organism codes, we introduced T numbers for specifying genomes including metagenomes.

When the user is interested in only one organism, it is efficient to jump to the corresponding GENOME page of interest. Clicking the "mmu", for instance, in the KEGG Organism page (**Fig. 9a)** takes the user to the GENOME page specific for mouse *Mus musculus* (**Fig. 9b**). KEGG provides this type of pages for all registered organisms. The user can also reach to this page from the KEGG GENOME page (http://www.genome.jp/kegg/genome.html).

Note 3. KAAS automatic annotation

KAAS (KEGG Automatic Annotation Server) ***(9)*** has been used for annotating DGENES, EGENES and MGENES in KEGG. The public version of KAAS is available to annotate any groups of gene sequences, when the user wants to display the genes in the organism that is not still a member of the KEGG Organisms, or when the user has a set of sequences for which the corresponding IDs are not known. This service is of particular value when the user has a draft genome, EST, or the sequence sets obtained from microarray analysis. Note that KAAS uses BLAST search, therefore the user should examine the quality and the length of the input sequences just as when using BLAST. Multiple FASTA format is used as an input. KAAS accepts both nucleic and amino acid sequences; however, the two types of sequences should not be mixed in one file.

The user can jump to the KAAS page (http://www.genome.jp/tools/kaas/) by clicking one of the links in the KEGG sitemap (**Fig. 1b**). It is recommended that the user specify a set of organisms that are evolutionally close to the input organism, because the KAAS searches the similar sequences in KO. It may take a while depending on the data size or the status of the server, therefore an e-mail will be sent later to inform the URL to access the result page, containing the corresponding KO list. The automatically colored PATHWAY pages are obtained according to the result. It is recommended that the user download the result since they will be removed from KEGG server in a few days. The results can also be seen in the BRITE form, where the annotated functions such as enzymes, transcription factors and receptors are listed hierarchically to help understand the overview of the gene set.

Note 4. Find organisms more easily

KEGG has already included more than 1,000 organisms, which makes it hard for the user to find the organisms of interest. Therefore, KEGG provides some options by which the user limits only the organisms of interest (**Fig. 10**). Once the user selects this option, it keeps working as long as the cookie retains.

Note 5. KEGG Organism Groups

We also recently defined KEGG Organism Groups, combinations of organisms, enabling the analysis of the combined pathways generated as the results of symbiosis or pathogenesis. The combined pathways can be obtained using the search box located in the middle of the KEGG GENOME page (http://www.genome.jp/kegg/genome.html; **Fig. 11a**). For example, when the user inputs "hsa+pfa", meaning human (*Homo sapiens*) plus a pathogen (*Plasmodium falciparum 3D7*), this option provides the two-colored pathways. These two colors represent the gene products from the two organisms.

In fact, this option is not limited only for symbiosis and pathogenesis, but this accepts any combinations of genomes. For instance, the query "hsa+mmu+dme", which means human (*Homo sapiens*) + mouse (*Mus musculus*) + fruit fly (*Drosophila melanogaster*), provides the three-colored map (**Fig. 11b**) that is useful to compare the three pathways in a map.

Note 6. Accession ID conversion to the KEGG IDs.

KEGG entries have unique identifiers (KEGG IDs), which can be used for coloring the PATHWAY maps and the BRITE hierarchy (see **Section 3**). KEGG ID consists of the abbreviated name of the sub-database and the identifier of the entry connected with a colon (:), e.g., cpd:C00103, where "cpd" means the KEGG COMPOUND database, and "C00103" means the ID number of alpha-D-Glucose 1-phosphate. Another example is hsa:4357, where "hsa" means the KEGG Organism code (see **Note 2**) of human (or, in other words, the human-specific GENOME database), and "4357" means the GENES ID. The abbreviated name of the sub-databases in KEGG can be looked up at http://www.genome.jp/dbget/, and the format of the KEGG IDs can be seen at the KEGG Identifier page (http://www.genome.jp/kegg/kegg3.html).

The user needs KEGG GENES and COMPOUND IDs to color the PATHWAY maps. If the user only has the list of NCBI gene IDs or UniProt IDs, they can be converted to the corresponding KEGG IDs using the option in the KEGG Identifiers page (**Fig. 12**). Entry list style (**Fig. 12c**) is recommended because it can be simply pasted in the input box of the color objects page (**Fig. 4a**). KegArray (**Section 4**) also has an option to convert the external database IDs to the KEGG GENES IDs, which are necessary for mapping the array data to the KEGG resources such as pathway maps.

Note 7. Create new pathway maps that are not present in KEGG

Even though KEGG keeps incorporating novel pathways published recently, there is a good chance that the user finds a pathway that is not still present in KEGG. If this is the case, sending us a request is highly appreciated (See **Note 9**). In some cases, however, the user might need to create new pathway maps that are not present in KEGG. Such cases are divided into two types. In the first type of cases, the steps of the pathway are already described in a KEGG PATHWAY map, although they are not attached to the corresponding genes derived from an organism of interest. In the second type of cases, some (or all) of the steps are not described in the KEGG PATHWAY maps because they are still unpublished or unknown. KEGG provides KAAS to address the first type of cases, as explained in **Note 3**.

To address the second type of cases, PathPred ***(10)*** and E-zyme ***(11,12)*** are available (See **Fig. 1b**). When the user obtains a chemical structure of a metabolite for which the biosynthesis/biodegradation pathway is unknown, PathPred automatically suggests possible pathways. The suggested pathway includes the steps with the plausible EC numbers (enzyme classification IDs established by IUBMB), which are predicted by E-zyme. E-zyme is also available to suggest possible EC numbers for a given (partial) enzyme reaction equation. PathPred and E-zyme require chemical structures as input. If the chemical compounds are registered in KEGG, then the user can use the corresponding KEGG IDs. In the case the user wants to input the chemical compound that is not present in KEGG, or the user does not know the corresponding KEGG ID, we recommend to use KegDraw, a desktop application designed for drawing and searching chemical structures. This application has options to incorporate the chemical structures predefined in KEGG, as well as to edit

the structures. It is notable that this application is also capable of drawing glycan structures ***(13)***. The edited structures of compounds and glycans are also used as queries of the similarity search programs SIMCOMP / SUBCOMP ***(14,15)*** and KCaM ***(16)***, respectively.

Note 8. Retrieving several KEGG entries at a time

KEGG Identifiers page provides an option to retrieve a number of the KEGG entries at a time (**Fig. 13**). This is useful when the user is using a web browser. When retrieving more KEGG entries is preferred, go to the KEGG FTP site (http://www.genome.jp/kegg/download/) or try to use KEGG API (http://www.genome.jp/kegg/soap/).

Note 9. Feedback

We appreciate any suggestions, questions and comments on the KEGG data and tools. We intend that KEGG keeps incorporating more and more genomes, pathways, the BRITE hierarchies, *etc*. Suggesting something that should be added to KEGG is also greatly appreciated. Please send a message to the feedback form (http://www.genome.jp/feedback/).

Acknowledgements

The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University. The KEGG project is supported by the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency, and a grant-in-aid for scientific research on the priority area 'Comprehensive Genomics' from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res* 38:D355-D360.
2. Fujibuchi W, Sato K, Ogata H, Goto S, Kanehisa M (1998) **KEGG and DBGET/LinkDB: Integration of biological relationships in divergent molecular biology data**. In: *Knowledge Sharing Across Biological and Medical Knowledge Based Systems*, Technical Report WS-98-04, pp. 35-40, AAAI Press
3. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M (2002) **LIGAND: database of chemical compounds and reactions in biological pathways**. *Nucleic Acids Res* 30:402-404
4. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M (2008) **KEGG Atlas mapping for global analysis of metabolic pathways**. *Nucleic Acids Res* 36:W423-W426
5. Shigemizu D, Araki M, Okuda S, Goto S, Kanehisa M (2009) **Extraction and analysis of chemical modification patterns in drug development**. *J Chem Inf Model* 49:1122-1129
6. Masoudi-Nejad A, Goto S, Jauregui R, Ito M, Kawashima S, Moriya Y, Endo TR, Kanehisa, M (2007) **EGENES: Transcriptome-based plant database of genes with metabolic pathway information and EST indices in KEGG**. *Plant Physiol* 144:857-866
7. Wheelock CE, Wheelock AM, Kawashima S, Diez D, Kanehisa M, van Erk M, Kleemann R, Haeggstrom JZ, Goto S (2009) **Systems biology approaches and pathway tools for investigating cardiovascular disease**. *Mol Biosyst* 5:588-602
8. Wheelock CE, Goto S, Yetukuri L, D'Alexandri FL, Klukas C, Schreiber F, Oresic M (2009) **Bioinformatics strategies for the analysis of lipids**. *Methods Mol Biol* 580:339-368
9. Moriya Y, Itoh M, Okuda S, Yoshizawa A, Kanehisa M (2007) **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Res* 35:W182-W185
10. Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, Kanehisa M (2010) **PathPred: an enzyme-catalyzed metabolic pathway prediction server**. *Nucleic Acids Res* 38:W138-W143
11. Kotera M, Okuno Y, Hattori M, Goto S, Kanehisa M (2004) **Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions**. *J Am Chem Soc* 126:16487-16498

12. Yamanishi Y, Hattori M, Kotera M, Goto S, Kanehisa M (2009) **E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs**. *Bioinformatics* 25:i79-i86
13. Hashimoto K, Kanehisa M (2008) **KEGG GLYCAN for integrated analysis of pathways, genes, and structures**. In: Taniguchi N, Suzuki A, Ito Y, Narimatsu H, Kawasaki T, Hase S (eds) *Experimental Glycoscience*, 441-444, Springer
14. Hattori M, Okuno Y, Goto S, 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
15. Hattori M, Tanaka N, Kanehisa M, Goto S (2010) **SIMCOMP/SUBCOMP: chemical structure search servers for network analyses**. *Nucleic Acids Res*. 38:W652-W656
16. Aoki KF, Yamaguchi A, Ueda N, Akutsu T, Mamitsuka H, Goto S, Kanehisa M (2004) **KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains**. *Nucleic Acids Res* 32:W267-W272

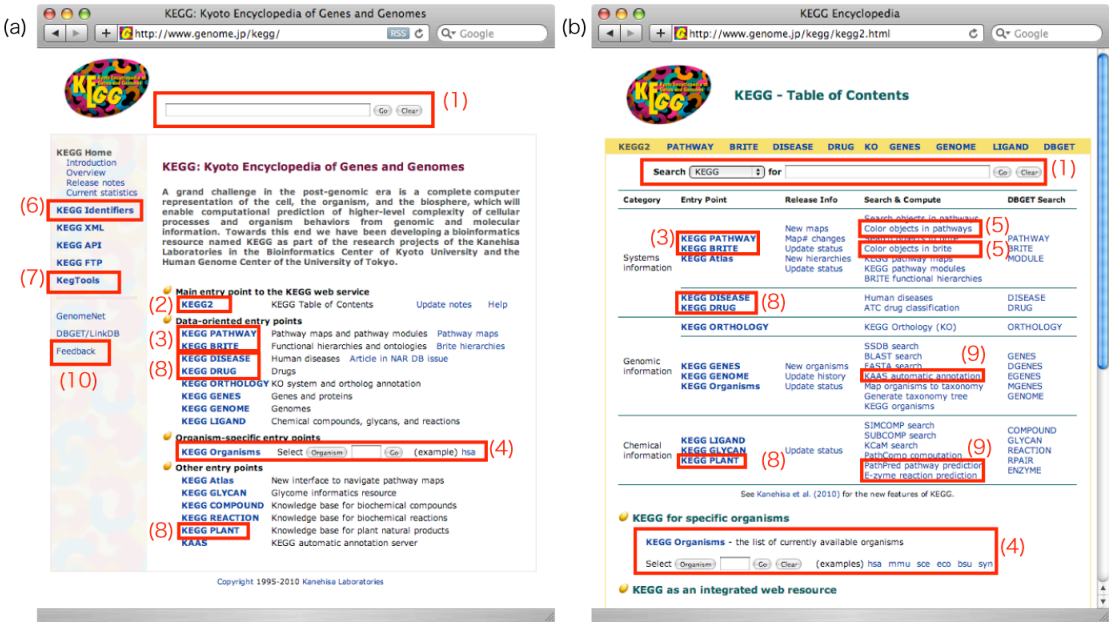


Figure 1. Overview of the KEGG homepage and sitemap. (a) KEGG homepage. (b) KEGG2: sitemap. (1) Search boxes. (2) Link to KEGG2. (3) KEGG PATHWAY/BRITE (4) KEGG Organisms: entry points for the genome-sequenced organisms (See **Note 2**). The user can limit the search only in an organism of interest (See **Note 4**). (5) Tools to customize PATHWAY/BRITE, with which the user can color the objects of interest (See **Section 3**). (6) KEGG Identifiers. The gene accession numbers from the outside databases can be converted to the corresponding KEGG IDs from here (See **Note 6**). The users can also obtain the multiple KEGG entries simultaneously (See **Note 8**). (7) KegTools: Desktop applications, KegHier, KegArray and KegDraw, can be downloaded from here (See **Section 2**, **Section 4** and **Note 7**, respectively). (8) KEGG DISEASE/DRUG/PLANT (9) KAAS, PathPred and E-zyme tools to create new pathways (See **Note 3 & 7**). (10) Feedback: Any questions or comments are appreciated (See **Note 9**).

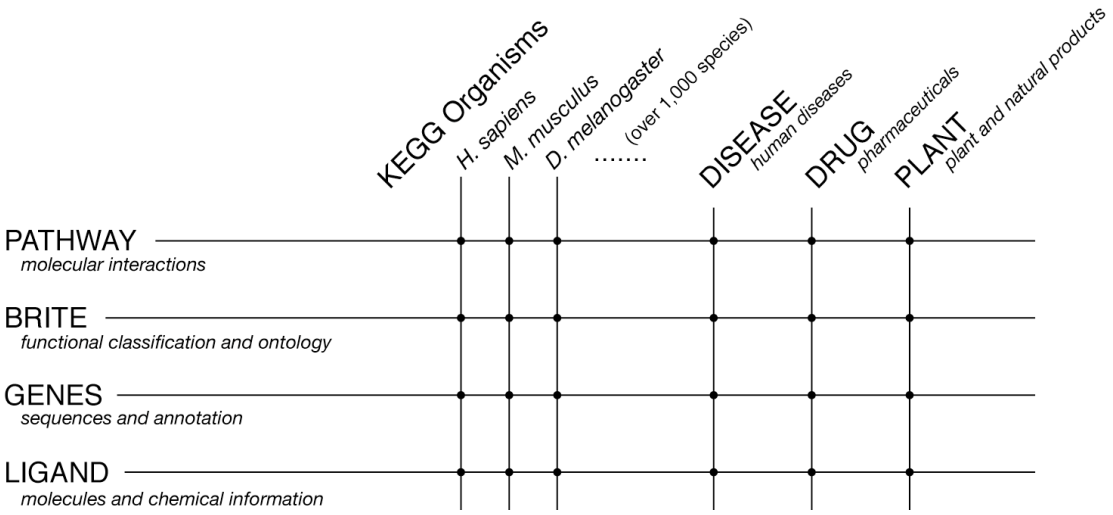


Figure 2. Grid-shaped structure of the KEGG data. KEGG has a variety of entry points from which the user can start searching or analyzing data, depending on the various perspective. For example, PATHWAY contains molecular interaction data such as metabolic or regulatory pathways throughout all the genome-sequenced organisms, which we refer to as "reference pathways" (**Fig. 3a**). The user can also limit the pathway for only a specified organism (see **Note 2**), or can compare the pathways in different organisms (see **Section 2**). The DISEASE category of the PATHWAY database (or the PATHWAY category of the DISEASE database) can be regarded as the human pathways that are perturbed by diseases. The DRUG and PLANT categories of the PATHWAY database are the collections of pathway maps specialized for pharmaceuticals and plants, respectively. These relationships also apply for other databases such as BRITE, GENES and LIGAND. This figure is illustrated simply for the explanation: the actual structure is a little more complicated. For example, chemical compounds in LIGAND are also hierarchically classified in BRITE. Similarly, GENES are grouped by KO (KEGG Orthology), which is also hierarchically classified in BRITE.

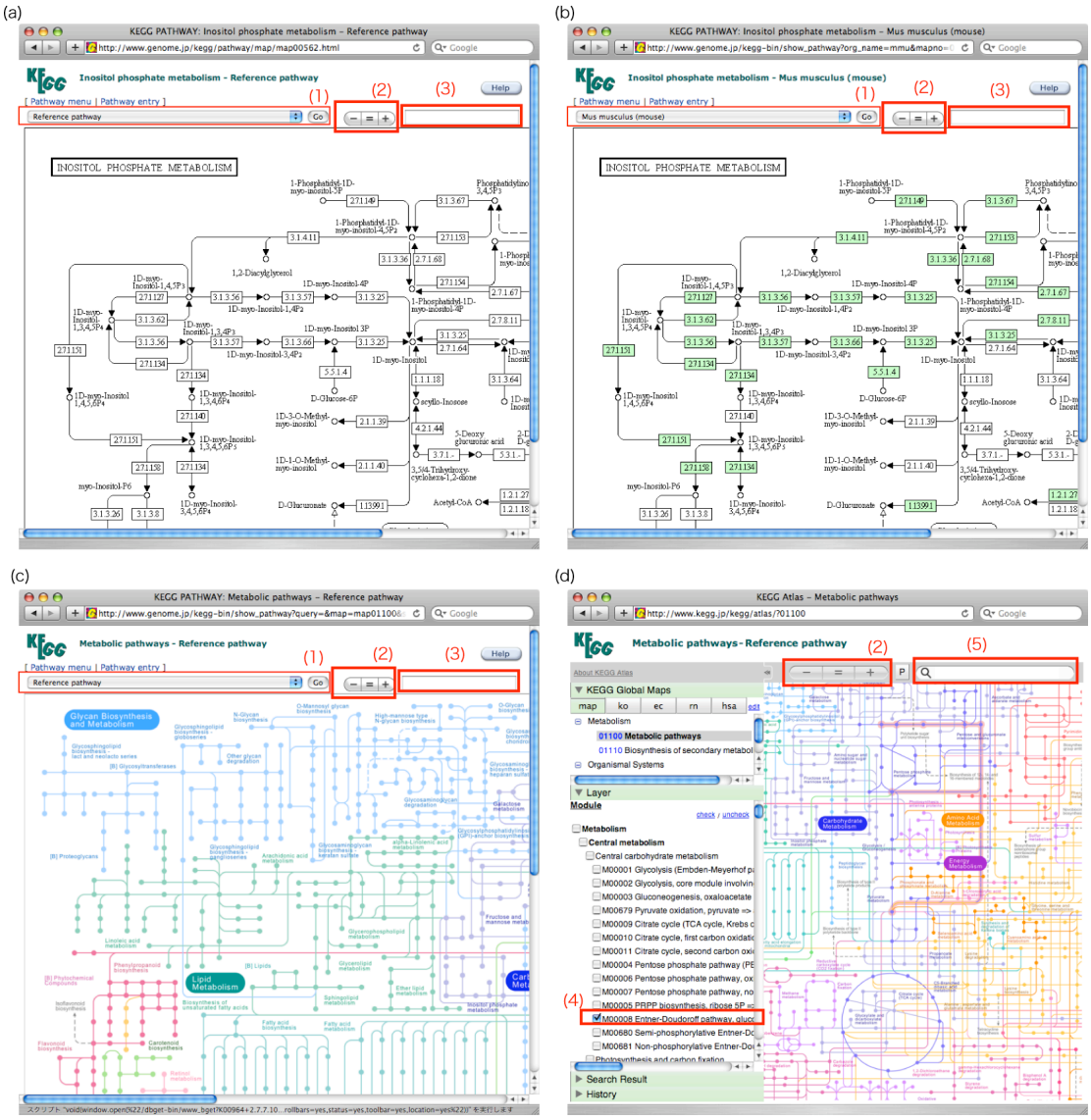


Figure 3. KEGG PATHWAY and Atlas. (a) KEGG PATHWAY map of inositol phosphate metabolism as a reference pathway. Chemical compounds are represented as circles, and gene products (such as enzyme proteins) are represented as rectangles. (b) The same map with the genes information deduced from mice genome. (c) An example global map. Chemical compounds are represented as dots, and enzyme reactions are represented as lines. Different categories of pathways are drawn in different colors in a map. (d) KEGG Atlas. (1) The pull-down menu to choose an organism. If the user selects "reference pathway" in the menu, the rectangles provide the links to other objects that are not specific to an organism, such as enzymes, reactions and KO (KEGG Orthology). The user can customize the selection of organism in the menu (See **Note 4**). (2) The graphics can be zoomed in or out by clicking these buttons. (3) Input any term in this search box, and the corresponding objects are highlighted, if any. (4) KEGG Modules, manually defined tighter functional units for pathways and protein complexes, can be selected to emphasis the part of the global map of interest. (5) Search box accepting any term to navigate the Atlas.

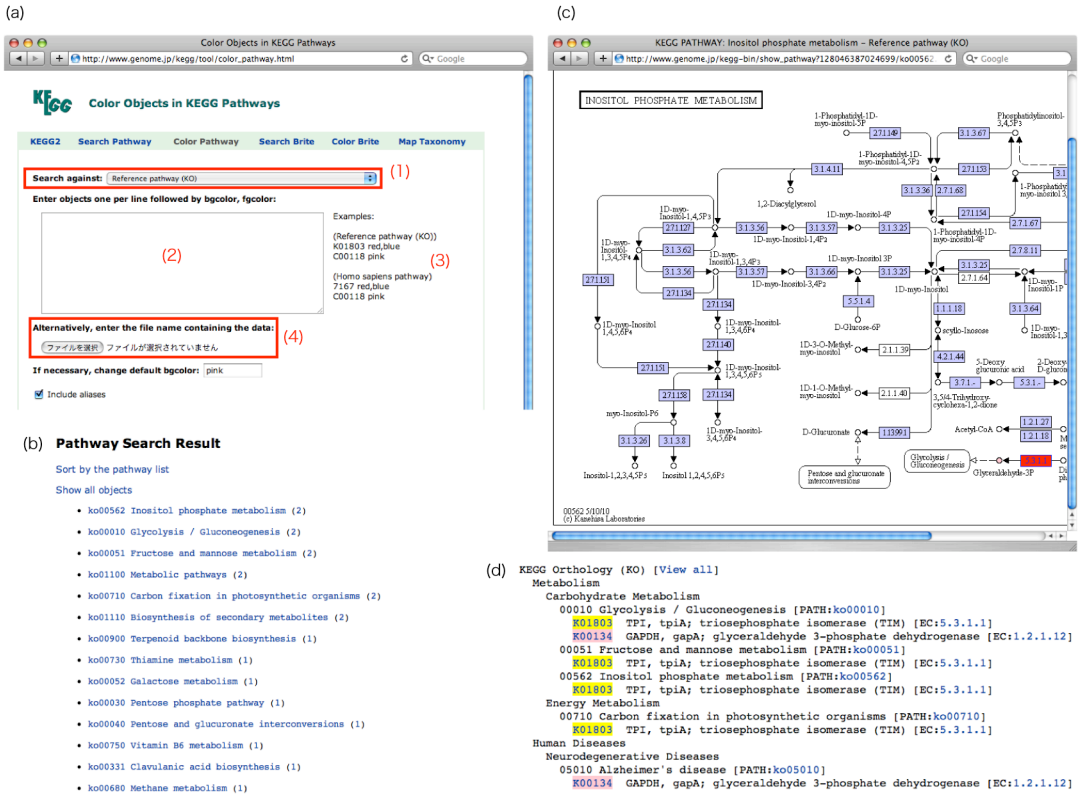


Figure 4. Color objects in PATHWAY/BRITE. (a) The page for coloring the KEGG pathways. (1) An organism or a reference pathway has to be specified in this menu. (2) Input the list of the genes by KEGG IDs and colors for them. (3) Examples of the inputs are shown here. (4) The input data can be also uploaded from here. (b) After clicking the "Exec" button, the list of the PATHWAY maps containing the input objects is displayed. (c) One of the pathways derived from the resulting list. The graphics of the maps are automatically generated as gif files, which will be removed from the KEGG server within few hours. If the user wants to preserve the graphics, they should be downloaded to the local computer. (d) An example result of coloring the BRITE functional hierarchy. The user can grasp the genes of interest at a sight, with using different colors for different groups as the user wants.

(a)

#organism: syn						
#ORF	x	y	Control-sig	Control-bkg	Target-sig	Target-bkg
slr1485	1	2	1037.13	502.62	1593.30	695.25
slr1119	1	3	1261.63	494.72	2685.37	742.87
sll0708	1	4	922.97	561.38	1598.37	727.28
sll1120	1	6	2152.80	560.96	2591.23	771.07
sll1734	1	7	1918.47	574.57	5968.97	823.66
:	:	:	:	:	:	:
:	:	:	:	:	:	:

(b)

#organism: syn			
#ORF	X	Y	Ratio
slr1485	1	2	0.610282
slr1119	1	3	2.360655
sll0708	1	4	0.842321
sll1120	1	5	0.769038
:	:	:	:
:	:	:	:

(c)

# COMPOUND	ratio
C00668	1.2
C00221	0.5
C01172	2.2
C00118	1.0
:	:
:	:

Figure 5. Example input files for KegArray. All lines beginning with the "#" character (other than the '#organism:' or '#source:' line) are regarded as comments and skipped by KegArray. The organism information is necessary to identify the ORFs. The organism should be provided by the three-letter (or four-letter) KEGG Organism code (See **Note 2**). The lines in tab-delimited format below the #ORF section contain gene expression profile data. **(a) Table representing intensity values:** First column represents the KEGG GENES ID, the unique identifier of the ORF in the organism. The second and third columns are for specifying the location (X- and Y-axis coordinates, respectively) of the ORF on the DNA microarray. The fourth and fifth columns are the signal intensity and the background intensity of the control channel, respectively. The sixth and seventh columns are the signal intensity and the background intensity of the target channel, respectively. **(b) Table representing ratio values:** The first column is for the KEGG GENES ID. The second and the third columns are X- and Y-axis coordinate information of the ORF on the microarray, respectively. The fourth column describes the ratio value between control channel and target channel. **(c) Table representing metabolome data:** The first column represents KEGG COMPOUND ID, and the second column represents the relative amount of the target compound compared with the control.

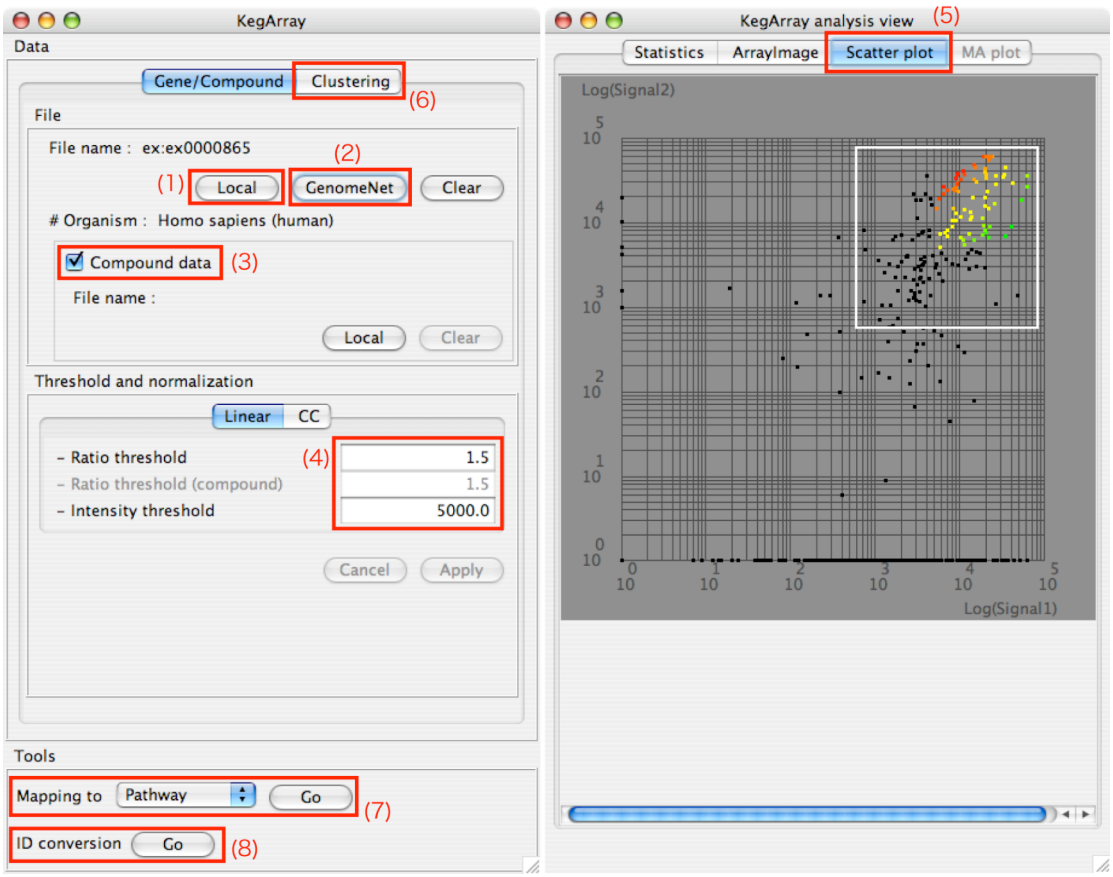
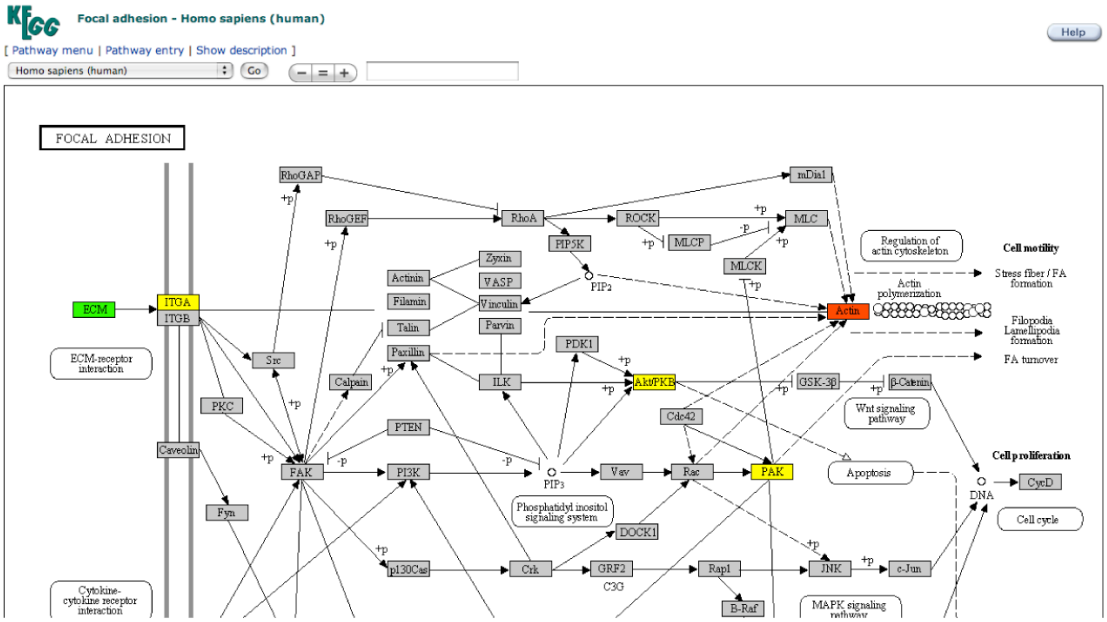


Figure 6. Screenshots of the KegArray control panels. (1) The "Local" button opens a pop-up window to select a data file on your local disk. The data file should comply with the format described in **Figure 5**. (2) The "GenomeNet" button opens a pop-up window to retrieve the data stored in the GenomeNet EXPRESSION database. Available entry IDs are listed in the window, and once you select one, its description will be displayed. (3) The "Compound data" box should be checked (default) for loading metabolome data. (4) There are three input boxes to specify the parameters for the confidence lines discriminating the regulated genes/compounds from unregulated ones. (5) The scatter plot of the data is shown in this pane. The colors of spots represent levels of increase or decrease of the target gene expressions against the control. The coloring scheme can be changed in the preference menu. (6) The "Clustering" pane. (7) Mapping to PATHWAY, Genome Map and BRITE. (8) ID conversion tool (See **Note 6**).

(a)



(b)

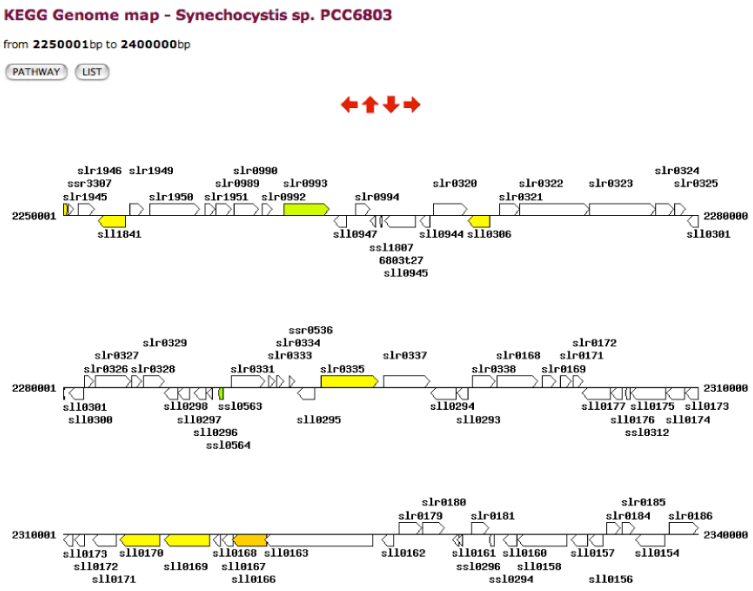


Figure 7. Mapping microarray data onto PATHWAY/GENOME/BRITE. KegArray has options to visualize the up- or down-regulated genes on various KEGG objects, *i.e.*, (a) PATHWAY, (b) GENOME, and BRITE. The input data does not have to be from microarray experiments; KegArray can be used as a visualization tool of gene functions as long as the data complies the format described in **Fig. 5**.

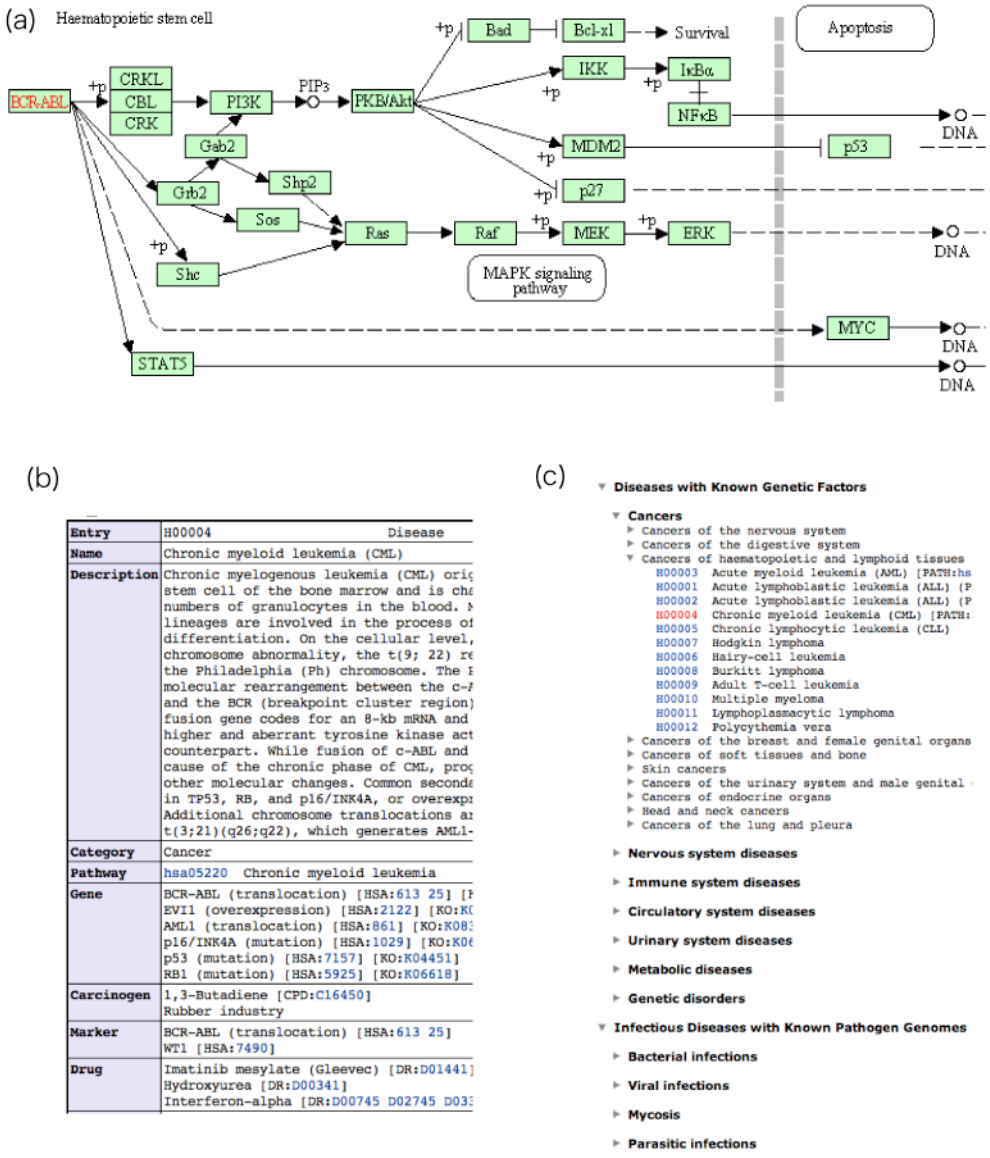


Figure 8. KEGG DISEASE. KEGG DISEASE describes human diseases in computable forms. This figure illustrates chronic myeloid leukemia in the following three representations. (a) Human diseases are described as perturbed states of human molecular network. If some genes are known to be related with the disease, they are marked in red. The user can look up the genes by clicking the corresponding rectangles. (b) Even if the mechanism is not known, the list of the known information, such as mutated genes, is still valuable. The user can obtain further information by clicking the links. (c) Diseases are organized and classified in the BRITE hierarchy, where the disease in question is marked in red. The user can view the detail of the disease by clicking the accession number (*e.g.,* H00004), and look up diseases in other categories by clicking the triangles.

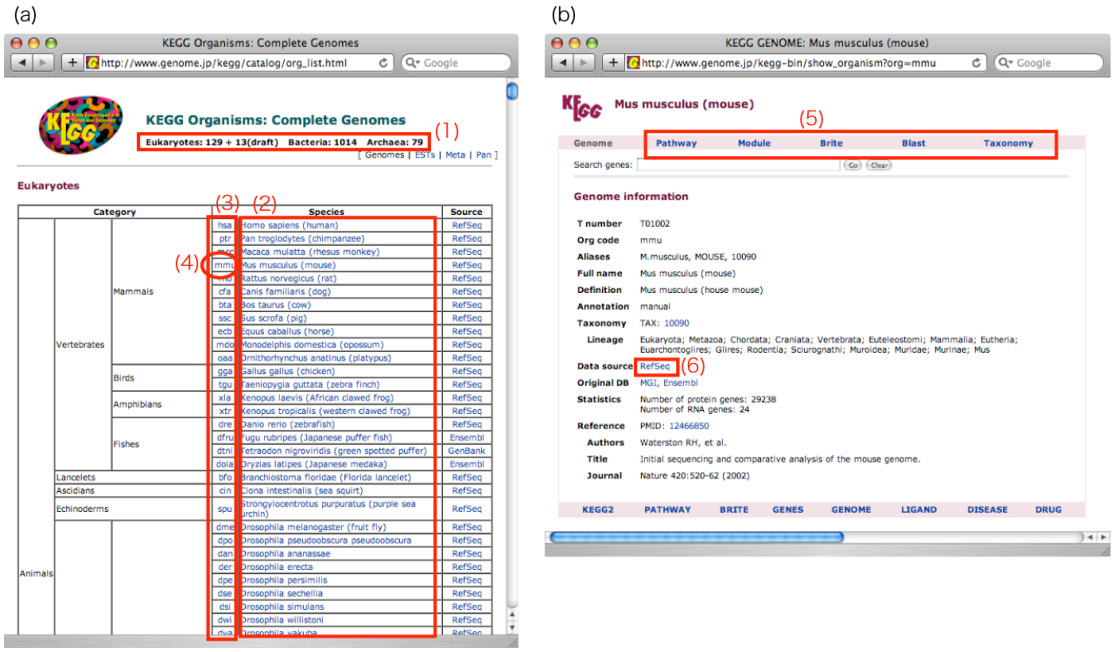


Figure 9. KEGG Organisms and GENOME. (a) KEGG Organism page. (1) Statistics of the genome sequences registered in KEGG. (2) The scientific names and common names of organisms, providing the links to the corresponding search pages for GENES. (3) KEGG Organism codes, providing the links to the corresponding GENOME pages. (4) Clicking this link leads the user to the GENOME page of mouse genome. (b) An example GENOME page. (5) Links to the organism-specific pathways, modules, BRITE hierarchies, BLAST searches, and taxonomy information. (6) The sequence data is downloadable from the link at the "Data source".

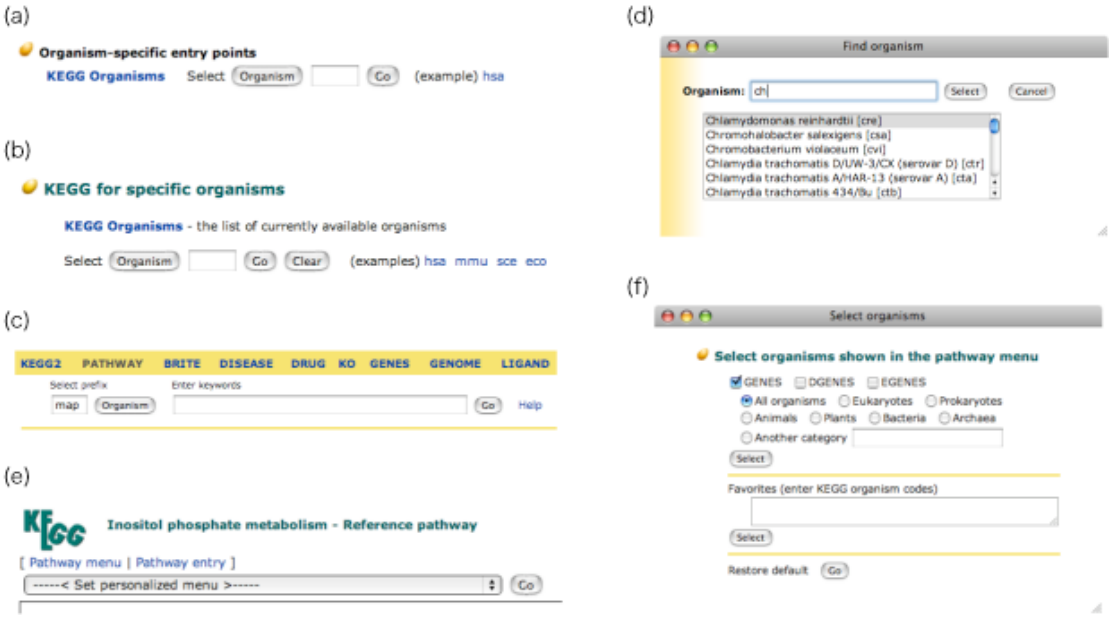
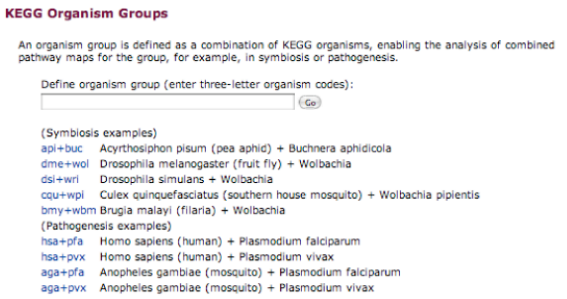


Figure 10. Finding or limiting organisms. Organism search options are located in various pages such as (a) the KEGG homepage (**Fig. 1**), (b) the KEGG sitemap (**Fig. 2**), and (c) the KEGG PATHWAY page. If the user knows the KEGG Organism code for the organism of interest, input the code in the box to reach the GENOME page (**Fig. 9b**). In the case the user does not remember the code, click the "Organism" button to pop up the "Find organism" window. (d) This window can be used as a dictionary, and also a reverse dictionary, of the scientific name of organisms and the corresponding KEGG Organism codes. The user needs not complete the spell of organism names; the search engine complements the name, as shown in this figure. This window works even after other web pages are closed, so this can still be used for looking up the organisms. (e) Every PATHWAY page (**Fig. 3a**) has a pull-down menu to select an organism from more than 1,000 organisms with complete genomes. For the user feeling difficulty in finding an organism of interest, there are options to sort organisms in alphabetical order, and to generate the personalized menu. Select "< Set personalized menu >" and click "Go", and the "Select organism" window pops up. (f) The user can generate the personalized menu by specifying organisms of interest. These settings are preserved in the user's browser, and are used next time.

(a)



(b)

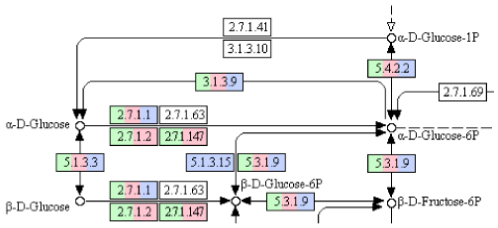
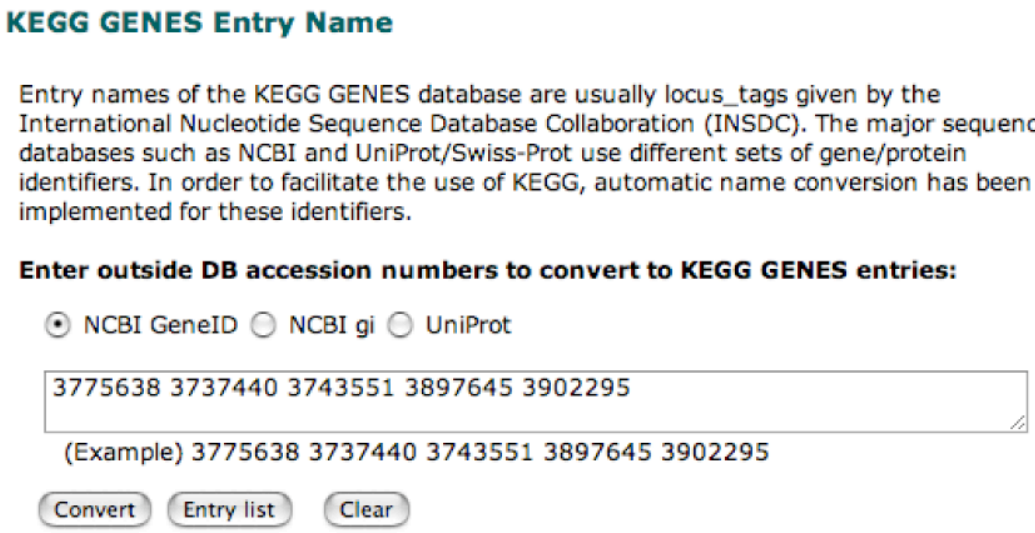
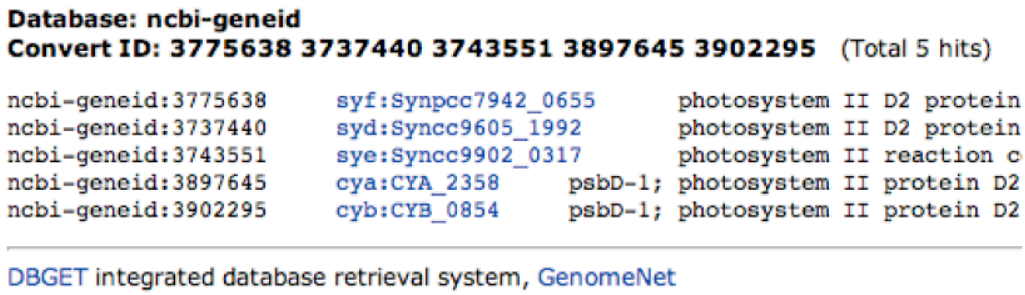


Figure 11. KEGG Organisms groups. (a) The option to specify two or more organisms in the middle of the KEGG GENOME page. (b) Using the option provides multicolor pathway maps representing the gene products from the specified organisms.

(a)



(b)



(c)

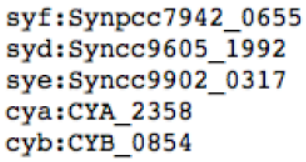


Figure 12. The accession ID conversion tool. The user can see the KEGG Identifiers page (http://www.genome.jp/kegg/kegg3.html) by clicking one of the links of the KEGG homepage (**Fig. 1a**). (a) In the middle of the page, the accession numbers from outside databases can be converted to the corresponding KEGG entries. (b) Click the "Convert" button to obtain this page, showing external-DB IDs, the corresponding KEGG IDs and brief annotations. (c) Click the "Entry list" button, and obtain the list that can be directly used as an input of coloring the KEGG objects (see **Section 3** and **Fig. 4a**).

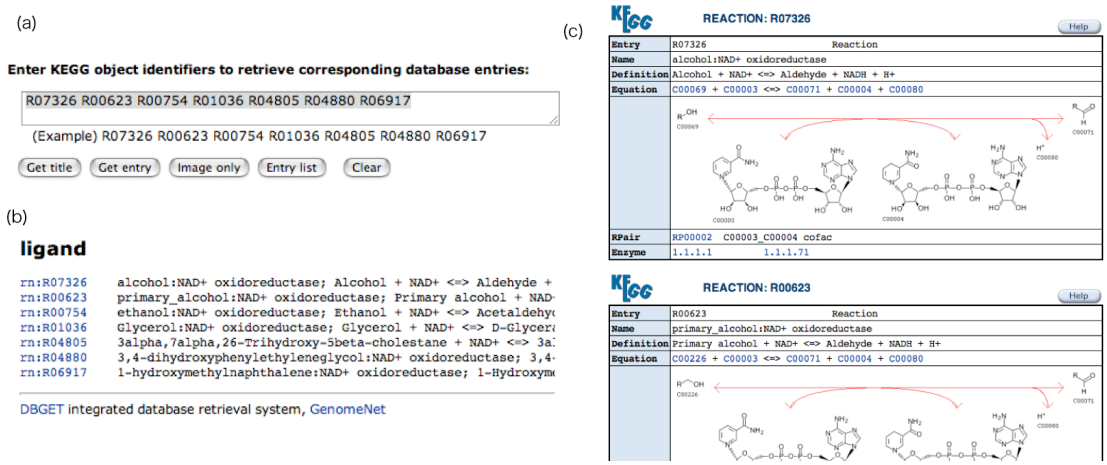


Figure 13. Retrieving multiple KEGG entries simultaneously. We provide a convenient way to simultaneously view a number of objects indicated by KEGG identifiers. (a) In the middle of the KEGG Identifiers page, there is an input form. (b) Input some KEGG IDs and click the "Get title" button, and the user can obtain the list of IDs and the corresponding titles (descriptions or annotations). (c) Click the "Get entry" button, and the user can obtain the corresponding entries simultaneously in a page.