

Using the KEGG Database Resource

UNIT 1.12

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a bioinformatics resource for understanding biological function from a genomic perspective. It is a multispecies, integrated resource consisting of genomic, chemical, and network information with cross-references to numerous outside databases and containing a complete set of building blocks (genes and molecules) and wiring diagrams (biological pathways) to represent cellular functions (Kanehisa et al., 2004).

In this unit, protocols are described for using the five major KEGG resources: PATHWAY, GENES, SSDB, EXPRESSION, and LIGAND. The KEGG PATHWAY database (see Basic Protocols 1 to 4) consists of a user-friendly tool for analyzing the network of protein and small-molecule interactions that occur in the cells of various organisms. KEGG GENES (see Basic Protocols 5 and 6) provides access to the collection of gene data organized so as to be accessible via text searches, from the PATHWAY database, or via cross-species orthology searches. The KEGG Sequence Similarity Database (SSDB; see Basic Protocols 7 to 9) consists of a precomputed database of all-versus-all Smith-Waterman similarity scores among all genes in KEGG GENES, enabling relationships between homologs to be easily visualized on the pathway and genome maps or viewed as clusters of orthologous genes. The KEGG EXPRESSION database (see Basic Protocols 10 to 14) contains both data and tools for analyzing gene expression data. User-defined data such as microarray experiments may also be uploaded for analysis using the tools available in KEGG EXPRESSION. Finally, KEGG LIGAND (see Basic Protocols 16 to 19) is a database of small molecules, their structures, and information relating to the enzymes that act on them. All KEGG databases are heavily cross-referenced, providing a truly integrated view of biological processes.

THE KEGG PATHWAY DATABASE: GETTING STARTED

BASIC
PROTOCOL 1

This protocol provides an introduction to the KEGG Pathway database. The KEGG Pathway Database is a collection of manually drawn reference diagrams, or maps, each corresponding to a known biological network of functional significance. Each map is a union among multiple species, but species-specific pathways can be viewed by colorizing the genes of a given species.

The items on the pathway map are represented by various symbols (Fig. 1.12.1). Most of the nodes of a pathway map are rectangles that represent gene products, usually proteins. Small circles represent chemical compounds and other molecules. Large ovals represent links to other pathway maps, and a cluster of rectangles represents a protein complex. There are also a number of interaction types—or edges—on the pathway map, including (de)phosphorylation, ubiquitination, and glycosylation, among others. Thus the KEGG pathway maps provide a comprehensive view of the biological network.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

Using Biological
Databases

Contributed by Kiyoko F. Aoki and Minoru Kanehisa

Current Protocols in Bioinformatics (2005) 1.12.1-1.12.54

Copyright © 2005 by John Wiley & Sons, Inc.

1.12.1

Supplement 11

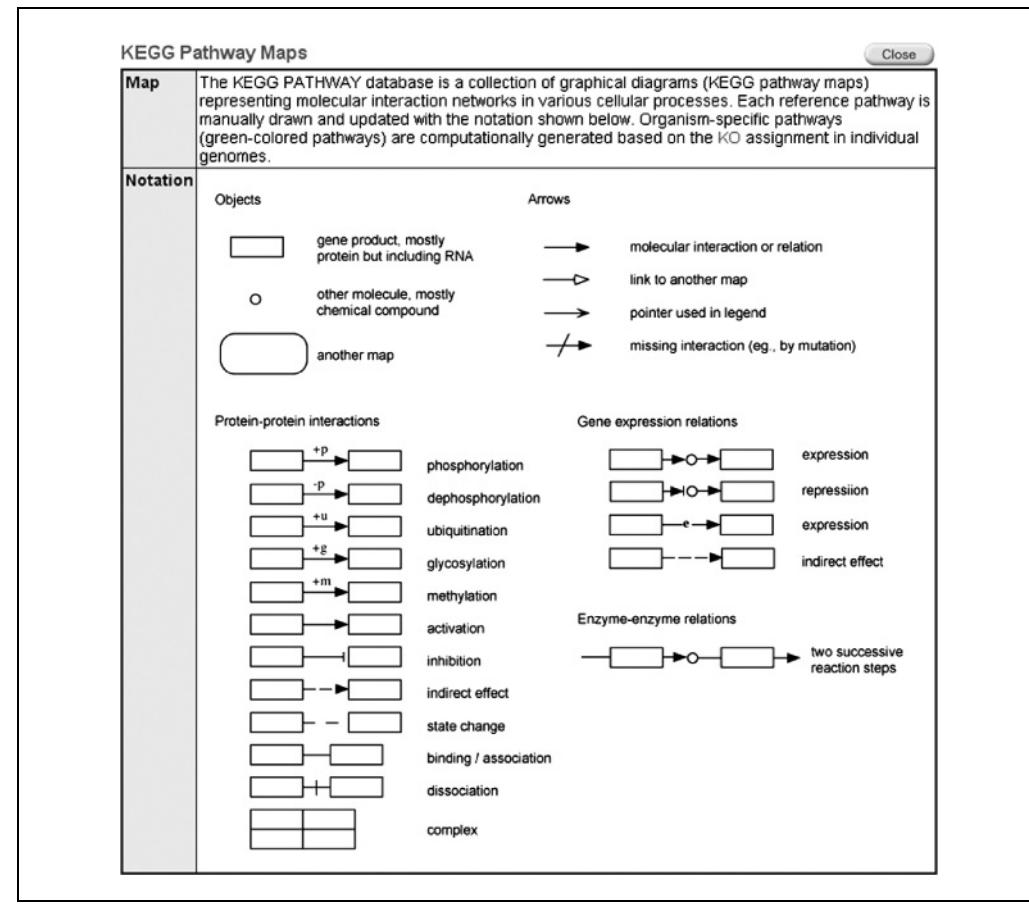


Figure 1.12.1 Symbols on the KEGG pathway map and their descriptions.

1. Open the KEGG GenomeNet Web page at <http://www.genome.jp/> and click on the KEGG2 link located in the navigation bar at the top of the page.
2. This brings up the KEGG Table of Contents (Fig. 1.12.2) which is the entry point to all of the KEGG databases.
3. Click on KEGG PATHWAY” in the “Database column of the table near the top of the page.
4. This brings up a list of KEGG’s numerous pathways organized in a hierarchical manner (Fig. 1.12.3). Each level of the hierarchy corresponds to an increasingly specific view of the biological network. For example:
 - a. Click the “1. Metabolism” link to display the KEGG metabolic pathway data from a bird’s-eye view. This makes it easy to see the overall picture of how the pathways interact with one another in a network (Fig. 1.12.4).
 - b. Return to the previous page and click on the “1.1 Carbohydrate Metabolism” link for an example of a second-level pathway (Fig. 1.12.5). The various components of carbohydrate metabolism are displayed.
 - c. Return to the previous page and click on the “Glycolysis/Gluconeogenesis” link to view the most detailed level of a pathway (Fig. 1.12.6). At the top, a pull-down menu of organisms can be used to colorize the parts of the pathway that are known to exist for any given species. Select a different organism to display the genes involved for that organism as colored nodes. The “Current selection” button will display the list of organisms available in the pull-down menu, and the Select button to its right can be used to edit this list.



KEGG - Table of Contents

KEGG2 PATHWAY GENES LIGAND BRITE XML API DBGET

Generalized KEGG

Content	Database	Search & Compute	DBGET Search
Pathway information	KEGG PATHWAY	Search objects in KEGG pathways Color objects in KEGG pathways KEGG pathways in XML	PATHWAY
Genomic information	KEGG GENES	BLAST search against GENES/GENOME FASTA search against GENES/GENOME KEGG EXPRESSION	GENES DGENES / EGENES GENOME
Chemical information	KEGG LIGAND	Search similar compound structures Search similar glycan structures Predict reactions and assign EC numbers Generate possible reaction paths	COMPOUND GLYCAN REACTION RPAIR LIGAND ENZYME
Binary relations and hierarchies	KEGG BRITE	KEGG Orthology (KO) Therapeutic category of drugs	KO

Specialized KEGG

- KEGG for specific organisms**

Enter KEGG organism code: (examples) hsa mmu sce eco bsu syn
Or use the list of KEGG organisms
Customize the organism menu with selected organisms
Show currently selected organisms (Archaea in GENES)
- KEGG for selected research areas**

KEGG EXPRESSION for microarray data analysis
KEGG GLYCAN for glycane informatics
- KEGG for software development**

KGML - XML representation of KEGG pathways

Figure 1.12.2 The KEGG Table of Contents, via which all of KEGG's numerous resources can be easily accessed. To view the Pathways, click on the KEGG PATHWAY link in the Database column of the table, as indicated.

- d. Looking at this map more closely, one can see the sequences of reactions involved in this pathway. For example, one can see that the circled enzyme 5.4.2.1, which is phosphoglyceric acid mutase, catalyzes the reversible reaction of 3-PGA to 2-PGA.
5. Click on the node of the enzyme 5.4.2.1 (shown circled in Fig. 1.12.6) to see the information for this enzyme entry (not shown). Scroll down to the Genes section of this page and click on entry HSA:5223 to view the detailed information page, called the GENES Entry, available for this gene as shown in Figure 1.12.7. The details of the available information in the GENES Entry will be covered later in Basic Protocol 5 and 6.

Alternatively, one can obtain the GENES entry page directly from the pathway map shown in Figure 1.12.6 by selecting the organism to be "Homo sapiens" in the drop down menu at the top of the page, clicking Go and then clicking on enzyme 5.4.2.1 on the resulting pathway map page.

Using Biological Databases

1.12.3

Figure 1.12.3 The main KEGG Pathway Database page from which dozens of pathway maps can be referenced. Each level in the hierarchy of maps provides different views of the pathways, providing a comprehensive and user-friendly resource for studying the molecular interaction networks of cells and organisms.

BASIC PROTOCOL 2

KEGG PATHWAY: KEGG MARKUP LANGUAGE (KGML)

The KEGG Markup Language, or KGML, has been developed so that graph objects in KEGG may be easily transferred, reconstructed, and displayed in a system-independent manner. All metabolic pathways and some regulatory pathways such as signal transduction are now available in KGML.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

Using the KEGG Database Resource

1.12.4

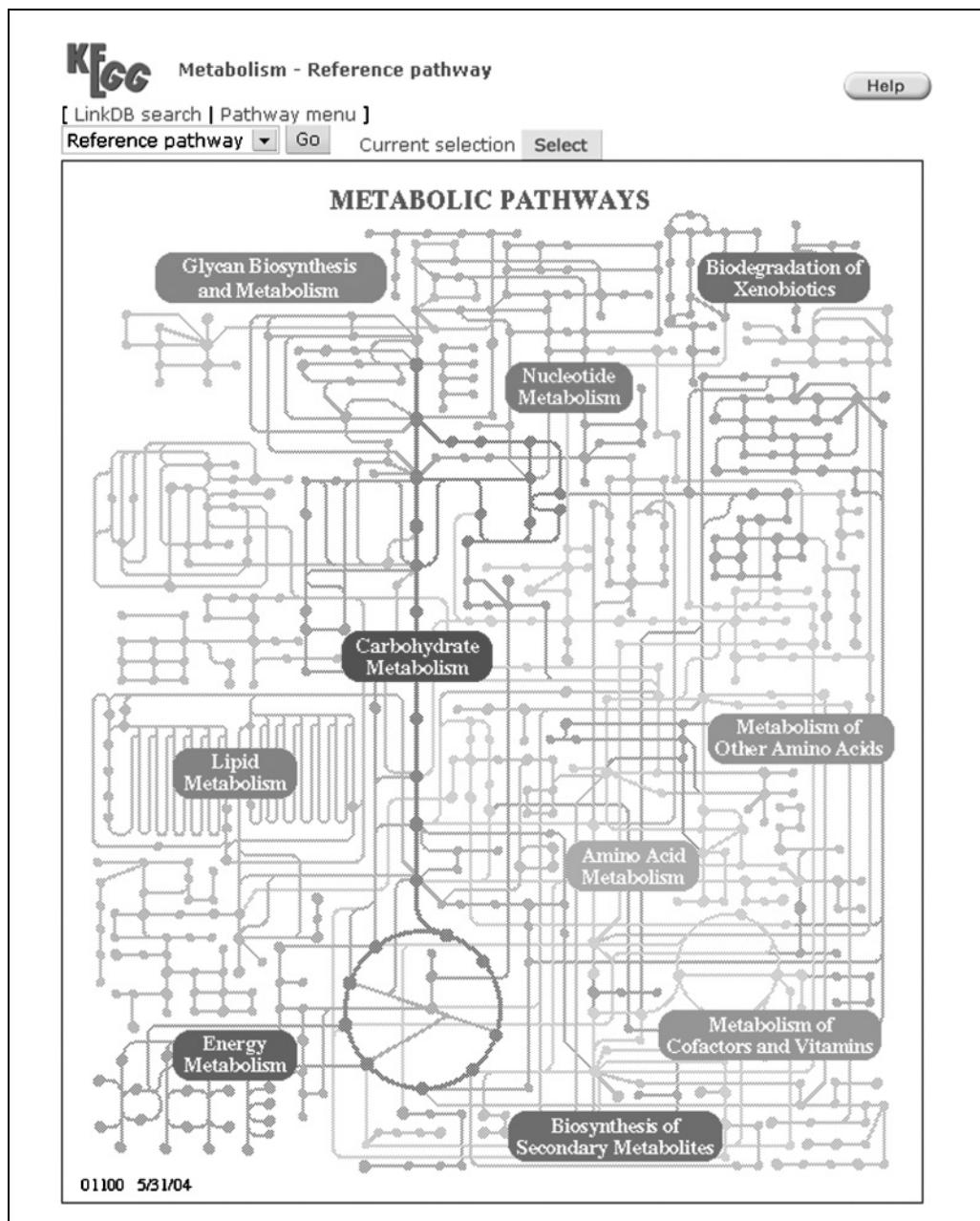


Figure 1.12.4 Reference pathway map for all metabolic pathways in general.

Files

XML file in KGML format (optional)

1. Access the KGML Web page at <http://www.genome.jp/kegg/xml/> to obtain pathway information in KGML format. Links to all metabolic pathways can be downloaded in XML format, as well as in HTML format and as an image file. For this example, click on “KEGG reference metabolic pathways.”

Alternatively, it is possible to access the KGML Web page by first accessing the KEGG Table of Contents from the GenomeNet home page as described in Basic Protocol 1, and then clicking the XML link on the toolbar need the top of the page.

Using Biological Databases

1.12.5

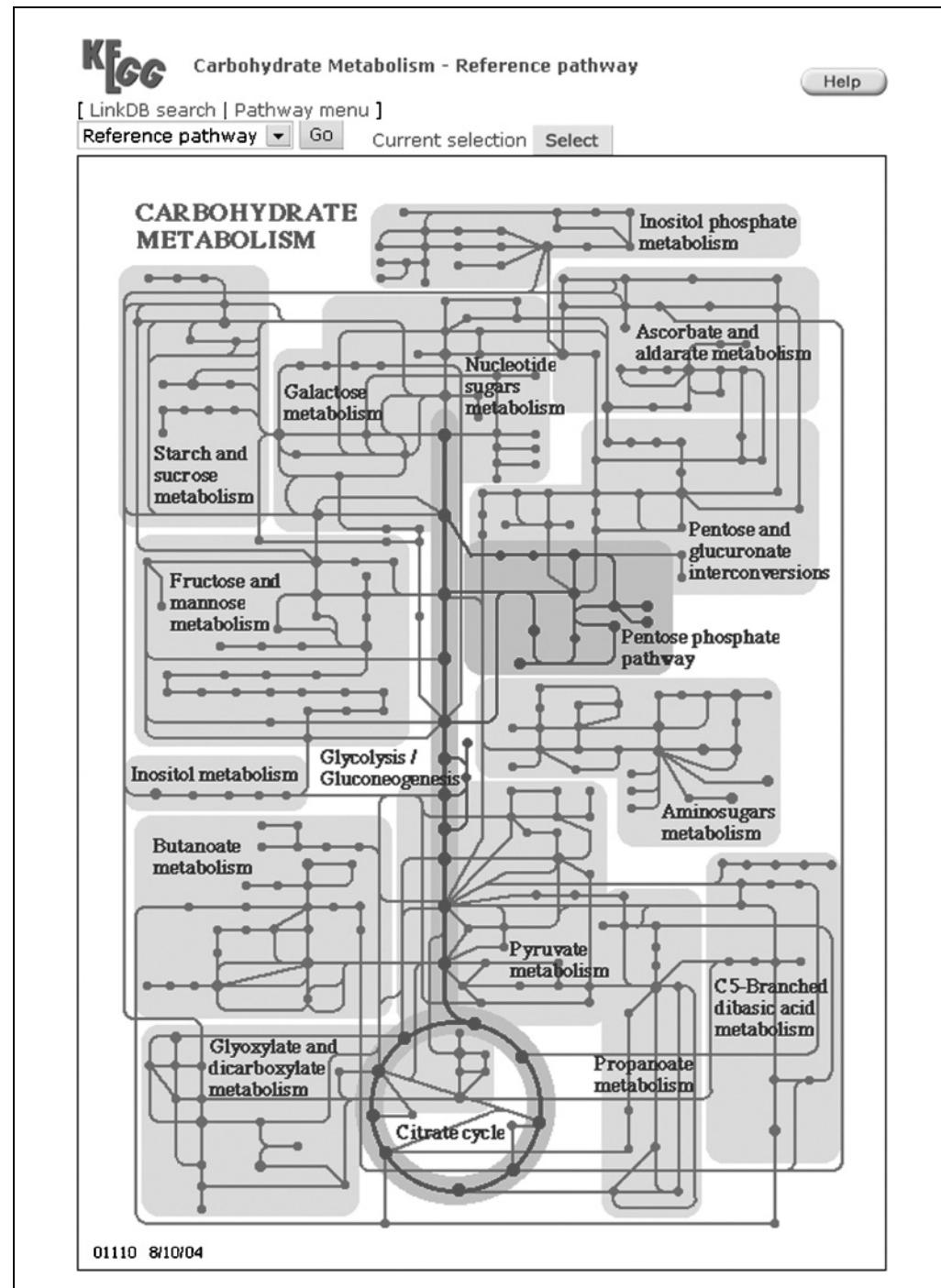


Figure 1.12.5 Carbohydrate metabolism map.

2. Select a pathway to download. For this example, scroll down to the “Metabolism of Cofactors and Vitamins” section and select the “graphics” link of map00770, Pantothenate and CoA biosynthesis.
3. This will display the pathway in the KGML Pathway viewer, as displayed in Figure 1.12.8. This is a user-friendly Java applet viewer. Individual nodes may be double-clicked to view the corresponding DBGET entry, whether it be a compound, enzyme, reaction, or gene.

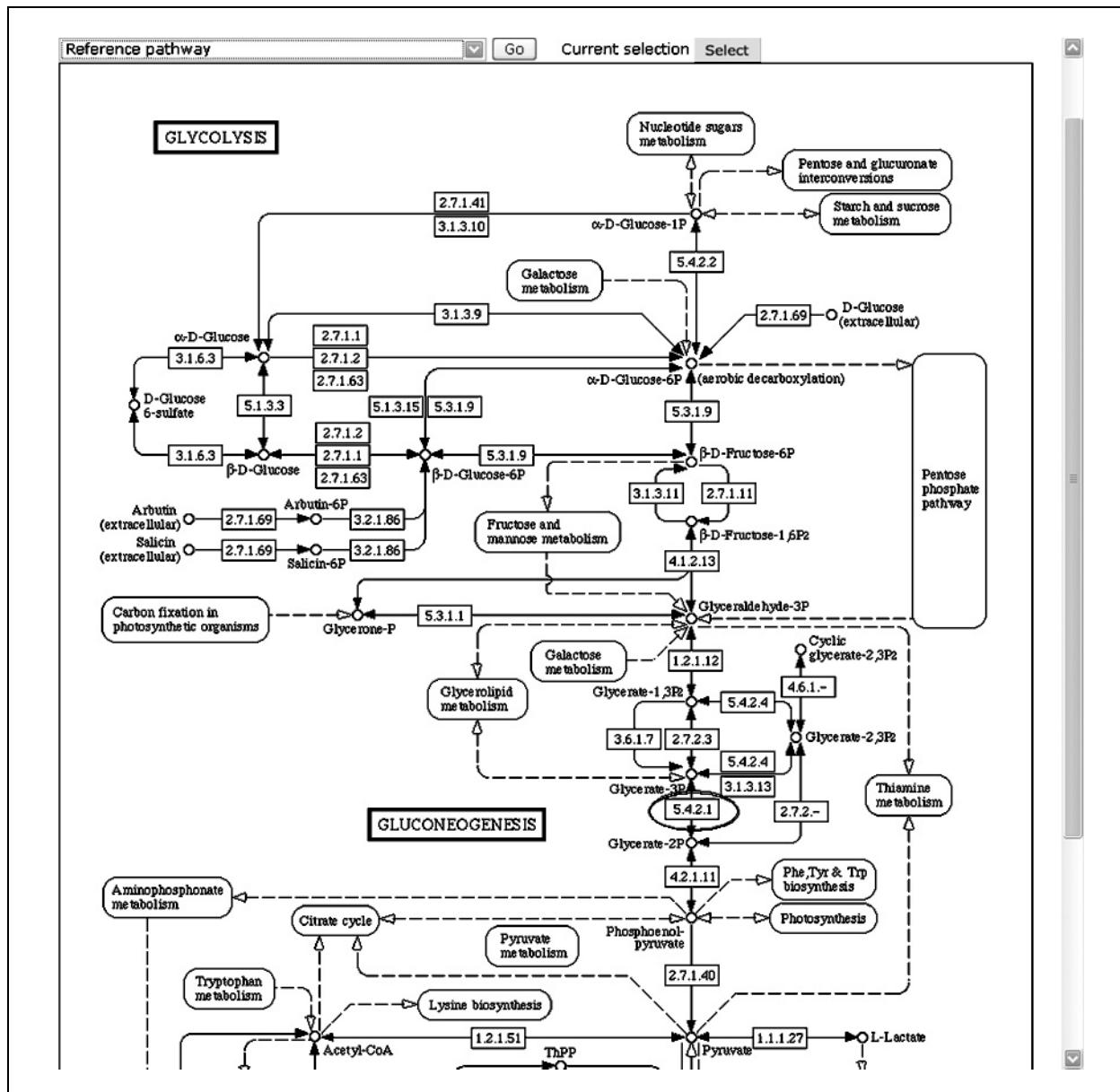


Figure 1.12.6 Glycolysis/gluconeogenesis pathway map at the most detailed level. Different views based on organism may be selected via the pull-down menu at the top. The circled node indicates the location of phosphoglyceric acid mutase.

KEGG PATHWAY: COLORING EC NUMBERS

The KEGG Pathways provides a comparative genomics analysis tool whereby genes from multiple pathways may be visualized simultaneously. This protocol describes how genes in the pathway maps can be colorized so that it is possible to graphically view both the location in the pathway map and the correspondence of the given genes among several organisms. For example, given several genes whose function has been hypothetically determined, one may easily examine the pathways for organisms containing the genes.

**BASIC
PROTOCOL 3**

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

Using Biological Databases

1.12.7



Homo sapiens: 5223

Help

Entry	5223	CDS	H.sapiens
Gene name	PGAM1, PGAMA		
Definition	phosphoglycerate mutase 1 (brain) [EC:5.4.2.1 5.4.2.4 3.1.3.13]		
KO	KO: K01088 bisphosphoglycerate phosphatase KO: K01834 phosphoglycerate mutase KO: K01837 bisphosphoglycerate mutase		
	OC search OC viewer		
Pathway	PATH: hsa00010 Glycolysis / Gluconeogenesis		
Class	Gene catalog		
SSDB	Ortholog Paralog Gene cluster		
Motif	Pfam: PGAM PROSITE: PG_MUTASE Motif		
Other DBs	GDB: 120530 OMIM: 172250 NCBI-GI: 4505753 NCBI-GeneID: 5223 UniProt: P18669		
LinkDB	PDB All DBs		
Position	10q25.3		
AA seq	254 aa	AA seq	BLAST
	MAAYKLVLIRHGESAWNLENRFGWYDADLSPAGHPEAKRGQALRDAGYEFDICFTSVQ KRAIRTLWTVLDADIDQMLPVVRTWRLNERHYGGLTGLNKAETAALKGEAQVKIWRRSYD VPPPFMEMPDHFFYSNISKDRRYADLTDQLPSCESLKDTIARALPFWNEEIVPQIKEGKR VLIAAHGNSLRGIVKHLEGLSSEAIMELNLPPTGIPIVYELDKNLPIKPMQFLGDEETVR KAMEAVAAQGKAKK		
NT seq	765 nt	NT seq	+upstream <input type="text" value="0"/> nt +downstream <input type="text" value="0"/> nt
	atggccgcctacaaactgggtgctgtccgcacggcgagagcgcatggaaacctggagaac cgcttcagcggtgttgcacggccgacctggccggccggccacaggaggcgaaacgc ggccggcaggcgctacggatgtggctatggatgtttgcacatgtctcacattcagtgcag aagagacgttgcggcccttggacagtgttagatgtccatgttagatgtggctgcca gtgggtggacttggccctcaatggatgtggccactatgggtcttaaccggcttcaataa gcagaaactgtgtcaaaaggcatgttgaggcccaggtaagatctggaggcgcttcatgt gtccccaccacctccatggagggccgaccatcttctacagcaacatcagaataggatcgc aggatgtgcacactcacagaatcatgtactaccctctgtggaggtctgaaggatactatt gccagagactgtcccttctggaatgaagaaatagtcccccagatcaaggaggggaaacgt gtactgttgcagccatggcaacaggctccgggcattgtcaagcatctgtggaggtctc tctgaagaggcttatgtggactgtggatcttccattgtctatgtatggat		

Figure 1.12.7 The GENES Entry page showing detailed information for the selected gene.

1. Access the KEGG Table of Contents as described in Basic Protocol 1. Click on the “Color objects in KEGG pathways” link under “Search & Compute” in the KEGG PATHWAY Database row. This will bring up a search window (Fig. 1.12.9).

Note that the “Search objects in KEGG pathways” also provides similar functionality.
2. Assuming that one has a list of EC numbers of enzymes collected while studying *Anabaena*, and that one wishes to compare these to those in the *Synechocystis* pathway, carry out the following example search. Select the *Synechocystis* sp. PCC6803 pathway from the pull-down menu under “Search against” and copy-and-paste the list of EC numbers into the large text field.

Note that it is also possible to upload a file of EC numbers using the upload field located below the large text field (labeled “Alternatively, enter the file name containing the data.”).

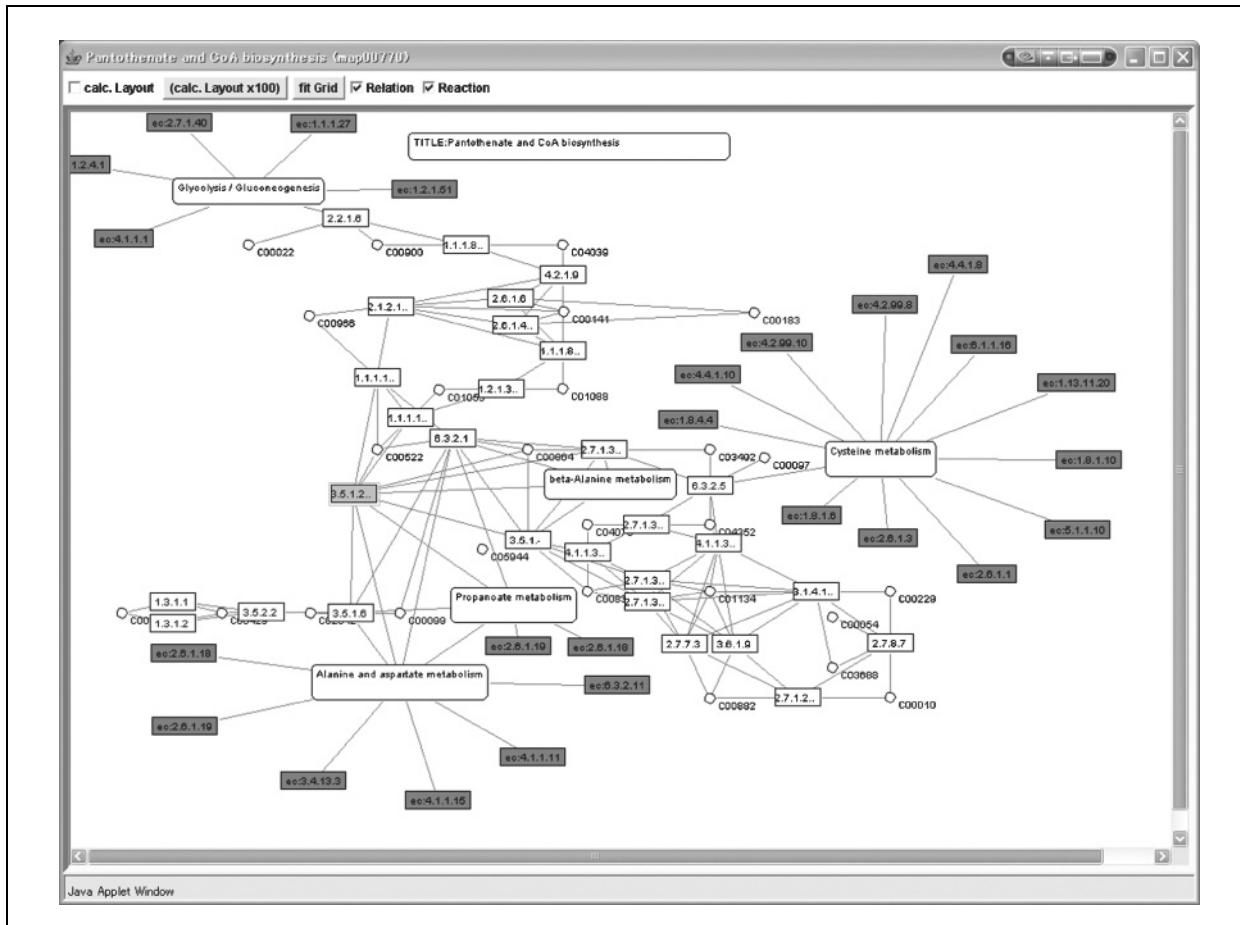


Figure 1.12.8 KGML Viewer. Individual nodes may be double-clicked to view its corresponding DBGET entry, whether it be a compound, enzyme, reaction, or gene. This viewer gives the user a quick and convenient overview of all the relationships available in KEGG, which can be stored or transferred in a machine-independent manner.

3. Within this set, assume that there are a few genes in which one is particularly interested. In order to specify the color in which the corresponding search hits are to be displayed, type the text `red` next to EC number 3.6.1.7, for example.
4. Click the Exec button to see the results (Fig. 1.12.10). At the top of the page is a list of those objects not found, followed by a textual listing of all of the pathways that contain the EC numbers and organism specified. Clicking on “Pyruvate metabolism” brings up the page illustrated in Figure 1.12.11, which shows the genes found only in the *Synechocystis* pathway colored green, genes common to both *Anabaena* and *Synechocystis* colored pink, and the specified genes (see step 3) in red. From here, various items may be inspected to see the GENES entries or other pathways, or chemical compounds involved with these genes in this network.

KEGG PATHWAY: USING THE KEGG ORTHOLOG TABLES

The Pathway maps illustrate possible reaction paths and interaction networks retrieved from experimental data in the literature over all organisms. However, gene functions discovered by wet-lab experiments are organism-specific, and genes whose function has been proven experimentally in the strict sense are extremely few compared to those found by genomic analysis. It is expected that this gap will widen in the future. Homology-based

BASIC PROTOCOL 4

Using Biological Databases

1.12.9

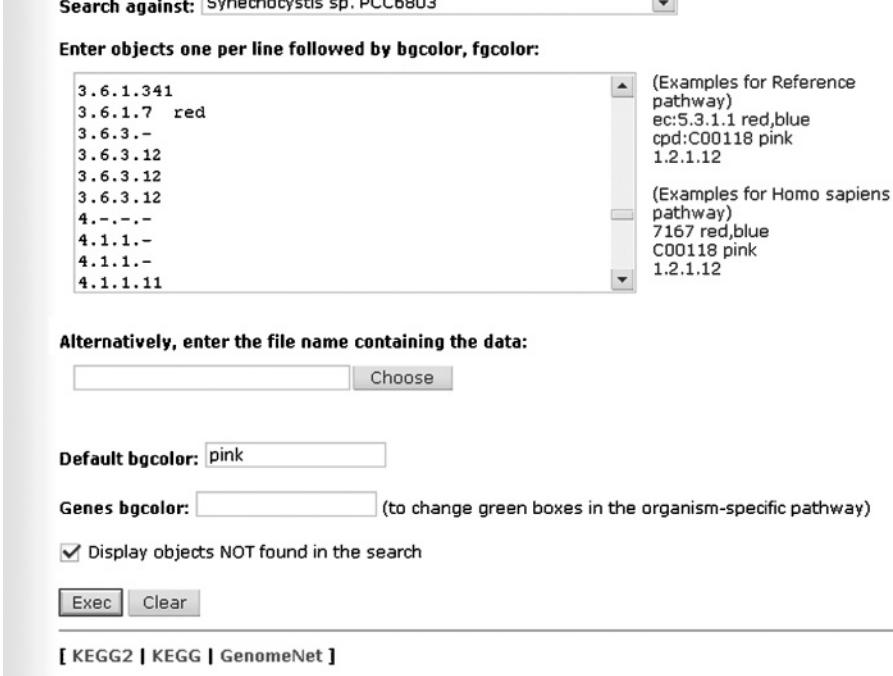


Figure 1.12.9 Pathway coloring form.

hypotheses of gene function are imperfect. There are orthologous genes with the same function but with low sequence similarity, and genes with high sequence similarity in the same paralog group with significantly different functions. KEGG provides ortholog tables to represent information in both a genomic context and a pathway-based context, providing the user with a tool to visualize the context in which a gene function is expressed. Thus, this protocol can be used to easily investigate all functionally related genes, as well as the corresponding pathways in which they occur, in a single table.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

1. Access KEGG Table of Contents, then bring up the list of pathways, as described in Basic Protocol 1. Access the KEGG Ortholog Tables by clicking the Ortholog link for a given pathway. Alternatively, click on the Ortholog Table link above the organism pull-down menu at the top of any pathway map (see Basic Protocol 1). For example, by clicking the Ortholog Table link at the top of the Glycolysis/Glucogenesis reference pathway page shown in Figure 1.12.6 (at <http://www.genome.jp/kegg/pathway/map/map00010.html>) one is taken to the Ortholog Table shown in Figure 1.12.12.

• syn00620 Glucuronidation
EC 2.4.1.-
EC 2.4.99.-
• syn00603 Globoside metabolism
EC 2.4.1.-
EC 2.4.99.-
EC 3.2.1.52 beta-N-acetylhexosaminidase; hexosaminidase; beta-acetylaminodeoxyhexosidase; N-acetyl-beta-D-hexosam
• syn00604 Ganglioside biosynthesis
EC 2.3.1.-
EC 2.4.1.-
EC 2.4.99.-
• syn00620 Pyruvate metabolism
EC 1.1.1.28 D-lactate dehydrogenase; lactic acid dehydrogenase; D-specific lactic dehydrogenase; D-(-)-lactate de
EC 1.1.1.37 malate dehydrogenase; malic dehydrogenase; L-malate dehydrogenase; NAD-L-malate dehydrogenase; malic
EC 1.1.1.38 malate dehydrogenase (oxaloacetate-decarboxylating); 'malic' enzyme; pyruvic-malic carboxylase; NAD-
EC 1.2.1.1 formaldehyde dehydrogenase (glutathione); NAD-linked formaldehyde dehydrogenase; formaldehyde dehydroge
EC 1.2.1.3 aldehyde dehydrogenase (NAD+); CoA-independent aldehyde dehydrogenase; α -methylbenzaldehyde dehydroge
EC 1.2.1.22 lactaldehyde dehydrogenase; L-lactaldehyde:NAD oxidoreductase; nicotinamide adenine dinucleotide (NAD)
EC 1.2.4.1 pyruvate dehydrogenase (acetyl-transferring); MtPDC (mitochondrial pyruvate dehydrogenase complex); py
EC 1.2.7.1 pyruvate synthase; pyruvate oxidoreductase; pyruvate synthetase; pyruvate:ferredoxin oxidoreductase;
EC 1.8.1.4 dihydrolipoamide dehydrogenase; LDH-Glc; LDH-Val; dihydrolipoate dehydrogenase; dihydrolipop
EC 2.3.1.12 dihydrolipoyllysine-residue acetyltransferase; acetyl-CoA:dihydrolipoamide S-acetyltransferase; dihyd
EC 2.7.1.40 pyruvate kinase; phosphoenolpyruvate kinase; phosphoenol transphosphorylase pyruvate kinase (phospho
EC 2.7.2.1 acetate kinase; acetokinase
EC 2.7.9.2 pyruvate, water dikinase; phosphoenolpyruvate synthase; pyruvate-water dikinase (phosphorylating); PE
EC 3.1.2.6 hydroxyacylglutathione hydrolase; glyoxalase II; S-2-hydroxyacylglutathione hydrolase; acetoacetylgl
EC 3.6.1.7 acylphosphatase; acetylphosphatase; 1,3-diphosphoglycerate phosphatase; acetic phosphatase; Ho 1-3; G
EC 4.1.1.31 phosphoenolpyruvate carboxylase; phosphopyruvate (phosphate) carboxylase; PEP carboxylase; phosphoen
EC 4.1.1.-
EC 4.1.3.-
EC 4.4.1.5 lactoylglutathione lyase; methylglyoxalase; aldoketomutase; ketone-aldehyde mutase; glyoxylase I
EC 6.2.1.1 acetate-CoA ligase; acetyl-CoA synthetase; acetyl activating enzyme; acetate thiokinase; acyl-activat
EC 6.4.1.2 acetyl-CoA carboxylase; acetyl coenzyme A carboxylase
• syn00621 Biphenyl degradation
EC 4.1.3.-
• syn00622 Toluene and xylene degradation
EC 4.1.3.-
• syn00624 1- and 2-Methylnaphthalene degradation
EC 1.1.1.1 alcohol dehydrogenase; aldehyde reductase; ADH; alcohol dehydrogenase (NAD); aliphatic alcohol dehyd
EC 1.2.1.-
EC 1.14.13.-
EC 2.3.1.-
EC 4.1.1.-
EC 4.2.1.-
• syn00625 Tetrachloroethene degradation
EC 1.1.1.-
EC 1.18.6.1 nitrogenase
EC 4.2.1.-
• syn00626 Nitrobenzene degradation
EC 1.2.1.-
EC 1.14.13.-
EC 2.1.1.-
• syn00627 1,4-Dichlorobenzene degradation
EC 1.14.13.-
EC 3.1.1.45 carboxymethylenebutenolidase; maleylacetate enol-lactonase; dienelactone hydrolase; carboxymethylene
EC 4.1.3.-
• syn00628 Fluorene degradation

Figure 1.12.10 Textual listing of EC coloring results.

In the Ortholog Table, rows correspond to organisms and columns correspond to ortholog sets. Column names beginning with the letter E are KO (KEGG Orthology) identifiers. In this example, these identifiers correspond to the EC numbers on the pathway map. Each cell is color-coded to illustrate its genomic context. Genes that are next to one another on the genome, such as operon structures, have a high probability of being functionally located such that they are transcriptionally regulated simultaneously. Thus, they are colored similarly.

2. There are many analyses that can be done with the Ortholog Table. For example, under the Organism column, each entry has links to “P | G | T” corresponding to Pathway, Genome, and Title. “P” links to the pathway map, “G” to the genome map, and “T” to a list of gene functions. Clicking “P” will invoke a pathway map with the boxes of the proteins colored according to the colors of those genes appearing in the ortholog table. Therefore, it is possible to confirm those genes located consecutively on the genome as those appearing consecutively on the pathway. Clicking on “G” will invoke the genome map (introduced later in Basic Protocol 6), which will have red lines indicating the genes that are in the Ortholog Table in the overall view and red boxes around names of those genes in the detailed view. The listing of genes provides a convenient interface for analyzing the genes selected. Next, take a look at the genes aligned down the columns of the ortholog table. Click on the column name (i.e., the KO name, the designation beginning with “E” under the number at the head of each column) to display a list of all of the genes in the column, as shown in Figure 1.12.13 for KO E2.7.2.3.

Using Biological Databases

1.12.11

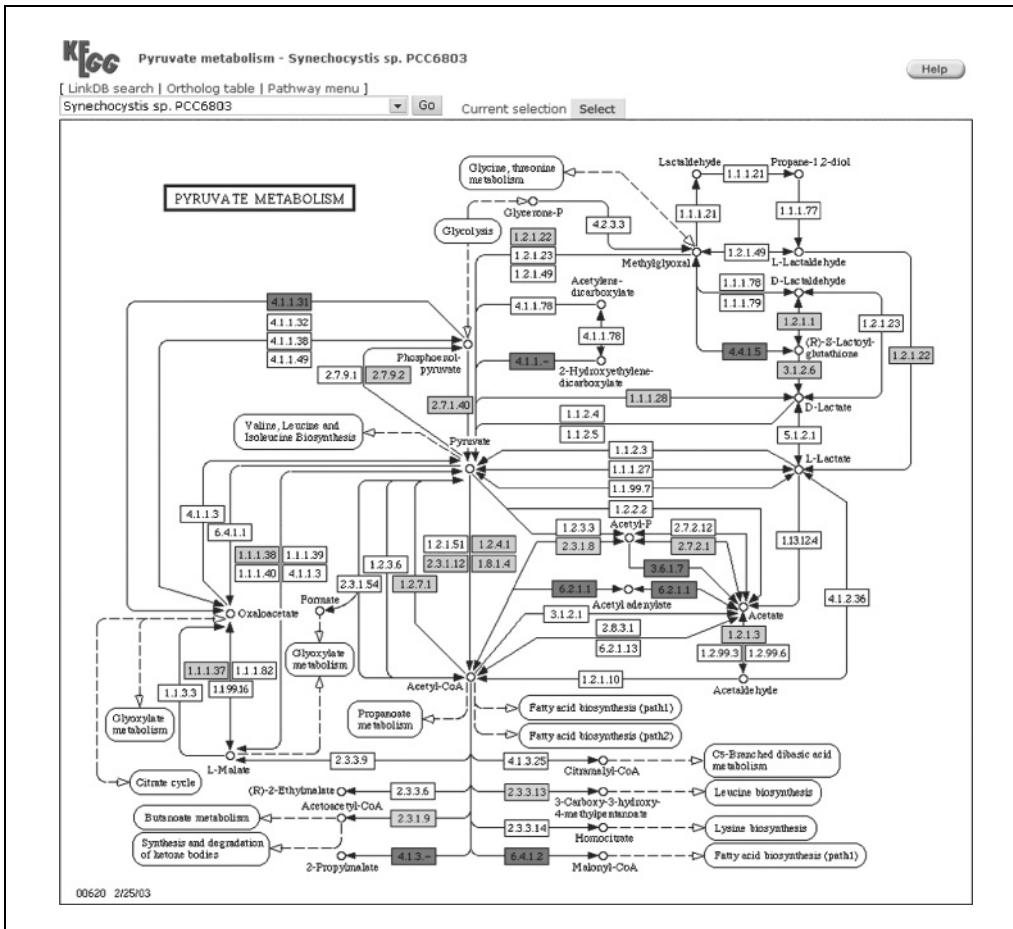


Figure 1.12.11 The resulting pathway colored with the values specified. Note the genes in green which are only found in *Synechocystis*, the ones in pink which are common to both *Anabaena* and *Synechocystis*, and the ones in red whose color was specified. For the color version of this figure go to <http://www.currentprotocols.com>.

From here, various operations can be performed, such as multiple alignments from the amino acid sequences in FASTA format, dendrograms of relationships between genes based on the SSDB clustering results, and searching common motifs in the selected genes, which are the same operations available for the SSDB search results. See the section on SSDB (Basic Protocols 7 to 9) for details on these procedures.

- Because the ortholog tables are manually generated and updated, the default tables contain only those organisms that are entered in the GENES database. For tables containing entries for organisms not listed, go back to the ortholog table page (shown in Fig. 1.12.12) and click on the Select button under the ortholog table title to open a new window that allows organism selection.. Here, specific organisms may be selected with All Species, Eukaryotes, Bacteria, or Archaea. Moreover, a user-defined list of organisms may be created and appended.

Note that the latter function cannot be performed if cookies are disabled in the browser.

- As an example, click the Select button above the table and select Bacteria instead of “All organisms.” Note that a user-defined list of organisms may also be entered in the text box on the bottom half of this dialog box. Next, click Select to update the ortholog table (Fig. 1.12.14).

Glycolysis / Gluconeogenesis											
See also: Pyruvate/oxoglutarate oxidoreductases											
Current selection: Select											
Organism	1	2	3	4	5	6	7	8	E6.4.1.2		
	E2.7.1.11	E4.1.2.13	E5.3.1.1	E1.2.1.12	E2.7.2.3	E5.4.2.1	E4.2.1.11	E2.7.1.40	Acetyl-CoA carboxylase (See also: [PATH:tab0061])		
	Phospho-fructokinase	Aldolase	Triose-phosphate isomerase	Glyceraldehyde-3P dehydrogenase	Phosphoglycerate kinase	Phosphoglycerate mutase	Enolase	Pyruvate kinase	alpha	beta	BcoP
class II	class I										
eco [P G T]	b3916(pfkA) b1723(pfkB)	b2925(fba)	b2097(dhnA)	b3919(tpfA)	b2927(epd) b1779(gapA) b1417(gapC_1) b1416(gapC_2)	b2926(pgk)	b0755(gpmA) b1395(gpmB) b3612(yfbO)	b2779(eno)	b1854(pykA) b1676(pykF)	b0185(accA)	b2316(accD) b3255(accB)
sty [P G T]	STY1785 STY3809	STY3226	STY2370	STY3789	STY3228(1.2.1.-) STY1825	STY3227	STY0804 STY4091 STY4932	STY3081	STY1744 STY2096	STY0255	STY2597 STY3559
ypo [P G T]	YP00920 YP03960	YP00078	YP00085	YP00922(1.2.1.-)	YP00921	YP00044 YP00455 YP01133	YP03376	YP02064 YP02393	YP01060	YP02768	YP03659
sfl [P G T]	SF3994 SF1507	SF2910	SF2159	SF3997	SF2912(1.2.1.-) SF1444 SF1795 SF02157	SF2911	SF0549 SF3651 SF4427	SF2794	SF1705 SF1864	SF0175	SF2392 SF3293
buc [P G T]	BU305	BU451		BU307	BU298	BU450	BU304	BU417	BU319		
wbr [P G T]					Wbr0496	Wbr0018	Wbr0462	Wbr0556	Wbr0412	Wbr0640	Wbr0387 Wbr0456
hin [P G T]	HIO982	HIO524		HIO678	HIO0001	HIO525	HIO757	HIO932	H11573	H10406	H11260 H10971
pmu [P G T]	PM0069	PM1861	PM1311	PM1640	PM0924	PM1860	PM1906 PM0634	PM1871	PM0533	PM0292	PM0636 PM1092
xfa [P G T]	XF0274		XF0826	XF0303	XF0457	XF0823	XF1893	XF1291	XF0824	XF0203	XF1467 XF0048
xac [P G T]	XAC3438		XAC3344	XAC2707	XAC3347	XAC2874	XAC1719	XAC3345	XAC3345	XAC1405	XAC2715 XAC0532
vch [P G T]	VC2689	VC0478		VC2670	VC0476(1.2.1.-) VC1069 VC2000	VC0477	VC0336	VC2447	VC0485 VC2008 VC0708	VC2244	VC1000 VC0296
vvu [P G T]	VV11257	VV11541		VV11343	VV11540(1.2.1.-) VV1141 VV13140	VV11540	VV11281	VV11579	VV10644 VV10922 VV2026	VV11876	VV11993 VV11235
vpa [P G T]	VP2855	VP2599		VP0239	VP2601(1.2.1.-) VP2157 VP2970	VP2600	VP2829	VP2561	VP0256 VP2039 VPA0823	VP2302	VP2189 VPA0795 VPC280
pae [P G T]	PA0555			PA4748	PA0551(1.2.1.-) PA3195	PA0552	PA5131	PA3635	PA1498 PA4329	PA3639	PA3112 PA4047
ppu [P G T]	PP4960			PP4715	PP4964(1.2.1.-) PP1009	PP4963	PP5056	PP1612	PP1362 PP4301	PP1607	PP1996 PP0559
sen [P G T]	SO0933			SO1200	SO0931(1.2.1.-) SO0538 SO2345	SO0932	SO0049	SO3440	SO2491		SO0511
nme [P G T]	NMB1869			NMB8187	NMB0207 NMB2159	NMB0010	NMB1604	NMB1285	NMB0089 NMB1177 NMB1139	NMB0679	NMB1860
rso [P G T]	RS04892			RS03629	RS00105	RS04894	RS03320 RS05023	RS04624	RS04893 RS05095	RS04560	RS03554 RS05392
hpv [P G T]	HP0176			HP0194	HP1346 HP0921	HP1345	HP0974	HP0154		HP0557	HP0950 HP0371
cje [P G T]	CJ0597			CJ1401c	CJ1403c	CJ0434	CJ1672c	CJ0392c	CJ0443	CJ0127c	CJ1291c
mlo [P G T]	mlf5025	mlf7273	mlr3754	mlr7275 mlr0610	mlr3750	mlr3753	mlr4643 mlr0406	mlr3078	mlr3819	mlr3576	mlr5075 mlr206
sme [P G T]	SMb20199 SMb21192	SMc03983	SMc01023 SMc01614		SMc03979	SMc03981	SMa5027 SMc02838	SMc01028	SMc04005	SMc00690	SMc02764 SMc01344
ahf [P G T]				dhv174n	dhv169n	dhv1727	dhv1720	dhv146q	dhv1729	dhv169n	dhv1720n Abu1331

Figure 1.12.12 Ortholog Table for Glycolysis/Gluconeogenesis. The numbers above each column correspond to the numbers in the pathway, illustrating the relationship of these orthologs in the pathway map. Each cell is color-coded to illustrate its genomic context. For the color version of this figure go to <http://www.currentprotocols.com>.

In this way, it is possible to see patterns among various organisms. For example, for the organisms at the top of this list, it is apparent that the enzymes of E4.1.2.13 class II, E1.2.1.12, and E2.7.2.3 have a larger number of shaded cells than the others, indicating that this path of reactions is shared in common and may be considered a closely related ortholog group. The cells are shaded in different colors to distinguish between different ortholog groups.

THE KEGG GENES DATABASE: GETTING STARTED

The KEGG GENES database is an annotated collection of more than 930,000 genes in 304 organisms, derived from GenBank as well as the literature. Additional statistical information on the composition of the KEGG suite of databases can be found at <http://www.genome.jp/kegg/kegg1.html>. An introduction to the various features of the KEGG GENES database is given in this protocol, along with instructions for performing a simple search for a specific gene.

BASIC PROTOCOL 5

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

Using Biological Databases

1.12.13

Entries in E2.7.2.3

Check Clear Select operation Exec

[eco:b2926](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [sty:STY3227](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [ype:YPO0921](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [sf1:SF2911](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [buc:BU45Q](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [wbr:Wbr0462](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [hin:HIO525](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [pmu:PM1860](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [xfa:XF0823](#) phosphoglycerate kinase [EC:2.7.2.3]
 [xac:XAC3347](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [vch:VC0477](#) phosphoglycerate kinase [EC:2.7.2.3]
 [vvu:VV11540](#) phosphoglycerate kinase [EC:2.7.2.3]
 [vpa:VP2600](#) phosphoglycerate kinase [EC:2.7.2.3]
 [pae:PA0552](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [ppu:PP4963](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [son:SO0932](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [nme:NMB0010](#) phosphoglycerate kinase [EC:2.7.2.3]
 [rso:RSc0571](#) pgk, RSO4894; probable phosphoglycerate kinase protein [EC:2.7.2.3]
 [hpy:HP1345](#) phosphoglycerate kinase [EC:2.7.2.3]
 [cje:Cj1402c](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [mlo:mir3753](#) phosphoglycerate kinase [EC:2.7.2.3]
 [sme:SMe03981](#) pgk; probable phosphoglycerate kinase protein [EC:2.7.2.3]
 [atu:Atu3739](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [bme:BMEI0309](#) phosphoglycerate kinase [EC:2.7.2.3]
 [bms:BR1729](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [bjr:b111522](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [ccr:CC3249](#) phosphoglycerate kinase [EC:2.7.2.3]
 [bsu:BG11062](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [bha:BH3559](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [oiih:OB2437](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [sau:SA0728](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [sep:SE0558](#) phosphoglycerate kinase [EC:2.7.2.3]
 [lmo:lmo2458](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [lla:L0010](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]
 [spy:SPy1881](#) pgk; phosphoglycerate kinase [EC:2.7.2.3]

Figure 1.12.13 KO list for E2.7.2.3 where various operations can be performed for analyzing the relationships between these genes.

1. Before delving into the protocol, it is first necessary to define the fields available in a KEGG GENES Entry. Using Figure 1.12.7 as an example, the discussion in this protocol refers to:
 - a. GENE ID at the top of the figure (i.e., 5223) as the accession number listed in the Entry field.
 - b. The KEGG ID is the GENE ID preceded by the three-letter organism code (e.g., hsa:5223). The other fields are as follows:
 - c. The taxon name of *H. sapiens* is hyperlinked to its DBGET entry as if a text-based search was performed for this genome. This search result contains links to such information as the publications from which it was defined and its GenBank entry.
 - d. The Gene name line lists the general gene names by which this entry is known.
 - e. The Definition line describes this entry in human-readable terms.
 - f. The KO line contains a hyperlink to the KO (KEGG Ortholog) entry to which this gene belongs. It also contains links to its corresponding “OC search” and “OC viewer,” which provide useful ortholog information for the given entry. These tools are described in detail in Basic Protocol 4.
 - g. The Pathway line lists pathway maps in which this gene is involved.

Organism	E2.7.1.11	E4.1.2.13		E5.3.1.1	E1.2.1.12	E2.7.2.3	E5.4.2.1	E4.2.1.11	E2.7.1.40	E6.4.1.2	
	Phospho-fructokinase	Aldolase		Triose-phosphate isomerase	Glyceraldehyde-3P dehydrogenase	Phosphoglycerate kinase	Phosphoglycerate mutase	Enolase	Pyruvate kinase	Acetyl-CoA carboxylas (See also: [PATH:tab000]	
		class II	class I		b2927(epd) b1779(gapA) b1417(gapC_1) b1416(gapC_2)	b2926(pgk)	b0755(gpmA) b439(gpmB) b3612(ybO)			b0185(ackA) b2316(ackD)	b3155
eco [P G T]	b3916(pfka) b1723(pfkb)	b2925(fba)	b2097(dhnA)	b3919(tpiA)	b2927(epd) b1779(gapA) b1417(gapC_1) b1416(gapC_2)	b2926(pgk)	b0755(gpmA) b439(gpmB) b3612(ybO)	b2779(enol)	b1854(pykA) b1676(pykF)	b0185(ackA) b2316(ackD)	b3155
ecj [P G T]	JW1712 JW3887	JW2092	JW2064	JW3090	JW1413 JW14 JW1768 JW2894	JW2893	JW0738 JW3587 JW4358	JW2750	JW1666 JW1843	JW0180	JW2313
ece [P G T]	Z2752 Z5460	Z4263 Z5667	Z3260	Z5464	Z2304 Z2818 Z4266	Z4265	Z0925 Z5039 Z5997	Z4094	Z2704 Z2906	Z0197	Z3578
ecs [P G T]	ECs2429 ECs4841	ECs3795 ECs5069	ECs2900	ECs4844	ECs2022 ECs2488 ECs3708	ECs3797	ECs0783 ECs4490 ECs5353	ECs3639	ECs2383 ECs2564	ECs0187	ECs3200
ecc [P G T]	c2121 c4067	c3503 c4483	c2623	c4871	c1843 c2184 c3505	c3504	c0831 c438 c5482	c3344	c2071 c2268	c0223	c2861
sty [P G T]	STY1785 STY3809	STY3226	STY2370	STY3789	STY3228(1.2.1.-) STY1825	STY3227	STY0804 STY4091 STY4932	STY3081	STY1744 STY2096	STY0255	STY2597
stt [P G T]	t2106 t3557	t2987	t0715	t3537	t1169 t2989	t2988	t2115 t3815 t4624	t2853	t0989 t1246	t0233	t0498
spt [P G T]	SPA1518 SPA3905	SPA2939	SPA0711	SPA3924	SPA1554 SPA2941	SPA2940	SPA1980 SPA3556 SPA3935	SPA2809	SPA0981 SPA175	SPA0239	SPA0498
stm [P G T]	STM1326 STM4062	STM3068	STM2141	STM4081	STM1290 STM3070	STM3069	STM0772 STM3704 STM4585	STM2952	STM1378 STM1386 STM1392 STM1808	STM0232	STM2366
ype [P G T]	YPO0078 YPO0920 YPO3960			YPO0085	YPO0922(1.2.1.-) YPO2157	YPO0921	YPO0064 YPO0455 YPO1133	YPO3376	YPO2064 YPO2393	YPO1060	YPO2768
ypk [P G T]	y0059	y3307 y3859		y0052	y2165 y3309	y3308	y0077 y3048 y3724	y0814	y1944 y2246	y3119	y1601
ypm [P G T]	YPO080	YP3323 YP3520		YPO089	YP1957 YP3518	YP3519	YPO064 YPO1255 YPT3728	YPO310	YP1907 YP2180	YP2790	YP2396
yps [P G T]	YPTB0074	YPTB3195 YPTB3803		YPTB0081	YPTB2083 YPTB3197	YPTB3196	YPTB0050 YPTB0598 YPTB1156	YPTB0755	YPTB2047 YPTB2306	YPTB2987	YPTB2616
sfI [P G T]	SF3994 SF1507	SF2910	SF2159	SF3997	SF2912(1.2.1.-) SF1444 SF1795 SF1796	SF2911	SP0549 SF3651 SF427	SF2794	SF1705 SF1864	SF0175	SF2392
sfx [P G T]	S1624 S1624	S3110	S2285	S3750	S1559	S3111	S0557 S4117	S2988	S1838 S4117	S0178	S2527

Figure 1.12.14 Ortholog table after selecting various organisms. Note how there seems to be a clustering of similarly functioning genes according to the coloring in the middle of the table. For the color version of this figure go to <http://www.currentprotocols.com>.

- h. The Class line lists the hierarchical classification of this gene as listed in the Gene Catalog. It can be seen that this gene is involved in the Glycolysis/Gluconeogenesis pathway when the Gene Catalog button is clicked.
- i. The SSDB line contains links to search results for paralogs, orthologs, motifs, and gene clusters as calculated and stored in the KEGG SSDB database. This database and its protocols are described in detail under Basic Protocols 7 to 9.
- j. The Motif line lists known motifs that have been documented in this entry's sequence. Each link will display the corresponding motif database entry.
- k. The Other DBs line provides links to outside databases that contain relevant information corresponding to this entry.
- l. The LinkDB line provides links to PDB and a summarized listing of all other popular databases around the world, related to this gene. Some of the databases currently referenced include GenBank, EMBL, GDB, and OMIM, among others. For select organisms, links are also available to the GenomeNet Community Databases (BSORF and CYORF).
- m. The Position line indicates the location of this gene on the genome, and the Genome Map link will display a view containing the position on the chromosome map where this gene is located, with the gene name emphasized. The Genome Map tool is described in Alternate Protocol 5.
- n. The AA seq and NT seq lines contain links to the FASTA format amino acid and nucleotide sequences corresponding to this gene. These are useful for performing sequence analysis in that they can simply be cut-and-pasted.

Using Biological Databases

1.12.15

2. Access the KEGG GENES database from the KEGG GENES Database link on the KEGG Table of Contents (accessed as in Basic Protocol 1), or via the URL <http://www.genome.jp/kegg/genes.html>. The KEGG Genes database page is shown in Figure 1.12.15.

*KEGG GENES is actually a composite database where each organism is one database and each gene within a database is one entry. The KEGG nomenclature is defined such that either a database (organism) name, e.g., *Homo sapiens*, or the corresponding three-letter code, e.g., hsa, can be searched. An entry (gene) name can be searched by its accession number from its corresponding genome project database, such as b0002, or by its common gene name, such as *thrA*. However, note that, while accession numbers and primary gene names are unique, general gene names may result in a number of matching entries.*

- To search for *E.coli* genes involved in citrate synthase, enter eco and citrate synthase as the query terms at the “Search KEGG organism” bullet point, then click Go. This will return a results page showing genes b0333 and b0720. Note that this query will be considered as a Boolean search for citrate AND synthase. Other types of searches such as citrate or synthase only, a gene name (e.g., *gltA*), or an EC number (e.g., 4.1.3.7) may be performed.
- On the page that now appears, click “b0720” to see the KEGG GENES Entry for this gene.

Genomes in the taxonomy	Gene catalog	Remark
KEGG Organisms	GENES	High-quality genomes with manual KO assignment
	DGENES	Draft genomes with automatic KO assignment
	EGENES	EST contigs with automatic KO assignment
Complete viral genomes	VGENES	Available only in DBGET
Complete mitochondrial genomes		
Complete plastid genomes	OGENES	
Complete nucleomorph genomes		

Figure 1.12.15 Searching the KEGG GENES database by organism.

KEGG GENES: USING THE GENOME MAP

Biological meaning can often be derived from the physical location of genes on the genome. Examples of this can be found in the operons of prokaryotes and in the synteny of eukaryotes. KEGG provides the Genome map browser, where genetic information can be obtained from physical positions on the genome.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

- From the KEGG Table of Contents page (accessed as in Basic Protocol 1), enter *bsu* in the text box next to “Enter KEGG organism code” in the Specialized KEGG section. Click Go. This searches for the organism *Bacillus subtilis*. The displayed organism list shows two Genome headings, Genome Map and Genome Info, at the upper right. Genome Info links to the KEGG GENOME Database entry containing all of the information regarding the selected genome, such as the data sources, literary references, and sequence sources. Click on Genome Map to display the Genome map browser for *Bacillus subtilis*, as shown in Figure 1.12.16.
- There are two views of the Genome map, a global view shown in Figure 1.12.16 and a local, more detailed view as shown in Figure 1.12.17. The global view can be used to find genes everywhere in the genome by entering a gene accession number. As the mouse is moved over the gray bar representing the chromosome in the global

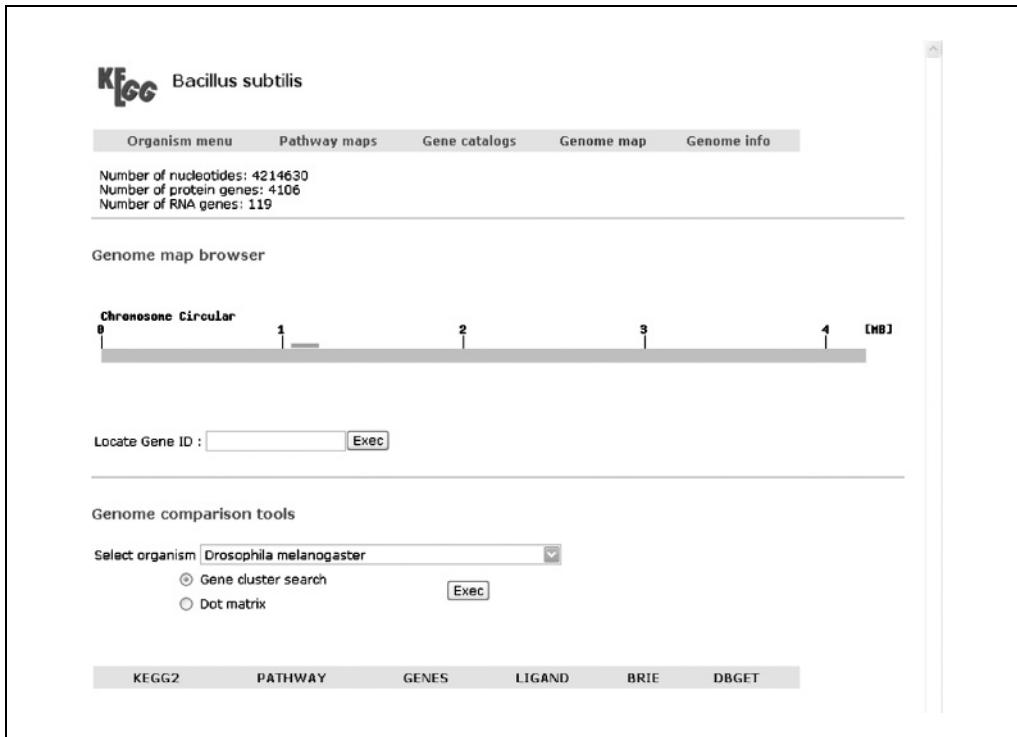


Figure 1.12.16 Global view of genome map for *Bacillus subtilis*. By clicking on a green-highlighted box in the global view (which follows the mouse as it is moved over the map), a detailed view of the map will be displayed, as in Figure 1.12.17.

Using Biological
Databases

1.12.17

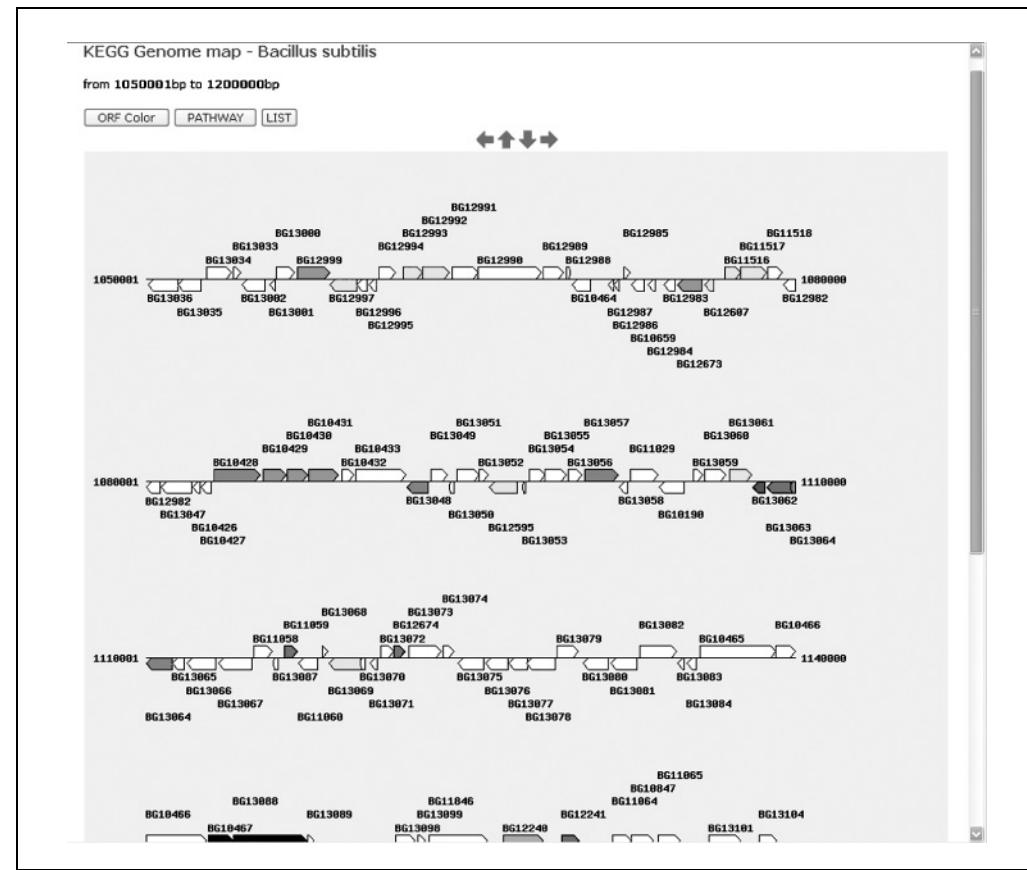


Figure 1.12.17 Detailed view of genome map of *Bacillus subtilis*. If a KEGG gene has been searched for in the global view, the corresponding gene ID will be boxed in red in this view.

map (Fig. 1.12.16), a green box follows the mouse to indicate which portion of the whole map can be examined in detail. Click on a green-highlighted box in the global view to display a detailed view of the genome map, as in Figure 1.12.17. In the case that a gene has been searched for in the global view, the gene name in the detailed view will be highlighted to emphasize its location. In the detailed view, the arrows indicate the genes in the forward and backward strands, and the corresponding gene names are linked to their KEGG GENES entries. Each gene is colored according to its KO classification. The arrows at the top and bottom of the screen allow the user to traverse the genome map in the corresponding direction. The down and right arrows move the viewing window right by 30,000 and 150,000 base pairs respectively; the up and left arrows move the viewing window to the left by the corresponding amounts.

Thus, the genome map provides a comprehensive, yet clearly visualized, illustration of the genes under study and their positions on the genome. Combined with the other tools in KEGG, such as the ortholog and paralog clusters described in Basic Protocol 9, the understanding of related genes and their functions can be analyzed with a click of the mouse.

BASIC PROTOCOL 7

Using the KEGG Database Resource

1.12.18

all of the protein-coding genes in one genome against all of the genes in another genome. This is done for all pairs in all known genomes. Not only does this produce sequence similarity scores for all pairs of genes in all known genomes, but this also simplifies the analysis of orthologs and paralogs.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

Searching orthologs and paralogs with KEGG SSDB

1. Access the KEGG SSDB functionality via the URL <http://www.genome.jp/KEGG/genes.html>.
2. Scroll down to the “Search orthologs:” field (under Precomputed Sequence Similarities) and enter the gene name. The format of this field is the three-letter species code, followed by a colon, followed by gene name (e.g., for this example, enter hsa:5921). Select “forward best” from the neighboring pull-down menu, then click Go. The results of a search will produce a screen similar to that shown in Figure 1.12.18.

To further refine the search, the species to search against may be specified by selecting or deselecting the species listed under “Search against” on the page illustrated in Figure 1.12.18. If no species to search against are selected, by default, all species will be searched against. Other search parameters include Threshold, which is the threshold for the Smith-Waterman score. SSDB is limited to all genes whose similarity scores are at least 100, so a threshold of 100 will retrieve all homology information. The “Select operation” pull-down menu provides a variety of options for performing ortholog and paralog searches of the query gene (described further in the on-line link).

The top half of the results page displays the query information, including the gene and species, its corresponding KEGG ID and its GENES definition. The GFIT link will open a new window to the results of the Gene Function Identification Tool (GFIT) system, which lists the results with gene annotation information (Bono et al., 1998), and the similar genes returned by SSDB are listed below all this header information.

In the search results list, along with each gene’s KEGG ID and GENES definition, the length of the common amino acid sequence (len), the Smith-Waterman score (SW-score), the bit score (bits), the percent of amino acid identity between the homologous portions (identity), and the amino acid length of the homologous portions (overall) are listed. Finally, the arrows (or lack thereof) in the right-most column, “best(all),” illustrate the similarity type of the hit in relation to the query. These are described in Table 1.12.1.

3. It is possible to run the search by modifying the “Show,” “Sort by,” “Search against,” and “Threshold” parameters and clicking the Exec button. The “Sort by” option, which appears after the initial search, allows one to sort the results by Smith-Waterman score, by species, or by a combination of the two.
4. Use the “Select operation” pull-down menu to integrate information from the search results. To use this facility, check one or more of the listed genes, select the desired operation, and click Select. Note that some operations may take longer with more selected genes. For illustration purposes, the following examples have been performed on the top eight resulting genes.

SSDB Forward Best Search Result

KEGG ID : hsa:5921 (1047 a.a.) [GFIT] [OC]
Definition: RAS p21 protein activator (GTPase activating protein) 1
Update status: H.sapiens (bfr,bma,ddi,ehi,ngo,osa,tbr,vfi : calculation not yet completed)

Show : Best-best Forward best Reverse best Paralogs Gene clusters
Sort by : SW-score SW-score by species KEGG-species
Search against: Bacteria Archaea Eukaryotes
Threshold: 100

Check **Clear** **Select operation** **Select**

Search Result : 2

Entry **len** **SW-score** **bits** **identity** **overlap** **best(all)**

<input checked="" type="checkbox"/> mmu:218397 RAS p21 protein activator 1	813	5365	1229	0.989	813	<-> 672
<input checked="" type="checkbox"/> dme:CG3209-PB vacuolar peduncle	954	2746	632	0.470	921	<-> 645
<input checked="" type="checkbox"/> cel:ZK899.8a GTPase-activating protein	1207	738	174	0.291	564	-> 382
<input checked="" type="checkbox"/> sce:YKL092C GTPase-activating protein (GAP) for Rsr1p/B	1104	401	97	0.227	449	-> 28
<input checked="" type="checkbox"/> ago:ABR021W GTPase-activating protein (GAP) for Rsr1p/B	1084	395	96	0.221	417	<-> 45
<input checked="" type="checkbox"/> cal:orf19.5219 GTPase activating protein	2643	276	69	0.250	296	-> 40
<input checked="" type="checkbox"/> mbo:Mb1096c PE_PGRS family protein	671	214	55	0.344	151	-> 165
<input checked="" type="checkbox"/> mtc:MT1096.1 PE_PGRS family protein	667	214	55	0.344	151	-> 150
<input checked="" type="checkbox"/> mtu:Rv1067c PE_PGRS-family protein	667	214	55	0.344	151	-> 133
<input checked="" type="checkbox"/> spo:SPBC646.12c GTPase-activating protein; no apparent	766	208	53	0.240	334	-> 27
<input checked="" type="checkbox"/> sco:SC05273 hypothetical protein	620	207	53	0.331	178	-> 235
<input checked="" type="checkbox"/> ath:At2g28670 disease resistance-responsive family prot	447	187	48	0.397	121	-> 208
<input checked="" type="checkbox"/> xft:PD1851 endo-1,4-beta-glucanase	614	186	48	0.286	185	-> 10
<input checked="" type="checkbox"/> cef:CE2654 hypothetical protein	603	180	47	0.263	228	-> 23
<input checked="" type="checkbox"/> nfa:nfa0180 hypothetical protein	428	179	47	0.395	114	-> 160
<input checked="" type="checkbox"/> sma:SAV3556 hypothetical protein	903	179	47	0.263	285	-> 210
<input checked="" type="checkbox"/> bce:BC4725 hypothetical membrane spanning protein	1303	176	46	0.292	171	-> 19
<input checked="" type="checkbox"/> mpa:MAP3821 hypothetical protein	890	176	46	0.316	155	-> 99
<input checked="" type="checkbox"/> bli:BLI00800 hypothetical protein	1292	176	46	0.331	163	-> 7
<input checked="" type="checkbox"/> bli:BL03072 hypothetical galactose-binding like	1094	176	46	0.331	163	-> 7
<input checked="" type="checkbox"/> bms:BRA0366 trbL protein	547	176	46	0.278	187	-> 17
<input checked="" type="checkbox"/> bca:BCE3739 hypothetical protein	1147	175	46	0.312	154	-> 18
<input checked="" type="checkbox"/> bja:bjl0521 hypothetical protein	745	175	46	0.355	155	-> 69
<input checked="" type="checkbox"/> bpa:BPP2368 putative membrane protein	561	175	46	0.331	124	-> 55
<input checked="" type="checkbox"/> milo:mil10460 hypothetical protein	388	174	45	0.341	170	-> 44
<input checked="" type="checkbox"/> Thra:Rh0304 hypothetical protein	717	173	45	0.305	174	-> 29

Figure 1.12.18 Search results for gene hsa:5921 using the “best” hits option.

Table 1.12.1 Description of the Arrows Listed Under the “best(all)” Column of the SSDB Search Results Page

Arrow	Definition
<->	The two genes are best-best hits
->	The listed gene is the best hit to the query gene
<-	The query gene is the best hit to the listed gene
No arrow	This gene scored higher than the threshold.

a. Draw alignment

A multiple sequence alignment of all checked genes will be displayed. The query sequence is displayed in green, the aligned subsequences are displayed in red, and the unaligned subsequences are displayed in blue. The GENES Entry for any of the listed genes may be viewed simply by clicking on the corresponding gene name.

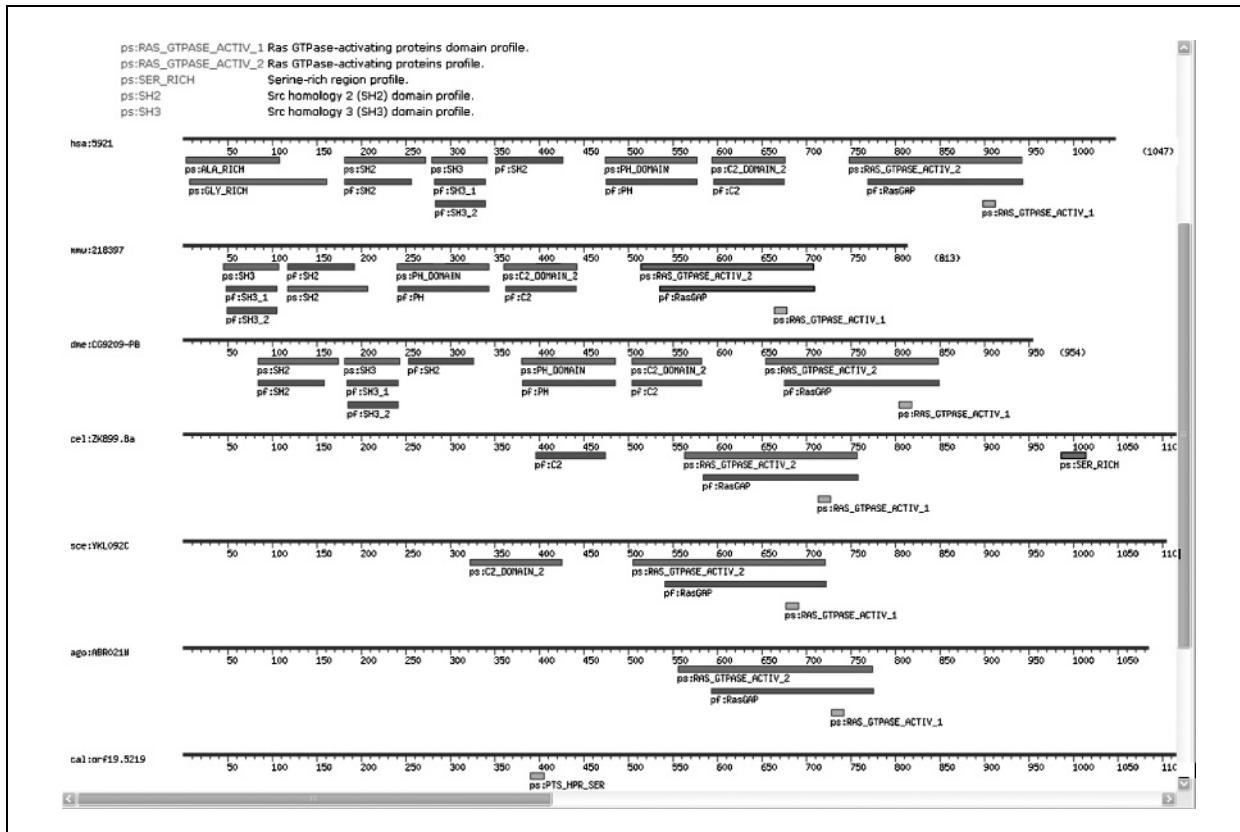


Figure 1.12.19 Common motifs of selected sequences from SSDB query. These are the motifs as stored in the GENES database for all amino acid sequences encoded by all genes in all species.

b. Search common motifs

The common motif sequences contained within the amino acid sequences encoded by the selected genes will be displayed, as in Figure 1.12.19. These are the motifs as stored in the GENES database for all amino acid sequences encoded by all genes in all species. This tool is the same as that available in SSDB, described in Alternate Protocol 6. The blue lines correspond to motifs from the Pfam database and the green lines correspond to those in the PROSITE database. Red lines correspond to those motifs common to all amino acid sequences encoded by the selected genes.

Note how the motifs found tend to correspond with the multiple sequence alignment, as described in the previous section. For example, the sequences of *mtu:Rv1067c*, *mtc:MT1096.1* and *mbo:Mb1096c* all align very well, as is also reflected in the motifs that are common to all three. On the other hand, *sco:SPBC646.12c* did not align well with most of these other sequences selected in this example, and, correspondingly, the motifs found in this sequence also partially align with the query as well as the sequences of *mmu:218397* and *sce:YKL092c*, which all contain the RasGAP motif. Thus, based on the motifs found, the sequence alignments in this example may actually be grouped into two sets, depending on whether or not they contain certain motifs.

c. List definitions

A listing of the selected genes and their definitions in the GENES database will be displayed. The individual GENES Entry, containing the amino acid sequence and other information, may be viewed by clicking on the corresponding gene's name. Further operations may be performed via a pull-down menu at the top of the page.

KEGG SSDB: SEARCHING SEQUENCE MOTIFS

This protocol briefly illustrates the functionality of KEGG that searches for common motifs among a group of sequences. There are actually three ways in which sequence motifs can be searched in the KEGG SSDB database. The main form for performing these searches is on the bottom half of the page at the URL <http://www.genome.jp/kegg/ssdb/>, as in Figure 1.12.20. Each search type is described in this section.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

Search motifs in a given sequence

This search method takes a gene name and displays all motifs found in its corresponding protein.

1. Go to <http://www.genome.jp/kegg/ssdb/>, scroll down to the Motif Search section, and enter hsa:5921 in the “Search motifs in a given sequence” text box. Click the Go button. The result is a textual list and graphical representation of the motifs in the given sequence, as displayed in Figure 1.12.21.

The upper half of this page is the textual results list, where the motifid begins with pf for a result from Pfam, tigrfams for a result from TIGR protein families, and ps for a result from PROSITE. The locations of the beginning and ending positions of the motifs are given in the From and To fields, respectively. Also listed are the definition, E-value, and significance scores of the resulting motifs. The corresponding graphical representation of this list is given on the lower part of the page, where the blue lines correspond to the motifs from Pfam, the cyan lines to TIGR, and the green lines to PROSITE. Detailed information for each motif can be viewed by clicking on the motif ID.

The screenshot shows the "Motif Search" section of the KEGG SSDB interface. It includes three main search options:

- Search motifs:** (enter keggid in the form of org:gene, e.g., eco:b0002) in Pfam PROSITE (pattern) PROSITE (profile)
- Search common motifs:** (enter multiple keggid's, e.g., eco:b0002 eco:b3940 eco:b4024) in Pfam PROSITE (pattern) PROSITE (profile)
- Search sequences with given motifs:** (enter one or more motif identifier, e.g., pf:DnaJ ps:DNAJ_2)

Below these, there is a "Search against:" section with radio buttons for "All species" and "KEGG species" (with a text input field for a three-letter code like "hsa") and a "Help" link. At the bottom are "Go" and "Clear" buttons.

Figure 1.12.20 Form for searching motifs in SSDB. Three methods are available: search motifs within a given sequence (indicated by a KEGG ID), search common motifs in given sequences (indicated by multiple KEGG IDs), and search sequences with given motifs (given by MOTIF identifiers).

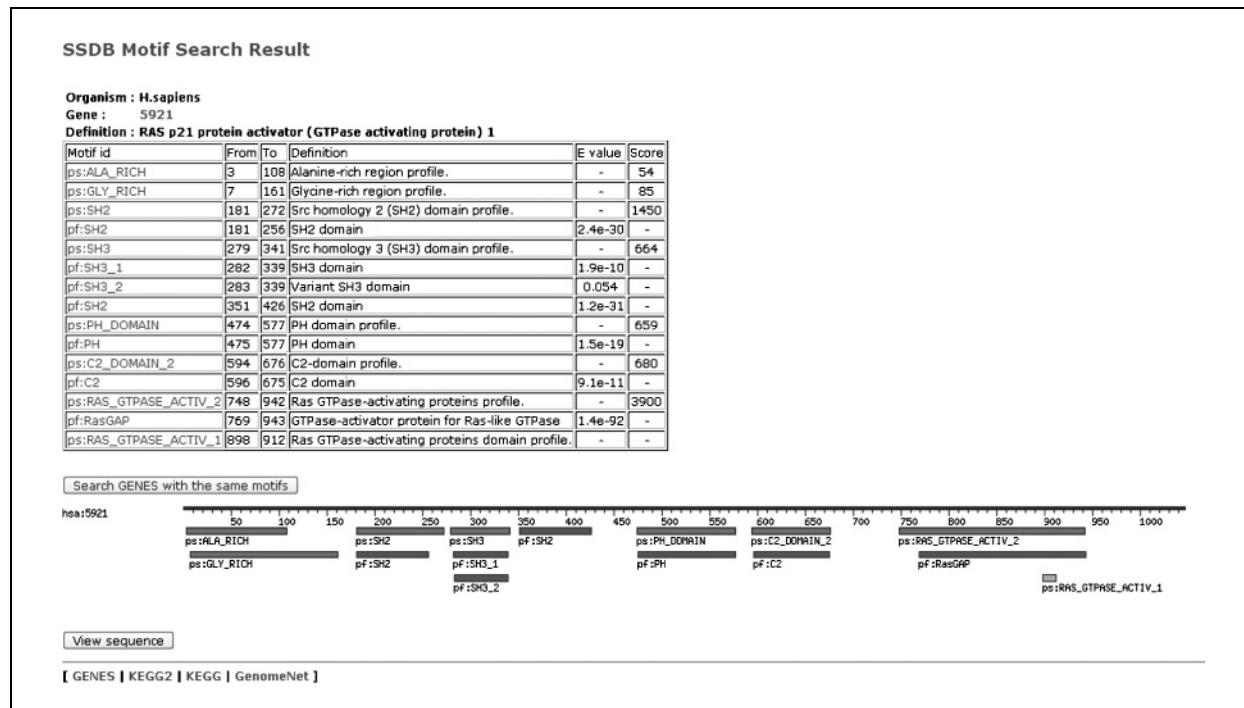


Figure 1.12.21 The results page for a search of motifs in the gene hsa:5921. The motif labeling system and other columns are discussed in Basic Protocol 8.

- Click on the “Search GENES with the same motifs” button to produce a Sequence Search Result window (Fig. 1.12.24), which displays the list of motifs found for the search.
- Alternatively, click the “View sequence” button, back on the Motif Search Results page (Fig. 1.12.21), to display a window where the actual motif sequences can be viewed in correspondence with the original amino acid sequence (Fig. 1.12.22).

Search common motifs in given sequences

This will find motifs shared by two or more named genes.

- This option provides searches for the given motifs in a given set of sequences. Enter a list of KEGG IDs in the given text field. In this example enter the example KEGG IDs of eco:b0002 eco:b3940 eco:b4024 in the text box.
- From among the check boxes below the text box, select the motif libraries within which to search, or leave unchanged to search all libraries.
- Click the Go button. A matrix as in Figure 1.12.23 will be displayed for the multiple KEGG IDs given as input. This matrix indicates which genes contain which motifs with an asterisk (*).

Search sequences with given motifs

This will look up sequences that contain one or more named motifs. From the SSDB home page (<http://www.genome.jp/kegg/ssdb/>), type in the IDs of one or more motifs in the “Search sequence with given motif” text box. A radio button selection allows one to search against all species or a single one using its three-letter abbreviation.

- The genes that contain any of a given set of multiple motif sequences will be displayed with this operation. For this search method, either all species or one particular species may be specified with the “Search against” radio buttons, as on the bottom of Figure 1.12.20. When specifying a species, the text field takes the three-letter

Using Biological Databases

1.12.23

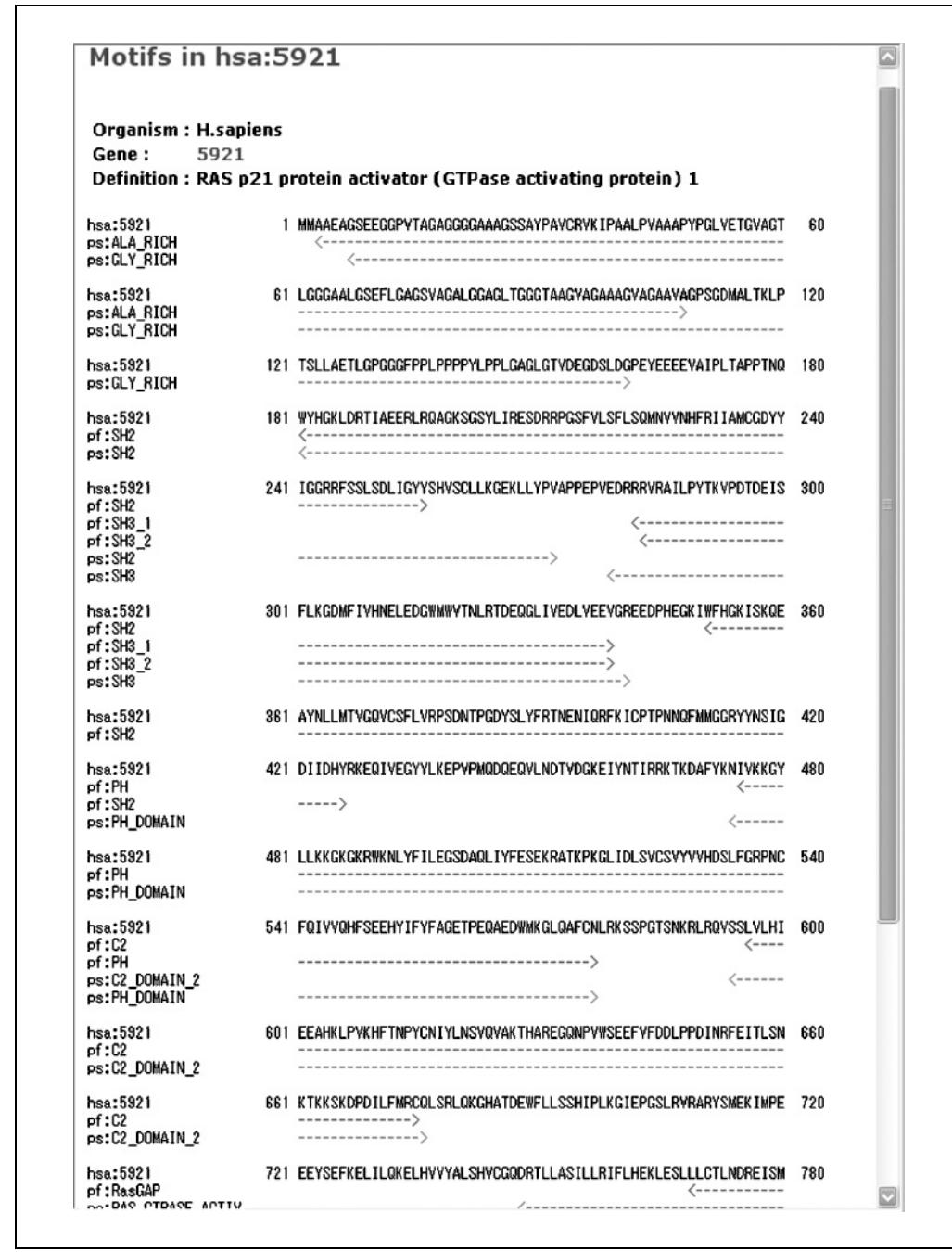


Figure 1.12.22 The motifs found in the given sequence.

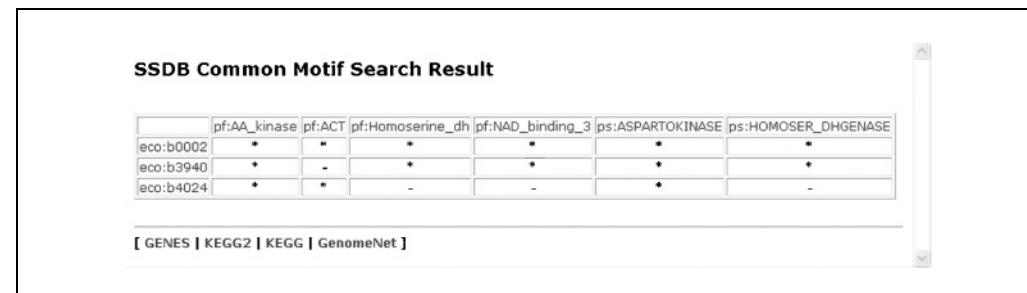


Figure 1.12.23 Result of searching for common motifs from among a group of genes in KEGG SSDB. This matrix indicates the actual motif sequences in correspondence with the original amino acid sequences.

SSDB Sequence Search Result

Organism : All species
Motifs : pf:AA_kinase Amino acid kinase family
 ps:ASPARTOKINASE Aspartokinase signature.
 tfam:asp_kin_monofn asp_kin_monofn: aspartate kinase, monofunctional class

Search Result : 178 hits

KEGG ID	Definition
<input checked="" type="checkbox"/> aae:aq_1152	aspartokinase
<input checked="" type="checkbox"/> aci:ACIAD1252	aspartate kinase
<input checked="" type="checkbox"/> ana:alr8644	aspartate kinase
<input checked="" type="checkbox"/> atc:AQR_L_1857	aspartate kinase
<input checked="" type="checkbox"/> ath:Atig31230	bifunctional aspartate kinase/homoserine dehydrogenase / AK-
<input checked="" type="checkbox"/> ath:Atg02020	aspartate kinase, lysine-sensitive, putative
<input checked="" type="checkbox"/> ath:At4g19710	bifunctional aspartate kinase/homoserine dehydrogenase, puta
<input checked="" type="checkbox"/> ath:At4g19710_1	bifunctional aspartate kinase/homoserine dehydrogenase, puta
<input checked="" type="checkbox"/> ath:At5g13280	aspartate kinase
<input checked="" type="checkbox"/> ath:At5g14060	aspartate kinase, lysine-sensitive
<input checked="" type="checkbox"/> atu:Atu4172	aspartate kinase, alpha and beta subunit
<input checked="" type="checkbox"/> baa:BA_2815	amino acid kinase family
<input checked="" type="checkbox"/> baa:BA_4408	amino acid kinase family

Figure 1.12.24 Search results for sequences with given motifs. The Motifs section lists the query motifs' IDs and their descriptions. The Search Result section lists the number of genes whose sequences contain one of the given motifs, and includes a listing of the resulting KEGG IDs along with their descriptions.

code for the organism as defined by KEGG. As an example, enter `pf:aakinase ps:ASPARTOKINASE tigrfams:asp_kin_monofn` as a query and select “All species.” Click the Go button.

8. The result is shown in Figure 1.12.24. The Motifs section lists the matching query motifs' IDs and their descriptions. The Search Result section lists the number of genes whose sequences contain one of the given motifs, along with a listing of the resulting KEGG IDs and their descriptions.
9. A “Select operation” pull-down menu is located above the “Search Result” section, where various operations may be performed on the genes that are checked in the list. These operations are “View motifs,” “List definitions,” and “Search common motifs.” The “View motifs” operation will display the sequences of the selected genes and the motifs within them. *Note that selecting more genes will take more time in order to process the request.* This is the same view as the graphical result of the “Search motifs in a given sequence” search method. The “List definitions” operation will list the KEGG IDs and corresponding definitions of the selected genes. This list provides links to the detailed sequence information for the listed genes. Finally, the “Search common motifs” operation is the same as the “Search common motifs in given sequences” search method.
10. On the original results page (Fig. 1.12.24), detailed information for a motif may be viewed by clicking on its ID.

KEGG SSDB: ORTHOLOG AND PARALOG CLUSTERS

A recent addition to KEGG is an automatic procedure based on a graph analytical method to computationally generate ortholog clusters (OCs) and paralog clusters (PCs) from the entire SSDB graph network, currently containing 200 million edges. The resulting ortholog clusters may be examined by clicking on the “OC search” link in the GENES Entry page. This section describes how to take advantage of this cluster information.

Necessary Resources*Hardware*

Computer with Internet access

Software

Web browser

1. Go to the GENES home page <http://www.genome.jp/kegg/genes.html> (this page is illustrated in Fig. 1.12.15). Scroll down to the Search KEGG OC (Ortholog Cluster) text entry box (near the bottom of the page) and enter hsa:5921. Click Go to display a page that links to a list of all ortholog clusters of this gene. Click on the link labeled “list” to display the results in tabular form (Fig. 1.12.25). This table provides links to GENES entries, to OC and PC matrix displays, and to outside data regarding each gene in the same OC as hsa:5921. An EPS graphics file of this OC may also be downloaded from this page.

From this list, one can see that this gene, which is a GTPase activating protein, clusters extremely well with other similar genes in C.elegans and D.melanogaster.

The screenshot shows a table titled "List of genes for KEGG OC" with the following details:

- OC number: 1332.2 (18 genes)
- Keyword: hsa:5921
- Link: >> Display whole view of the OC:1332.2 cluster or download EPS graphics for all the OC:1332.* clusters (69008 bytes)

Table columns include: GENES OC/PC viewer, KO, OT, COG, GO, GT, TC, and EC. The table lists 18 genes, many of which are orthologs of hsa:5921, such as K05767 (IQGAP), K04352 (RASGAP), and GO:0005099.

GENES OC/PC viewer	KO	OT	COG	GO	GT	TC	EC
[OC PC] ago:AER025C							
[OC PC] cal:orf19.5219							
[OC PC] hsa:10788	K05767 (IQGAP)						
[OC PC] hsa:128239	K05767 (IQGAP)						
[OC PC] hsa:4763	K04352 (RASGAP)			GO:0005099			
[OC PC] hsa:5921	K04352 (RASGAP)			GO:0005099			
[OC PC] hsa:8826	K05767 (IQGAP)						
[OC PC] hsa:9462							
[OC PC] mmu:19415							
[OC PC] mmu:29875	K05767 (IQGAP)						
[OC PC] mmu:320484							
[OC PC] rno:192117							
[OC PC] rno:192126							
[OC PC] rno:24592	K04352 (RASGAP)			GO:0005099			
[OC PC] rno:25676	K04352 (RASGAP)			GO:0005099			
[OC PC] sce:YBR140C							
[OC PC] sce:YKL092C							
[OC PC] sce:YOL081W							

Figure 1.12.25 List of orthologs in KEGG for the selected gene, hsa:5921. This table provides links to GENES entries, to OC and PC matrices, and to outside data regarding each gene in the same OC as hsa:5921. An EPS graphics file of this OC may also be downloaded here.

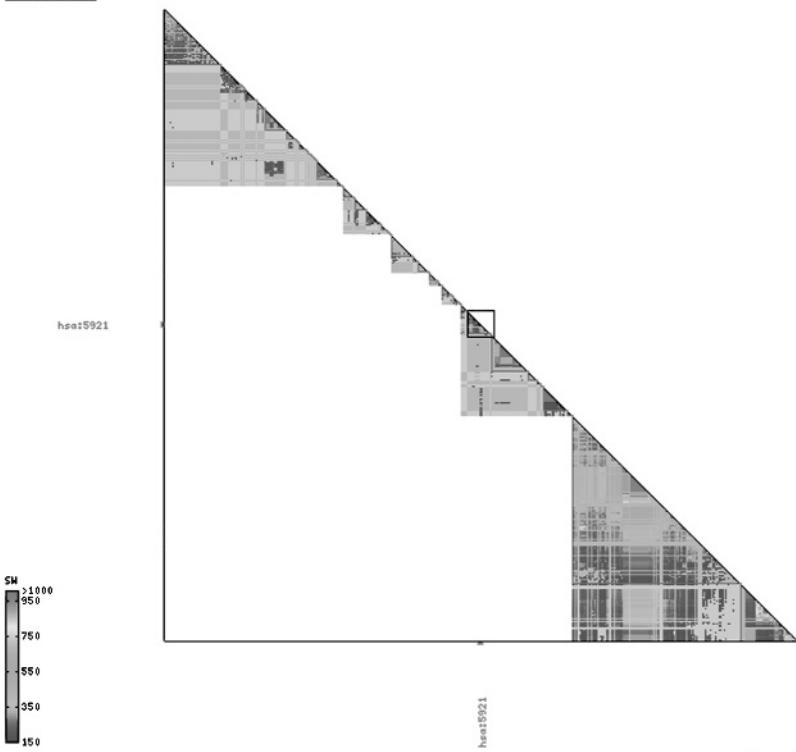
KEGG OC Viewer

Whole view

KEGG ID : hsa:5921

Definition : RASA1; RAS p21 protein activator (GTPase activating protein) 1

<UpperLeft



- Click a point near the diagonal to see a close-up view.

Display section : OC1325.1 - OC1333.2

Figure 1.12.26 Whole view of the ortholog genes to hsa:5921. The ortholog viewer provides a visual of the similarity between a selected gene and its orthologs as defined by KEGG. For the color version of this figure go to <http://www.currentprotocols.com>.

2. Click on the Display button to display two windows consisting of a graphical whole-view matrix of the ortholog clusters listed. This view may also be obtained by clicking on “OC viewer” in the gene’s GENES Entry page. For example, click on the OC link for hsa:5921 in the left-most column of the table illustrated in Figure 1.12.25. The close-up view will be displayed. Next, click on the “Whole view” link at the top of this view to display a detailed view of where the close-up can be found. Figure 1.12.26 shows an overview of the entire ortholog cluster, and Figure 1.12.27 is a close-up of the region containing the selected gene.
3. The ortholog viewer provides a visual depiction of the similarity between the selected gene and its orthologs as defined by KEGG. The selected gene is highlighted in red, and the coloring scheme is given in the legend at the bottom-left of either viewer. The more similar genes are red while the less similar genes are blue. Thus, it is easy to visualize the similarity relationships between hundreds of genes. From this figure, one can further see that this gene corresponds well with a Ras GTPase activating protein in *D. melanogaster*, confirming the similarity in function across these species. Entire sub-triangular regions represent highly similar clusters of genes.

Using Biological
Databases

1.12.27

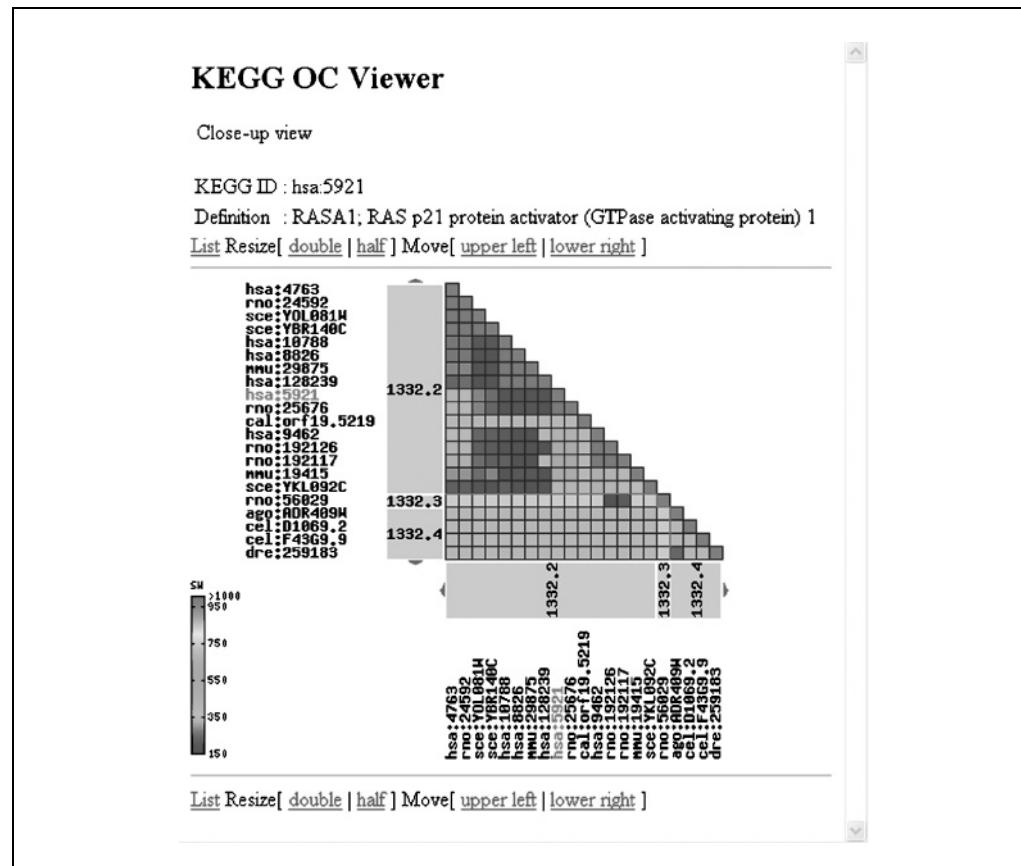


Figure 1.12.27 Close-up view of genes in a KEGG Ortholog Cluster, corresponding to the boxed portion of the whole view in Figure 1.12.26. For the color version of this figure go to <http://www.currentprotocols.com>.

BASIC PROTOCOL 10

KEGG SSDB: GENE SEARCH CLUSTER

SSDB also contains the positional correlations of genes on the chromosome. The GFIT table is a preprocessed table for each organism, containing the information about top-scoring genes (best-best hits or best hits) in other organisms together with the information about the order of genes on the chromosome. The gene cluster search involves a search against these GFIT tables (Bono et al., 1998). This is performed using the Gene Cluster Search tool available in the middle of the SSDB search page. This tool reconstructs a complete functional unit from a set of genes.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

1. Go to the SSDB home page <http://www.genome.jp/kegg/ssdb/>, scroll down to the Gene Cluster Search section, and enter a KEGG ID (for this example, `bsu:BG10898`) in the “Search conserved gene clusters” field.
 2. Modify the search criteria as desired. Select the radio button for “Table view.”

Using the KEGG Database Resource

1.12.28

SSDB Gene Cluster Search Result							
KEGG ID : bsu:BG10898 (511 a.a.) (GFT) Definition: phosphoglycerate mutase							
Include: <input checked="" type="checkbox"/> best hits		Gap size: <input type="text" value="3"/> <input type="button" value="Help"/>					
Sort by: <input checked="" type="radio"/> SW-score <input type="radio"/> KEGG-species							
Search against: <input checked="" type="checkbox"/> Bacteria <input checked="" type="checkbox"/> Archaea <input type="checkbox"/> Eukaryotes							
Threshold: <input type="text" value="100"/> <input type="button" value="Go"/>							
<input type="button" value="Help"/>							
bsu	BSU10898(3504)	BSU10898(3505)	BSU10897(3506)	BSU11062(3507)	BSU10827(3508)	BSU14085(3509)	
blu	BLU03661(3579)	BLU03662(3580)	BLU03663(3581)	BLU03664(3582)	BLU03665(3583)	BLU03666(3584)	
blu	BLU03468(3594)	BLU03467(3595)	BLU03465(3596)	BLU03465(3597)	BLU03464(3598)	BLU03463(3599)	
gka	GK3054(3167)	GK3055(3168)	GK3056(3169)	GK3057(3170)	GK3058(3171)	GK3059(3172)	
bca	BCE5239(5367)	BCE5239(5368)	BCE5240(5369)	BCE5241(5370)	BCE5242(5371)	BCE5243(5372)	
bce	BC5135(5123)	BC5136(5124)			BC5140(5127)	BC5141(5128)	
bcc	BTZK4824(4906)	BTZK4825(4907)	BTZK4826(4908)	BTZK4827(4909)	BTZK4828(4910)	BTZK4829(4911)	
blk	BT9727_4815(4902)	BT9727_4815(4903)	BT9727_4816(4904)	BT9727_4817(4905)	BT9727_4818(4906)	BT9727_4819(4907)	
bat	BAS4985(5056)	BAS4986(5057)	BAS4987(5058)	BAS4988(5059)	BAS4989(5060)	BAS4990(5061)	
baa	BA_0223(223)	BA_0224(224)	BA_0225(225)	BA_0226(226)	BA_0227(227)	BA_0228(228)	
bar	GBAA5363(5273)	GBAA5365(5274)	GBAA5366(5275)	GBAA5367(5276)	GBAA5369(5277)	GBAA5370(5278)	
ban	BA5364(5272)	BA5365(5273)	BA5366(5274)	BA5367(5275)	BA5369(5276)	BA5370(5277)	
bha	BH3556(3656)	BH3557(3657)	BH3558(3658)	BH3559(3659)	BH3560(3660)	BH3561(3661)	
bc	ABC3018(3098)	ABC3019(3099)	ABC3020(3100)	ABC3021(3102)	ABC3022(3103)		
oih	OB2434(2525)	OB2435(2526)	OB2436(2527)	OB2437(2528)	OB2438(2529)	OB2439(2530)	
lin	lin2549(2618)	lin2550(2619)	lin2551(2620)	lin2552(2621)	lin2553(2622)	lin2554(2623)	
lmf	LMOF2365_2426(2506)	LMOF2365_2429(2507)	LMOF2365_2430(2508)	LMOF2365_2431(2509)	LMOF2365_2432(2510)	LMOF2365_2433(2511)	
lmo	lmo2455(2530)	lmo2456(2531)	lmo2457(2532)	lmo2458(2533)	lmo2459(2534)	lmo2460(2535)	
ser	SERPO446(4468)	SERPO445(4467)	SERPO444(4466)	SERPO443(4465)		SERPO441(4463)	
sep	SE0561(554)	SE0560(563)	SE0559(562)	SE0558(561)		SE0556(559)	
sau	SAU0731(769)	SAU0730(768)	SAU0729(767)	SAU0728(766)		SAU0726(764)	
sav	SAV0776(793)	SAV0775(792)	SAV0774(791)	SAV0773(790)		SAV0771(788)	
sas	SAS0742(773)	SAS0741(772)	SAS0740(771)	SAS0739(770)		SAS0737(768)	
sam	MW0738(750)	MW0737(749)	MW0736(748)	MW0735(747)		MW0733(745)	
sac	SACOL0842(833)	SACOL0841(832)	SACOL0840(831)	SACOL0839(830)		SACOL0837(828)	
sar	SAR0832(838)	SAR0831(837)	SAR0830(836)	SAR0829(835)		SAR0827(833)	
cpe	CPE1299(1343)	CPE1301(1345)	CPE1302(1346)	CPE1303(1347)	CPE1304(1348)	CPE1305(1349)	
cac	CAC0713(756)	CAC0712(755)	CAC0711(754)	CAC0710(753)	CAC0709(752)	CAC0708(751)	
ctc	CTC00382(339)	CTC00381(338)	CTC00380(337)	CTC00379(336)	CTC00378(335)	CTC00377(334)	
sth	STH149(266)	STH148(265)	STH147(264)	STH146(263)	STH145(262)	STH144(261)	
poy	PAM284(300)	PAM285(301)	PAM170(176)	PAM174(180)	PAM175(181)		
mpe	MPYE3750(399)	MPYE3740(398)	MPYE3730(397)	MPYE3616(847)	MPYE3170(648)		
mf1	MF1468(484)	MF1502(520)	MF1504(522)	MF1577(606)	MF1578(607)		

Figure 1.12.28 Table view of Gene Cluster Search results for bsu:BG10898. The query gene is highlighted in the red cell. For the color version of this figure go to <http://www.currentprotocols.com>.

- Click the Exec button. The “Table view” option will display a window containing a table as in Figure 1.12.28. Note that the query gene is highlighted in red and that the genes from the same organism will correspond to those that appear next to this gene in the KEGG Ortholog Table (see Fig. 1.12.14).
- Click on the Graphics link at the bottom of the page illustrated in Figure 1.12.28 to see the corresponding Graphics View of this table. This view uses the coloring scheme employed by KEGG Orthology and illustrates the direction and relative position on the chromosome of other organisms where similar genes may be found in relation to the query gene. Figure 1.12.29 illustrates this view.

The Graphics View can also be immediately retrieved by selecting the Graphics View option in the original query.

- Modify the table or graphic views by modifying the options for the Smith-Waterman similarity calculation at the top of the results page (i.e., the radio buttons SW-score and KEGG-species; see Fig. 1.12.28 and Fig. 1.12.29) and clicking the Go button.

The Gene Cluster search results are originally sorted according to the SW-score, where the most similar ortholog genes are listed first. The KEGG-species option will list the species according to the ordering the KEGG Table of Contents.

*Note that the Ortholog Table of Figure 1.12.14 corresponds to the pathway map for Glycolysis. It is thus possible to compare these Gene Cluster results to the pathway map by coloring it with the genes in *Bacillus subtilis* in this pathway (by selecting this organism in the pull-down menu at the top of the pathway).*

Using Biological Databases

1.12.29

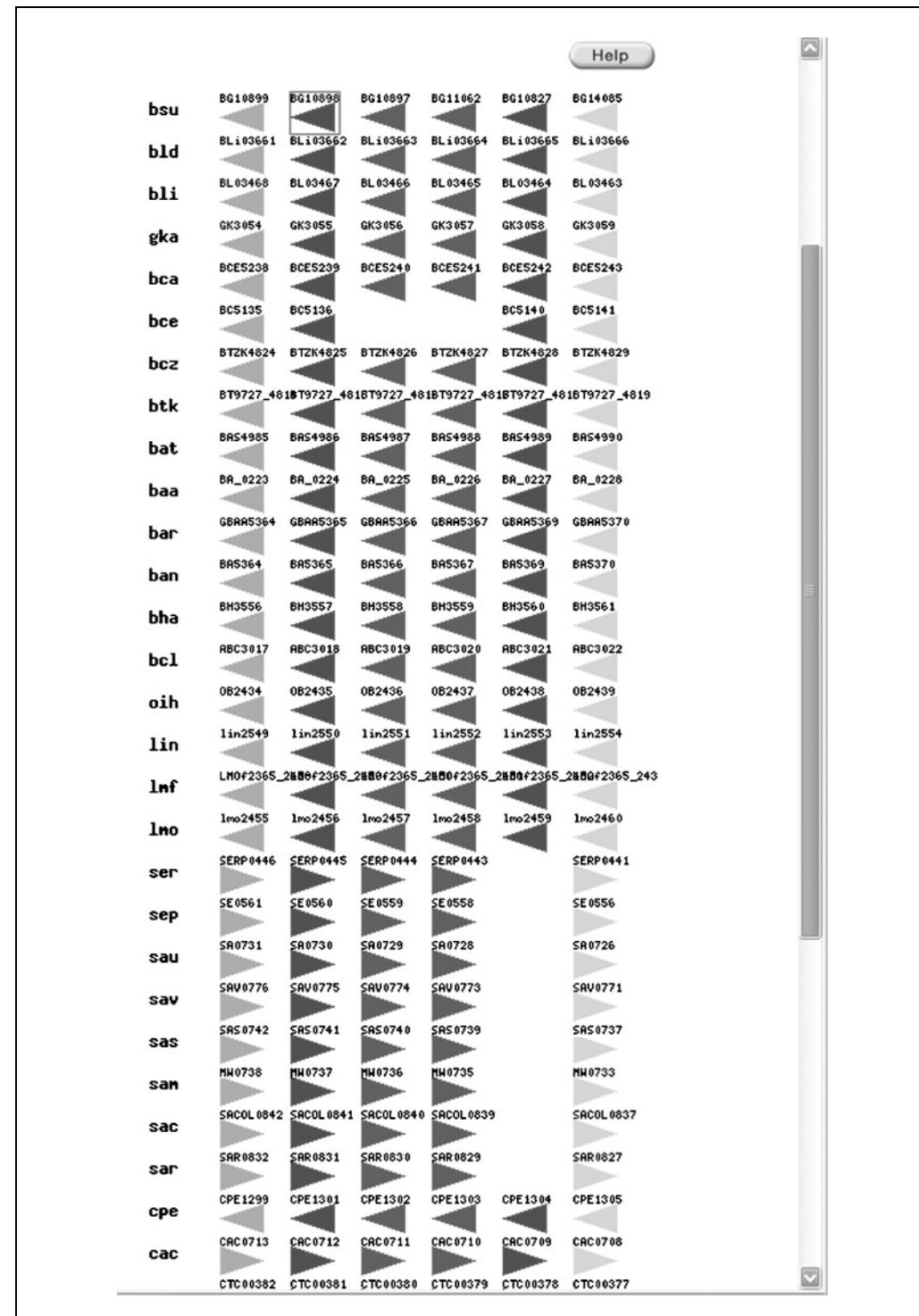


Figure 1.12.29 Graphics view of the Gene Cluster results corresponding to Figure 1.12.28.

BASIC PROTOCOL 11

Using the KEGG Database Resource

1.12.30

THE KEGG EXPRESSION DATABASE: GETTING STARTED

KEGG EXPRESSION is a database system for integrated analysis of gene expression profile data. It can be used together with KEGG pathway data and genome sequence data. It currently contains microarray gene expression profiles for *Synechocystis* PCC6803, *Bacillus subtilis*, and *Escherichia coli*, obtained by the Japanese research community. It may also be used to analyze expression data from an individual investigator's own laboratory.

Readers should be aware that KegArray, which is installed on a local computer, is an alternative program for analyzing microarray data. This is available for download at the KEGG EXPRESSION home page (<http://www.genome.jp/kegg/expression/>). In addition to the functionality provided by KEGG EXPRESSION, KegArray contains other tools, most notably those for handling systematic bias (Dudoit et al., 2000). Because of overlapping functionality with KEGG EXPRESSION, KegArray is not described further here.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser, or KegArray application for Windows XP or Mac OS X,
downloadable from <http://www.genome.jp/download/>

Files

Text file of microarray expression data, available from the KEGG EXPRESSION Web site at <http://www.genome.jp/kegg/expression/> (optional)

The screenshot shows the KEGG Expression Browser interface. At the top, there is a logo for "KEGG" and the title "KEGG Expression Browser". Below this is a detailed table of gene expression data:

ENTRY	ex0000260
ACCESSION	0000260
DEFINITION	B.subtilis DegU regulon, degU overexpression
CONTROL	strain: TT291/pDG148-degU trpC2 leuC7 DdegS aprE'-lacZ (Cmr)/pDG148-degU medium: DSM temperature: 37 deg C, harvest point: OD600=1.25 0.5 h after the end of logarithmic phase
TARGET	strain: TT291/pDG148-degU trpC2 leuC7 DdegS aprE'-lacZ (Cmr)/pDG148-degU medium: DSM+IMM IPTG temperature: 37 deg C, harvest point: OD600=1.12 0.5 h after the end of logarithmic phase
CONTACT	Yasutaro FUJITA
REFERENCE	Ogura M, Yamaguchi H, Yoshida Ki, Fujita Y, Tanaka T
AUTHOR	DNA microarray analysis of <i>Bacillus subtilis</i> DegU, ComA and PhoP regulons: an approach to comprehensive analysis of <i>B. subtilis</i> two-component regulatory systems.
TITLE	Nucleic Acids Res. 2001 Sep 15;29(18):3804-13
JOURNAL	11557812 [PubMed - in process]
PMID	11557812 [PubMed - in process]
FTP	ftp://ftp.genome.ad.jp/pub/kegg/expression/bsu/ex0000260.dat
DATA	ex0000260.dat
DATE	2000-09-27 00:00:00
ORGANISM	<i>B. subtilis</i>

Below the table are several input fields and checkboxes:

- Intensity threshold: [Examine distribution](#)
- Ratio threshold:
- Array image
- Scatter plot (Select plot type:

At the bottom, there is a button labeled "When ready, initiate the query by Exec:" followed by two buttons: "Exec" and "Reset".

At the very bottom of the window, there is a footer bar with links: [KEGG Home | GenomeNet Home | DBGET Links Diagram]

Figure 1.12.30 Expression entry ex0000260 for *Bacillus subtilis* (Ogura et al., 2001).

Using Biological Databases

1.12.31

1. Go to the KEGG EXPRESSION page via the EXPRESSION link on the main KEGG Table of Contents (accessed as in Basic Protocol 1) or go directly to the URL <http://www.genome.jp/kegg/expression/>. Click on “List of experimental data available” to view the list of all currently available expression datasets.
2. Click on “Ogura et al. (2001)” under the “Bacillus subtilis” heading. Three sets of expression data will be displayed.
3. Click on “ex0000260” to display the window shown in Figure 1.12.30. Each entry contains a description of the data, a link to the expression data, and a launcher for the Java-based applet for data analysis.

As displayed in the figure, the ENTRY and ACCESSION numbers are KEGG database IDs. These are followed by a brief description of the experiment under DEFINITION. CONTROL and TARGET describe the control and target biomaterials hybridized on the microarray. CONTACT is the name of the experimenter. If this experiment has been published, information regarding its publication will be listed under REFERENCE, including the author(s), title, journal name, and PMID. The FTP field provides the link to the actual data download site address. DATA is the filename for the data within KEGG. DATE is either the date of the experiment or the date on which these data were entered into the database. ORGANISM is the name of the organism corresponding to this data.

**BASIC
PROTOCOL 12**

KEGG EXPRESSION: USING THE JAVA APPLETS TO DISPLAY EXPRESSION DATA

The EXPRESSION entry page provides two Java applets for data analysis. The Array image applet reconstructs a symmetrical microarray image based on the control expression data and gene spot information. The Scatter plot applet plots the control expression data on the x axis and target expression data on the y axis. The Scatter plot applet allows one to plot all genes, or a subset based on categories.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

1. Continuing from step 3 of Basic Protocol 11, at the bottom of the page illustrated in Figure 1.12.30 is a form for executing a Java applet for data analysis. Two parameters are available: “Intensity threshold,” the minimum expression value which will be allowed for analysis, and “Ratio threshold,” the minimum value for the ratio between control and target. Only expression values that fall within the range specified by the ratio threshold for control/target or target/control will be allowed in the analysis. To aid in the selection of the appropriate intensity threshold, the “Examine distribution” link will display a histogram of intensity values for both the control and experimental data.
2. Check the applets (“Array image” and/or “Scatter plot”) to run. Both can be viewed simultaneously. Click the Exec button. This will launch new windows for the selected applets.

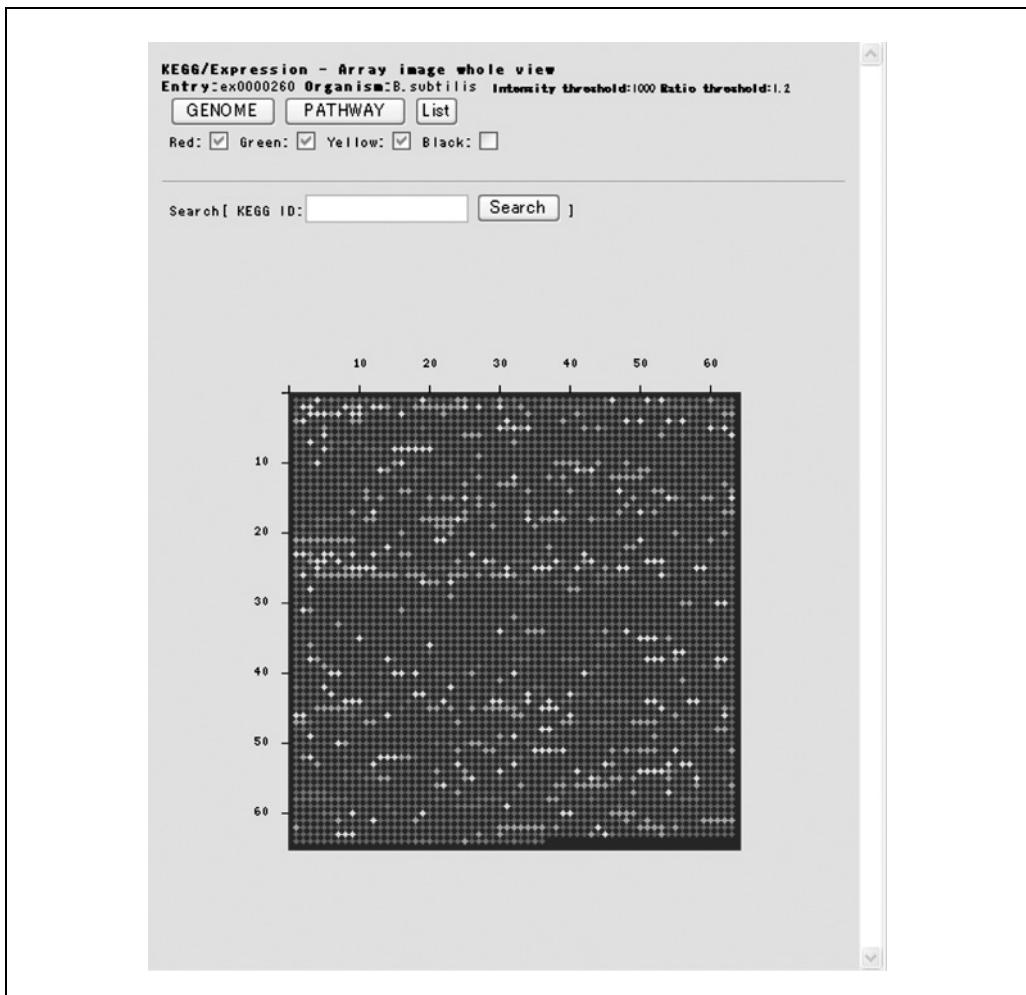


Figure 1.12.31 Whole array image. For the color version of this figure go to <http://www.currentprotocols.com>.

3. View the array image applet (Fig. 1.12.31):

- The initial array image represents the whole array. The black color represents data where both the control and target expression values are below the threshold value. Green spots represent data where the ratio of the control value to the target value is at least the value of the ratio threshold. Red spots indicate that the ratio of the target value to the control value is at least the value of the ratio threshold. Yellow spots represent data where neither the control-to-target nor the target-to-control ratio values exceed the ratio threshold. In a typical experiment, the green spots are where the target expression values are down-regulated, the red is where they are up-regulated, yellow is where there is no significant change, and black is where there are either no significant data to analyze or no valid experimental data to be retrieved.
- Click the mouse on a spot in the whole-array image to see a close-up view of that area of the image displayed in a new window. In the close-up view, the location and name of the gene/ORF at that spot will be displayed when the mouse is placed over it. Clicking on a spot will display the GENES Entry corresponding to the gene at that spot.

Using Biological
Databases

1.12.33

Specific genes may be searched for on the whole image view by entering the KEGG ID in the text field at the top and clicking the Search button. The corresponding close-up view will be updated to zoom in on the location on the array where the query gene is found. It will also have the query spot outlined in blue.

- c. There are three buttons at the top of the whole image view, below which are Red, Green, Yellow, and Black check boxes. Click on the GENOME button to display the Genome Map (see Basic Protocol 6), where the locations of the genes on the array may be viewed on the genome. The colors on the genome map correspond to the colors on the microarray, and those not colored are drawn in white.
 - d. The PATHWAY button will display all available pathways where the genes on the microarray may be found. Click on one of these pathways and the pathway map will display the corresponding genes colored according to the color on the microarray.
 - e. Finally, click the List button to generate a list of the genes whose colors are checked.
4. View the scatter plot applet (Fig. 1.12.32). The scatter plot shows the intensity values of each probe under experimental and control conditions of all genes in the microarray dataset. Click on the scatter plot image to display a new window with a close-up view of that area of the scatter plot. Individual spot information can be displayed moving over the spot, and the corresponding GENES Entry can be displayed by clicking on the spot.

The scatter plot plots the log values of the control data on the x axis and the log values of the target data on the y axis. The further away a spot is from the diagonal, the larger the difference in expression of the gene at that spot. The coloring of the scatter plot is the same as that for the array image.

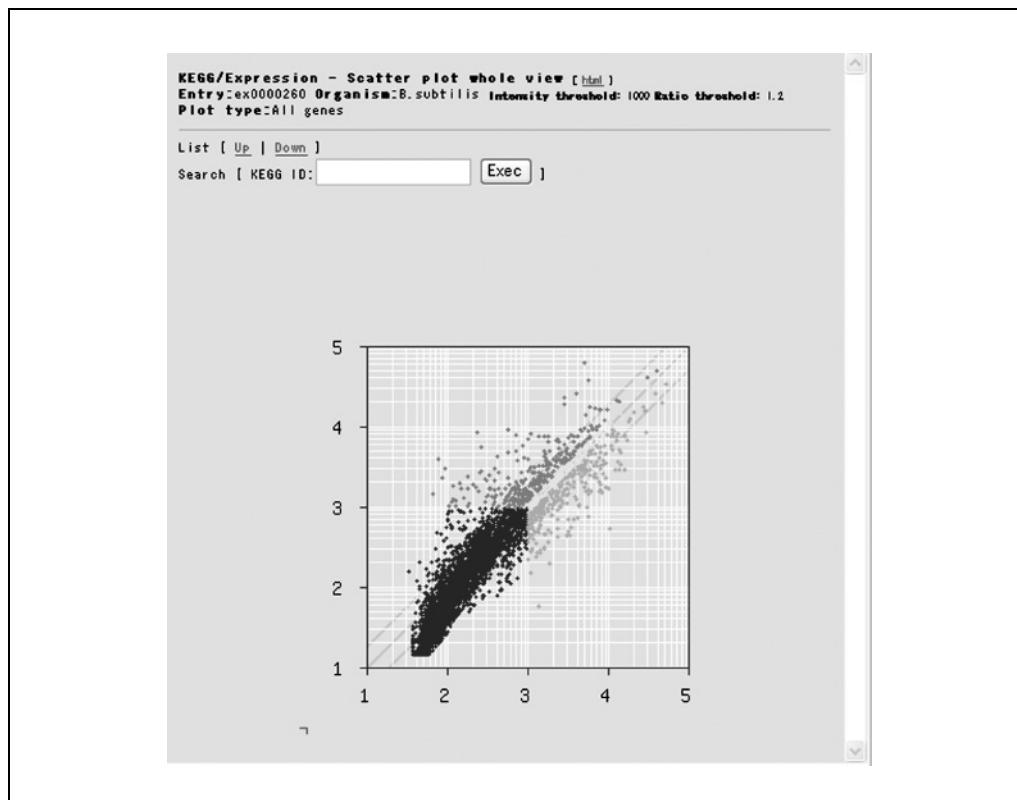


Figure 1.12.32 Scatter plot of all genes, with the logged values of the control data on the x axis and the logged values of the target data on the y axis. For the color version of this figure go to <http://www.currentprotocols.com>.

KEGG EXPRESSION: CLASSIFYING EXPRESSION DATA SETS BY KEGG NETWORK CATEGORY

BASIC PROTOCOL 13

This protocol describes how the expression data can be analyzed based on genes in a particular KEGG network category. When a scatter plot is executed on the main page with the “Classified—based on KEGG” option specified, the scatter plot whole view will be displayed plotted with only those genes in a particular KEGG network category, such as Carbohydrate Metabolism. In this way, it is easy to analyze the expression pattern of genes based on a category. Different categories may be selected using the Class pull-down menu at the top. A particular gene may be queried using the Search text field.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

- Starting from the view in Figure 1.12.30, check the “Scatter plot” option, select the “Classified—based on KEGG” option in the pull-down menu to the right of this, and click the Exec button. A page resembling Figure 1.12.32 will appear. Enter a KEGG ID (e.g., `bsu:BG10898`) in the displayed KEGG ID text entry box and click the neighboring Exec button. This displays a new Result field with a pull-down menu

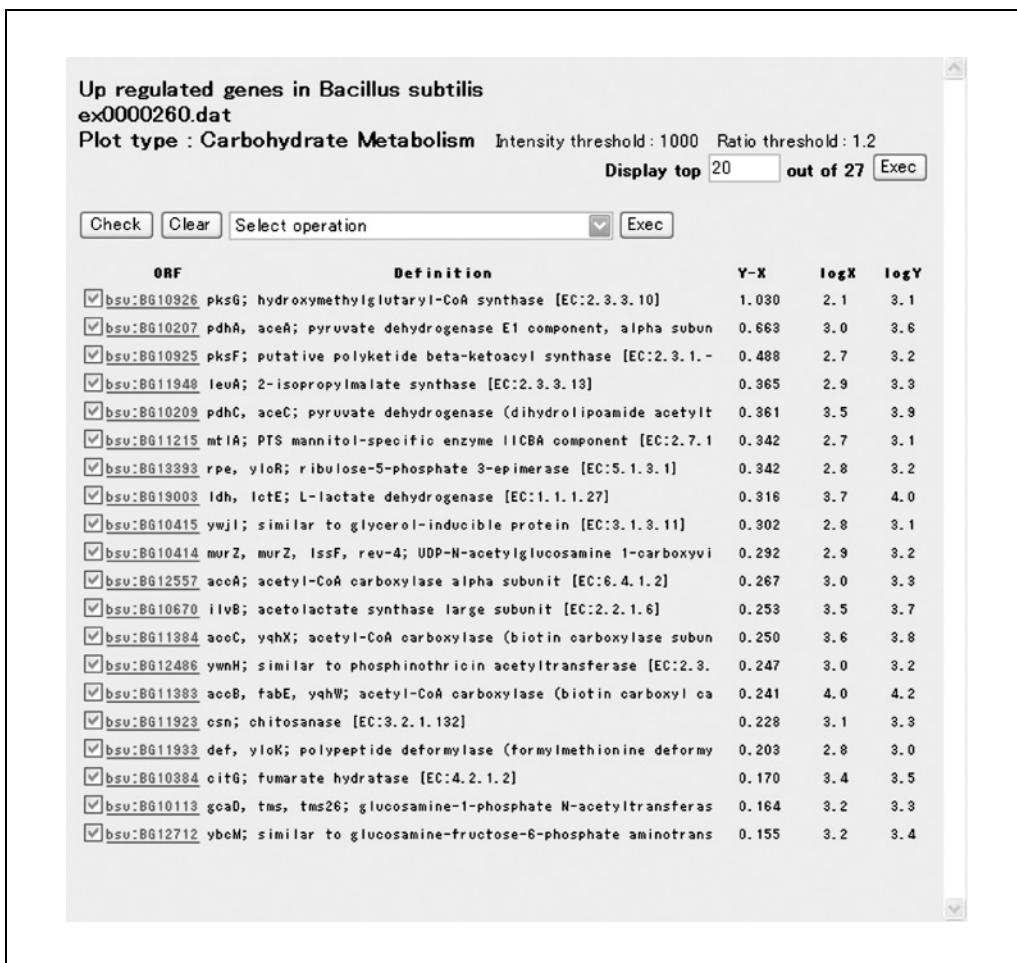


Figure 1.12.33 Depiction of 26 up-regulated genes in Carbohydrate Metabolism scatter plot.

Using Biological Databases

1.12.35

containing all the categories in which the queried gene may be found and the number of such genes found in parentheses. In addition, the close-up view and whole view will have the queried gene highlighted in purple.

2. An “html” link is located at the top left of the whole scatter plot view. Click on this to display a new window with all the scatter plots displayed by category. This window can thus be used to save or print in a convenient manner.
3. There are also two links near the top of the scatter plot view called Up and Down. The Up link will display a window of all genes that are up-regulated in relation to the control, listed in order of decreasing up-regulation. Correspondingly, the Down link will display a window of all genes that are down-regulated compared to the control. Click Up as an example, and a listing resembling that in Figure 1.12.33 will be displayed. In this example, it can be seen that there are a total of 26 genes that were up-regulated in Carbohydrate Metabolism. More or fewer genes may be displayed by modifying the value (20 in the example) and clicking the Exec button. In the list, the logX column lists the logged value of the control, the logY values are the logged target values, and Y-X is the value when logX is subtracted from logY. The list is sorted by this decreasing log-ratio value. Clicking on any of the ORF links will display the GENES Entry corresponding to the ORF. Finally, the check boxes next to each ORF can be used to select genes to examine using the tools under “Select operation.” Thus genes may be easily examined in the genome or pathway maps and sequences may be retrieved.

KEGG EXPRESSION: CLUSTER ANALYSIS

Multiple arrays are often analyzed simultaneously, as in time-series experiments, systematic knockout studies, and overexpression studies. In such analyses, clustering is one of the most popular tools. With clustering, it is possible to see which groups of genes tend to regulate together in a time series, for example, or even to see which genes tend to regulate oppositely. The KEGG clustering tool is described in the protocol below.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

1. From the KEGG EXPRESSION homepage <http://www.genome.jp/kegg/expression/>, click on the link to “List of experimental data available.” Click on the “Cluster analysis” link next to the “Bacillus subtilis” section under the list of all experimental data. This will display a listing resembling that in Figure 1.12.34.

The “prefix” indicates the code for this species. The “clustering type” is the type of hierarchical clustering to perform; currently single (minimum distance) and complete (maximum distance) linkage clusterings are available. The “distance definition” is the definition of distance between two expression profiles. “1.0-correlation” is calculated as 1 minus the Pearson correlation coefficient. In this case, the smaller the correlation coefficient (i.e., the less similar the expression profiles), the larger the distance, in order to obtain true coexpression clusters. “1.0+correlation” is 1 plus the correlation coefficient. The larger the coefficient, the larger the distance, in order to obtain clusters of opposing coexpression. The “threshold” is the same as for the array image and scatter plot, where only values that exceed this threshold value are used in the cluster analysis. Those arrays that are checked will be used to cluster based on the distance definition setting.

Bacillus subtilis

EXEC

prefix: bsu
 clustering type: single complete
 distance definition: 1.0-correlation 1.0+correlation
 threshold:

Yoshida et al. (2001), Bacillus subtilis glucose repression
 ex0000258 B.subtilis glucose repression
 ex0000259 B.subtilis CcpA-independent glucose repression

Ogura et al. (2001), Bacillus subtilis DegU, ComA and PhoP regulons
 ex0000260 B.subtilis DegU regulon, degU overexpression
 ex0000261 B.subtilis ComA regulon, comA overexpression
 ex0000262 B.subtilis PhoP regulon, phoP overexpression

Ogura et al. (2002), Bacillus subtilis ComK regulon
 ex0000659 B.subtilis ComK regulon, comK disruption, experiment1
 ex0000660 B.subtilis ComK regulon, comK disruption, experiment2
 ex0000661 B.subtilis ComK regulon, comK disruption, experiment3

Kobayashi et al. (2001), Bacillus subtilis two-component systems
 ex0000263 B.subtilis CitT regulon, citT overexpression
 ex0000264 B.subtilis DesR regulon, desR overexpression
 ex0000265 B.subtilis LytT regulon, lytT overexpression
 ex0000266 B.subtilis YbdJ regulon, ybdJ overexpression
 ex0000267 B.subtilis YcbB regulon, ycbB overexpression
 ex0000268 B.subtilis YcbL regulon, ycbL overexpression
 ex0000269 B.subtilis YccH regulon, yccH overexpression
 ex0000270 B.subtilis YclJ regulon, yclJ overexpression
 ex0000271 B.subtilis YdbG regulon, ydbG overexpression
 ex0000272 B.subtilis YdfI regulon, ydfI overexpression
 ex0000273 B.subtilis YesN regulon, yesN overexpression
 ex0000274 B.subtilis YfiK regulon, yfiK overexpression
 ex0000275 B.subtilis YhcZ regulon, yhcZ overexpression
 ex0000276 B.subtilis YkoG regulon, ykoG overexpression
 ex0000277 B.subtilis YrkP regulon, yrkP overexpression
 ex0000278 B.subtilis YtsA regulon, ytsA overexpression
 ex0000279 B.subtilis YufM regulon, yufM overexpression
 ex0000280 B.subtilis YvcP regulon, yvcP overexpression

Figure 1.12.34 Cluster analysis options for *Bacillus subtilis*.

2. Execute the cluster analysis with the default settings, using single-linkage clustering, defining distance as 1.0-correlation, and setting the threshold at 30000. Click the EXEC button at the top to see the cluster diagram as illustrated in Figure 1.12.35. The genes are aligned vertically, with the genes with the smallest distances connected more closely. Distance is measured in the horizontal direction.
3. Place the mouse cursor over any of the nodes (the black dots) of the diagram. The distance value calculated at that point will be displayed both as a balloon on the mouse cursor as well as in the upper-right area of the window.
4. Place the mouse cursor above any of the gene names to color the name in red and display it in the upper-right area.

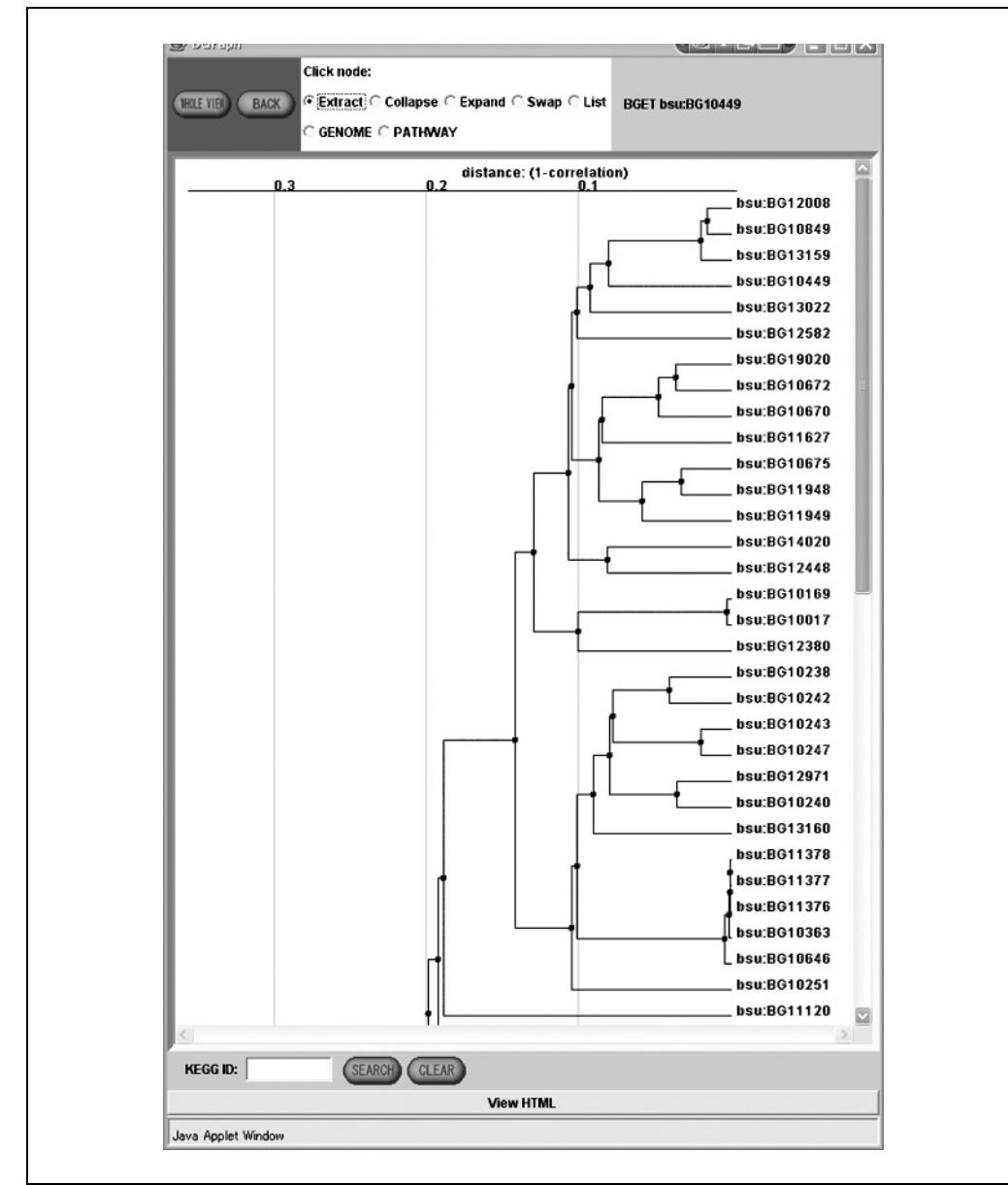


Figure 1.12.35 Dendrogram for cluster analysis of selected microarray data. The genes are aligned vertically, with the genes with the smallest distances connected more closely. Distance is measured in the horizontal direction.

5. Click on a name to display its GENES Entry. Also, just as in the other applets described earlier, a gene may be searched by entering its KEGG ID in the text field at the bottom of the window and clicking Search. Note that the ID as displayed in the window, beginning with “bsu,” for example, should be entered. The query gene will be displayed in blue, and the window will scroll such that it will be displayed in the center.
6. Finally, click the View HTML button at the bottom to produce an HTML view of the displayed dendrogram (including coloring). In this way, it is convenient to print and save the clustering to disk.
7. At the top of the window, a toolbar contains various radio buttons for manipulating the diagram. The Extract radio button will display the subtree below a selected node in a new window. The Collapse button will combine a subtree into one node; the Expand button will redisplay a collapsed subtree; the Swap button will exchange the

two subtrees under a node with one another; the List button will display the genes in a subtree in a list with links to other database resources corresponding to the selected genes; the GENOME option will display the genome map with the genes in the selected subtree highlighted; and PATHWAY will display all pathways on which the genes in the selected subtree exist.

- Click on one of these resulting pathways to display the pathway map with the selected genes highlighted. Furthermore, the WHOLE VIEW button at the top left will open a new dendrogram window, such that multiple views of the same dendrogram may be examined at once. Finally, the BACK button performs the “undo” function such that the last action taken may be undone.

KEGG EXPRESSION: UPLOADING PERSONAL DATA FOR ANALYSIS

This protocol covers a very useful feature of KEGG EXPRESSION where the user’s own microarray data may be analyzed and compared with KEGG’s pathway, genomic, and ortholog information.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

Files

Expression data in format specified by KEGG (see http://www.genome.jp/kegg-bin/Expression_Upload/help for format details)

The screenshot shows a web interface for uploading microarray data. At the top, there's a logo for KEGG and the title "Microarray expression profiles". Below that, it says "Your data: array7.txt (organism: E.coli)". There are input fields for "Intensity threshold" (set to 5000) and "Ratio threshold" (set to 1.2). Under these, there are two checked checkboxes: "Array image" and "Scatter plot". A dropdown menu next to "Scatter plot" is set to "All genes". Below these controls, there's a section titled "When ready, initiate the query by Exec:" with two buttons: "Exec" and "Reset". At the bottom, there are links for "GenomeNet", "KEGG", and "DBGET".

Figure 1.12.36 Uploaded data page where microarray data analysis may be performed on the user’s own data. Using KEGG to analyze such array data, it is very straightforward to analyze resulting clusters on KEGG’s pathways and compare clusterings with KO.

Using Biological
Databases

1.12.39

1. Access the KEGG EXPRESSION Web page as described in Basic Protocol 11. Click on “Upload and analyze personal data.” This will bring up a page for entering the name of a file on the user’s local computer to analyze.
A “help” link to the right of the data entry field (pointing to http://www.genome.jp/kegg-bin/Expression_Upload/help) explains the file format.
2. Type the filename or navigate to the file by using the Browse button. If the first line of the text file does not contain a line specifying the organism, select the organism from the pull-down menu under the text field.
3. Click the Exec button. This will bring up a page similar to Figure 1.12.36. From here, the same data analyses may be performed as explained in Basic Protocols 12 to 14 for existing data. By using KEGG to analyze such array data, it is very straightforward to analyze resulting clusters on KEGG’s pathways and to compare clusterings with KO.

KEGG LIGAND: THE COMPOUND DATABASE

KEGG LIGAND consists of four databases: COMPOUND, GLYCAN, REACTION, and ENZYME (Goto et al., 2002). COMPOUND is a database of chemical structures of most known metabolic compounds and some pharmaceutical and environmental compounds; GLYCAN is a database of carbohydrate structures; REACTION is a database of reaction formulas for enzymic reactions; and ENZYME is a database of enzyme nomenclatures. Both the COMPOUND and GLYCAN databases store their structures in a text-based format called KEGG Chemical Function, or KCF. This format for representing molecular structures is defined in Hattori et al. (2003).

All chemical structures in the KEGG COMPOUND database are manually entered, computationally verified and continuously updated. It contains over 12,800 entries, each identified by the prefix “C.” The COMPOUND database is searched via DBGET, which is a text-based keyword search of the annotation information of the compounds in LIGAND and is similar to the DBGET family of GENES and MOTIFS.

KEGG LIGAND also provides a tool for generating possible pathways between two compounds. All of the LIGAND databases, as well as this tool, will be described in this protocol.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

Files

Compound structures in MOL file format (optional)

1. Access the KEGG Table of Contents as described in Basic Protocol 1 (or go directly to <http://www.genome.jp/kegg/kegg2.html>; page is illustrated in Figure 1.12.2). Click the LIGAND link on the toolbar at the top of the page to bring up the KEGG LIGAND Database search page (Fig. 1.12.37).

2. The search page may be used to make simple queries by Entry (compound ID prefixed by “C”), compound name, formula, or a range for molecular weight. Any of these queries may be individually executed by clicking on the corresponding Go button. All of these queries are case-insensitive. For the Formula search, for example, either c6h7no or (C6H7NO may be input. To search for multiple entries at once, keywords may be entered separated from one another by a space (e.g., C00010 C00022).

Currently, it is not possible to search using multiple fields at once, e.g., combined with the Boolean operators AND or OR.

Compound structures in MOL files may be employed as queries by using either the SIMCOMP or SUBCOMP tools under the Computational Tools section of Ligand. SIMCOMP attempts to find the maximal matching compounds to the query, whereas SUBCOMP attempts to find compounds that contain the query compound entirely.

3. As an example, search for all compounds by name that contain the benzoic acid fragment. Select Name from the pull-down menu under the “Search COMPOUND” heading, and type benzoic acid in the text box. Click the Go button to execute the query. The results of this query will resemble Figure 1.12.38. This page lists the number of records retrieved in total, and a page-by-page listing of the results in

Figure 1.12.37 KEGG LIGAND Database search window.

Using Biological Databases

1.12.41

Compound Data Search Result

[Top](#)

 Number of entries in a page Hide structure

 Page : 1 Go of 2 Items : 1 - 20 of 40 [Top](#) [Previous](#) [Next](#) [Bottom](#)

No	Entry	Structure	Name	Formula
1	C00108		Anthranilate Anthranilic acid o-Aminobenzoic acid Vitamin L1 2-Aminobenzoate	C ₇ H ₇ NO ₂
2	C00156		4-Hydroxybenzoate Hydroxybenzoic acid 4-Hydroxybenzoic acid Hydroxybenzenecarboxylic acid	C ₇ H ₆ O ₃
3	C00180		Benzoate Benzoic acid Benzencarboxylic acid Phenylformic acid Dracylic acid Benzoin acid (TN)	C ₇ H ₆ O ₂
4	C00196		2,3-Dihydroxybenzoate 2,3-Dihydroxybenzoic acid	C ₇ H ₆ O ₄
5	C00230		3,4-Dihydroxybenzoate 3,4-Dihydroxybenzoic acid Protocatechuate Protocatechuic acid	C ₇ H ₆ O ₄

Figure 1.12.38 Compound data search results page.

groups of 20. Each result entry gives the Entry ID, which is clickable, the known names for the entry, its formula, and a graphical representation of the entry (under Structure).

- To look at one of the entries in more detail, for this example, C00180, click on its ID. This will give the summary page shown in Figure 1.12.39. This page lists all information, such as reactions and enzymes, related to the selected compound, which are all linked to the corresponding REACTION, ENZYME, and other databases. By clicking on any of the available links, further details may be examined. The structure can be saved to disk as a MOL file by clicking on the “Mol file” button in the row labeled Structure. Another query using the selected structure can also be executed with the neighboring SIMCOMP button, which searches for compounds that maximally match the query (Hattori et al., 2003).

The entries in the row labeled Reaction are hyperlinked to a list of all of the reactions related to this compound. The entries in the row labeled Pathway link directly the KEGG metabolic PATHWAY maps (see Basic Protocols 1 to 4). Similarly, links are available in the Enzyme row for the enzymes related to this structure and in the Other DBs row to entries in other databases in which this structure may be found.

KEGG COMPOUND: C00180

Help

Entry	C00180	Compound
Name	Benzoate; Benzoic acid; Benzene carboxylic acid; Phenylformic acid; Diacrylic acid; Benzoic acid (JP14/USP); Benzoic acid (TN)	
Formula	C7H6O2	
Mass	122.0368	
Structure	 C00180	
	Mol file SIMCOMP	
Remark	Drug: 7319	
Reaction	R00819 R01295 R01419 R01420 R01421 R01422 R01423 R01424 R01425 R01426 R01427 R02606 R05590 R05591 R05621 R05622 R06727 R06728	
RPair	A00337 A00847 A01121 A01415 A01498 A01499 A01501 A01502 A01504 A01506 A02347	
Pathway	PATH: map00360 Phenylalanine metabolism PATH: map00362 Benzoate degradation via hydroxylation PATH: map00621 Biphenyl degradation PATH: map00622 Toluene and xylene degradation PATH: map00632 Benzoate degradation via CoA ligation PATH: map00960 Alkaloid biosynthesis II	
Enzyme	1.2.1.7 1.2.1.28 1.14.12.10 1.14.13.12 3.1.1.1 3.5.1.4 3.5.1.32 3.5.1.40 3.5.5.1 3.6.1.7 3.6.1.20 3.7.1.8 6.2.1.25	
Other DBs	CAS: 65-85-0 PubChem: 3480 ChEBI: 16150	
LinkDB	All DBs	

Figure 1.12.39 Entry C00180 from the KEGG COMPOUND database.

KEGG LIGAND: THE GLYCAN DATABASE

The KEGG GLYCAN database is the newest addition to the KEGG suite of databases. It contains carbohydrate sugar chains, or glycans, whose entry IDs are prefixed by the letter “G.” Currently there are over 11,000 entries, most of which originated from the now-defunct CarbBank database. Pathway diagrams on the metabolism of complex carbohydrates and complex lipids are now linked to these glycan entries. Just as for chemical compounds, glycans are represented in KCF format, but instead of atom names, monosaccharides are used as the base unit representing nodes.

BASIC PROTOCOL 17

Readers should be aware that KegDraw, installable onto a local computer, is an alternative program for performing structural queries on KEGG GLYCAN. This program is available for download at <http://www.genome.jp/download/> and contains the latest functionality to search for similar glycan structures.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser, or KegDraw application for Windows XP or Mac OS X,
downloadable from <http://www.genome.jp/download/>

Using Biological Databases

1.12.43

Files

Glycan structure in KCF file format (optional). A sample dataset for an N-linked glycan in KCF format is provided as `glycan.txt` on the Current Protocols Web site (<http://www.currentprotocols.com>). For more information on the KCF file format for glycans, go to the KCaM (KEGG Carbohydrate Matcher) main page at <http://www.genome.jp/ligand/kcam/kcam/faq.html>, click on the Docs option, then click on the link to “KCF format for glycans.”

1. Access the KEGG LIGAND Database page as described in Basic Protocol 16 or go directly to <http://www.genome.jp/kegg/ligand.html>.

The GLYCAN database may be accessed under the “Search GLYCAN” heading, shown in Figure 1.12.37. From here, the database can be searched, for example, by glycan ID, composition, and class name, by selecting from the pull-down menu and entering the search text in the text box.

2. To perform a search for similar glycan structures, scroll down the KEGG LIGAND database page to the “Search similar glycan structures” heading. Click on the KCaM link to view the KEGG Glycan search page.

From here a glycan structure in KCF format can be uploaded or entered as a query.

3. Click on the link to the KCaM Main Server (in the highlighted box on the right-hand side) to go to the KCaM Glycan Structure Search page (Fig. 1.12.40).
4. A query can be performed on either the latest KEGG GLYCAN database or on the original CarbBank database. The source database can be selected via the pull-down menu labeled “Database.” There are currently six types of search options available, based on two main algorithms: Approximate and Exact match. Both of these algorithms implement dynamic programming techniques based on the Smith-Waterman dynamic programming algorithm for sequence alignment (Aoki et al., 2003, 2004).

[Feedback Form | LIGAND]

Figure 1.12.40 The KCaM Glycan Structure Search page.

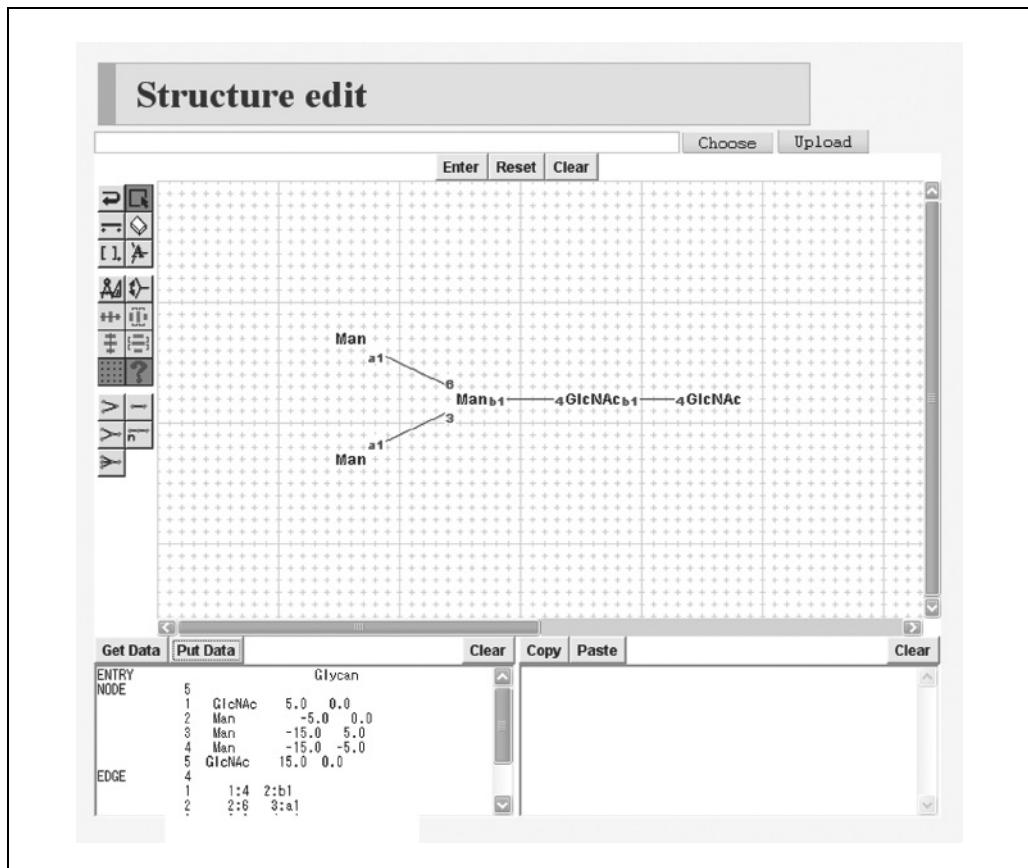


Figure 1.12.41 The Java-based applet for editing GLYCAN structures.

Thus, they both provide “global” and “local” options. Furthermore, the Exact matching algorithm provides the option to search linkages based on monosaccharide names only or based on the entire set of monosaccharide names, anomeric groups, and hydroxyl groups to which the linkage is attached on the monosaccharides. The Approximate matching algorithm aligns monosaccharides with one another and weights similarity of linkages based on preset parameters.

5. The structure editor for GLYCAN is a Java applet, with many features similar to ISIS/Draw. Click on the “Click here to specify...” box to bring up the editor as shown in Figure 1.12.41.

The buttons on the left provide tools to edit, erase, and add nodes on the main canvas. There are also templates of branches available below these set of tools. Structures entered on this canvas may be saved to disk in KCF format by clicking the Get Data button below the canvas. The contents of this canvas can then be copied-and-pasted into a KCF-formatted text file.

6. In this example, the structure for an N-linked glycan (see Necessary Resources) is uploaded into the structure editor. Go to the Current Protocols Sample Data Web site at the URL in Necessary Resources, find the file `glycan.txt` in the section of the Web site reserved for this unit, and copy and paste the contents of the `glycan.txt` file into the bottom-left text area in the structure edit window. Click on Put Data, and the structure appears in the window, as shown in Figure 1.12.42.

Alternatively a KCF format file can be uploaded into the file into the field above the canvas by using Choose and clicking Upload.

KCaM

Glycan Structure Search using KCaM

Complex carbohydrates vital to the functioning and development of multicellular organisms.

Main Tutorial F.A.Q. Docs

This form provides an interface for performing a structural analysis of glycans.

First time users of KCaM may refer to the tutorial for a quick-start guide to using KCaM, the web service for glycan structure analysis.

A FAQ is also available.

A paper on the algorithm used for performing glycan matching is also available here.

Database KEGG Glycan

Search type Approximate Match (Global)

[Feedback Form | LIGAND]

Figure 1.12.42 Glycan Search with query structure ready to go. The user may now select the Database and Search type to perform, then click Go to execute the query.

Glycan Data Search Result

Number of entries in a page : 20 Hide structure

Page : 1 Go of 33 Items : 1 - 20 of 649 Top Previous Next Bottom

No	Entry	Structure	Name	Composition	Class
1	G00311	 Similarity-Score : 500		(GlcNAc)2 (Man)3	Glycoprotein; N-Glycan
2	G06928	 Similarity-Score : 455		(GlcNAc)2 (Man)3	Glycoprotein; N-Glycan
3	G00286	 Similarity-Score : 400		(GlcNAc)2 (L-Fuc)1 (Man)3	Glycoprotein; N-Glycan
4	G00471	 Similarity-Score : 400		(GlcNAc)2 (Man)3 (Xyl)1	Glycoprotein; N-Glycan Neoglycoconjugate

Figure 1.12.43 Glycan structure search results with numerous analysis options available. See text for details.

7. Click on the Enter button to send the structure back to the GLYCAN Web page. The structure is now ready to be used as a query (Fig. 1.12.42) as described in substeps a and b, below.

- Choose KEGG Glycan for the database and “approximate match (global)” as the “Search type” to perform, then click Go (located at the lower right-hand corner of the inset window) to execute the query. The results will be displayed in a Glycan Search Result page, similar to that shown in Figure 1.12.43.
- Note how most of the classes returned using this Approximate matching search are N-Glycans. A search using Exact match will return only N-Glycan structures because of the restriction of no gaps.

KEGG GLYCAN: G00286

Help

Entry	G00286	Glycan
Composition	(GlcNAc) ₂ (Lfuc) ₁ (Man) ₃	
Mass	1057	
Structure		
Class	Glycoprotein; N-Glycan	
Remark	LECTIN: Rhizopus stolonifer lectin [PMID:12788923] LECTIN: wheat germ agglutinin [PMID:12361949]	
Reference	1 [PMID:2110822] Yamashita K, Inui K, Totani K, Kochibe N, Furukawa M, Okada S. Characteristics of asparagine-linked sugar chains of sphingolipid activator protein 1 purified from normal human liver and GM1 gangliosidosis (type 1) liver. <i>Biochemistry</i> . 29 (1990) 3030-9. 2 [PMID:2331705] Katoh H, Ohgi K, Irie M, Endo T, Kobata A. The structure of the asparagine-linked sugar chains of bovine brain ribonuclease. <i>Carbohydr. Res.</i> 195 (1990) 273-93. 3 [PMID:11937329] Kurahashi T, Miyazaki A, Murakami Y, Suwan S, Franz T, Isobe M, Tani N, Kai H. Determination of a sugar chain and its linkage site on a glycoprotein TIME-EA4 from silkworm diapause eggs by means of LC-ESI-Q-TOF-MS and MS/MS. <i>Bioorg. Med. Chem.</i> 10 (2002) 1703-10. 4 [PMID:12361949] Cipollo JF, Costello CE, Hirschberg CB. The fine structure of <i>Caenorhabditis elegans</i> N-glycans. <i>J. Biol. Chem.</i> 277 (2002) 49143-57. 5 [PMID:12788923] Oda Y, Senaha T, Matsuno Y, Nakajima K, Naka R, Kinoshita M, Honda E, Furuta I, Kakehi K. A new fungal lectin recognizing alpha(1-6)-linked fucose in the N-glycan.	

Figure 1.12.44 Glycan ID G00286. Various links to databases both within and outside of KEGG are available.

Using Biological Databases

1.12.47

The glycan results page provides a plethora of options. First, another structural query may be performed on this resulting data set or on the original KEGG GLYCAN or CarbBank databases by clicking on the image of the structure of interest. In this case, the second structure with which to query will be displayed in a new KEGG Glycan Search page. Second, different resulting pages may be viewed by either directly specifying the page to view or by traversing the pages using the Top, Previous, Next, and Bottom buttons. Finally, the details of each entry may be further examined by clicking on any glycan ID. The similarity score may be clicked to visually see the alignment that resulted in the given score.

8. To view the data on a specific glycan entry, click on that entry number. For example, to view the data for the third resulting glycan ID, G00286, click on the “G00286” glycan entry to display the view shown in Figure 1.12.44, which shows all related information.

This page lists all related information for this N-glycan, including PubMed IDs and any other information on compound and pathway; this information is also linked. This structure is the basic core-fucosylated N-glycan core structure. In addition to the various links to other database entries, the glycan entry page also provides the user with the option to use this structure, possibly edited, as the query for another search. This can be done by clicking the KCam button under the Structure heading.

KEGG LIGAND: THE REACTION DATABASE

The KEGG REACTION database contains reaction formulas for enzymic reactions, currently totaling 6127 entries. Each entry is identified by the prefix “R,” representing a unique reaction corresponding to sets of reactants and products represented by compounds or glycans from the COMPOUND and GLYCAN databases, respectively. This is in contrast to EC numbers, which may correspond to multiple reaction formulas.

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

1. Go to the main KEGG LIGAND database page at <http://www.genome.jp/kegg/ligand.html>. The REACTION database can be searched in several ways under the “Search REACTION” heading: reaction ID, name, reactant entry, pathway, or enzyme.
2. Under “Search REACTION,” select Name from the drop-down menu and enter benzoate in the text box. Click Go. The Reaction data search result page shown in Figure 1.12.45 will be displayed, which shows all entries in which the reactant benzoate is involved.

The entry number of the reaction and the reaction itself are displayed, along with the name of the enzyme involved in catalysis. The graphical representations of the reactions can be viewed by clicking on the “Show structure” button. One could alternatively first perform a compound ID search (see Basic Protocol 16) to obtain the compound ID, and then use this ID to perform a search for its reactions.

3. Click on a reaction ID and its related information will pop up in a new window. That window provides information on enzymes and details of the individual compounds involved in the reaction. A link to pathway information is also available, which displays the pathways within which this reaction may be found. Similarly, clicking on the DBGET link for a reaction on the reaction search results page will provide information related to the name of the reaction as found in the KEGG REACTION database.

Reaction Data Search Result				Top
Number of entries in a page [20 ▾]		Show structure		
Page :	1	Go	of 2	Items : 1 - 20 of 35
No	Entry	Definition	Name	
1	R00819	Benzoate + Oxygen + NADH <=> Catechol + CO ₂ + NAD+	Benzoate,NADH:oxygen oxidoreductase (1,2-hydroxylating, decarboxylating)	
2	R00821	2,3-Dihydroxybenzoate <=> Catechol + CO ₂	2,3-Dihydroxybenzoate carboxy-lyase	
3	R00990	Antranilate <=> Aniline + CO ₂	2-Aminobenzoate carboxy-lyase	
4	R01033	2-Chlorobenzoate + Oxygen + NADH + H+ <=> Chloride + Catechol + NAD+ + CO ₂	2-Chlorobenzoate,NADH:oxygen oxidoreductase (1,2-hydroxylating, dechlorinating, decarboxylating)	
5	R01238	4-Hydroxybenzoate <=> Phenol + CO ₂	4-Hydroxybenzoate carboxy-lyase	
6	R01295	Benzoate + Oxygen + NADPH <=> 4-Hydroxybenzoate + NADP+ + H ₂ O	Benzoate,NADPH:oxygen oxidoreductase(4-hydroxylating)	
7	R01296	4-Hydroxybenzoate + Oxygen + NADH <=> 3,4-Dihydroxybenzoate + NAD+ + H ₂ O	4-Hydroxybenzoate,NAD(P)H:oxygen oxidoreductase (3-hydroxylating)	
8	R01297	4-Hydroxybenzoate + Oxygen + NADH <=> Hydroquinone + CO ₂ + NAD+ + H ₂ O	4-Hydroxybenzoate,NADH:oxygen oxidoreductase (1-hydroxylating, decarboxylating)	
9	R01298	4-Hydroxybenzoate + Oxygen + NADPH <=> 3,4-Dihydroxybenzoate + NADP+ + H ₂ O	4-Hydroxybenzoate,NADPH:oxygen oxidoreductase (3-hydroxylating)	
10	R01299	4-Hydroxybenzoate + Oxygen + NADPH <=> Hydroquinone + CO ₂ + NADP+ + H ₂ O	4-Hydroxybenzoate,NADPH:oxygen oxidoreductase (1-hydroxylating, decarboxylating)	
11	R01300	ATP + 4-Hydroxybenzoate + CoA <=> AMP + Pyrophosphate + 4-Hydroxybenzoyl-CoA	4-Hydroxybenzoate:CoA ligase (AMP-forming)	
12	R01304	UDPGlucose + 4-Hydroxybenzoate <=> UDP + 4-(beta-D-glucosyloxy)benzoate	UDPGlucose:4-hydroxybenzoate 4-O-beta-D-glucosyltransferase	
13	R01306	4-Methoxybenzoate + Reduced acceptor + Oxygen <=> 4-Hydroxybenzoate + Formaldehyde + Acceptor + H ₂ O	4-Methoxybenzoate,hydrogen-donor:oxygen oxidoreductase (O-demethylating)	
14	R01307	4-Chlorobenzoate + H ₂ O <=> 4-Hydroxybenzoate + Chloride	4-Chlorobenzoate chlorohydrolase	
15	R01422	ATP + Benzoate + CoA <=> AMP + Pyrophosphate + S-Benzoate coenzyme A	Benzoate:CoA ligase (AMP-forming)	
16	R01504	ATP + 2,3-Dihydroxybenzoate <=> Pyrophosphate + (2,3-Dihydroxybenzoyl)adenylate	ATP:2,3-dihydroxybenzoate adenylyltransferase	

Figure 1.12.45 Results of reaction search for benzoate.

KEGG LIGAND: PATH COMPUTATION

As one function of pathway computation, given a starting and ending chemical compound structure, it is possible to calculate a path between the starting compound as an initial substrate and the ending compound as a final product, as long as a sequence of enzymatic reactions exists between these two compounds. With the option to specify the number of reactions in any path, many different reaction pathways may be produced. The basic concept is that, from among all the enzymic reaction data, if one compound that is generated in a reaction is the same as the initial substrate of a totally different reaction, even if they are in different pathways, these two reactions may be connected together.

BASIC PROTOCOL 19

Necessary Resources

Hardware

Computer with Internet access

Software

Web browser

Using Biological Databases

1.12.49



Generate Possible Reaction Paths

This program computes possible reaction paths between two compounds using the binary relations of substrates and products in known enzymatic reactions. The list of binary relations is generated from the enzyme list that corresponds to the KEGG reference pathways (all known enzymes) or that is found in an annotated genome (organism specific composition of enzymes).

Search against: Reference pathway (Reaction)

Enter initial compound (compound ID or keyword):

Enter final compound (compound ID or keyword):

Enter optional parameters:

Cut off length:

Query relaxation: No relaxation with level 2

Sort option: by path length by compound ID

Alternatively, use the pathway map interface

Select map: Glycolysis / Gluconeogenesis

Search against: Reference pathway

[KEGG2 | KEGG | GenomeNet]

Figure 1.12.46 The PathComp tool interface.

1. Go directly to <http://www.genome.jp/kegg/tool/pathcomp.html> or go to the KEGG Table of Contents as described in Basic Protocol 1 and follow the “Generate possible reaction paths” link. The path computation page which appears is shown in Figure 1.12.46.

The “Search against” field provides the option to select “Reference pathway (Reaction)” (default, meaning all available reactions) or a particular organism from which to select reaction data. The “Enter initial compound” and “Enter final compound” fields allow the user to enter keywords or accession numbers for the compounds to use as the starting and ending compounds, respectively.

2. Select “Escherichia coli K-12 MG1655” under “Search against” and enter pyruvate for initial compound and citrate for final compound. In this example, the query is designed to find paths between pyruvate and citrate in *E. coli* K-12. Click the Exec button. The substrate and product fields now become pull-down menus containing matching compound entries. Select “C00022 pyruvate” and C00158 citrate.”

If the compound IDs are entered directly from the start, this step is not necessary.

3. Enter a threshold for the maximum number of reactions to be allowed in the resulting paths in the “Cut off length” field, and select EC number hierarchy with level 3 for relaxation.

The default value is 5; if a value of 10 or larger is entered, the calculations may become extremely slow; therefore, it is recommended that 5 be set initially. If a satisfactory pathway is not generated, then this cut-off value can be increased little by little. In fact, results may be increased in another manner by using relaxation. Relaxation basically states that, for a particular organism, if it does not contain a particular enzyme, it is possible to use enzymes performing similar reactions instead. Specifically, the “Select hierarchy for relaxation” field can be used to select a standard relaxation level, or no

Result of Path Computation

Organism : E.coli
 Initial substrate : C00022 Pyruvate
 Final product : C00158 Citrate
 Cutoff length : 5
 Relaxation : No relaxation
 Number of Results : 60

[Show as Diagram]

```

2 C00022 <R03145> C00033 <R00362> C00158 [Known pathways] [Show compound structures]
2 C00022 <R00209> C00024 <R00351> C00158 [Known pathways] [Show compound structures]
3 C00022 <R00214> C00149 <R00472> C00024 <R00351> C00158 [Known pathways] [Show compound structures]
3 C00022 <R00214> C00149 <R00342> C00036 <R00362> C00158 [Known pathways] [Show compound structures]
3 C00022 <R00200> C00074 <R00345> C00036 <R00362> C00158 [Known pathways] [Show compound structures]
3 C00022 <R03145> C00033 <R00235> C00024 <R00351> C00158 [Known pathways] [Show compound structures]
3 C00022 <R00209> C00024 <R00235> C00033 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00214> C00149 <R00472> C00024 <R00235> C00033 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00214> C00149 <R00472> C00048 <R00479> C00031 <R01324> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00214> C00149 <R00342> C00036 <R00354> C00566 <R01323> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00220> C00065 <R00586> C00079 <R00897> C00033 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00212> C00058 <R00519> C00011 <R00345> C00036 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00212> C00058 <R00519> C00011 <R00267> C00311 <R01324> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00142> C05125 <R03270> C01136 <R02569> C00024 <R00351> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00782> C00097 <R04859> C00079 <R00897> C00033 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00200> C00074 <R00345> C00036 <R00354> C00566 <R01323> C00158 [Known pathways] [Show compound structures]
4 C00022 <R03145> C00033 <R00711> C00084 <R00228> C00024 <R00351> C00158 [Known pathways] [Show compound structures]
4 C00022 <R03145> C00033 <R00316> C05993 <R00236> C00024 <R00351> C00158 [Known pathways] [Show compound structures]
4 C00022 <R03145> C00033 <R00317> C00227 <R00230> C00024 <R00351> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00209> C00024 <R00472> C00149 <R00342> C00036 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00209> C00024 <R03991> C00010 <R00351> C00036 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00209> C00024 <R01323> C00566 <R00354> C00036 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00209> C00024 <R00230> C00227 <R00315> C00033 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00209> C00024 <R00236> C05993 <R00316> C00033 <R00362> C00158 [Known pathways] [Show compound structures]
4 C00022 <R00209> C00024 <R00228> C00084 <R00711> C00033 <R00362> C00158 [Known pathways] [Show compound structures]
5 C00022 <R00214> C00149 <R00472> C00024 <R003991> C00010 <R00351> C00036 <R00362> C00158 [Known pathways]
5 C00022 <R00214> C00149 <R00472> C00024 <R01323> C00566 <R00354> C00036 <R00362> C00158 [Known pathways]
5 C00022 <R00214> C00149 <R00472> C00024 <R00230> C00227 <R00315> C00033 <R00362> C00158 [Known pathways]
5 C00022 <R00214> C00149 <R00472> C00024 <R00236> C05993 <R00316> C00033 <R00362> C00158 [Known pathways]
5 C00022 <R00214> C00149 <R00472> C00024 <R00228> C00084 <R00711> C00033 <R00362> C00158 [Known pathways]
5 C00022 <R00214> C00149 <R00472> C00048 <R00479> C00311 <R01900> C00417 <R01325> C00158 [Known pathways]
5 C00022 <R00214> C00149 <R00342> C00036 <R00351> C00010 <R03991> C00024 <R00351> C00158 [Known pathways]
5 C00022 <R00214> C00149 <R00342> C00036 <R00345> C00011 <R00267> C00311 <R01324> C00158 [Known pathways]
5 C00022 <R00214> C00149 <R01082> C00122 <R00490> C00049 <R00355> C00036 <R00362> C00158 [Known pathways]
5 C00022 <R00214> C00149 <R01082> C00122 <R00412> C00042 <R00479> C00311 <R01324> C00158 [Known pathways]
5 C00022 <R00220> C00065 <R00586> C00079 <R00897> C00033 <R00235> C00024 <R00351> C00158 [Known pathways]
5 C00022 <R00220> C00065 <R00945> C00037 <R03425> C00011 <R00345> C00036 <R00362> C00158 [Known pathways]
5 C00022 <R00220> C00065 <R00945> C00037 <R03425> C00011 <R00267> C00311 <R01324> C00158 [Known pathways]
5 C00022 <R00226> C06010 <R04672> C05125 <R03270> C01136 <R02569> C00024 <R00351> C00158 [Known pathways]

```

Figure 1.12.47 Pathway computation results. In each row, information on path length, individual compound IDs and corresponding enzyme IDs are listed for each path found. In the case that relaxation was used, # marks will surround the enzyme IDs.

relaxation. If the EC number hierarchy is selected, it has four levels, so if one wishes to raise the relaxation from the default by one level, one would set the “with level” field to 3, meaning the program will use the top three levels of the EC number if they are the same. However, note that in the case where the Standard Dataset is selected in the “Search against” field, this relaxation option has no meaning. Finally, the “Sort option” radio buttons allow the results to be sorted by increasing path length or by compound ID. Thus, those who wish to see their results by path length may select the first option, and those who wish to see similar pathways clustered together may use the second option.

- Click the Exec button. The results page (Fig. 1.12.47) lists all paths found. It provides a variety of information in each row, including path length, followed by each compound ID along with its enzyme ID, consecutively. In the case that relaxation was used, the enzyme IDs will have pound signs (#) surrounding them.
- Click on the “Show compound structures” link at the end of any row to display a graphical view of the selected pathway, as in Figure 1.12.48 which displays the result for the first listed pathway. This is extremely useful for confirming how the reaction pathway came about. Also, the link for “Known pathways” near the end of each row will search the pathway maps for the enzyme IDs appearing in the selected reaction pathway. This allows the user to examine specifically where in the existing PATHWAY database the reaction pathways can be found.

Pathway with Compound Structures

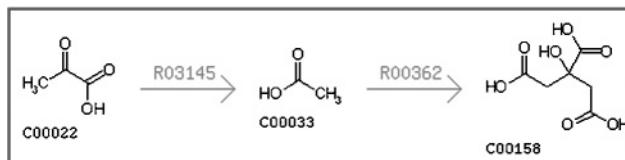


Figure 1.12.48 Graphical representation of a generated compound pathway, useful for visually confirming each pathway found.

COMMENTARY

Background Information

KEGG first began in 1995 under the Human Genome Program of the Ministry of Education, Science, Sports and Culture of Japan (Kanehisa, 1997). One objective of KEGG was to computerize the known knowledge of molecular and genetic pathways gained from experimental observations in genetics, biochemistry, and molecular and cellular biology. Currently, KEGG not only has metabolic pathways but a number of regulatory pathways available online. In addition, the pathways are organized such that organism-specific pathway information may also be visualized. Additional objectives were to maintain the gene catalog of every organism that has been sequenced as well as to map each component in the catalog onto the KEGG pathways. Yet another objective was to develop new informatics technologies associated with interactions and pathways. Not only have all of these objectives been realized, but KEGG continues to progress at an impressive rate, with chemical compound and carbohydrate structure search capabilities, gene clustering, ortholog and paralog cluster visualizations, and microarray expression analysis tools all easily available online.

The data in KEGG is not only updated through publications, but is also provided by a community of biologists who directly enter their latest discoveries to make them available for the community. BSORF and CYORF are examples of such databases. This data are then examined and transferred into KEGG. CarbBank may be considered a similar meta-database for KEGG's GLYCAN database of LIGAND. Thus, the latest information is available for the research community.

KEGG GENES database

The KEGG GENES database was started in June 1995, slightly before TIGR (The In-

stitute for Genomic Research) announced the completion of the full sequences of citrate *Haemophilus influenzae* Rd and *Mycoplasma genitalium* using shotgun sequencing. The genetic information contained in the genome is most fundamental for systematically understanding life. However, in any organism, there are still many genes whose functions are unknown. The KEGG GENES database initially collected all information from GenBank related to genes, and independent annotation of this data continues even today. There are a number of ways of using the KEGG GENES database, and the protocols in this unit constitute the most straightforward methods of querying the text-based database.

Orthologs and paralogs

Orthologs really cannot be defined without considering evolution, but they can conveniently be defined based on the relationships between reciprocal best hits as described below. Paralogs can be defined as those sequences within the same species whose similarity scores are higher than a specified threshold. Given gene x from genome A , and assuming that one wishes to compare it to all of the genes in genome B , if it is found that gene y of genome B has the highest sequence similarity score with gene x , y can be called the best hit for x . On the other hand, if gene y of genome B is compared against all of the genes of genome A , one may not necessarily get gene x as the best hit of gene y , because there may be several genes that are just as similar, if not more. However, if x does turn out to be the best hit of y , then one may claim that x and y are orthologs.

Approximation algorithms such as BLAST and FASTA are asymmetric. They produce different sequence similarity results for sequences x and y based on the sequence used

as the query. For this reason KEGG uses the SSEARCH program, which is an implementation of the rigorous Smith-Waterman dynamic programming algorithm that produces optimal alignments (Smith et al., 1981a, b). Because a Smith-Waterman score produced by the SSEARCH program is not asymmetric, one can use it as the sequence similarity score for two genes.

Critical Parameters and Troubleshooting

As all of KEGG's resources are maintained manually with advanced computational tools, the database continues to increase at a steady pace, but there may be areas which are lacking. The curators are furiously updating the information as quickly as possible, which is a daunting task, especially considering the amount of experimental data that continues to be published constantly. However, the manual curation of KEGG data ensures that the data are biologically accurate and more trustworthy than data generated via computerized techniques that may not account for the many exceptions known to occur in biology.

The most immediate change to be expected will be in the LIGAND database. The COMPOUND and REACTION databases may be renovated even before this documentation is published. The Chime plug-in will no longer be required to perform structure searches of chemical compounds. Instead, the standard MOL file format will be used, and the structures will be drawn as images. Detailed documentation of these updates will be made available as soon as possible. In addition, the search parameters for the GLYCAN database will also be updated in the near future in order to make them easier to use. The six options currently available will be parameterized, and an "Advanced options" section will be made available for those who need to fine-tune the search parameters.

As illustrated in this documentation, the KEGG data resource provides a comprehensive suite of intricate links that are presented in such a way that makes it easy for biologists and researchers to understand the workings of biological processes. From genetic to genomic to proteomic and metabolic information, to expression and chemical compound and carbohydrate data, the goal is to integrate all relevant networks both fully within all of KEGG's resources, as well as with links to popular databases outside of KEGG.

Suggestions for Further Analysis

The KEGG Application Programming Interface (API)

The KEGG Application Programming Interface (API) provides a valuable means for accessing the KEGG system, enabling users to search and compute biochemical pathways in cellular processes and to analyze the universe of genes in the completely sequenced genomes over the Internet. Users can access the KEGG API server using SOAP technology over the HTTP protocol. The SOAP server also comes with a Web Services Description Language (WSDL), which makes it easy to build a client library for a specific computer language. This enables users to write their own programs for many different purposes and to automate the procedure of accessing the KEGG API server and retrieving the results. The KEGG API documentation, WSDL file, Java library, and other information may be obtained by clicking on the API link at the top of the KEGG Table of Contents (Fig. 1.12.2) or by accessing the URL <http://www.genome.jp/kegg/soap/>. Further details are beyond the scope of this document, and interested readers are encouraged to access this Web page for the latest information.

Literature Cited

- Aoki, K.F., Yamaguchi, A., Okuno, Y., Akutsu, T., Ueda, N., Kanehisa, M., and Mamitsuka, H. 2003. Efficient tree-matching methods for accurate carbohydrate database queries. *Genome Inform.* 14:134-143.
- Aoki, K.F., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S., and Kanehisa, M. 2004. KCaM (KEGG Carbohydrate Matcher): A software tool for analyzing the structures of carbohydrate sugar chains. *Nucl. Acids Res.* 32:W267-W272.
- Bono, H., Ogata, H., Goto, S., and Kanehisa, M. 1998. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Res.* 8:203-210.
- Dudoit, S., Yang, Y.H., Callow, M.J., and Speed, T.P. 2000. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report #578, Statistics, University of California, Berkeley, August 2000. Available online at <http://www.stat.berkeley.edu/users/terry/zarray/TechReport/578.pdf>.
- Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. 2002. LIGAND: Database of chemical compounds and reactions in biological pathways. *Nucl. Acids Res.* 30:402-404.

Using Biological Databases

1.12.53

- Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. 2003. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* 125:11853-11865.
- Kanehisa, M. 1997. A database for post-genome analysis. *Trends Genet.* 13:375-376.
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. 2004. The KEGG resource for deciphering the genome. *Nucl. Acids Res.* 32:D277-D280.
- Ogura, M., Yamaguchi, H., Yoshida, Ki., Fujita, Y., and Tanaka, T. 2001. DNA microarray analysis of *Bacillus subtilis* DegU, ComA and PhoP regulators: An approach to comprehensive analysis of *B. subtilis* two-component regulatory systems. *Nucl. Acids Res.* 29:3804-3813.
- Smith, T.F. and Waterman, M.S. 1981a. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.
- Smith, T.F. and Waterman, M.S. 1981b. Comparison of biosequences. *Adv. Appl. Math.* 2:482-489.

Contributed by Kiyoko F. Aoki and
Minoru Kanehisa
Bioinformatics Center
Kyoto University, Japan