VERSION 9.5

User's Guide

PATHWAY TOOLS

Volume I: Navigator

Pathway Tools User's Guide, Version 9.5

We gratefully acknowledge contributions to Pathway Tools, used by permission, from:
Jeremy Zucker, Harvard Medical School
The Laboratory of Christos Ouzounis, European Bioinformatics Institute

SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025 U.S.A.
biocyc-support@ai.sri.com

Contents

GLOSSARY

EcoCyc

A Pathway/Genome Database for *E. coli* developed jointly by P. Karp, M. Riley, J. Collado-Vides, I. Paulsen, and M. Saier.

GKB Editor

An SRI software system for interactively browsing and editing DBs that are managed by the Ocelot frame knowledge representation system. For information on the GKB Editor, see **`http:// www.ai.sri.com/~gkb/user-man.html`.**

KB

See ‘knowledge base’.

DB developer

A user who modifies a Pathway/Genome Database, as opposed to users who view, but do not update, information in a PGDB.

Knowledge Base (KB)

A collection of frames and their associated slots, values, facets, and annotations. KBs can be saved permanently in Oracle or MySQL databases and in disk files. This document uses the terms ‘knowledge base’ and ‘database’ (DB) interchangeably. A Pathway/Genome Database is a type of knowledge base.

Lisp listener

A program that accepts statements written in the Lisp programming language, evaluates those statements, and prints the results. Pathway Tools can be invoked in a mode in which a Lisp listener is active, when the user wants to issue Lisp-based queries to PGDBs -- use the “-lisp” switch when starting Pathway Tools.

Listener Window

The short horizontal pane toward the bottom of the Pathway Tools main window. Pathway Tools prints messages and sometimes requests information in the listener window. Note that the listener window has nothing to do with a Lisp listener.

MetaCyc

A meta-metabolic database that describes pathways, reactions, and enzymes across a range of different organisms. MetaCyc is a PGDB and is jointly developed by SRI International and the Carnegie Institution of Washington Department of Plant Biology.

PGDB

See ‘Pathway/Genome Database’.

PathoLogic Pathway Predictor

A computer program that generates a new PGDB by analyzing the annotated genome of an organism to predict its metabolic pathways. PathoLogic is a component of the Pathway Tools software.

Pathway

In the context of the Pathway Tools software, we define a pathway as an interconnected set of biochemical reactions, where reactions are connected by sharing common reactants and products. In metabolic pathways, reactants and products are typically low-molecular-weight chemical compounds. In signaling pathways, reactants and products are typically proteins.

Pathway/Genome Database (PGDB)

An integrated database that employs the same database schema as does EcoCyc. Pathway/Genome Databases are typically created using the PathoLogic Pathway Predictor and represent value-added extensions of annotated genomes. Pathway/Genome Databases may be visualized, queried, and updated using SRI's Pathway Tools software. Each Pathway/Genome Database contains a collection of interconnected information describing the biochemical pathways and genome of an organism.

Pathway Tools

The Pathway Tools software system consists of three software components for manipulating Pathway/Genome Databases: (i) Pathway/Genome Navigator, (ii) PathoLogic Pathway Predictor, and (iii) Editing Tools. The Navigator is used to query, visualize, and analyze the information contained within a PGDB. It does not change this information in any way. To modify this information the user must use Editing Tools, for example, to update information about a particular gene or ORF or to enter information about a newly discovered pathway. The user may also use Editing Tools to link any piece of information within a PGDB to information contained within an external Web-based database, for example, a gene-expression relational database. The PathoLogic Pathway Predictor allows the user to computationally predict the pathways of an organism from its annotated genome. The predicted pathways and the description of the genome are then combined to create a new data set.

Reference Pathway/Genome Database

PathoLogic predicts the pathways of an organism by analogy to a reference PGDB.

SMILES

A language for writing chemical structures in terms of character strings [17]. The chemical substructure searcher within Pathway Tools accepts substructures in the SMILES language. See **Help -> Smiles** for more examples and a description.

Subject Organism

The organism for which a new PGDB is to be constructed using PathoLogic.

PREFACE

This document will familiarize you with the Pathway Tools software and the Pathway/Genome Databases (PGDBs) that are available from SRI International. Pathway Tools operates across one or more PGDBs, including the EcoCyc *E. coli* DB and the MetaCyc DB. Each database describes the biochemical pathways and genomes of a single organism. All PGDBs are managed by an object-oriented database system called 'Ocelot'.

Pathway Tools contains several component software modules including

Pathway/Genome Navigator, the primary tool by which users visualize, query, and analyze the information contained within a PGDB.

PathoLogic Pathway Predictor, which creates new PGDBs from annotated genomes.

Pathway/Genome Editors, which modify the information contained within PGDBs, such as allowing users to add in-house proprietary data about a pathway or gene of interest. The Editors include a general tool for database browsing and editing, called the GKB Editor.

Pathway Tools is most commonly provided in two different configurations:

BioCyc Configuration. Includes the Pathway/Genome Navigator and one or more PGDBs, such as EcoCyc and MetaCyc. This configuration is available for the Sun workstation running Unix, for the PC running Linux, and for the PC running Windows. The Sun and Linux versions can run as both an X-windows application and as a web server for the user's intranet, whereas the PC/Windows version can run as both a local application and a web server. In this configuration, PGDBs are physically included inside the binary executable program, and are available for read-only access. In addition, PGDBs that have been created using the full Pathway Tools configuration can be imported into the BioCyc configuration.

Full Pathway Tools Configuration. Includes the Navigator, PathoLogic, and Editors as well as one or more PGDBs. This configuration is available for the Sun workstation running Unix and for the PC running Linux, but not for the PC running Windows. The Navigator functionality can run as an X-windows application and as a web server for the user's intranet. This configuration allows the creation of new PGDBs, which can be stored either as disk files or in an Oracle or MySQL database server. Newly created PGDBs can be updated by users. Use of an Oracle or MySQL database is recommended if multiple users will be updating a given PGDB.

Volume I of the *Pathway Tools User's Guide* describes the Navigator. Volume II describes PathoLogic and the Pathway/Genome Editors.

The PGDBs present in your installation of Pathway Tools will depend on exactly which data sets your organization has ordered. Other PGDBs are available in addition to EcoCyc and MetaCyc. The data content of different PGDBs varies. The EcoCyc DB generally contains more extensive data than do other PGDBs. For example, EcoCyc contains extensive data on operons, promoters, transcription factors, transporters, and paralogous gene groups, which other PGDBs tend to lack.

In Volume 1 of this user's guide

Chapter 1 provides an overview of the pathway bioinformatics tools and databases: what they are, the information they contain, and some of the things they help you to do.

Chapter 2 describes how to invoke Pathway Tools and how to access the Pathway Tools web server.

Chapter 3 outlines the Pathway/Genome Navigator. It begins by providing some background information on using the mouse and some examples, to get you started quickly. It then outlines how to use the Navigator to select one or more databases and, subsequently, query, visualize, and analyze the information contained within these databases.

Chapter 4 covers the PGDB Sharing System.

Chapter 5 outlines what to do if you encounter problems with the functionality of any of the tools or with information contained within any of the provided databases.

ADDITIONAL PATHWAY TOOLS PUBLICATIONS AND WEB SITES

For more detailed information regarding different aspects of this software, consult the following resources (predominantly, EcoCyc related publications).

SRI Web site containing a variety of information on the Pathway Tools and its component software systems, including

- Pathway Tools information site containing samples of Pathway Tools file formats, example Lisp queries to PGDBs, and more, at **`http://bioinformatics.ai.sri.com/ptools/`**
- Generic Frame Protocol (the API for Pathway Tools data sets) specification at **`http:/ /www.ai.sri.com/~gfp/spec/paper/paper.html`**

GKB Editor User's Guide at **`http://www.ai.sri.com/~gkb/user-man.html`**

Scope and contents of the EcoCyc data, discussed in [11]

Ontology (representations) used in the Pathway Tools, discussed in [9, 10]

Pathway/Genome Navigator, described in [9]

Possible computational uses of the EcoCyc database, discussed in [7]

Ocelot database management system used in Pathway Tools, described in [9]

Many of these publications are available through the World Wide Web at **`http://www.ai.sri.com/pkarp/bio.html#SelPubs`**

1 INTRODUCTION TO THE PATHWAY TOOLS SOFTWARE

Bioinformatics has been largely sequence centric and focused on building databases to store the large amounts of sequence data being generated by the Human Genome Project and microbial genome initiatives. In addition, this discipline has involved the development of tools for visualizing, querying, and analyzing sequence data. Most bioinformatics analyses focus largely on identifying genes and predicting the function of encoded gene products. However, increasingly, the focus is moving from considerations of individual genes to consideration of how sets of genes and their corresponding gene products interact with each other and with other molecular players to form the pathways that are the biochemical basis of cellular behavior.

Pathway bioinformatics involves the creation of databases and algorithms for storing and computing with biochemical pathways. Pathway databases typically encode interactions between gene products and other compounds, and the organization of reactions into pathways. In addition, pathway bioinformatics involves the development of tools to query, visualize, and analyze this information. One of the primary advantages of pathway bioinformatics is that it provides descriptions of how the molecular components encoded in a genome interact with each other and with other molecular players to form the biochemical basis of cellular function.

1.1 DATABASES AND TOOLS

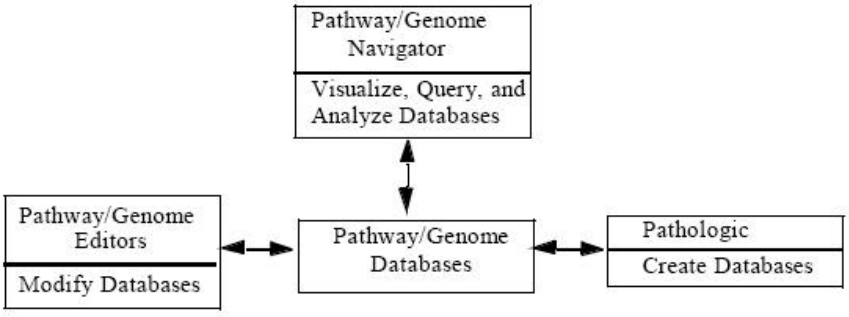


Figure 1-1 The Pathway Tools and related databases

Figure 1-1 provides an overview of the Pathway Tools software.

The software components operate on one or more Pathway/Genome Databases (PGDBs) (see the central rectangle). A PGDB is a collection of information that describes some of the biochemical pathways and genes of (typically) a single organism. The pathways might include, for example, signal transduction, transport, and metabolic pathways. The genes may constitute an entire genome, or a subset of genes of interest, for example, those genes that encode a set of pathways of interest.

1.1.1 Databases

The schema of a PGDB describes pathways in terms of five biological entities (see Figure 1-2 A):

Union of described pathways (the metabolic overview)

Individual pathways

Reactions that comprise these pathways

Compounds that participate in these reactions

Enzymes (a subset of the gene products) that catalyze these reactions

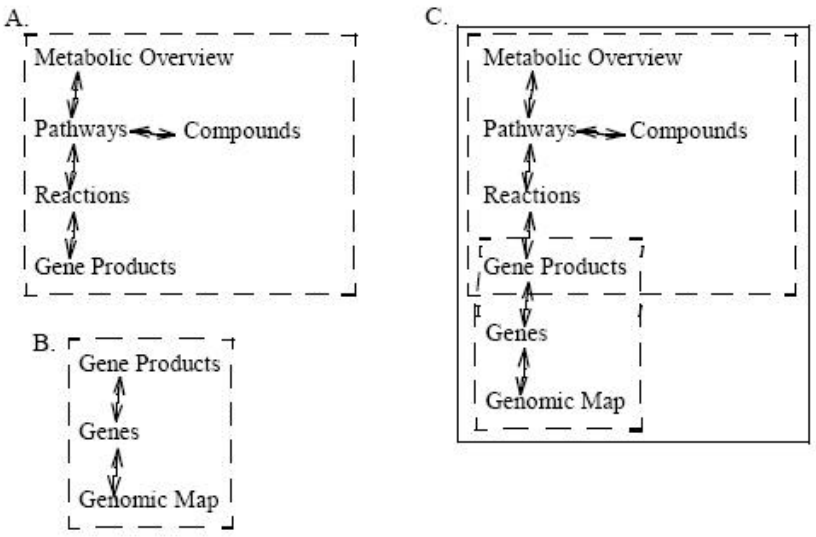


Figure 1-2 Representation of pathways and genomes

Genomes are described in terms of three biological entities (see Figure 1-2 B):

Genomic maps of sequenced genetic element(s)

Their constituent genes

Corresponding gene products (see Figure 1-2 B)

The gene products that are enzymes provide the primary link between the representation of the genomes and that of the metabolic pathways (see Figure 1-2 C). However, PGDBs typically describe all known or predicted gene products of the corresponding organism, not just those genes whose products are known or predicted to be enzymes—that is, the space of enzymes is a subset of the space of gene products. Incorporated pathways include those of biosynthesis, degradation, energy production, and intermediary metabolism, for compounds such as amino acids, carbohydrates, fatty acids, nucleotides, and enzyme cofactors. Different PGDBs vary as to the degree to which they contain transmembrane transport, genetic regulatory, or signal

transduction pathways (with the exception of the EcoCyc *E. coli* database, which contains all of the preceding pathway types).

1.1.2 Pathway Tools

The primary tool via which the user may visualize, query, and analyze the information contained within databases is the Pathway/Genome Navigator. The PathoLogic Pathway Predictor allows you to computationally predict the pathways of an organism from its annotated genome. The predicted pathways and the description of the genome are then combined to create a new PGDB. In addition, with special-purpose, intuitive, interactive editing tools that are not included in the configuration documented here you can modify existing databases, for example, update information about a particular gene or open reading frame (ORF) or enter information about a newly discovered pathway. You can also use these editing tools to link any piece of information within a given database to one or more pieces of information contained within a Web-accessible external database, for example, a gene expression relational database management system.

2 INVOKING THE PATHWAY TOOLS

This chapter provides the information necessary to get you started using the Pathway Tools. It is assumed that the Pathway Tools and database(s) have been successfully installed on your machine and are ready to run.

The Pathway Tools can be operated in two different modes. In the first mode, a user interacts with the Pathway Tools through the Unix X Window System (X Windows) or through Microsoft Windows. All of the Pathway Tools software components are available through the X Windows or Microsoft Windows operation mode: the Navigator, Editing Tools, and PathoLogic. In the second mode of operation, the Pathway Tools operate as a web server that can support simultaneous queries from many different users through the network. Only the Navigator operates in web mode.

For technical reasons, when the Pathway Tools are running in web mode, they nonetheless require access to the X Windows or Microsoft Windows system, meaning that a proper Unix X Windows or Microsoft Windows environment must be established even though users do not interact with the Pathway Tools through that environment. Fortunately for Microsoft Windows users, their environment is already set up to run Pathway Tools in web mode. Therefore, we begin with a brief overview of how to define the X Windows environment for use by the Pathway Tools.

Microsoft Windows users should read the installation guide for information on starting and stopping Pathway Tools. Sections 2.3.1, 2.3.2, and 2.3.3 are for all users. Otherwise, the remainder of this chapter is for Unix users.

2.1 X WINDOWS BASICS

X Windows is the name of the windowing system on Unix computers. If you're physically at the computer where Pathway Tools will run, you normally don't need to know anything about X Windows: it will simply work like it is supposed to. But, to be able to use X Windows when remotely logging into a Unix computer (named, say, **`my-ptools-server`**), your local computer needs to be able to forward X Windows.

To remotely login from another Unix computer (or from the optional X11 application on a Mac OS X computer) to **`my-ptools-server`**, use the following Unix command, which both logs you in and forwards X Windows:

ssh –X my-ptools-server

To remotely login from a Microsoft Windows computer to **`my-ptools-server`** you will need to install and configure a third-party X Server such as xFree86 or Hummingbird Exceed. Be sure to configure it to forward X Windows.

2.2 RUNNING PATHWAY TOOLS

To start Pathway Tools for X Windows operation on a Unix computer, type:

pathway-tools *args*

The text ***args*** consists of zero or more command-line arguments as defined under “Command Line Arguments” in section 2.2.1.

Either the **pathway-tools** command must be defined in the Unix **/usr/local/bin** directory, or the directory **aic-export/ecocyc/genopath/9.5/**, which contains the Pathway Tools executable file, must be on the user's search path.

For both X Windows and web server operation, a proper X Windows environment must have been established, as described in the previous section

To exit from the Pathway Tools, click on the **Exit** command in the **File** menu of the Navigator window.

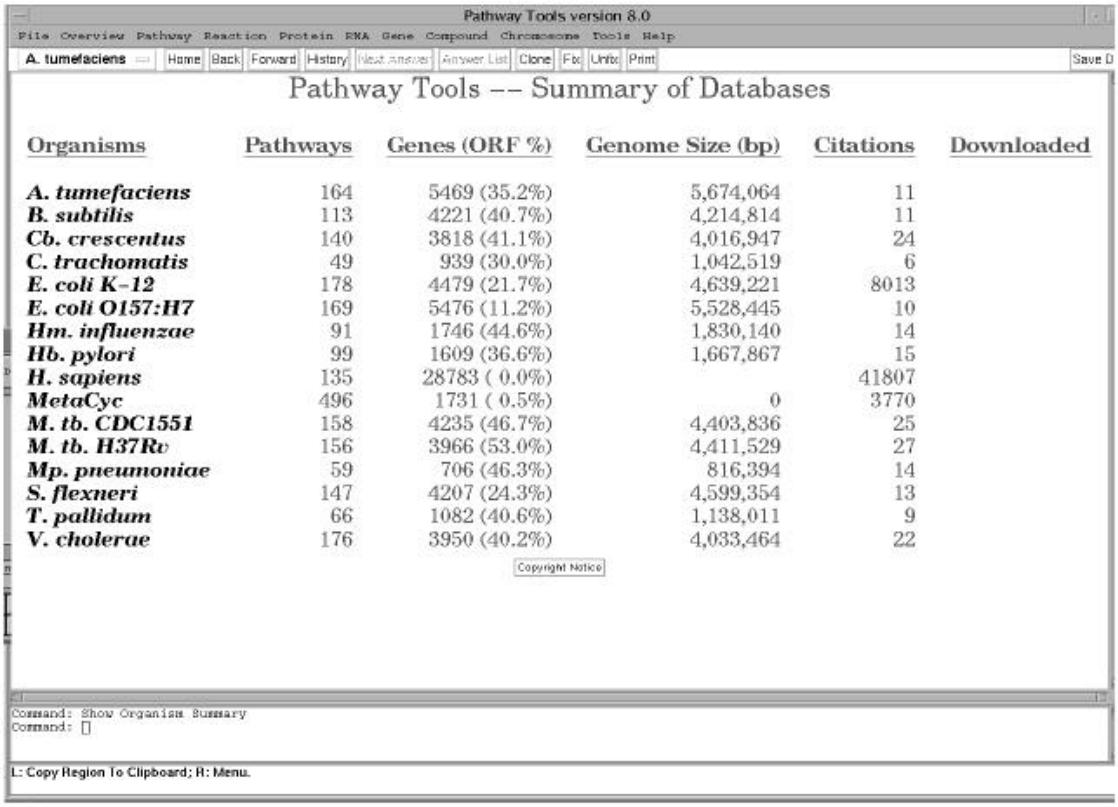


Figure 2-1 The All Organisms display

When you invoke the Pathway Tools for X Windows operation, it will create one large window for the Pathway/Genome Navigator that is divided into several regions, called panes (see Figure 2-1).

Different information is presented in the different panes. The main pane initially contains a listing of loaded databases; in this case, 16 databases. The horizontal pane at the bottom will print the names of commands that you select by clicking the mouse in the command menu. At times, informational messages will be printed in this pane. In the course of use, pop-up windows will also appear periodically with messages or to request information from you. At times, the

main pane will contain command buttons.

2.2.1 Command Line Arguments

Valid command-line arguments for the Unix Pathway Tools are:

`-id`

Instructs the Pathway Tools to print identifying version information for itself, and then exit.

`-patch`

Instructs the Pathway Tools to download and install all the latest patches from the public Pathway Tools patch site, and then exit.

`-lisp`

Instructs the Pathway Tools to start the Lisp interpreter instead of the Pathway/Genome Navigator interface so that the user may enter Lisp expressions, such as queries to Pathway Tools data. To exit from the Lisp interpreter back to Unix, type "(exit)". This option is not valid during web operation.

`-oratest`

When the Pathway Tools are configured for operation with an Oracle database, this option attempts to connect to the Oracle server, then lists those Pathway Tools data sets that are present, and then exits. This option is used for verifying that a proper connection to Oracle can be established.

`-org orgid`

Instructs the Pathway Tools to select the organism whose organism ID is as the current organism. In web mode, this organism will be the default organism selected on the Pathway Tools query page.

`-api`

This argument sets up the Pathway Tools to accept external queries from a package such as **perlcyc**. The perlcyc module allows users to write programs in perl that query a PGDB located on a Pathway Tools server running on the same machine as the perl program. Communication is by Unix file sockets. Consequently, this functionality is available only under Unix. The perlcyc module is not included in the Pathway Tools distribution. For more information, see **`http://www.arabidopsis.org/tools/aracyc/perlcyc/`**

`-linkdef, -dbdef`

These two arguments are used for bulk loading links to other external databases. See Pathway Tools User's Guide Volume II, Section 2.4.10.5 for more information.

The following command line arguments are applicable only to Pathway Tools when operating in web mode:

-www

Instructs Pathway Tools to operate in web mode.

-port NNN

This option is valid only when Pathway Tools is operating in web mode and is typically used in conjunction with the -user option. It specifies in NNN the TCP/IP port on which thePathway Tools web server will listen for requests. By default, port 1555 is used on Unix (Solaris and Linux), and port 80 is used on Microsoft Windows. See the -user argument for Unix security information.

-user NNN

This option is valid only when Pathway Tools is operating in web mode and is typically used in conjunction with the -port option. It specifies in NNN the Unix account that Pathway Tools should use to process web requests. Unix allows only root accounts to listen on TCP/IP ports numbered up to 1024, so if you specify -port 80, for example, then only root can start Pathway Tools, but Pathway Tools will switch to running as user NNN right after starting to listen on port 80. This option is neither available nor necessary on Microsoft Windows.

-proxy-port NNN

This argument is valid only when Pathway Tools is operating in web mode. It specifies in NNN the TCP/IP port on which another web server, such as Apache, will listen for requests and forward them to the port on which the Pathway Tools web server listens for requests. When using this argument, you also have to configure the other web server, such as Apache, to actually forward the requests. Typical use of this command-line argument is to specify **-proxy-port 80** to work around firewall restrictions; most firewalls do allow traffic to port 80.

-www-publish pubspec

This argument is applicable when Pathway Tools is running in web mode, and affects which of the available PGDBs are visible through the web server. This argument assists users in complying with the SRI license agreement provision that prevents users from publishing PGDBs owned by SRI or DoubleTwist (such as MetaCyc or BsubCyc) on their external web site. However, all available PGDBs may be published on a web site that is visible only within their institution.

Examples:

- Consider a user who has developed a new PGDB and wishes to publish it on their external web site without making other PGDBs such as EcoCyc visible. In this case, use the following command**:**
    - **pathway-tools -www -www-publish public**

  This command causes only those databases marked as “public” to be visible through the Pathway Tools web server, which includes all new PGDBs created by the user.
- Consider a user who wishes to publish all available PGDBs on their internal web

site. In this case, use command:

```
pathway-tools -www -www-publish all
```

The full set of possible **pubspec** arguments are:

- **public** -- make public PGDBs visible (excludes restricted PGDBs)
- **all** -- make all PGDBs visible
- **orgA+orgB+...+orgX** -- make the specified set of organisms visible. Each **orgX** is an organism ID.

-gene-link-db db

When a user site sets up a Pathway Tools database in conjunction with a previously existing database of genes for an organism, it is sometimes useful to have references to genes in Pathway Tools web pages link directly to gene pages in the user's pre-existing database, rather than to the gene pages generated by the Pathway/Genome Navigator. To accomplish this, the user must create a Database frame in the PGDB that contains information necessary for linking to the desired external database, and each gene frame in the PGDB must contain a link to the corresponding object in that database. Then, if the database frame ID is supplied as the value of this command line argument, pages generated by the Pathway Tools web server will substitute links to the external database anywhere it would normally link to a gene page.

-no-blast

Ordinarily the web query page contains an option to invoke the BLAST program. If access to BLAST from this page is not desired (e.g. if the BLAST program is not installed, or if such functionality is available elsewhere on a website), then supplying the **-no-blast** option causes it to be removed from the query page. Note that BLAST is disabled by default when the **-gene-link-db** argument is supplied.

-blast

Restore access to the BLAST program from the web query page if it was removed by virtue of the **-no-blast** or **-gene-link-db** arguments being supplied.

-no-google-text-search

By default, the Pathway Tools query page will contain a full-text-search query box, powered by Google™. This capability is only useful if the site can be indexed by Google. If your web server is running on an internal network or cannot be indexed by Google for other reasons, supply this argument to remove the search box.

-email support@site

Specifies the email address to which technical and content-related support questions should be addressed. This address will appear on the Pathway Tools web pages.

2.3 WEB SERVER OPERATION

After installation of Pathway Tools is complete, you can start the Pathway Tools web server on a Unix computer by typing the following. This command starts an active web server running on your computer. The server must be running during the time when you expect users to issue web requests to the server.

pathway-tools -www *args*

The text ***args*** consists of zero or more command-line arguments, as defined under " Command Line Arguments" in section 2.2.1. To specify the organism that will be selected by default in the server.html page, use the **-org** command-line argument. If you are running the web server on your intranet so that it is not accessible by users outside your organization (such as because of firewall protection), you almost certainly want to use the additional arguments **-www-publish all** to instruct the software to make all available PGDBs visible from the server. By default, some PGDBs will not be visible to facilitate compliance with the Pathway Tools license agreement. The preceding section describes the **-www-publish** argument in more detail.

Users can access the Pathway Tools web server at URL

http://hostname:1555/server.html

where hostname is the name of the computer at your site on which the web server is running. The page served by that URL provides a number of different ways of querying the Pathway Tools web server, and documentation on how to use the web server. To access this page with a particular organism preselected (i.e. not the default organism specified using the **-org** command line argument), access **http://hostname:1555/<*ORGID*>/server.html**, substituting the ID for the desired organism for **<*ORGID*>**.

Note that even when web server operation is desired, the Unix environment must still be properly configured for X Windows operation, as described under " X Windows Basics" in section 0. That is, the **DISPLAY** environment variable must have an appropriate value, and an xhost command must have been issued to establish appropriate access permissions. When running in web server mode, we recommend that the X Windows compute host and display host are the same computer. Thus, the X Windows system itself must be actively running on that computer.

When the Pathway Tools web server executes, it will create a window for the Pathway/Genome Navigator on the display host. That window may be minimized (iconified) during operation of the web server, but the window should not be destroyed, nor should you invoke the Exit command from the Navigator.

To stop the Pathway Tools web server type the command **(exit)** in the same window in which the original pathway-tools command was issued. The program will exit back to the Unix shell.

2.3.1 Web Server Log File

The Pathway Tools web server logs its requests in the file **aic-export/ecocyc/genopath/*/logfiles/server.log** or

ecocyc/genopath/*/logfiles/server.log on Microsoft Windows. You may wish to copy the logfiles in this directory to a more permanent location when installing a new version of Pathway Tools.

2.3.2 Customizing the Web Server Pages

The pages served by your Pathway Tools web server will in general look similar to those that appear on the BioCyc.org web site. There are however a few ways in which you can customize their content and/or appearance. For the purpose of these examples we assume that your httpd root directory is /home/htdocs/.

- To link to a page of release notes for a PGDB, create a directory <***ORGID***>cyc (substituting the ***ORGID*** for your PGDB) in your http root directory, if one does not yet exist. Create a file called release-notes.shtml and save it in this directory. For example, if your http root directory is **/home/htdocs/**, and your ***ORGID*** is TEST, you would save release notes to the file **/home/htdocs/testcyc/release-notes.shtml**. Users will see this page when they click on "History of updates to this dataset" in the server.html page.
- To include a small organism graphic to the left of the page header for each data page, save the image file as **/home/htdocs/<*ORGID*>cyc/organism-image.<gif/jpg/png>**, substituting the ***ORGID*** of your PGDB, and using the correct file suffix for your image type. Some web browsers can display text over an image when the user's mouse points to the image; if you want to provide this text for your organism graphic, put the text in file **/home/htdocs/<*ORGID*>cyc/organism-image-text.txt**.
- The style sheet for Pathway Tools web pages is stored in /home/htdocs/style.css, and may be edited. Note that the colors of generated graphics are set to show up best against a white background.
- You may add html text to the beginning and/or end of every generated page by creating files **banner.html** and/or **footer.html** in the directory **aic-export/ecocyc/genopath/9.5/misc/**.
- You may add html text under the title of the server.html page by creating a file **server-announcements.html** in the directory **aic-export/ecocyc/genopath/9.5/misc/**.

2.3.3 Setting up BLAST access

By default, the server.html query page provides the option of performing a BLAST search for some sequence within the genome of a single organism (see the description of the command-line argument **-no-blast** to disable this functionality). The Pathway Tools distribution does not include a copy of the BLAST software, but it can be downloaded from NCBI (see **http://www.ncbi.nih.gov/blast/).** Add the path of the directory containing the executable programs **blastall** and **formatdb** to the Unix PATH environment variable of the user

running the Pathway Tools web server. In addition, a file named **`.ncbirc`** in the home directory of the user running the Pathway Tools web server must contain the following two lines of text (note that the command contains backquotes):

[ncbi]

Data=`pwd`/../data

Once BLAST has been installed, the BLAST sequence databases for each organism must be created before web users can access them. From the **Tools** menu, select **Prepare Blast Reference Data => Both**.

2.3.4 Troubleshooting

If the Pathway Tools web server malfunctions, please check the following:

If the web server is not functioning at all:

- Be sure you have defined the X-windows environment properly
- Be sure not to exit from the Pathway/Genome Navigator window
- Be sure the **`/tmp`** directory is accessible on your computer; the web server writes some temporary files there

If some PGDBs are not visible in the "Select a dataset" selector:

- Please review the **`-www-publish`** argument in the preceding section. It determines which PGDBs are accessible via the web server.

3 PATHWAY/GENOME NAVIGATOR

The preceding chapter describes how to start Pathway Tools, including the Navigator. This chapter describes the basics of how to work with the Navigator and then leads you through several examples, which you should attempt to execute as you read. We will focus on navigating the information contained within the EcoCyc *E. coli* database.

3.1 USING THE MOUSE TO NAVIGATE AND ISSUE COMMANDS

User interaction with the Navigator is primarily mouse oriented. In general, the left mouse button is used to invoke specific commands and for hypertext navigation, whereas the right button is used to bring up menus of additional operations, where available. The middle button has some specialized uses that will be described later. Items that can be clicked on include elements of the command menu, buttons within display panes, pop-up menus, and textual and graphical elements in display windows. You can tell when an item is mouse sensitive: if you move the mouse pointer over a sensitive item, a bounding rectangle appears around the item. In some cases, a mouse-documentation line also flashes at the bottom of the window to explain what will happen if you click on the item. For example, move the mouse cursor over the **Reaction Mode** command of the command menu, and note that a rectangle comes up and surrounds these words. Note also that the following optional mouse operations appear in the bottom pane, **`L: Retrieve reactions; R: Menu`**. This indicates that use of the left mouse button allows you to retrieve one or more reactions from the current database (see below), while use of the right button will bring up a menu that indicates the available options for the current cursor position.

The Navigator may appear to be unresponsive for several reasons (see “Frequently Asked Questions” in Chapter 5). The most common reason occurs periodically and depends on how much memory is available for operation of Pathway Tools. At roughly half-hour intervals, the Navigator pauses to perform a memory management function called 'garbage collection’ that takes approximately 30 seconds to complete.

3.2 MENUS AND DIALOGS

The Pathway/Genome Navigator has a Menu Bar, Single-Choice Menus, Multiple-Choice Menus, and Dialogs, described below. You may encounter Single-Choice Menus, Multiple-Choice Menus, and Dialogs while using the Menu Bar or by right-clicking on the title of a biological object display.

3.2.1 Menu Bar

The Menu Bar at the top of the Navigator window contains a pull-down menu for each major biological object type -- Metabolic **Overview**, **Pathway**, **Reaction**, **Protein**, **RNA**, **Gene**, **Compound**, **Chromosome** -- as well as **File**, **Tools**, and **Help** menus containing general commands. For example, clicking on the **Reaction** menu will reveal a list of query options that

apply specifically to reactions.

3.2.2 Single-Choice Menus

This type of pop-up menu lets you select a single item from a list. As soon as you left-click on an item, the selection is made and the menu vanishes.

3.2.3 Multiple-Choice Menus

This type of pop-up menu allows you to select *one or more* items from a list. Click on as many items as you want; when you have finished selecting, click on the **OK** button at the bottom. If you change your mind and want to deselect an item before you have clicked **OK**, simply click on the item a second time.

3.2.4 Dialogs

A Dialog allows you to answer several questions in one window. Alternative answers are presented, with default answers in boldface. Click on a different answer to override the default. When all questions are answered, click **OK** at the bottom of the menu.

3.2.5 Aborting out of Menus and Dialogs

If the Pathway/Genome Navigator is requesting input from you via a Menu or a Dialog and you want to abort out of it, try one of the following actions:

Click Cancel or No Select.

- Type ^Z (i.e., hold down the control key and press the “Z” key while the control key is still held down).
- Click the top of the Menu or Dialog.
- Click outside the Menu or Dialog.

3.3 ORGANISM SUMMARY DISPLAY

When you invoke the Navigator, the main window contains the Organism Summary display (see Figure 3-1).

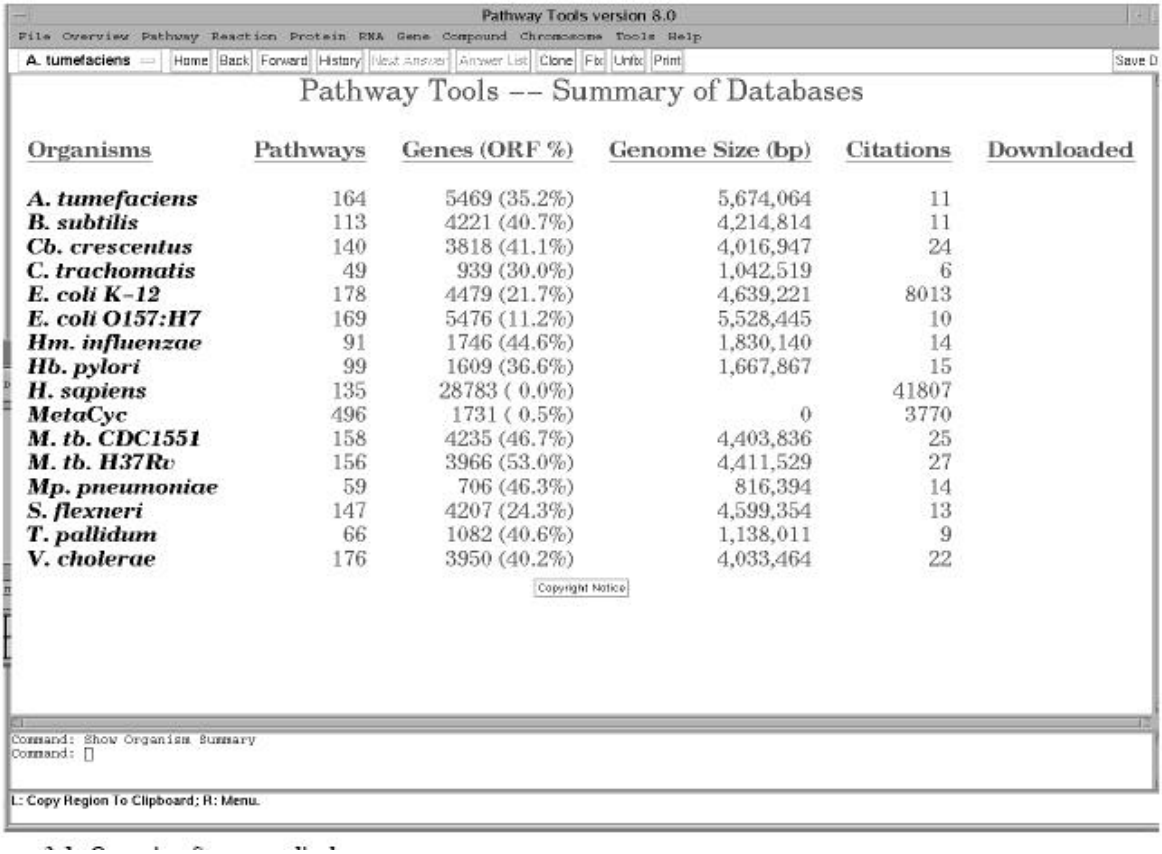


Figure 3-1 Organism Summary display

At the top of the Organism Summary display is the Pathway Tools banner, followed by a list of available databases. The first column lists the names of organisms for which databases are available. To the right are additional columns containing statistics on the database content. From left to right, each row gives the number of pathways (computationally predicted and user created), the number of genes (and the percentage of these that represent open reading frames (ORFs)), the genome size in base pairs, the number of literature citations, and a download date, if available. At the top of the command menu, the Current Organism is listed. By default, the Current Organism is Escherichia coli K12. At the bottom of the page is the *Copyright Notice* command.

To return to the Organism Summary display from any Navigator window, use the command **File -> Summarize Databases** or the **Home** button.

3.4 SINGLE ORGANISM DISPLAY

The Single Organism display summarizes the content of the database for an individual organism (see Figure 3-2).

Figure 3-2 Single Organism display

The Single Organism display for an organism is accessible from the All Organisms display by left-clicking the name of this organism in the *Organisms* column of the All Organisms display. The Single Organism display for the selected organism then replaces the All Organisms display in the main display window. In addition, this organism becomes the Current Organism.

The top of the Single Organism display gives the organism name and the specific strain whose genome and predicted metabolic pathways are available as a database. Directly below is the **Summarize Pathway Evidence** command (available from the Single Organism display of PGDBs, but not for EcoCyc). Clicking on this option takes you to a series of HTML pages that outline the genomic evidence for the metabolic pathways of this organism. The **Replicon** column lists sequenced replicons (chromosomes and plasmids) of the organism; for example, in the case of *S. cerevisiae* (yeast), all 16 chromosomes and the mitochondrial genome are listed. For each such element, summary information on its genes is provided in additional columns, that is, the number of mapped genes, and a breakdown of this number between protein coding and RNA coding genes. The size (in base pairs) is also listed. Clicking on an element name takes you to a window display of the genomic map of that element (see "Genome Browser" in Section 3.8.10). For example, clicking on Chromosome XVI of the *S. cerevisiae* summary display takes you to a displayof the genomic map for chromosome XVI.

3.5 NOTION OF CURRENT ORGANISM

Pathway Tools always keeps track of what it considers to be the current organism for processing. The identity of the current organism is important in two respects. First, all command-mode queries are directed against the database for this organism and, second, for comparative analyses this database serves as the reference for comparison with one or more other user-specified databases. However, when you follow hypertext links within the Navigator (such as by clicking on the name of an enzyme within a pathway drawing), the next object displayed (the enzyme) is always from the same organism as the organism containing the previous object (the pathway). This organism is not necessarily the current organism.

When you first enter the Navigator, by default the current organism is *E. coli*. This means that all queries will be directed against the database for this organism. To change the current organism, select a database for a different organism from the All Organisms display (see above) or go to the top of the command menu where the name of the current organism is listed within the Organism Selector. Left-click once on this name to bring up a listing of all available databases and left-click once on the name of an organism to make it the new current organism. The last organism selected in this way or via the All Organisms displayis the current organism. Its name is displayed in the Organism Selector at the top of the command menu. An asterisk ("*") next to the name of an organism in the Organism Selector means that the DB for that organism has unsaved changes.

To explore the pathway/genome information space of a given organism, it must be the current organism. To follow the examples inthe rest of this chapter, set *E. coli* as the current organism. All the Navigator commands outlined below for navigating, querying, visualizing, and analyzing EcoCyc may also be used to explore any one of the other available databases, except MetaCyc (which does not contain information on objects that are not relevant, for example, replicons).

3.6 EXAMPLES

In the following examples it is assumed that the main Navigator window is now visible on your monitor and that EcoCyc is current. These examples are designed to introduce you to some basic Navigator features that allow you to retrieve and view information from EcoCyc.

Example 1
Invoke the menu command **Reaction -> Search by EC#**, which allows you to look up a reaction by its Enzyme Commission number. A pop-up window appears: type "5.3.1.9" and click **OK**. (If nothing happens when you type, you may have to move the mouse pointer into the pop-up window, and/or click on its title bar).

A display window for the reaction catalyzing the first step of glycolysis will appear. A number of items in this display are mouse sensitive, including the name of the enzyme that catalyzes the reaction, the gene that encodes this enzyme, the pathways containing the reaction, and the compounds in the reaction equation. Click on any of these items to see displays of the respective objects, such as the glycolysis I pathway.

Example 2
Invoke the command **Pathway -> Search by Class**. A pop-up menu gives a classification

hierarchy for metabolic pathways.

Select a class, for example **Amino Acids -> Individual amino acids** (under **Biosynthesis**). A new pop-up menu lists a number of individual pathways for amino acid biosynthesis (the instances of this class). Click on one of the pathways.

A drawing of the selected pathway appears. Virtually every item in the drawing is mouse sensitive. For example, you can click on a reaction arrow to see a display of that reaction, and you can click on an enzyme name to see a displayof that enzyme.

Example 3
Invoke the command **Protein -> Search by Substring**. In the pop-up menu, type "pyruvate" and click **OK**. A second pop-up menu lists all proteins (including enzymes) contained in EcoCyc whose name (or one of its alternate names) contains the substring pyruvate. Click on any protein name, and then click on **OK** at the bottom of the menu, to display that protein.

Scroll downward in the pane for this protein (scroll by clicking on the small arrow at the bottom of the scroll bar). Click on one of the gene names. When the gene display appears, click on the map position of the gene (assuming your gene has a map position).

When the genome browser appears, you can zoom in on a region of interest by using the navigation bar (see "Genome Browser" in Section 3.8.10). You can obtain detailed information on a gene by left-clicking on a gene name.

3.7 QUERY FACILITIES

When searching for a particular piece of information about *E. coli* metabolism, we can usually try two different strategies to find that information: a direct and an indirect approach.

Using the direct approach, we issue queries for the entity we seek. For example, imagine that we seek information on the *hisA* gene, such as its map position and the name of the enzyme it encodes. The Navigator allows you to call up an information window for a gene by its common name (or by other names by which that gene is known).

Using the indirect approach, we bring up the information window for an object by first issuing a direct query for a related object, and by then navigating to the object of real interest. For example, imagine that we had forgotten the name of the *hisA* gene, but we knew it encoded the enzyme that catalyzed the last step in the biosynthesis of histidine. We could use a direct query to display the biosynthetic pathway for histidine, and then click on the name of the enzyme catalyzing the last reaction in the pathway. The resulting information window for that enzyme names the gene (*hisA*) encoding the enzyme. Clicking on the gene name displays the information window for *hisA*.

In summary, by using the Pathway/Genome Navigator we can traverse many paths through EcoCyc to arrive at the same information.

3.7.1 Direct Queries

To query a given type of object, you must select from the menu associated with that class of object. The object type menus are

Overview

Pathway

Reaction

Protein

RNA

Gene

Compound

Chromosome

Each menu contains a set of type-specific predefined queries. For example, from the compound menu, you can query compounds by any of these criteria:

An exact compound name or ID

A substring within a compound name

A chemical substructure specified using the SMILES chemical notation [17] (all compounds containing that substructure will be returned as the result of the query)

A compound class chosen from a menu-based classification hierarchy of compounds (such as amino acids, carbohydrates, and nucleotides)

An advanced search that includes some combination of name, molecular weight, chemical formula and substructure.

Note that most commands that allow you to query objects by their exact name allow you to enter in several names within one pop-up window, separated by commas, for example, "hisA, hisB, hisC." The exception is the compound menu, because many compounds have commas within their names.

3.7.1.1 Queries with Multiple Answers

Some queries return more than one object as the result. For example, most modes allow you to query objects by a substring search, such as searching for all proteins with "pyruvate" in their names. In this instance, the Navigator creates a menu that lists the proteins satisfying the query. The menu allows you to select one, some, or all of the proteins — click on individual protein names to select them, or click on **Select All** to select all of them (clicking on an already-selected name will deselect it). When you have finished selecting, click on **OK**.

The Navigator immediately displays one or more of the proteins you selected (see "Pane Layout" in section 3.9.11.1 to find out how to divide the main display window into multiple panes; see "History and Answer Lists" in section 3.9.11.9 for information on how to control the number of objects displayed by **Next Answer**). It remembers the others on a list called the Answer List.

When you want to see the next protein on the Answer List, click on **Next Answer** in the command menu. For more information about the Answer List, see section 3.9.8.6.

3.7.2 Indirect Queries: Navigation

The information window for each object usually lists a number of related objects. For example, a gene display shows the product of that gene; if the product is an enzyme, then the reaction(s) catalyzed by the enzyme is listed, as are the pathways in which that reaction occurs. Similarly, a compound display lists the reaction(s) that produce and consume that compound, and the pathway(s) in which the compound is found.

Each of these related objects is "live" in the sense that clicking on the object displays an information window for that object. Objects are color coded by type to make their relationships more evident, and to make it more obvious which visual elements within a complex display are mouse sensitive. A bold-face font is used instead for mouse-sensitive objects, when monochrome monitors are used.

On occasion, nothing happens when you click on a related object. The reason is most likely that EcoCyc has no information about that object, although that incompleteness should be remedied in a future version.

3.7.3 History List

The Navigator keeps a list of the last few objects that you have displayed. This list is called the *history list.* You can return to a previously displayed object by clicking on the **Back** button in the command menu. For example, if there are three items on the history list, clicking on **Back** three times returns you to the display you started with. Clicking on **Forward** moves you through the history list in the other direction. To select one or more arbitrary objects from the history list, click on the **History** button in the command menu. You are presented with a multiple choice menu of every item on the history list. To see the current history list, select **Tools -> History -> Show on Console**.

By default, the history list contains as many as fifty items. This length can be changed (see "History and Answer Lists" in section 3.9.11.9).

When you exit the Navigator, your history list is saved in a file called **`.ecocyc-history`** in your home directory. This feature enables you to easily start up again at the same place where you left off.

3.7.4 Programmatic Queries

The examples below of different complex queries against Pathway/Genome Databases are written in the Common Lisp programming language. To write programmatic queries, you must understand a number of aspects of Pathway/Genome Database schemas, such as class and slot names. The current schema is described in Appendix A.

The Preface lists a variety of additional reference sources relevant to writing Lisp queries, including a longer set of example queries that are available through the SRI Web site.

One convenient way to examine the answer to a query is to put the result of the query on the Answer List of the Pathway/Genome Navigator, and to look at each answer using the **Next Answer** command. The first query below shows how to do so.

```
;; Find genes located between 20 and 30 centisomes on the map
(loop for gene in (get-class-all-instances '|Genes|)
      for pos = (get-slot-value gene 'centisome-position)
      when (and pos (> pos 20) (< pos 30))
      collect gene)


;; The preceding query returns a list of genes. To run the
Pathway/ ;; Genome Navigator with those genes on the Answer
list,
;; evaluate the following. The "*" means "the result returned by
;; the last expression evaluated".
(eco :answer-list *)


;; Find reactions involving pyruvate as a substrate
(loop for rxn in (get-class-all-instances '|Reactions|)
      when (member 'pyruvate (get-slot-values rxn 'substrates))
      collect rxn)


;; Find all genes whose products catalyze a reaction involving
;; pyruvate as a substrate
(loop for rxn in (get-class-all-instances '|Reactions|)
      for genes = (genes-of-reaction rxn)
      when (member-slot-value-p rxn 'substrates 'pyruvate)
      append genes)


;; Find all enzymes that use pyridoxal phosphate as a cofactor
;; or prosthetic group
(loop for protein in (get-class-all-instances '|Proteins|)
      for enzrxn = (get-slot-value protein 'enzymatic-reaction)
      when (and enzrxn
                (or (member-slot-value-p enzrxn
                           'cofactors 'pyridoxal_phosphate)
                    (member-slot-value-p enzrxn
                           'prosthetic-groups
```

```
'pyridoxal_phosphate)
                    ) )
      collect protein)
```

3.8 OBJECT DISPLAYS AND QUERIES

Specific queries can be issued in each command mode and result in different object displays.

3.8.1 Shared Display Characteristics

Certain aspects of the object displays are shared by most or all of the different object classes.

3.8.1.1 Gene-Reaction Schematic

The many-to-many relationships among genes, enzymes, and reactions can be complex. An enzyme composed of several subunits might catalyze more than one reaction, and a given reaction might be catalyzed by multiple enzymes. The *Gene-Reaction Schematic* depicts the relationships among a set of genes, enzymes, and reactions (see Figure 3-3 and

http://ecocyc.org/new-image?type=REACTION-IN-PATHWAY&object=DAHPSYN-RXN).

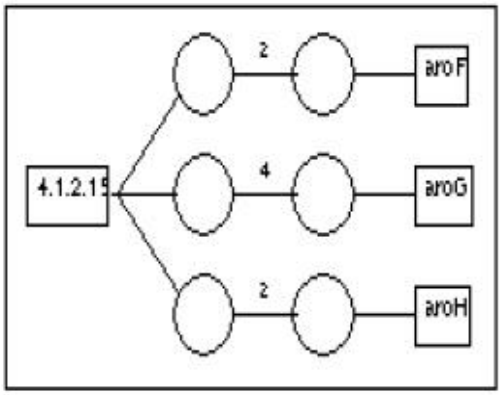


Figure 3-3 A gene-reaction schematic

It is drawn in reaction displays, protein displays, and gene displays. It is generated by starting with the object that is the focus of the current display (which is highlighted in the schematic), and then recursively traversing database relationships from that object to related objects, such as from a gene to its product, or from a reaction to the enzyme(s) that catalyzes it. The schematic summarizes these complex relationships succinctly, and also constitutes a navigational aid: click on an object in the schematic to cause the Navigator to display that object.

In gene-reaction schematics, the boxes to the left represent reactions, the boxes on the right represent genes, and the circles in the middle represent proteins. The lines indicate relationships among these objects. For example, the schematic in Figure 3-3 means that the aroF gene encodes a polypeptide (the circle to the left of the box for the aroF gene) that forms a homodimer (the

next circle to the left — the 2 indicates two copies) that in turn catalyzes reaction 4.1.2.15. The situation for the aroG gene is similar except that its product forms a homotetramer. Gene-reaction schematics that depict heteromultimers show more than one gene-product circle connected to a single circle for a protein complex. Gene-reaction schematics also include modified forms of a protein (or tRNA) when relevant. For example, the schematic for the acyl carrier protein shows a yellow circle for the unmodified form of the protein, and it shows 19 orange circles, which represent different modified forms of the protein.

Gene-reaction schematics should *not* be confused with pathway diagrams; although both diagrams are graphs, they mean different things.

3.8.1.2 Citations and Comments

Because citations and comments are found in all types of objects within organism databases, we begin by discussing them.

Comments authored by database curators are found in several locations in object displays. In some cases, citations are presented in the same manner as are comments. Consider a line from an enzyme display such as

Inhibitors (allosteric) [3]: NADH [4], succinate [2,Comment1]

This line is identifying NADH and succinate as allosteric inhibitors of the enzyme being shown. The **[4]** indicates that EcoCyc contains a citation that pertains to the fact that NADH is an allosteric inhibitor of the enzyme. When the mouse is over the "4", the pointer documentation window at the bottom of the screen shows the start of the citation information. Clicking on the "4" navigates to where the full citation information is displayed in the References section at the bottom of the page. Clicking on the full citation information displays the citation in a pop-up window or, if the citation is available through PubMed, in a Web browser window. Analogously, a comment and citation pertain to the role of succinate as activator or inhibitor: **[2,Comment1]**. Passing the mouse over the "Comment1" shows the start of the comment in the pointer documentation window, and clicking on it brings up the full comment in a pop-up window. The **[3]** is a more general citation about the inhibition of the enzyme that does not pertain precisely to NADH or succinate alone.

Citations of general relevance to an object (as opposed to citations that pertain to a particular property or data value) are shown on a separate line in the object display window, such as

Citations: [1,2,3,4]

Note that citation indicators can be either numeric or mnemonic. To choose the style of citation indicator you prefer, use the command **Tools -> Preferences -> Citation Reference Style**.

3.8.1.3 Database Links

Objects within EcoCyc contain links to a variety of other databases. For example, some *E. coli* polypeptides are linked to both SwissProt and PDB entries; some *E. coli* genes are linked to entries in the Coli Genetic Stock Center database. These databases are not part of, or provided together with, Pathway Tools and may require a license from their owners or distributors to

access their contents.

In general, databases contain two types of links: *unification* and *relationship* [4]. Unification links are links to descriptions of the *same object* in the other database, and are displayed in a line that lists one or more links, where each link displays the name of the other database, and the unique identifier of the target object in that database, such as

```
Unification links: SwissProt:P34554
```

Relationship links refer to a *related object* in the remote database. Because a number of possible relationships might exist between the source object in EcoCyc and the target object in the other database, the line describing relationship links displays the name of the foreign database, the ID of the target object, and the name of the relationship to the target object. For example, some *E. coli* polypeptides are linked to PDB homologs if PDB does not describe the structure of this exact *E. coli* protein:

```
Relationship links: PDB:Homolog:P34554
```

When you click on the target identifier, the Web browser displays the target object (clicking on the target identifier will invoke Netscape and then cause Netscape to display the target object).

3.8.1.4 Classes

EcoCyc contains a number of taxonomic hierarchies. Reactions are classified according to the Enzyme Nomenclature system [16]. Genes are classified according to a system devised by Riley [13]. Compounds and pathways are also classified. All the classification systems are multilevel, and involve a number of instances (such as reactions) that are assigned to a number of classes (such as those defined by the Enzyme Nomenclature system). Many of these classes in turn are assigned to superclasses (which may recursively be assigned to additional superclasses).

Each object display shows the parent class(es) of that object. Clicking on a parent class displays the parent class. The display window for a class lists both its parent classes and its instances, if any. You can also click on these parent classes, or instances, to display them. In this manner, the user can navigate the *E. coli* taxonomic hierarchies.

3.8.2 The Cellular Overview

The Overview diagram is a representation of all metabolic pathways and reactions, signaling pathways, membrane proteins and transporters defined for the current organism. In this diagram, each icon (circle, square, ellipse, etc.) represents a single metabolite. The shape of the icon encodes the chemical class of the metabolite, as listed in Table 3-1.

Icon Shape	Compound Class
square	carbohydrate
triangle	amino acid
upside down triangle	cofactor
ellipse (horizontal)	purine
ellipse (vertical)	pyrimidine
diamond	protein
T-shape	tRNA
circle	all other compound types

Table 3-1 Compound shapes used in the Overview Diagram

The shading of the icon indicates the phosphorylation state of the compound: shaded compounds are phosphorylated; unshaded compounds are unphosphorylated.

Each thick line in the Overview diagram represents a single bioreaction. Neither the icons nor the lines are unique in the sense that a given metabolite or a given reaction may occur in more than one position in the diagram.

The "barbells" along the right side of the diagram represent individual reactions that have not been assigned to a particular pathway. They are presented as single reactions because their direction and role are determined by the metabolic condition of the cell. The barbell region also contains some reactions of macromolecule metabolism, such as DNA metabolism. In the region to the left of the barbells, the glycolysis and the TCA cycle pathways in the middle separate predominately catabolic pathways on the right from pathways of anabolism and intermediary metabolism on the left. The existence of anaplerotic pathways prevents rigid classification. The majority of the metabolic pathways operate in the downward direction. Pathways are generally grouped by class, and the extent of a class is indicated by background shading.

The border drawn around the Overview depicts the cytoplasmic membrane, and contains embedded transport proteins. Transported substrates use the same shape codings as for metabolic substrates. Where possible, transporters are positioned in the membrane so as to be near some of the metabolic reactions into which their substrates feed.

In the EcoCyc Overview, both the inner and outer membranes are shown. Periplasmic reactions and proteins are depicted in the space between the two membranes at the right of the diagram.

You can interrogate the Overview in several ways. To identify a compound within the Overview, move the mouse pointer over a geometric figure in the diagram — the Navigator will print the name of the metabolite and the name of the containing pathway at the bottom of the screen. To identify a reaction, move the mouse pointer over a thick line — the Navigator will print the

equation of that reaction, and the name of the containing pathway. If the mouse pointer is moved over a shaded but blank region, the Navigator will print the name of the pathway class.

If you left-click on a compound or a reaction in the Overview, that object is displayed in its own display window.

Conversely, if you are looking at another display that contains a link to a compound, reaction or pathway, you can right-click over the link to bring up a short menu of operations. One of those operations is "Show compound/reaction/pathway in overview". Selecting that menu item causes the Overview diagram to be drawn, with the designated entity highlighted.

If you middle-click on a reaction in the Overview, its enzymes and genes are listed in the listener window. The enzyme and gene names are also left-clickable, to allow you to navigate to the displays for those objects.

3.8.2.1 Overview Menu

- **Show Overview**: Draw the Overview diagram for the current organism.
- **Show Key:** Displays a pop-up window containing a key for the Overview diagram that explains what compound classes are denoted by each node shape, and explains what the different highlighting colors represent.
- **Show/Hide Transport Links**: Toggle whether or not faint lines are shown that connect transported substrates to the pathway(s) in which they participate. By default, these links are hidden.
- **Highlight:** You can request that some entity be highlighted in the Overview. You can request that a compound, a reaction, a pathway, an enzyme (i.e., the reaction(s) catalyzed by the enzyme) or a gene (i.e., the reaction(s) catalyzed by the product of a given gene) be highlighted. The object to be highlighted can be specified in a number of ways (e.g., by name, by substring, by EC number) using a set of cascading menus that reproduce many of the query capabilities present in the other command modes. Specifically, the highlighting commands are
  - **Species Comparison** (see "Comparative Operations" in Section 3.10)
  - **Pathway**
    - By Name or Frame ID
    - By Substring
    - By Class
    - All by Class (colors all pathways according to their role, for example, all amino-acid biosynthetic pathways are in one color)
    - By Genome Clustering (colors pathways according to the clustering within the genome of the genes that encode their enzymes — the accompanying pop-up window describes the color scheme in more detail)
  - **Reaction(s)**
    - By Enzyme Name
    - By Enzyme Substring

- By EC Number (e.g., reaction with EC number 1.2.3.4)
- Using EC Class Hierarchy (e.g., all reactions in class 1.2)
- All by Top-Level EC Class (colors the entire diagram to reflect the chemical type of each reaction)
- From File of Reaction Designators (takes as input a file containing EC#s, one per line)
- All without EC Numbers (of which there are many in EcoCyc, because the Enzyme Commission has yet to assign EC numbers to many reactions)
- By Pathway (e.g., all reactions that occur in the TCA cycle)
- By Substrates (this option allows you to specify a full or partial list of the reactants and products of the reaction)
- By Effects of Compound(s) on Enzyme Activity (this option allows you to highlight reactions according to modulation of the enzyme(s) that catalyze the reaction, e.g., to highlight all reactions whose enzyme is activated by ADP)
- By Enzyme Cellular Location (highlights reactions whose enzymes are known to be located in a user-selected cellular location)
- All with Multiple Isozymes
- All in Multiple Pathways (highlights reactions that occur in more than one pathway)

o **Gene**

- By Name or Frame ID
- By Substring
- By Class
- By Genetic Regulator Protein (allows you to select a transcription factor, and highlights all reactions whose genes are in operons that are regulated by that transcription factor)
- Gene List From File (the file should contain a list of gene names, one per line)
- All by Replicon (colors reactions according to the replicon — chromosome or plasmid — on which their genes are located)

o **Compound(s)**

- By Name or Frame ID
- By Substring
- By SMILES Structure
- Using Class Hierarchy (e.g., all amino acids)

o **Undo**: Unhighlights the last item or set of items that were highlighted in the Overview.

o **Redo**: Rehighlights the last item or set of items that had just been removed via the Undo command.

o **Clear All**: Unhighlights all items in the Overview.

o **Save to File**: Saves a given pattern of overview highlighting on to a file.

- **Load from File**: Inputs a file created using the Save to File command. This restores the specific pattern of overview highlighting previously saved to this file.

- **Omics Viewer: Overlay Experimental Data from:** Experimental data, such as gene expression, proteomics, reaction flux or metabolomics data, is read from a file, and reactions and/or compounds are colored according to the experimental values (absolute or relative) associated with the corresponding genes, proteins, reactions or compounds. See section3.8.2.3 on viewing experimental data using the Omics Viewer. The two options are:
  - **Text File**
  - **SAM Output File**

- **Update:** Regenerate the Cellular Overview diagram to reflect any changes in the database. This command is disabled if the database cannot be modified. Note that this operation can take quite a long time -- from several minutes up to an hour or so, depending on the speed of the computer and the number of pathways in the database.

Additional operations are accessible through the right mouse button. Right-clicking on a compound gives you the choice of accessing either a menu for the compound or a menu for the pathway. Right-clicking on a reaction gives you the choice of accessing a menu for the reaction, for the pathway, or for any of the compounds involved in the reaction (including those that are not displayed in the overview because they are side compounds). Right-clicking on any object also gives you the opportunity to zoom in or out of the overview. You may choose from one of several predefined magnification levels or you may specify your own. Zooming changes the scale of the display, but does not cause any additional information to be printed.

The compound, reaction, and pathway menus are all described below.

Right-Button Compound menu:

- **Display compound information in main display**: Displays the selected compound in the main display window.
- **Display compound information in pop-up window**: Displays the selected compound in a new pop-up window.
- **Highlight all reactions of this compound**: Highlights all reactions that contain this compound as either a main substrate or a side substrate. This command is a more complete way of finding reactions of a compound than is the next command.
- **Highlight this compound everywhere it appears as a main**: Highlights all occurrences of this compound as a main substrate only.
- **Display all connections for this compound**: Highlights all occurrences of this compound as a main substrate, and draws dim lines connecting the clicked-on compound to all other occurrences of that compound, including situations in which the compound is a side substrate of a reaction (in this latter case, there is no icon for the selected compound, so the line is drawn to some other main compound icon in the relevant reaction). The lines are removed when the highlight operation is cleared or undone. These

new lines are drawn slightly brighter than those used to link identical compounds to show flow of material between pathways.

- **Show**: Invokes the usual right-button Show submenu for the compound. This submenu is described in section 2.4.3 in volume II of the Pathway Tools User Guide.
- **Edit**: Invokes the usual right-button Edit submenu for the compound. This submenu is described in section 2.4.3 in volume II of the Pathway Tools User Guide.

Right-Button Reaction menu:

- **Display reaction information in main display**: Displays the selected reaction in the main window.
- **Display reaction information in pop-up window**: Displays the selected reaction in a new pop-up window.
- **Highlight this reaction everywhere it appears**: Highlights all occurrences of the selected reaction in the Overview.
- **Show enzymes and genes of this reaction in listener window**: Prints the names of the enzymes that catalyze this reaction, and the genes that encode those enzymes.
- **Display all connections for substrates of this reaction**: Draws dim lines connecting the substrates of the clicked-on reaction to all other occurrences of those compounds.
- **Display all connections for reactants of this reaction**: Draws dim lines connecting the reactants of the clicked-on reaction to all other occurrences of those compounds.
- **Display all connections for products of this reaction**: Draws dim lines connecting the products of the clicked-on reaction to all other occurrences of those compounds.
- **Highlight reactions involving genes in same operon/regulon**: Highlights in one color all genes that are in the same operon as the gene whose enzyme catalyzes the selected reaction; highlights in a second color all genes that are in the same regulon as the gene whose enzyme catalyzes the selected reaction. If the selected gene is in more than one regulon, you are asked to select the transcription factor defining the regulon of interest (a regulon is defined as the set of operons regulated by a specified transcription factor).
- **Show**: Invokes the usual right-button Show submenu for the reaction. This submenu is described in section 2.4.3 in volume II of the Pathway Tools User Guide.
- **Edit**: Invokes the usual right-button Edit submenu for the reaction. This submenu is described in section 2.4.3 in volume II of the Pathway Tools User Guide.

Right-Button Pathway menu:

- **Display pathway information in main display**: Displays the pathway containing the selected compound or reaction in the main window.
- **Display pathway information in pop-up window**: Displays the pathway containing the selected compound or reaction in a new pop-up window.
- **Highlight this pathway**: Highlights the pathway containing the selected compound or

reaction.

- **Display all connections for compounds in this pathway**: Draws faint lines connecting most compounds in the pathway to all other occurrences of those compounds, including situations in which the compound is a side substrate of a reaction (in this latter case, there is no icon for the selected compound, so the line is drawn to some other main compound icon in the relevant reaction). Compounds in the pathway for which these lines are not drawn are either common small molecules (e.g., phosphate, water) or compounds whose function in the pathway is purely to act as a donor or acceptor of some small moiety (e.g., ATP as a donor of phosphate, glutamine as a donor of nitrogen, acetyl-CoA as a donor of CoA). The lines are removed when the highlight operation is cleared or undone. Lines from main compounds in the pathway are drawn slightly brighter than those used to link identical compounds to show flow of material between pathways; lines from side compounds in the pathway are drawn slightly dimmer.
- **Show**: Invokes the usual right-button Show submenu for the pathway. This submenu is described in section 2.4.3 in volume II of the Pathway ToolsUser Guide.
- **Edit**: Invokes the usual right-button Edit submenu for the pathway. This submenu is described in section 2.4.3 in volume II of the Pathway Tools User Guide.

3.8.2.2 Displaying Reactions Corresponding to a Set of Genes

To highlight a set of reactions corresponding to some gene set (such as reactions catalyzed by a set of essential genes or knockout genes), select menu **Overview -> Highlight -> Gene -> Gene List from File**. This menu selection highlights all metabolic reactions catalyzed by the product of a gene listed in the file.

3.8.2.3 The Omics Viewer: Using Overview to View Experimental Data

Note: In earlier versions of the software, the Omics Viewer was known as the Expression Viewer. It has since been extended to display many other types of experimental data, not just gene expression data, so has been renamed.

The Pathway Tools Omics Viewer uses the Cellular Overview for an organism to illustrate the results of high-throughput experiments in a global metabolic pathway context. Genes (in the case of a gene expression experiment) and proteins (in the case of a proteomics experiment) that are involved in metabolism are mapped to reaction steps in the Cellular Overview, and the range of data values in a given experimental dataset is mapped to a spectrum of colors. Reaction steps in the Cellular Overview are colored according to the corresponding data value. Similarly, for metabolomics experiments, compound nodes are colored according to the data value for the corresponding compound. This facility enables the user to see instantly which pathways are active or inactive under some set of experimental conditions.

The Omics Viewer can be used for:

- **Microarray Gene Expression Data:** Reaction lines (and protein icons, where present) are color-coded according to the relative or absolute expression level of the gene that codes for the enzyme that catalyzes that reaction step. The Omics Viewer allows a

scientist to interpret the results of gene-expression experiments in a pathway context.

- **Proteomics Data:** Reaction lines (and protein icons, where present) are color-coded according to the concentration of the enzyme that catalyzes that reaction step.
- **Metabolomics Data:** Compound icons are color-coded according to the concentration of the compound.
- **Reaction Flux Data:** Reaction lines are color-coded according to reaction flux values.
- **Other Experimental Data:** Any experiment, high-throughput or otherwise, in which data values are assigned to genes, proteins, reactions or metabolites can be viewed in a pathway context using the Omics Viewer.

The Omics Viewer can show absolute data values (such as the concentration of a metabolite or protein, or the absolute expression level of a gene), or it can be used to compare two sets of experimental data by computing a ratio and mapping the ratios onto a color spectrum. Multiple sets of experimental data can be superimposed on the same overview diagram so that users can, for example, combine gene expression and metabolomics in the same figure, or view the results of two different microarray experiments together. When combining multiple datasets, users should be careful to assign color schemes that avoid ambiguity.

The superposition of multiple sets of experimental data on the metabolic overview can also be animated to show, for example, how gene expression levels of enzymes change with time over the course of an experiment. The animation can be exported to HTML so that it can be published on the Web.

After displaying Omics data on the Cellular Overview, navigating to any pathway display will show the Omics data superimposed on the individual pathway. If a particular reaction step has multiple isozymes then, rather than just choosing one value as is done on the Cellular Overview, all values are shown. Some colors may not show up as well against the pathway background (white by default) as they do against the gray Omics Viewer background – if this is a problem, either customize your colors to choose those that show up well against white, or use the Preferences menu to choose a gray background instead. To remove Omics data from individual pathway displays, select **Overview -> Highlight -> Clear All**.

3.8.2.4 Omics Dataset File Format

Experimental data is imported from a file that is provided by the user and is stored on the user's computer. Each line of the file contains data for a single gene, protein, reaction or metabolite, and is of the form

<name-or-ID> <data-column1>... <data-column N>

Columns are separated by the **Tab** character. The first column contains the name,a known synonym for the name, or the unique identifier of a gene, protein, reaction or metabolite. Gene IDs from sequencing projects (such as the *E. coli* B-numbers) are generally acceptable and unambiguous. For protein or reaction data, EC numbers may be used.

The numbers in the data columns can represent either absolute or relative data values. If the numbers represent absolute numbers, then the software can compute relative values from two

data columns. An entry (a row of data for an entity) may contain any number of data columns, but only the specified (one or two for a single experiment display, more for an animation) data columns can be visualized at one time. Reusing the same file in consecutive viewings improves display performance, as the file needs to be read only once.

Lines that start with either **#** or **;** are taken to be comments and are ignored by the program. The software uses the first row of data (i.e., the first line that is not a comment line) to determine the number of data columns to process. For example, if the first row contains five columns, only the first five columns of each subsequent row will be processed. Thus, even if not all fields for the first row contain data, you must make sure that it contains the appropriate number of **Tab** characters.

3.8.2.5 Using gene expression Data from a SAM spreadsheet

The Omics Viewer can import gene expression data from a spreadsheet generated by the SAM (Significance Analysis of Microarrays) Microsoft Excel plug-in (see **http://www-stat.stanford.edu/~tibs/SAM/**). This package combines multiple expression experiments to produce a list of statistically significant positively and negatively regulated genes. The Omics Viewer displays the positively regulated genes in one color, and the negatively regulated genes in another color. In order to import data generated using SAM, the SAM results sheet must be saved out from Microsoft Excel in text format (not in Excel format). You can then select the SAM option for input to the Omics Viewer, and supply the filename.

3.8.2.6 Color Scales

By default, the color scale used depends on the type and range of the data. Thus, for example, a particular color may correspond to one gene expression level for one data set, and a different gene expression level for another data set, depending on the range of values in each data set. Alternatively, you can specify a color scheme to use, either by simply specifying a maximum color cutoff, or by supplying a list of cutoff values and their corresponding colors. By default, we use the spectrum from yellow/green to red, with yellow representing the lowest values or ratios in the data set, and red representing the highest. Reactions for which no data was provided are shown in the Overview Key, which pops up automatically when experimental data is displayed (if it is not already visible).

For absolute data values, the spectrum is mapped evenly along a log scale between the log of the smallest value in the data set to the log of the largest value in the data set. (The key shows actual values, however, not their logs, unless the supplied data is in log form to begin with.) For relative data values, the ratio is fixed at 1 (or 0, for log values) at the center of the scale.

In many cases, several genes or enzymes, each with its own expression level or concentration, will map to a single reaction. This is because the reaction might be catalyzed by an enzyme complex made of several gene products, or the reaction might be catalyzed by several isozymes, each with its own gene or genes. Since a reaction can be shown in only a single color, the following method is used to choose which of several values to display. For absolute data values, the maximum value is displayed; for relative data values, the value whose log has the greatest deviation from zero is displayed, under the assumption that the user is primarily interested in identifying the genes whose expression levels differ most between the two data sets.

3.8.2.7 Usage

To create an experiment view, select **Overview -> Omics Viewer: Overlay Experimental Data from**, and specify whether data will be from a text file or from a SAM output file. A dialog box pops up with the following fields (assuming data is from a text file — only a filename is needed for SAM output):

- **Experiment Title:** Enter a few words to describe the data being viewed. The title is displayed in the appropriate section of the Overview Key. This field is optional.
- **Type of display:** Select either **Single Experiment** or **Animation**.
- **Type of data:** Select either **Relative** or **Absolute**. Some of the later options in the dialog box, as well as the color scale used, will depend on this choice.
- **File:** Click on the button to select a file containing the experimental data. The **Help** button describes the required file format.
- **Reload?:** This option is available only if the specified file has already been loaded (i.e., if you previously displayed data from the same file). If not selected, the previously cached data is used. Select this option if the file has changed since it was last read.
- **Column zero contains:** The data file format requires that the first column (column zero) contain either gene, protein, reaction or compound names or IDs. Specify which of these your data file contains. If your data file contains multiple types of data, you may select the last option, which allows column zero to contain any of the above types of data. Care must be taken when mixing different data types in the same file, however, as data values for one type of object may not be directly comparable to data values for a different object type. In addition, names which are unambiguous when looking only for a gene, for example, may become ambiguous when the object can be either a gene, protein or compound.
- **Superimpose on previous display?:** This item is available only when the Omics Viewer is already being used to display another dataset. If this box is checked, then the new experiment view will be superimposed on the old one. If this box is unchecked (the default), then the previous display will be cleared before the new data is shown.
- **Use data from:** This item is available only when relative data is displayed. If the ratios themselves appear in a column in the data file, select **single data column**. If you want the ratios computed from columns of absolute data values, select **ratio of 2 data columns**.
- **Use expression data from data column:** (when using a single column) Enter the column number containing the data to be viewed. The column containing the gene or protein names is considered to be column 0, so the data column numbering starts at 1. When displaying an animation, you should enter a list of columns, one on each line. Alternatively, you can enter a range — for example, 1-10 — to indicate that all columns in that range should be used.
- **Compute relative data values as the ratio of data column __ / column __:** (when using two columns) Enter the two column numbers containing the data to be compared. When displaying an animation, for the denominator you can specify either a single

column, which will serve as the denominator for each numerator column, or you can specify a different denominator column for each numerator column.

- **Are data values log values, or do they use a zero-centered (as opposed to 1-centered) scale?:** Check this box when the data values are logs, and leave it unchecked when the data values are linear values. Also check this box if the scale comprising the data is centered at zero (as opposed to 1), for example when displaying reaction flux data in which the sign of each datum indicates reaction direction. If linear values are used, any zero or negative values in the data are ignored.
- **Highlighting Color Scheme:** Here, you can elect to use either the default color scheme or one you specify.
  - Default color scheme with the maximum bin cutoff computed from the data: This option is the simplest to specify. It is useful when you do not know much about the range of values in your data file or you want to see the color spectrum divided evenly across the full range of the data. Since this option will produce a different color scheme for every experiment, it is not useful for comparing figures generated across multiple experiments (though every time point in a single animation will use the same color scheme, of course).
  - Default color scheme with a specified maximum value bin cutoff: All values above the maximum cutoff (or below the corresponding minimum cutoff) are displayed in the same color, and the full color spectrum is divided evenly over the space between the maximum and minimum cutoffs. You can specify a maximum cutoff either by selecting one from the menu of commonly used possibilities or by typing in a value yourself. If you specify the same maximum cutoff for multiple experiments, all will be displayed using the same color scheme.
  - Specify color value cutoffs, assign colors automatically: To exercise more control over the display, you can provide the full list of cutoff values. This is useful, for example, if you are interested in grouping the data into only a few broad categories (e.g., 2x over/ under-expressed, 10x over/under-expressed) and are not interested in finer gradations. Enter a list of numbers, one per line (e.g., 10, 2, 0.5, 0.1).
  - Specify color value cutoffs, assign colors manually: This option gives you maximum control over the appearance of the resulting display. After providing the value cutoff numbers, click on **Assign colors manually**. A color selector box appears containing a scale computed from your supplied cutoff values, and a color spectrum. Assign colors to cutoff bins by clicking on the button corresponding to a bin and then clicking on the desired color. Click **OK** when done.
  - Once you have created a color scheme, you can elect to save it to a file, so you can later retrieve and reuse it. Use the **Save Color Scheme Parameters** and **Retrieve Saved Color Scheme Parameters** buttons for these purposes.

There may be a pause while the data is read and processed. Once processing is complete, the Overview is redrawn to show the experimental data, and a report window pops up containing a few statistics about the data and reporting any problem rows in the data file. Note that in whole genome microarray experiments, typically a large fraction of a given genome does not code for enzymes and therefore will not have any corresponding reaction in the Overview. Statistics are

provided both for the data set as a whole and for the subset of data shown painted onto the Overview. For an animation, a new window pops up containing the animation display, along with buttons to start, stop, and step through the animation. No reports are generated for animations.

The Overview Key, in addition to the legend for mapping colors to data values, also includes a histogram showing the distribution of data values across the range, by color. The range is broken down into 50 subranges. Histogram bars to the left of the central axis count the genes or other entities that actually appear in the Overview. Bars to the right of the central axis count the genes in the remainder of the data set (i.e.,those not in the Overview).

For gene expression experiments, to see exactly what expression values correspond to a colored reaction, middle-click on the reaction. The expression value is displayed in parentheses after each gene name in the listener window. This capability is particularly useful when a reaction is catalyzed by several isozymes, to see the expression level for each individual isozyme.

3.8.3 Pathways

The Pathway/Genome Navigator produces automated drawings of biochemical pathways that are familiar to biochemists — the drawings that mimic those found in biochemistry textbooks. Because the power of computer graphics exceeds that of the printed page in several respects, the automated drawings provide more flexibility than those found in textbooks by allowing pathways to be expanded, contracted, and combined, and by adding additional information (e.g. regarding regulation) to pathway drawings.

Pathway drawings show a set of interconnected chemical reactions, the enzymes that catalyze those reactions, and the reacting substrates, as in Figure 3-4.

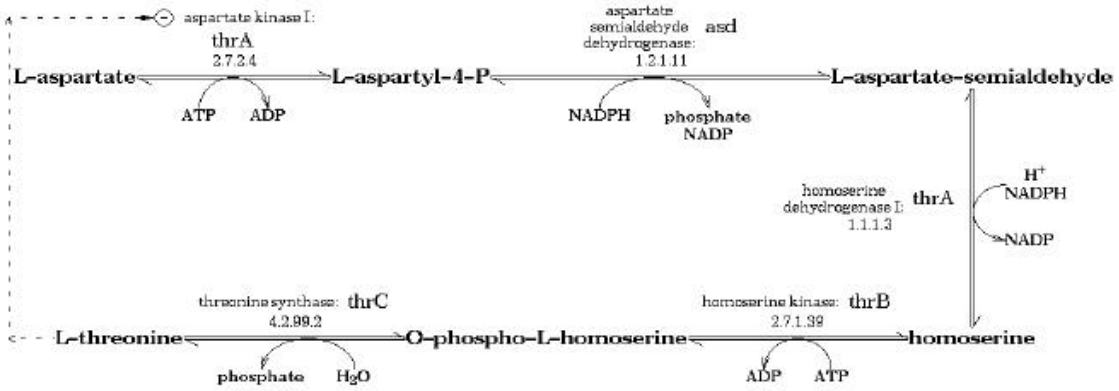


Figure 3-4 EcoCyc pathway: threonine biosynthesis

Substrates are drawn in two ways. The compounds that are shared between subsequent reactions lie along the backbone of the pathway; they are called main compounds or simply *mains*. *Side* compounds are drawn adjacent to the reaction arrow, with a curved arrow showing whether they are consumed or produced by the reaction. Enzyme names are drawn on the other side of the reaction arrow.

Pathway diagrams may include arrows showing regulatory interactions among the substrates and enzymes of a pathway. A dashed arrow leads from a substrate to a "+" or "–" sign adjacent to the enzyme whose activity the substrate modulates. The "+" or "–" indicates whether the effect on enzyme activity is positive or negative. Some enzymes are modulated by additional compounds that are not substrates in the reaction. Clicking on the "+" or "–" displays a list of all compounds that activate and inhibit the enzyme, respectively. The display page for the enzyme shows a more detailed breakdown of the activators and inhibitors into different classes (e.g., allosteric, competitive). Click on the enzyme name to see that page.

One type of relationship among pathways is shown within a pathway display under the headings Superpathways" and Subpathways." EcoCyc contains superpathways that are defined as connected aggregations of smaller pathways. For example, when the pathway for tryptophan biosynthesis is displayed, the Superpathway subheading lists a superpathway called "superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis." Clicking on the name of this superpathway navigates to it to show the synthesis of all three aromatic amino acids from chorismate.

All pathway drawings are computed automatically using pathway-layout algorithms devised by SRI's Bioinformatics Research Group. Although SRI continually improves these algorithms to produce more intuitive and informative displays, be aware that the algorithms sometimes produce unintuitive results.

A somewhat novel aspect of the pathway displays is the use of branching reaction arrows to represent complex relationships among reactions, as shown in Figure 3-5. For example, (a) depicts a situation in which one reaction converts two reactants *A* and *B* to the product *C*; two different reactions produce *A* and *B*, as shown by the in-pointing reactions. Five different reactions transform *A* to *B* in situation (f). The situation in (g) involves two reactions that convert *A* to *X*, but only one of those reactions involves the reactant *B*.

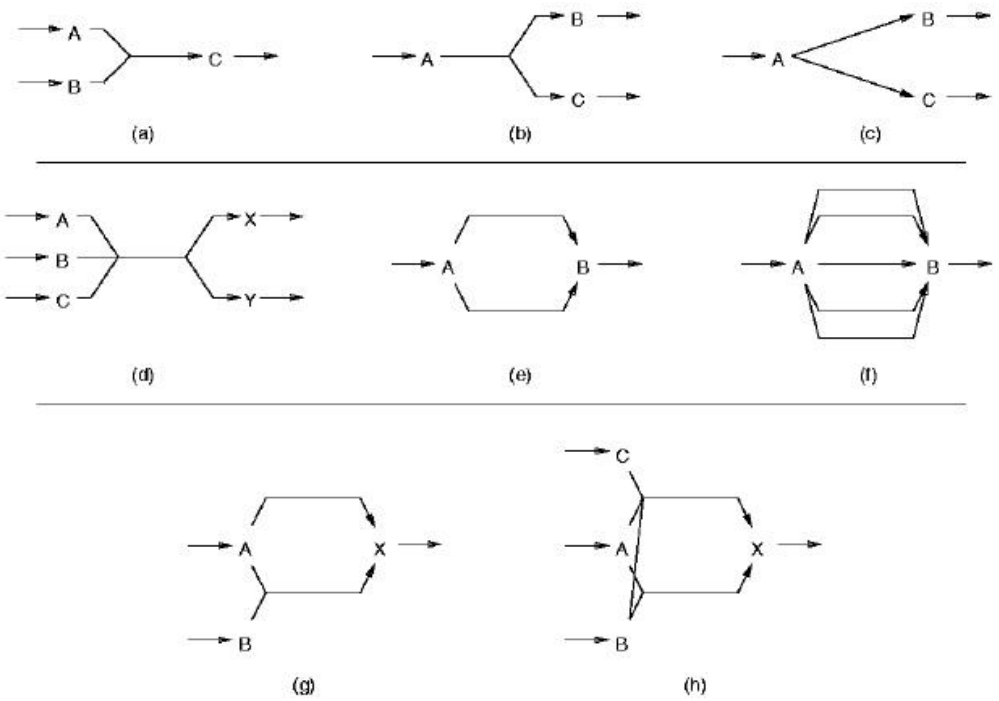


Figure 3-5 Examples of branching reaction arrows

Another aspect of our pathway displays that is not typical of textbook pathway drawings is the depiction of polymerization steps. A dashed line indicates that two compound names are in certain situations meant to represent the same species. For example, most textbooks depict saturated fatty acid elongation as a spiral, where each turn of the spiral adds two carbons to the backbone. Our representation shows the pathway as a cycle, using generic rather than specific names for the compounds involved. At the "beginning" of the cycle is acyl(n)-ACP, which undergoes several reactions producing acyl(n+2)-ACP. A dashed line is drawn between these two names to indicate that the (n+2) species becomes the (n) species for the next iteration of the cycle. We also use the dashed line when showing equivalence between a specific name for a compound (such as a starting or ending compound for a series of polymerization reactions) and the generic form. Using this scheme, we can compactly represent polymerization pathways as cycles of generic compounds, with specific compounds as inputs and/or outputs.

A small circle at the bottom of the pathway display depicts the positions of the genes that encode the enzymes within the current pathway on the *E. coli* genomic map. When you move the mouse over a given gene, its name and map position are printed at the bottom of the display window, and all reactions in the pathway involving the gene's enzyme are highlighted; clicking on the gene displays a window for that gene. Also included is a graph showing which transcription factors affect transcription of the genes in the pathway.

3.8.3.1 Pathway Commands

- **Search by Name or Frame ID**: (see "Direct Queries" in Section 3.7.1 for a general description of the query by name command)

- **Search by Substring** (similar to other substring queries)
- **Search by Class**: You choose one or more pathways by first selecting one class from a menu of pathway classes, and then one or more pathways from a menu of all *E. coli* database pathways in that class.
- **Search by Species**: You query pathways according to the species in which they occur. You are first asked to select one or more species from a menu of all species defined in MetaCyc. Your can select one or more pathways from a menu of all pathways that are known to occur in those species. This command is enabled only for MetaCyc, since other PGDBs pertain to only a single species.
- **Search by Substrates**: Your can search for pathways according to the compounds that participate in their component reactions. As many as 20 compounds can be specified, and each compound can optionally be constrained to be an input to the entire pathway, a net product of the entire pathway, a reactant in any component reaction, or a product of any component reaction. Pathways are retrieved only if *all* specified criteria are met.

3.8.3.2 Command Buttons

- **More Detail/Less Detail** The Pathway/Genome Navigator can customize pathway drawings in a variety of respects by filtering more or less information from the drawings. For example, EC numbers and gene names can be displayed or hidden, compound structures can be drawn, and pathways can be drawn as a skeletal overview that shows only those compounds at the exterior of the pathway, and at branch points. Although you can specify preferences (see "User Preferences" in Section 3.9.11) to provide fine control over pathway drawings, these command buttons provide a fast and easy way to increase or decrease the amount of detail shown in a pathway drawing.
- **Enzyme View** MetaCyc contains enzymes from many different organisms. By default, a pathway display in MetaCyc includes associated enzymes from any of the species for which the pathway is listed as having data available. Alternatively, your can choose to display enzymes from only a single organism by selecting that organism for the Enzyme View.

3.8.4 Reactions

A reaction display shows the class(es) containing the reaction within the classification of reactions. It shows the enzyme(s) that catalyze the reaction, the gene(s) that codes for the enzymes, and the pathway that contains the reaction. The displays show the EC number for the reaction and the reaction equation. Note that there exists a one-to-one mapping between EC numbers and reactions, but not between EC numbers and enzymes [16], which is why we label reactions, and not enzymes, with the EC number. The standard change in Gibbs free energy of the reaction is listed when known.

The direction in which the reaction is drawn depends on the setting of a user preference. The default behavior is for the reaction to be drawn in the direction in which the reaction is defined by the Enzyme Commission, or the direction in which the reaction is stored in the database, for reactions that do not have assigned EC numbers. The alternative behavior is for the reaction to

be drawn in the direction in which it occurs in a pathway.

Links to the ENZYME and LIGAND databases by EC number are shown.

Many of the preceding items are mouse sensitive. For example, if you click on the name of an enzyme, gene, substrate, or pathway, the Navigator displays that object.

3.8.4.1 Reaction Menu

- **Search by Name or Frame ID**: (see "Direct Queries" in Section 3.7.1 for a general description of the query by name command) Reactions typically do not have names, although in some cases the name of the enzyme that catalyzes a reaction can be used to retrieve a reaction. This command is most useful for calling up a specific reaction by its frame ID.
- **Search by Substring**
- **Search By EC#**: Allows you to call up a reaction by its EC number.
- **Search By Class**: You can choose one of the reaction classes defined by the Enzyme Nomenclature committee [16] from a menu of all such classes. You are presented with a menu of all reactions within that class; your selected reaction is displayed.
- **Search by Pathway**: You choose a pathway by first selecting from a menu of pathway classes, and then from a menu of all *E. coli* pathways in that class; then a third menu lists all reactions within that pathway. Your selected reaction is displayed.
- **Search by Substrates**: You specify one or more desired reactants and/or products, and you can choose from a list of reactions meeting these criteria. Because all reactions are considered reversible for these purposes, there is no real distinction between reactants and products. However, if the "Constrain compounds to specified sides?" box is checked (the default), then, in order for a reaction to meet the criteria, compounds specified in the reactants section must all be on the same side of the reaction, and on the opposite side to all compounds specified in the products section. If this box is unchecked, there is no distinction between compounds specified as reactants and compounds specified as products—any reaction that contains all specified compounds as either reactants or products meets the criteria.

3.8.5 Proteins

Protein displays are fairly complicated because of the many-to-many relationship between enzymes and reactions (one enzyme can catalyze multiple reactions, and one reaction may be catalyzed by multiple enzymes). Furthermore, each catalytic activity of an enzyme may be influenced by different sets of cofactors, activators, and inhibitors. Also, many genes can code for subunits of a protein complex. The protein display is usually divided into sections to address these complexities. (See [10] for details of our representations of enzymes and activators.)

The first section of the display lists general properties of the protein, such as synonyms, molecular weight, pI, cellular location, and subunit structure. If the protein is itself a substrate in one or more biochemical reactions, those reactions are listed, sorted by the pathways in which

they occur. If EcoCyc records that the protein is modified by some chemical group, a drawing of the protein coupled to the appropriate structure is shown.

Subsequent sections of the display describe each catalytic activity of the protein, if it is an enzyme (see Figure 3-6 and Figure 3-7).

Each activity section lists a reaction catalyzed by the enzyme, and the enzyme name (and synonyms) for that activity. The substrate specificity of the enzyme is described in some cases by listing alternative compounds that the enzyme will accept for a specified substrate. The cofactor(s) and prosthetic groups required by the enzyme are listed next (see Appendix A.5.6), along with any known alternative compounds for a specified cofactor. Activators and inhibitors of the enzyme are listed, qualified as to the mechanism of action, when known. In addition, this section indicates which of the listed activators and inhibitors are known to be of physiological relevance, as opposed to whether the effects are known purely because of *in vitro* studies.

The direction in which the reaction is drawn (i.e., A + B = C + D vs. C + D = A + B) depends on the setting of a user preference. The default behavior is for the reaction to be drawn in the direction in which the reaction is defined by the Enzyme Commission, or in the direction in which the reaction is stored in the database, for reactions that do not have assigned EC numbers. The alternative behavior is for the reaction to be drawn in the direction in which it occurs in a pathway, if known, or the direction in which the database indicates that the current enzyme tends to catalyze the reaction, if specified.

Additional sections of the display list each component (subunit) of the protein; when those components are polypeptides, the gene encoding the polypeptide is shown. These sections, and the first section, also list the molecular weight and pI of the subunits and the protein complex, when known.

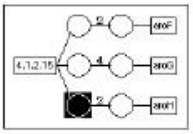
If the protein is a transcription factor, the list of known transcription units (operons) controlled by the transcription factor is displayed in the protein display. See “Transcription Units” in section 3.8.9 for more information about interpreting displays of transcription units.

***E. coli* Enzyme: 2-dehydro-3-deoxyphosphoheptonate aldolase**

Superclasses: **protein-complexes**

Component composition: **AroH x 2**

Gene-Reaction Schematic:



Enzymatic reaction of: 2-dehydro-3-deoxyphosphoheptonate aldolase

Synonyms: phospho-2-keto-3-deoxyheptonate aldolase, DHAP synthase, DHAPS, KDPH synthetase, tryptophan sensitive 3-deoxy-D arabino-heptulosonate 7-phosphate synthase, 3-deoxy-D-arabinoheptulosonate-7-phosphate synthetase (trp)

3-deoxy-D-arabino-heptulosonate-7-phosphate + phosphate $\overset{4.1.2.15}{\rightleftharpoons}$ **phosphoenolpyruvate + erythrose-4-phosphate + H_2O**

Reaction direction: REVERSIBLE

In pathways: **chorismate biosynthesis**

Comment: The presence of three isozymes provides coli with the capability for tight, multivalent regulation of the first step toward aromatic amino acid biosynthesis, while allowing sufficient residual enzyme activity in the presence of excess aromatic amino acids to provide for the synthesis of the other aromatic compounds. The aroH DAHP synthase contributes only about 1% of the total activity. [1] Although catalyzing the same reaction, each isozyme is feedback-regulated by a different aromatic amino acid. The three genes are widely separated on the coli chromosome. [2]

Citations: [3,4,5,2,6]

Cofactor binding comment: ferrous iron

Activators (mechanism undefined): **Fe^{+2}**

Inhibitors (mechanism undefined): **L-tryptophan**

Primary physiological regulators of enzyme activity: **L-tryptophan**

Figure 3-6 Upper portion of protein display window for 2-dehydro-3-deoxyphosphoheptonate aldolase

Subunit: AroH

Synonyms: AroH

Gene: **aroH**

Molecular weight (kdaltons, from nucleotide sequence): 38.721

Isozyme sequence similarity [7]:
AroG: YES,
AroF: YES [8]

Citations: [2]

Unification Links: **Entrez.P00887, SWISS-PROT.P00887**

Comment: The aroH gene has two promoters. One is regulated by the trp repressor and is favored by growth on minimal media. The other promoter is activated under conditions of growth in rich medium by an unknown mechanism. The presence of a second promoter that is active during growth in the presence of high levels of aromatic amino acid could allow aroH to escape from repression and ensure a low level of metabolic flux through the shikimate pathway for the biosynthesis of aromatic vitamins not present in the growth medium. Of the three isozymes, DAHP synthase (Trp) is only moderately feedback-inhibited and will function despite high levels of intracellular tryptophan [3] . In wild-type cells grown in minimal medium, the aroG isozyme makes up about 80% of the total DAHPS activity, the aroF isozyme makes up 20%, and the aroH isozyme makes up about 1%.

Figure 3-7 Lower portion of protein display window for 2-dehydro-3-deoxyphosphoheptonate aldolase

3.8.5.1 Protein Menu

- **Search By Name or Frame ID**: (see "Direct Queries" in Section 3.7.1 for a general description of the query by name command)
- **Search By Substring**: The program prompts you for one or more substrings, and then searches for proteins whose common name or synonyms contain all the substrings you entered.
- **Search by Pathway**: You choose a pathway by first selecting from a menu of pathway classes, and then from a menu of all *E. coli* database pathways in that class; then a third menu lists all enzymes within that pathway. The chosen enzyme is displayed.
- **Search by SwissProt ID**: You query proteins by their Swiss-Prot unique identifier. Usually PGDBs include the Swiss-Prot identifier of the form P12345.
- **Search by Weight, pI**: You query proteins by molecular weight and pI value. Proteins are retrieved only if they match all specified criteria, but fields left blank are treated as unspecified and are ignored.
- **Search for Enzyme by Modulation**: You query enzymes according to the activators, inhibitors, or cofactors that modulate their activity. The dialog box for this command allows you to select the type of modulation (e.g., activation, inhibition), and then allows you to select the compound of interest.

With a button labeled **Protein Sequence** within the protein display, you can retrieve the amino-acid sequence of a polypeptide. The amino-acid sequence is computed on demand by translating the nucleotide sequence stored for each gene. This button appears multiple times within the protein display for a heteromultimer, to allow you to retrieve sequences for each polypeptide chain within the multimer.

3.8.6 RNAs

RNA displays contain comments and citations for an RNA, a button to display the RNA sequence, and a link to the gene that encodes it.

3.8.6.1 RNA Menu

The RNA menu items are analogous to the menu items by the same name in the other object type menus.

Search by Name or Frame ID

Search by Substring

Search by Class

3.8.7 Genes

The gene display lists information for a gene such as its map position on its chromosome, the functional class(es) assigned by Riley [13], and the direction of transcription. The gene product is listed (when known); when the product is an enzyme, the display shows the equation(s) of the reaction(s) catalyzed by the enzyme, and the pathways that contain those reactions.

The gene map position is mouse sensitive: clicking on this number brings up the genome browser zoomed and centered on the region containing the gene (see Section 3.8.10 for a description of the genome browser).

If this gene is known to be interrupted (by a stop codon), the display prints a message to this effect.

The PGDBs for some organisms contain definitions of paralogous groupings of genes within the organism, which are usually computed using sequence-clustering methods. If a gene is known to be a member of one or more paralogous groups (genes containing multiple domains can be part of multiple groups), a message to this effect is printed, along with the name(s) of the paralogous group(s). Clicking on the name of a group displays all genes in that group; the genes within that list are themselves clickable. The display of a paralogous gene group also shows the chromosomal locations of all genes within the group.

The bottom of the gene display shows the local context of the gene in its chromosomal location. The displayincludes the upstream and downstream open reading frame, the transcription unit, transcription start sites, transcription factor binding sites and terminators (if known). If the transcription unit(s) containing the gene is (are) known, they are displayed below the local context display. See “Transcription Units” in Section 3.8.9 for more information about interpreting displays of transcription units.

In eukaryotic organisms, if the gene contains introns, a graphic shows their locations. Alternative splice forms are displayed.

3.8.7.1 Gene Commands

- **Search By Name or Frame ID**: (see “Direct Queries”, section 3.7.1, for a general description of the query by name command)
- **Search By Substring**: The program prompts you for one or more substrings, and then searches for genes whose common name or synonyms contain all the substrings you enter.
- **Search by Class**: You first choose a functional class from a menu of gene classes, and then choose one or more genes from a list of all the genes in that class.

Three buttons at the top of the gene display provide access to sequence information:

- **Nucleotide Sequence** retrieves the nucleotide sequence of the gene (coding region).
- **Nucleotide Sequence Neighborhood** retrieves an arbitrary nucleotide sequence with endpoints specified by the user but defaulting to the gene boundaries.

- **Protein Sequence** retrieves the amino-acid sequence of the gene product. The amino-acid sequence is computed on demand by translating the nucleotide sequence stored for each gene. If the gene codes for multiple proteins (due to alternative splice forms), you must select which splice form to view.

3.8.8 Compounds

A compound display lists the common name and synonyms for a compound, plus its parent class or classes within the classification of compounds. It shows a compound's two-dimensional structure, plus its empirical formula, molecular weight, and pK_A when known. The display lists all reactions in which the compound appears, sorted by the pathways that contain each reaction.

The display of some chemical structures within compound displays uses a concept called superatoms, which is a hierarchical structuring of chemical structures. For example, when displaying the structure for succinyl-CoA, the structure is initially displayed with the word "CoA" in place of the structure of the CoA moiety. If you click on the word CoA, however, the full structure of that moiety is displayed.

3.8.8.1 Compound Menu

The compound menu can be used to query for small molecules such as pyruvate, D-glucose, and ATP.

- **Search By Name or Frame ID**: Usually several synonyms are recorded for each compound (and for other objects) — you can retrieve the compound by any of these known names (see "Direct Queries", section3.7.1, for a general description of the query-by-name command).
- **Search By Substring**: The program prompts you for one or more substrings, and then searches for compounds whose common name or synonyms contain all the substrings entered by the user.
- **Search by SMILES substructure**: You are prompted to enter a structure in SMILES format, and can then choose one or more compounds from a menu of compounds containing the target structure (the Navigator online help system describes the SMILES format).
- **Search by Class**: You choose one class from a menu showing all compound classifications, and then select one or more compounds from a menu of all compounds in the chosen class.
- **Advanced Search**: This command displays a dialog box that allows you to specify criteria for compound name, molecular weight, chemical formula, and substructure. Compounds are retrieved if they match all the specified criteria. Fields left blank are ignored.

3.8.9 Transcription Units

Some PGDBs contain information about the clustering of genes into transcription units, the transcription factors that control a transcription unit, and the location of transcription start sites and transcription factor binding sites within a transcription unit. We define a transcription unit as the set of genes, DNA control sites, and transcription factors associated with one transcription start site. When a set of genes is transcribed from more than one transcription start site, those genes are part of more than one transcription unit.

The display of a transcription unit shows the transcription start site, transcription factor binding site(s), gene(s), and transcription terminator(s) associated with that transcription unit, when known. Directly below the transcription start site, the base pair position of the transcription start site is printed, as is the direction of transcription ("+" for clockwise). The transcription factor binding sites are displayed in two different colors (which depend on the current color preferences). By default those colors are red when binding of the transcription factor results in inhibition and green when binding stimulates transcription.

The drawing of the transcription unit can be used for navigation within the PGDB. Clicking on a gene displays the gene display for that gene. Clicking on a transcription factor binding site displays the protein display for the transcription factor, which lists all other transcription units controlled by that transcription factor. Clicking on the transcription start site produces a transcription unit display window such as that shown in Figure 3-8, which displays detailed information about each site within the transcription unit, including its nucleotide position and sequence, evidence for the site, and literature citations for the site. Each site is numbered to unambiguously identify multiple sites for the same transcription factor.

Currently, there exists no menu for querying transcription units directly — they can be found only by starting with transcription factors that control them, or genes contained within them.

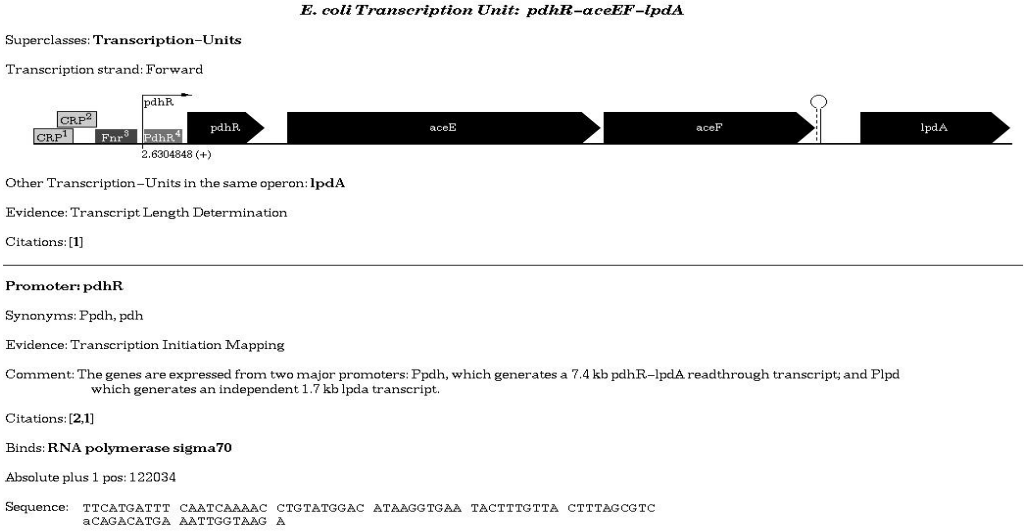


Figure 3-8 A transcription-unit display window

3.8.10 Genome Browser

Starting with Pathway Tools version 9.0, the previous vertical map browser was replaced by a new genome browser that displays horizontally, to better take advantage of available screen real estate. The genome browser can be used to examine one replicon (chromosome or plasmid) at a time. It can be invoked by clicking on a replicon listed on the single organism display, from a gene display by clicking on the basepair coordinates mentioned on the map position line, or from the menu item **Chromosome -> Select & Browse Chromosome/Replicon**.

At the top of the display, the full length of the chromosome is shown at low resolution. A region of the chromosome can be selected for display at much higher magnification in the lower part of the screen. The selected region will be drawn using as many lines as will comfortably fit on the screen, often 5 lines. The full chromosome view at the very top indicates the magnified region by means of a red, rectangular cursor.

Selection of the magnified region can be achieved by the following methods:

- Clicking on a position within the full chromosome line at the top will show the immediate neighborhood of that position. In the desktop version, one can click anywhere on the full chromosome. However, through the WWW, only the tickmarks are clickable. The tickmarks in the magnified region can also be clicked on, to recenter the region around the selected tickmark quickly.
- Start and end basepair positions can be entered in the corresponding text entry boxes; clicking the Go button displays that region.
- The region around a gene can be shown by entering the gene name in the corresponding text entry box and clicking on the Go button. The selected gene will be visually

highlighted.

? The panel of navigation arrows can be used for moving to a nearby region. The panel allows lateral translation to the left or right, and zooming in or out.

The magnified section indicates the transcription direction of genes by rectangular blocks with an arrow at one end, pointing from the 5' to the 3' end. ORFs for actual or inferred proteins have symmetrical arrowheads (with the arrow apex in the center), whereas RNA genes have an asymmetrical arrowhead (with the apex at the top edge). Pseudo-genes are crossed out with a big, diagonal X. When a gene wraps across more than one line, a zigzag at the end of the line indicates that the gene continues on the next line. Clicking on a gene brings up the corresponding gene description page.

Gene arrows filled with solid colors have transcription unit (operon) information available. All the adjacent genes that are part of a given operon are assigned the same color. Genes that have not been assigned to any transcription unit are not colored.

Additionally, transcription-units are indicated by a grey background area behind the genes, spanning the entire region of the operon.

Moving the mouse-cursor over the genes reveals their product name and the length in basepairs of the intergenic regions between the chosen gene and its neighboring genes to the left and right. If the number of basepairs carries a minus sign, the genes overlap by that many bases. As an example:

```
Gene: xdhB

Product: putative xanthine dehydrogenase subunit, FAD-binding domain

Intergenic distances (bp): xdhA< +11 xdhB -3 >xdhC
```

This means that there are 11 bp to the left of xdhB before xdhA is reached, but to the right, xdhC overlaps with xdhB by 3 bp.

If the overlap between adjacent genes is more than a small amount, the shorter gene is drawn above the longer gene to avoid visual clashes.

When zooming in to a great level of detail, transcription start sites and terminators are drawn. Transcription start sites are indicated by small arrows that point towards the 3' end of the transcript. Moving the mouse-cursor over the transcription start sites reveals the operon they are part of. The transcription factors controlling the operon are also shown, with a plus sign meaning activation and a minus sign meaning inhibition. Clicking on a transcription start site brings up the corresponding transcription unit description page.

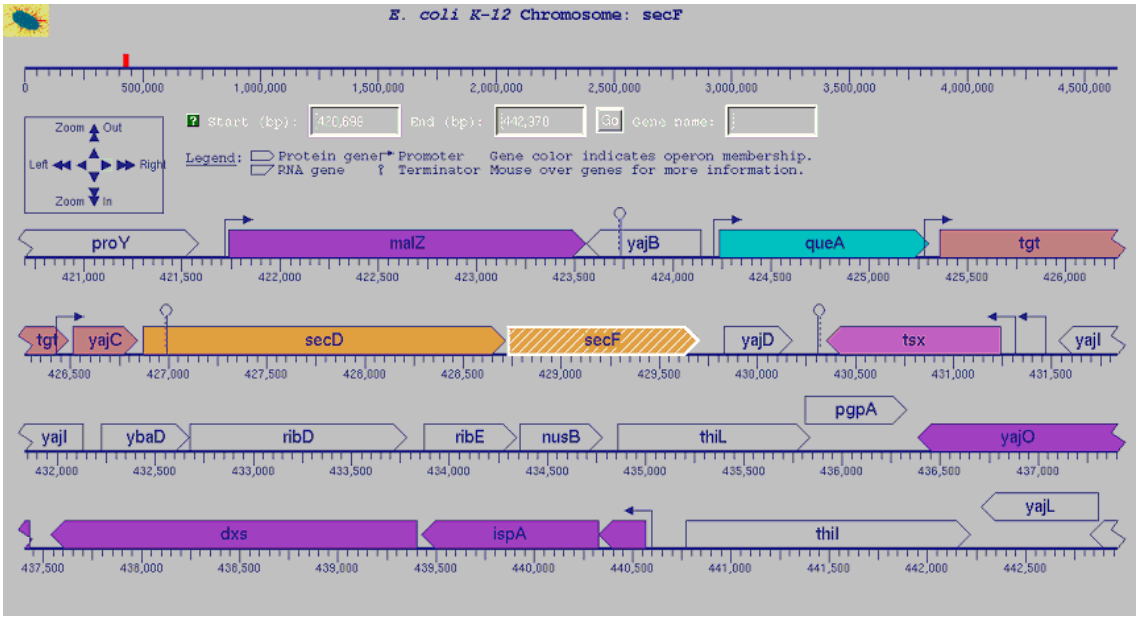


Figure 3-9 Genome browser display

3.8.10.1 Chromosome Menu

- **Select & Browse Chromosome/Replicon**: For organisms with multiple chromosomes or plasmids, you can select the current chromosome or plasmid for display. Clicking on a chromosome within an organism-summary page also selects that chromosome.
- **Show Sequence of a Segment of Chromosome:** Retrieves the nucleic acid sequence of a region of the current chromosome, or the reverse-complement of a specified region.

3.8.11 Comparative Genome Browser

The comparative genome browser can be used to examine several replicons (chromosomes or plasmids) simultaneously side-by-side. This allows easy visual comparison of related organisms to observe similarities and differences in their gene arrangements. For the alignment to work, Database Links of the relationship type "Ortholog" need to exist among genes of the organisms to be compared. Such Database Links can also be dynamically loaded from a MySQL database, with an additional setup.

The comparative genome browser is usually entered from a page describing a gene. In the section that shows the Database Links, there are 2 buttons that can be clicked on, one called "Align in Multi-Genome Browser" and the other "Select Organisms". When the button invoking the comparative genome browser is clicked for the first time in a session, the user will be asked in a popup menu to select the set of organisms to consider in the alignment. Thereafter, the selected set of organisms is remembered by the Navigator session until the user changes the selection with the corresponding button. The selected organism set also filters the ortholog links shown on the gene page itself. When the Pathway Tools work as a WWW server, the organism selection is stored as the so-called cookie named BIOCYC-ORGIDS in the user's WWW browser, for the duration of the session.

When the comparative genome browser is invoked from a gene page, that gene and its organism orchestrate the rest of the alignment. In the display, the top-most replicon is the reference, against which the comparisons are made by following the ortholog links for every gene of the top replicon in its visible section. The selected gene that is the focus of the comparison is highlighted on each replicon by a thick outline and a slanted hashed background. These selected genes are lined up at the center position of their lengths.

The navigation in the magnified region can be achieved by the following methods:

- The region around a gene can be shown, by entering the gene name in the corresponding text entry box, and by then clicking on the "Go" button. This substring search will show candidates of the lead replicon, and orthology links from it are followed to position the other replicons.
- The gene description page carries a button called "Align in Multi-Genome Browser", if there are ortholog links, which when clicked will bring up the browser centered around the gene in question.
- The panel of navigation arrows can be used for moving to a nearby region. The panel allows lateral translation to the left or right, and zooming in or out.

Genes with solid colors have links to orthologs. All the corresponding orthologs are assigned the same color, out of a set of a dozen colors that will be reused repeatedly. Genes for which no ortholog links were found in the PGDB are not colored.

The other display features are the same as described for the regular genome browser.

3.9 MISCELLANEOUS COMMANDS AND TOOLS

Various tools not specific to any type of biological object are available from the File or Tools menu or as buttons in the main window.

3.9.1 Home

The **Home** button takes you to the Organism Summary display (see "Organism Summary Display", on page 3-24).

3.9.2 Back

The **Back** button returns you to the object displayed before the one you are currently viewing. You can use the Back button to go many steps backward in this manner, or you may instead use the History button (see "History" on page 3-53).

3.9.3 Forward

If you have used the **Back** or **History** buttons to view a previously viewed object, you can advance forward to more recently viewed objects one step at a time using the **Forward** button.

3.9.4 History

The history list tracks the objects you have recently displayed. You can move forward or backward in this list one step at a time, or select a specific object from the list. See " History List" in section 3.7.3, for more details.

3.9.5 Next Answer

When a query such as a substring search returns multiple answers, the first answer is displayed and the rest are placed on the Answer List. This button displays the next object on the Answer List.

3.9.6 Clone

It is sometimes useful to capture one or more object displays for future reference, such as to allow information about two different objects to be visually compared (see "Parallel Comparison" in section 3.10.3). The **Clone** button causes a new window to be created as a copy of an existing pane in the main display window (the command prompts you to select the pane to clone, if more than one pane is visible). You can navigate in the cloned window as usual, so that several tracks of inquiry may be pursued at the same time.

3.9.7 Print

Select menu **File -> Print** to print any Pathway/Genome Navigator pane to a PostScript file.

3.9.8 Tools Menu

The Tools menu provides access to the following commands.

3.9.8.1 Pathologic

This command invokes the PathoLogic module (see Volume II: Editors and PathoLogic).

3.9.8.2 Preferences

You can customize the displays produced by the Pathway/Genome Navigator. See "User Preferences" in section 3.9.11 for more information.

3.9.8.3 Pane

The **Clone** command provides one way of looking at displays of more than one object at a time (see "Clone" in section 3.9.6). A second approach is to change your preferences so that the Navigator has more than one display pane active at a time (see "Pane Layout" in section 3.9.11.1). When multiple panes are active, the Navigator normally displays the next object in the least recently used pane. However, you can use the **Fix** button to fix a given display pane so that it is not overwritten. **Fix** asks you to click on the window that you want to fix, if more than one

display pane is active in the Pathway/Genome Navigator. Similarly, the **Unfix** button prompts you to click in the window that you no longer wish to remain fixed.

3.9.8.4 Instant Patch

Selecting menu item **Tools -> Instant Patch -> Download and Activate All Patches** downloads and applies the latest Pathway Tools software patches to both the Pathway Tools installation and the current running Pathway Tools application. There is usually no need to restart the Pathway Tools application to incorporate patches. Select **Activate Installed Patches** if you have installed a patch manually since Pathway Toolswas last started and you want to load it into your running session (all locally installed patches are automatically loaded when a new session starts up).

3.9.8.5 History

The history list tracks the objects you have recently displayed. You can move forward or backward in this list one step at a time, or select a specific object from the list. See " History List" in section 3.7.3, for more details.

3.9.8.6 Answer List

When a query such as a substring search returns multiple answers, the first answer is displayed and the rest are placed on the Answer List. The **Next** command displays the next object on the Answer List, the **Select** command lets you choose lets you choose one or more objects from the Answer List and the **Show on Console** command displays the Answer List in your original terminal window.

3.9.8.7 Ontology Browser

Invokes the GKB Editor hierarchy viewer for examining the taxonomy of classes and instances in the PGDB. From this viewer you can perform various editing operations and can invoke other GKB Editor viewers, such as the relationships viewer.

3.9.8.8 Prepare Blast reference Data

If you want users to be able to blast a query sequence against your organism's genome, you must (a) make sure that the **blast** and **formatdb** programs (available for download from **`http://www.ncbi.nlm.nih.gov/BLAST/)`** are installed and on your path, and (b) build protein and nucleotide blast databases for your organisms using the commands in this menu.

Note that blast searches are available only from a running Pathway Tools Web server, not in the stand-alone configuration.

3.9.8.9 Browse Downloadable PGDBs

Select this command to download and install optional PGDBs from a list. Some of these databases are provided by SRI, and some are provided by third parties. See section 1, "Database Sharing" if you're interested in making yours publicly available via this feature.

If you select any number of PGDBs to download, you may see a click-through license. You must agree to the terms of this license to gain access to the downloadable PGDBs.

Important Notes:

Commercial users: Your license specifies which PGDBs you may use with Pathway Tools; you may download these PGDBs only if allowed by the license.

Microsoft Windows users: Perform these steps prior to browsing or downloading PGDBs:

Install patches (Pathway Tools menu: **Tools -> Instant Patch -> Download...**). This will not be required beginning with Pathway Tools 9.5.

Download unxutils from **http://unxutils.sourceforge.net/**, unzip it, move its usr/local/wbin directory into your ptools installation directory (the one containing ecocyc, systyem-dlls, and uninstall), and rename "wbin" to "winutils". To double-check, your ptools installation directory should contain folders named ecocyc, systyem-dlls, uninstall, and winutils.

If you attempted to use PGDB Sharing before installing unxutils, PGDB sharing will likely fail, and you may need to remove tar and tar.gz files from your Windows temporary folder, the location of which depends upon which version of Windows you have as well as how Windows is configured. An example temporary folder is:

c:\Documents and Settings\smith\Local Settings\Temp\

Download speed: Downloading may take hours depending on the size of the database, your Internet connection speed, the database provider's FTP server speed, and the database provider's Internet connection speed.

Omitted PGDBs: Two types of shared PGDBs are omitted from your list:

PGDB versions you've already installed

any version of a PGDB that was built into your Pathway Tools

- **Firewall:** The PGDB sharing system uses FTP to transfer files. This requires you to be able to connect to ftp.ai.sri.com on port 21 as well as all ports in the range 1024 to 65535. Your computer need not LISTEN on any port. Consult your network administrator for further assistance.

3.9.8.10 Publish PGDBs

See section 1, "Database Sharing" for information on using this command.

3.9.8.11 Upgrade Schema of Current DB

When you upgrade to a new version of Pathway Tools, you must invoke this command for each PGDB that you have created or imported from elsewhere to support any schema or other important changes required by the software.

3.9.9 Help

The **Help** menu contains a small number of help topics for the Pathway/Genome Navigator.

3.9.10 Exiting Pathway Tools

Selecting menu **File -> Exit** terminates the Pathway/Genome Navigator. If the Navigator is the only part of Pathway Tools you were running, then selecting this menu item also exits Pathway Tools.

3.9.11 User Preferences

You can customize the displays produced by the Pathway/Genome Navigator. Commands to change user preferences are found in the **Tools -> Preferences** menu. If you change preferences during a session, then before you exit the session you will be asked if the changes should be saved. Preferences are saved in the file `.ecocyc-prefs` in your home directory. That file is loaded when the Pathway/Genome Navigator starts, so the program is automatically configured to your own preferences.

The options within the preferences menu are as follows:

3.9.11.1 Pane Layout

One to four display panes can be present simultaneously. These panes are arranged on the screen in tiled fashion, so the size of each pane depends on the total number of panes present. You can choose the number of panes from the **Pane Layout** menu. Note that some object types look better displayed on panes of certain sizes. For example, the complex graphical displays of pathways and genetic maps generally look better when only one or two display panes are present, so that the display can cover the entire screen width. Simple or primarily textual displays such as for genes or compounds (especially those with simple structures) do not suffer from being displayed on a smaller pane, and it might be advantageous to be able to display several objects at a time.

3.9.11.2 Color

Several color palettes, which assign specific colors to the window background and each type of object, have been predefined and are named in the **Colors** menu. Because different monitors show colors differently, you are encouraged to try out several of these color combinations until you find one you like. Note that for monochrome monitors, only two color palettes are available: **Black on White** and **White on Black**. On color monitors, palettes are available with black, white, gray, or blue backgrounds.

3.9.11.3 Text Font Size

To select a font size for the Navigator display panes, go to the **Tools -> Preferences -> Text Font Size** menu and click on the preferred size. Text embedded in graphics is controlled not here but rather in the object displays listed below .

3.9.11.4 Citation Reference Style

You can use this preference to choose whether references in page displays show up in numeric form (e.g. [1]), in short-hand mnemonic form (e.g. [Smith95]), or in full or abbreviated APA style (e.g. [Smith & Jones, 1995]).

3.9.11.5 Overview Display

You can apply a scale factor to alter the size of the Overview. Enter an integer that will be treated as a percentage — a value of 100 is the default scale factor.

3.9.11.6 Pathway Display

This menu determines what elements are included in pathway diagrams, and how those elements are drawn. The options **None**, **Most,** and **All** for structure-drawing preferences are the same as for reaction displays, although for pathways the default is not to show structures. Most of the preference options should be self-explanatory.

The preferences available for pathway displays are:

- **Show structures for main compounds:** Main compounds are those that run along the main backbone of the pathway.
- **Show names when structures are shown:** If *No* and compound structures are drawn, the compound names will be omitted.
- **Show side compounds:** Side compounds are those that do not run along the main backbone of the pathway. They can be omitted from the display completely, if only a general overview of the pathway is desired.
- **Show side structures:** Whether or not structures are drawn for side compounds.
- **Show enzyme names:** Like side compounds, enzyme names can be either omitted or included in the pathway diagram.
- **Font size for mains:** Five logical font sizes are available. Smaller font sizes mean that pathways can be displayed more compactly, whereas larger font sizes may be easier to read (especially, for example, when the display is to be converted to a slide or transparency).
- **Font size for sides and enzymes:** Side compounds and enzymes can be displayed using either the same size font as for main compounds, or one size smaller.
- **Reaction arrow emphasis:** Two options are available for reaction arrows. The **reversibility** option draws double arrows to indicate that a particular reaction is reversible (a reaction is assumed to be reversible unless it is known to be irreversible). The **pathway-flow-direction** option draws single arrows for each reaction, in a direction to indicate the typical flow of the pathway. The former option provides more information, but the latter option may be clearer to read, particularly in complex branching pathways.
- **Layout for linear pathways:** Linear pathways are laid out in snake fashion by default, to

enable as much of the pathway as possible to fit on the viewport. The other options are to draw the pathways in a single horizontal or vertical line. Displays of branched or cyclic pathways cannot be customized in this fashion.

- **Show pathway graph only, without title or text:** When this option is selected, only the pathway graph itself is displayed. This might be useful, for example, when preparing figures for publication or slides.

3.9.11.7 Reaction Display

Preferences in this menu control how the reaction participants are drawn within the reaction display. Three possible layouts can be used for reactant and product compounds: drawing all participants in a single horizontal line, drawing the reactants in a vertical row to the left and the products in a vertical row to the right, or drawing the reactants in a horizontal row on top and the products in a horizontal row underneath. By default, the layout is chosen automatically to ensure that the reaction will fit within the window viewport and still meet the default size constraints. However, you can choose a particular layout for constant use, in addition to choosing the various scaling constraints as for compounds.

By default, most compounds are drawn with structures. However, there is a set of compounds for which we choose not to show structures. These are typically common cofactors such as ATP and NADH, whose structures, if drawn, are likely to distract the user from the principal transformation occurring. You can change this default to show either all or no structures. Be aware that structures for some compounds are not currently available in MetaCyc and other PGDBs.

You can add to and remove from the set of compounds for which structure drawing is suppressed under the **Most** option by right-clicking on the compound in question. A menu of commands appears if the drawing of a compound structure has not been suppressed. The menu contains the item **Show name only for this compound**, which turns off structure drawing for the compound on this and future reaction displaysIf structure drawing is currently suppressed for the compound, the menu item reads **Show structure for this compound**.

The preferences available for reaction displays are:

- **Scale:** Scale can be either variable or fixed, as for compounds.

  For variable scale:

  - **Minimum scale factor:** The minimum allowable average bond length in pixels. A large minimum scale factor means that some reactions will not fit entirely within the viewport. Small minimum scale factors may make compound structures difficult to read.
  - **Maximum scale factor:** The maximum allowable average bond length in pixels.

  For fixed scale:

  - **Scale factor:** The average bond length in pixels when structures are to be drawn at a fixed scale.

- **Layout:** Four options are presented for layout of reaction equations, as described above. The **Variable** option allows the Navigator to automatically choose the best of the other three options.

- **Show structures:** This determines whether structures are drawn for compounds in the reaction equation. The options are **Most**, **None,** and **All** and are described above.
- **Display reaction direction:** Reactions can be drawn either in the direction specified by the Enzyme Nomenclature Commission (option **enzyme nomenclature**, the default), or in the direction in which the reaction appears in metabolic pathways (option **direction in pathway(s)**). If a particular reaction appears in different directions in different pathways, then the reaction is always displayed in the Enzyme Nomenclature direction.

3.9.11.8 Compound Display

User preferences control how compound displays, and primarily compound structures, are drawn. By default, compound structures are automatically scaled to be as large as possible and still fit entirely within the designated window, while remaining within a predefined size range. The font size used for the atoms is selected to look appropriate in comparison to the average bond length. At times, however, this might not be the case. For example, some machines do not have large font sizes available to them. Atoms drawn in the smaller fonts may look silly against the long bonds produced by the scaling algorithm. In this case, you might want to decrease the maximum allowable scale factor to ensure that all compounds are drawn in proportion. In another situation, you might want all compound structures to be drawn to the same scale so that, for example, hardcopy outputs of different compounds are consistent. In this case, you can switch from a variable to a fixed scale, and assign a suitable scale factor.

The preferences available for compound displays are:

- **Scale:** Scale can be either variable or fixed, as described above.

  For variable scale:

  - **Minimum scale factor:** The minimum allowable average bond length in pixels. A large minimum scale factor means that some larger compound structures will not fit entirely within the viewport. Small minimum scale factors may make large compound structures difficult to read.
  - **Maximum scale factor:** The maximum allowable average bond length in pixels. Large maximum scale factors may suffer from the font size problem mentioned above.

  For fixed scale:

  - **Scale factor:** The average bond length in pixels when structures are to be drawn at a fixed scale.

- **Verbose mode:** Some compound information is intended only for developers and is not normally displayed in compound displays. When this preference is set to **Yes**, however, all slots in a compound frame appear in the compound display.

3.9.11.9 History and Answer Lists

This dialog box allows you to alter the length of the history list. A smaller history list means that fewer items are stored, but the list is faster to cycle through. To change the length of the history

list, click on the number currently displayed. It disappears, and you can enter a new number. If you change your mind and do not want to enter a new number, simply press return, and the previous value will reappear.

Another part of the History/Answer List dialog box allows you to control what happens when the **Next Answer** command is invoked. By default, each time **Next Answer** is invoked, new objects are displayed in all the unfixed display panes in the main window. However, you can change the default so that only one new object at a time is displayed.

3.9.11.10 Reverting and Saving User Preferences

Select **Tools -> Preferences -> Restore Saved Preferences** to revert to the set of preferences previously stored in your **`.ecocyc-prefs`** file. Unless you have saved a new set of preferences during the current session, these were the preferences in effect when you started your session.

Select **Tools -> Preferences -> Save** to save the current set of preferences to your **`.ecocyc-prefs`** file. These preferences will be loaded in your next session. Note that if you change preferences but do not save them, or if you change them again after saving them, then upon exiting you are automatically asked if you want to save the new preferences.

Select **Tools -> Preferences -> Restore Defaults** to revert to the "factory settings" for the preferences.

3.10 COMPARATIVE OPERATIONS

Information is stored in databases using a consistent format compatible with the Pathway/Genome Navigator. Consequently, the Pathway/Genome Navigator can be used not only to navigate the information contained within a given database but to also compare the information contained in two or more databases; that is, the Navigator supports comparative analyses of the pathways and genomes of two or more databases.

This section is relevant to you only if you currently have access to more than one database. For the purposes of demonstrating some of the available comparative operations, we provide some examples centered around a number of PGDBs, some of which may not be included in your distribution of the Pathway Tools and databases. If this is the case, simply replace one or more of the named databases with one or more of your own databases (e.g., proprietary databases developed using the PathoLogic Pathway Predictor).

Two distinct types of comparative analysis are supported:

- Global comparisons of the metabolic networks of a user-defined set of databases
- Comparisons of specific instances of biological objects across two or more databases

Global analyses are available from within the Overview Mode of the Current Organism with the results painted onto the Overview diagram for this organism. Comparisons of specific instances of biological objects across databases are available from within each of the seven command modes.

3.10.1 Global Comparative Analyses

The metabolic overview of the current organism can be used to highlight all reactions of the Overview shared or not shared with any or all members of a user-specified group of organism databases. This highlighting allows you to compare the metabolic overview of the current organism against those of one or more other organisms. For this reason, these comparison are referred to as *global comparative analyses*. If your interest lies in developing antimicrobial drugs, these kinds of analyses provide a convenient means of computationally predicting the spectrum (i.e., across the organisms for which databases are provided) of an antimicrobial agent designed to target a specific metabolic enzyme/reaction(s).

To perform such an analysis:

1. Select an organism as the current organism and invoke **Overview -> Show Overview.**
2. Remove any past highlighting using the **Overview -> Highlight -> Clear All** command.
3. Invoke **Overview -> Highlight -> Species Comparison**.
4. Specify the nature of the comparison using the **Shared/Not-Shared** and **Any/All** options.
5. Select a set of organisms for comparison.

To select the organism to be compared against other selected organisms (the reference organism), make it the current organism. After you display the overview and invoke the **Overview -> Highlight -> Species Comparison** command, the **Species Comparison Dialog** box (see Figure 3-10) will appear.

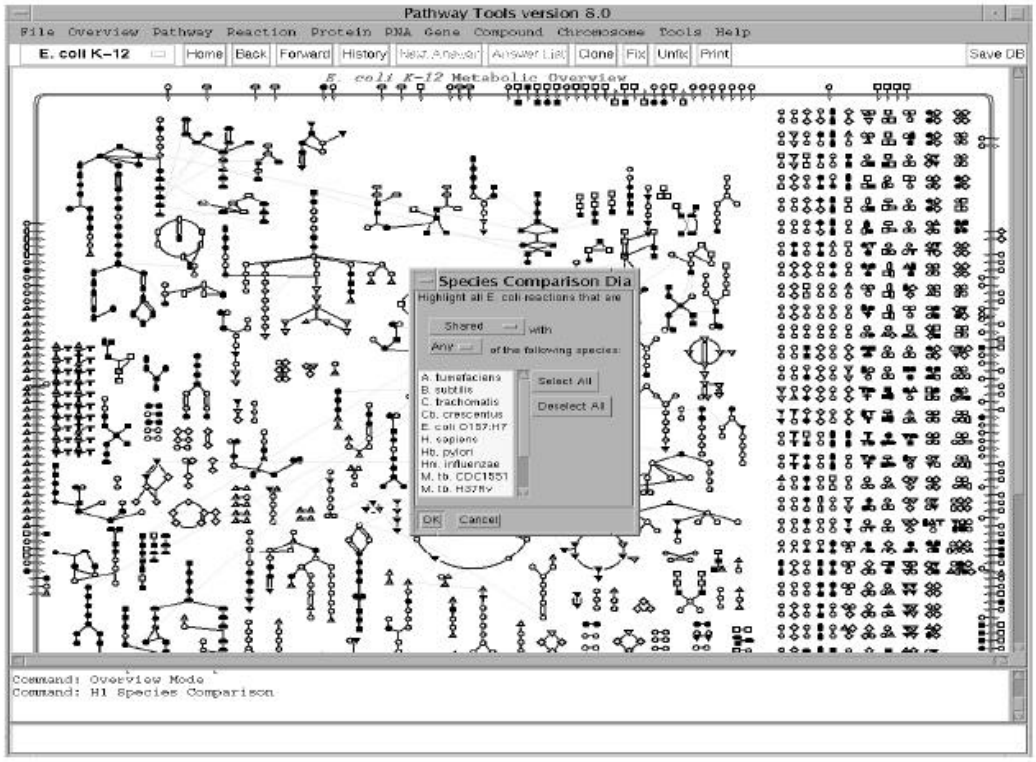


Figure 3-10 Global comparative analysis

This dialog box is used to select both the nature of the comparison and the organisms to be compared against the Current Organism. Specify the type of comparison to be performed by selecting the **Shared** or **Not-Shared** option. By default the **Shared** function will be used. This option defines the nature of reactions highlighted, that is, those shared among or not shared among the Current Organism and any (one or more) or all of the members of a user-defined set of organisms. Next, specify how the comparison is to be performed by selecting the **Any** or **All** option. Selecting **Any** means that reactions shared between the reference organism and *any one or more* or *all* of the organisms in the user-specified list will be highlighted. Selecting **All** imposes a stricter criterion; only those reactions shared between the reference organism and all listed organisms are highlighted. The final step is to define the set of databases to be used for comparison. Go to the box within the **Species Comparison Dialog** box that lists the available databases (except the reference organism). Select one or more of these databases by left-clicking on the **Select All** button. Alternatively, you can select a subset of this list. To create such a subset, select each desired database by left-clicking on its name, which will then become shaded. To deselect a database, left-click once on its shaded name (it will become unshaded). To deselect all selected organisms, left-click on the **Deselect All** button. Once a set of one or more organisms has been selected, left-click on the **OK** button.

Depending on the number of databases selected, the exact nature of the comparison, and the type

of machine on which you are running, the comparison may take up to several minutes. The clock face icon indicates that the analysis is under way. Once the analysis is finished, the results are painted onto the metabolic overview of the reference organism. All reactions that satisfy the specified conditions are highlighted in a given color and the number of highlighted reactions is listed in the LISP listener pane.

Additional species comparisons can be performed with the results superimposed upon the existing highlighting using a different color. Any number of comparisons can be superimposed. Specific colors are used for successive comparisons. When a Gray/Black Background is selected in the User Preferences (see "User Preferences" in section 3.9.11) the colors yellow, green, cyan, orange, magenta, Navaho-white, red, purple, spring-green, and deep-pink are used in succession. Alternatively, with a White Background selected in the User Preferences (see "User Preferences" in section 3.9.11) the colors green, orange, magenta, turquoise, hot-pink, blue-violet, light-sky-blue, red, and fire-brick are used. Once all colors have been used, the Navigator recycles through the given list of colors in the order stated above. Any overlap between sets of highlighted reactions is shown in white. To identify the specific combination of overlapped colors that resulted in a given reaction being highlighted white, move the mouse cursor over the reaction; the stoichiometric equation for this reaction and the name(s) of overview pathway(s) in which it occurs will appear in the bottom pane. A set of colored bars appears to the left of the reaction equation. These denote the overlapped highlight colors for this reaction. Invoking the **Overview -> Show Key** command brings up the general key for the Metabolic Overview. At the bottom of this is a specific key for the highlight colors painted onto the Overview. For each color, the nature of the corresponding analysis is summarized.

The key is updated dynamically with successive highlights. However, if you undo the previous highlight (see below) the corresponding key entry is removed and will not be recovered should you choose to Redo the last highlight (see below). The **Overview -> Highlight -> Undo** command removes the highlighting introduced by the last species comparison performed. It can be used consecutively to remove superimposed highlighting colors. The **Overview -> Highlight -> Redo** command rehighlights only the last set of reactions unhighlighted. However, as noted above, the color key is not updated to reflect this. The **Overview -> Highlight -> Clear All** command removes existing highlighting.

You can unhighlight the results of a previous comparison before your perform another one. Alternatively, you can clear the overview of all previous highlightings by using the **Clear All** command. Use multiple highlighting to perform complex comparative analyses such as those relevant to designing a desired spectrum for an antimicrobial drug. For example, you could identify all reactions that are shared by all members of a target set of organisms (e.g., those that commonly cause a specific infectious disease) but that are not predicted to occur in another set of organisms (e.g., members of normal gut flora).

3.10.2 Sequential Comparison

When browsing through a PGDB, you might find that more detailed information about a given biological object may be available from EcoCyc than from other PGDBs. This occurs because the information content of the other PGDBs is largely computationally derived, while that for *E. coli* is experimentally derived. It is therefore sometimes informative to navigate from an object display to the comparable EcoCyc display. It should not be assumed *a priori* that objects

with identical names in the computationally derived and EcoCycs are functionally identical since, for example, "equivalent" enzymes may differ across organisms in terms of their regulation and subunit composition. However, unless reliable experimental data is available about the object in the computationally derived database (consult the primary literature to find such data), the *E. coli* homolog can serve as a working "model" for the computationally derived object (since it will generally contain more information).

3.10.2.1 Current Organism-Dependent Comparisons

One method that may be used to navigate from a computationally derived database object display to that for the *E. coli* homolog involves changing the current organism from one of the computationally derived databases to *E. coli* and then using the command mode query facilities to find the relevant homolog (assuming one exists). When you search, you can find homologs by using substrings of enzyme and pathway names. Be cautious when using this approach to find *E. coli* gene homologs because, in general, a consistent vocabulary has not been used for naming genes across organisms. The primary disadvantage of this approach is that it may require numerous mouse operations and by the time you navigate to the relevant *E. coli* display, you could have forgotten pertinent details of the other display. This can be ameliorated by either printing the individual window displays or making notes about each display for subsequent comparison. Alternatively, as outlined below, you can set display command options so that relevant object displays can be placed side by side for direct visual comparison.

3.10.2.2 Current Organism-Independent Comparisons

A second method for navigating from a computationally derived database object display to that for an *E. coli* homolog requires the use of pop-up windows. This option is available from compound, reaction, protein, and gene display windows. Specifically, when viewing a display for the given object within a specific organism, the comparable object, if present in another organism, can be accessed by first moving the mouse cursor over the object name in the display title (a rectangle should appear around it) and right-clicking the mouse button. This brings up a pop-up menu from which the **Show in Organism** option can be selected to display a list of available organism databases Selecting a given database results in the display window being replaced with a comparable display for the same object in the selected database (if the object does not exist in the selected database, then the following error message appears in the bottom pane (the listener pane): **"`Object does not exist in selected organism`"**). By alternately using the **Back** and **Forward** commands, you can compare the entries for the given object in the two databases. To view the same entry in a third database, go to the display window, repeat the above procedures, and select the name of a third database. Now use the **Back** and **Forward** commands to scroll between three distinct displays of the same object across the three selected databases. The above operations can be continued until displays of the given object in any number of available organisms have been accessed. Note that none of these operations changes the Current Organism. You should be cognizant of the nature of the Current Organism since all queries issued through the command menu operate upon the corresponding database.

Alternatively, from a given display window (e.g., Overview and pathway displays, you can directly navigate from a component of the display (e.g., a reaction or compound to a display of

that component as present in another user-specified database. For example, set the current organism to *Mycobacterium tuberculosis.* Using the **Pathway -> Search by Substring** command, bring up a display of the glycolytic pathway in the main display window, and then right-click on any compound in the pathway to bring up a menu of options. Select the **Show in Organism** option by either left- or right-clicking on it to bring up the **Choose species** menu, which lists available databases. Left-click on the name of the organism of interest, for example, *E. coli*. This brings up a display window for the compound in the specified database (assuming such an object exists). If this compound information cannot be found, then an error message to that effect appears in the bottom pane (listener pane): **"`Object does not exist in selected organism KB`"**. By using a comparable set of commands, you can also select a reaction, enzyme, or gene from a pathway display and navigate directly to a display of that object in a specified database (if present).

A comparable set of operations can be used on Overview displays. For example, select *Mycobacterium tuberculosis* as the current organism and use **Overview -> Show Overview** to bring up the Overview Display for this organism. Move the mouse cursor over a compound in the overview display. A left-click will take you to the display window for this compound. Alternatively, a right-click will bring up a menu the last two options of which are relevant to species comparison:

- Display compound information for selected species
- Display pathway information for selected species.

Selecting either one of these options causes the **Choose species** menu to appear, listing names of available databases. Selecting a database brings up the compound/pathway display for the selected database. This operation does not change the nature of the current organism. Use the **Back** command to return to the metabolic overview for the current organism.

Alternatively, the mouse cursor may be moved over a reaction in the database for the metabolic overview. A left-click on the reaction will take you to the display window for this reaction in the current organism. Alternatively, a right-click on the reaction in the overview will bring up a menu, the last two options of which are relevant to species comparison:

- Display reaction information for selected species
- Display pathway information for selected species.

Selection of either of these options brings up the **Choose species** menu, which lists available databases. Selecting a given database brings up the relevant reaction/pathway display. Use the **Back** command to return to the metabolic overview.

3.10.3 Parallel Comparison

The operations described above present sequential displays of specific instances of particular objects, for example, a particular pathway, as described in two or more distinct databases. However, user preferences can be set to support concurrent display of multiple display windows such that you can view equivalent instances of an object across two or more organisms; that is, each window provides a display for each organism.

For example, to compare the alanine biosynthesis pathways of *E. coli* K-12 and *Hm. influenzae,*

select **Tools -> Preferences -> Pane Layout -> 2 panes**. Set the Current Organism to *E. coli* K-12 and then invoke **Pathway -> Search by Name or Frame ID**. This brings up the Dialog Box into which you should type **`alanine biosynthesis-I`**. A display for this pathway appears in the top window. Click on the **Fix** command button and then click on the top window. Move the cursor over the words "alanine biosynthesis-I" as they appear in the display title and right-click the mouse button. This brings up a pop-up menu from which you should select the **Show in Organism** option to bring up a list of available databases. Selecting one of these — for example, *H.influenzae* — results in a display of the alanine biosynthesis pathway for the selected database in the main display of the bottom window. You can now visually compare the pathway across the two organisms (see Figure 3-11).

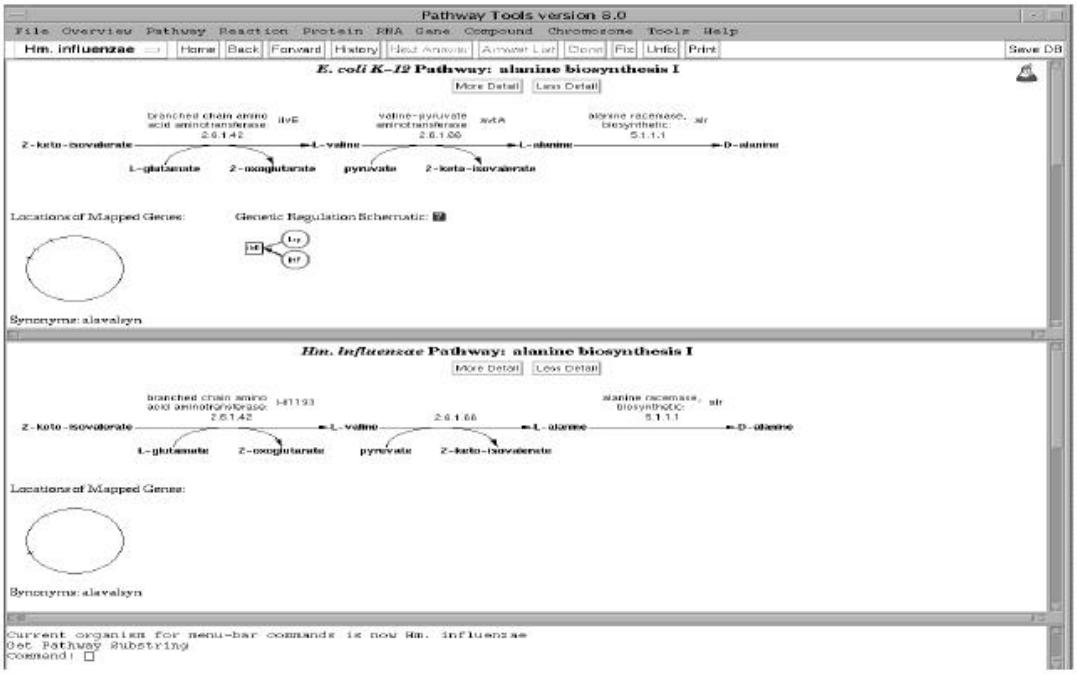


Figure 3-11 Global comparative analysis

By using a comparable set of commands, you can also compare other objects. For example, Figure 3-12 shows a comparison of the genomic maps of *E. coli* and *H.influenzae*.

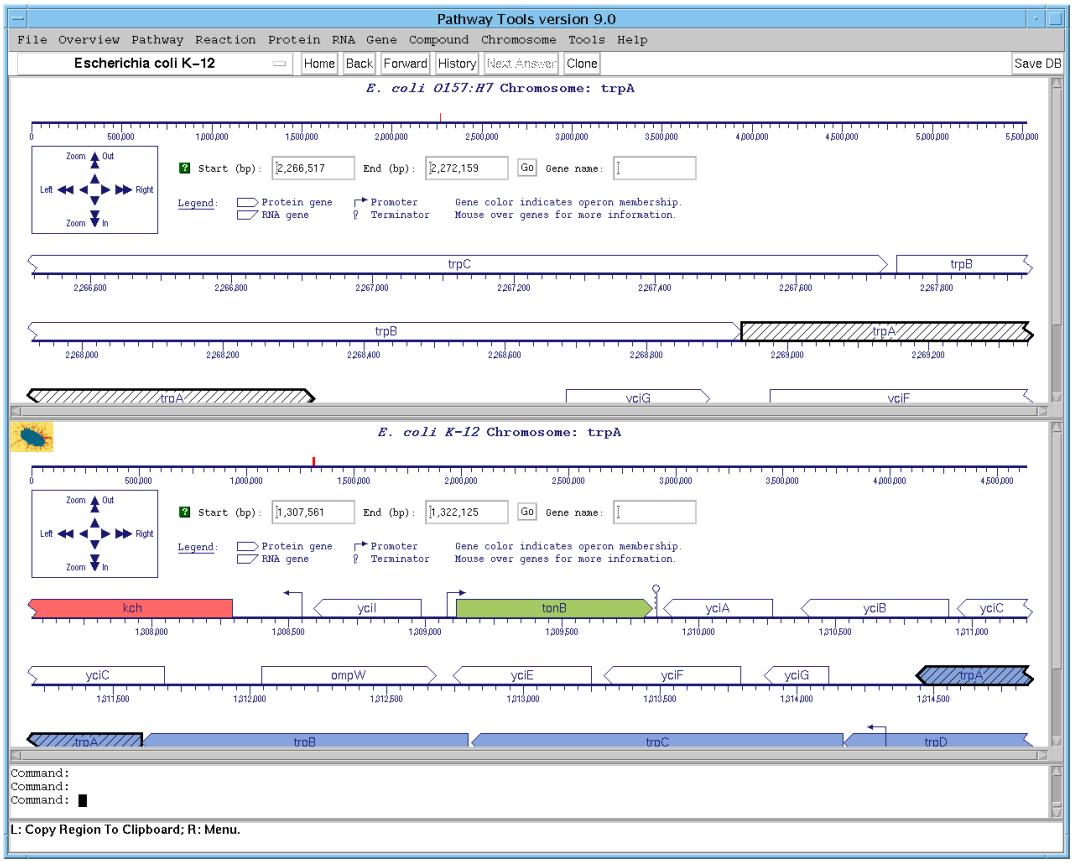


Figure 3-12 Global comparative analysis

The user preferences support incorporation of as many as four distinct windows in the display layout. By consecutive application of the **Fix** command, you can visually compare four instances of a given class. Note, however, that two or more metabolic overviews cannot be displayed using this approach.

To compare two metabolic-overview diagrams, use the **Clone** command (see "Clone", section 3.9.6). You can also use this option to compare instances of other classes, for example, pathways. Use the **Clone** command to duplicate any window display; you can clone any number of window displays. Cloned windows do not support the full range of menu options in the full Navigator window, but they do support history commands and clicking on objects within the window, such as for hypertext navigation or for editing. However, you do not have access to any of the command menu options. The history for the clone window is cleared at the time of cloning, but as you navigate between related displays, a local history is constructed for this window. Move back and forth between these displays by using the move back and forth in history commands available within the clone window.

To use the **Clone** command to compare two metabolic overviews, set the Current Organism to an organism of interest (e.g., *Hb. pylori*) and invoke **Overview -> Show Overview**. The overview for this organism appears in the main display window. Click on the **Clone** button to bring up the overview display window (see Figure 3-13).

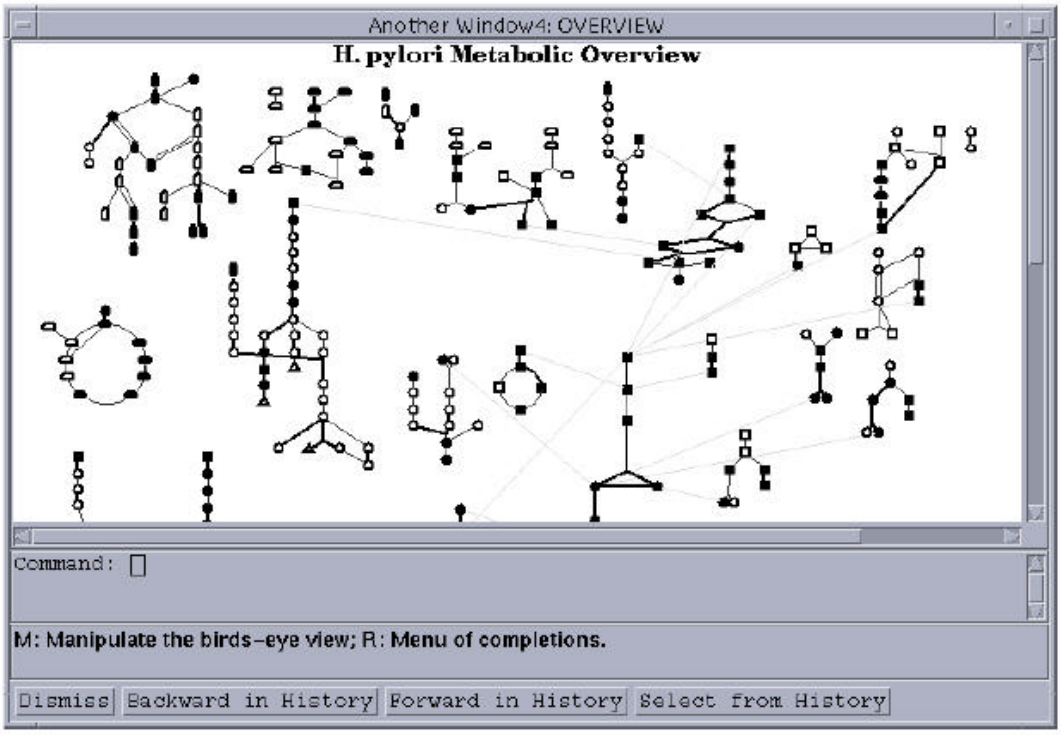


Figure 3-13 A clone window showing the Hb. pylori metabolic overview

Change the current organism to that of the organism to be compared and again invoke **Overview -> Show Overview**. The metabolic overview for the second organism appears in the main display window. Now position the clone and main display windows so that the respective overviews can be visually compared.

4 THE IMPORT/EXPORT FACILITY

While the native data storage format for PGDBs is an object-oriented, so-called Frame Representation System named Ocelot, there are many reasons for exchanging some or all of the data in other formats. Pathway Tools supports import and export in several types of file formats:

- Column-delimited formats are easy to manipulate with external Spreadsheet programs.
- Attribute-value formats are easy to parse with external text processing tools.
- BioPAX is an OWL RDF/XML-based format for exchange of pathway data. See http://www.biopax.org/
- SBML is an XML-based format for capturing models of biochemical reaction networks. See http://www.sbml.org/
- Genbank is a traditional format for exchange of gene annotations for an entire chromosome.
- Pathways can be easily exchanged between PGDBs as complete units, in a file format that can contain all the relevant pieces, including compounds and enzymes.
- Compound structures can be exchanged in the widely used MDL Molfile format. See the section "MDL Molfile Import/Export" in vol.II of the User Guide.

4.1 PATHWAY IMPORT/EXPORT

The import/export facility allows you to export selected pathways and related objects from a PGDB to a file, which can then be imported into another PGDB, possibly at another site. The following scenarios illustrate situations for which this facility is useful:

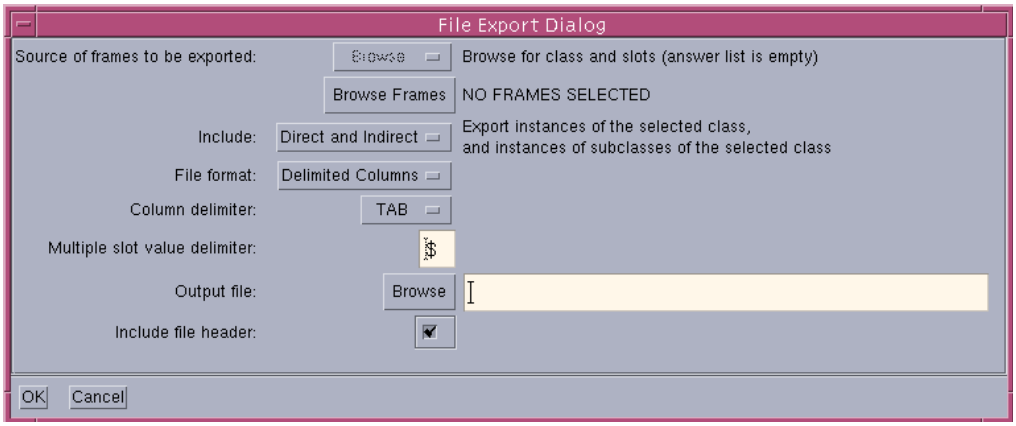


Figure 4-1 File Export dialogue

- You have created new pathways for your organism PGDB, and we would like to incorporate these pathways into the next release of MetaCyc. You would export your pathways to a file and send us the file.
- You want to exchange pathways you have created with another user who has been

developing a PGDB for a related organism.

- You want to import individual pathways that have appeared in a new release of MetaCyc, but you do not want to run the general pathway rescoring procedure.

To select a pathway for export, right-click on the pathway handle and choose **Edit -> Add Pathway to File Export List**. All pathways that you want to export to a single file should be selected in this fashion. When this is complete, select **File -> Export -> Selected Pathways to File…**. You will have an opportunity to edit the list of pathways to be exported and specify the file name. In addition to the pathway frames themselves, related objects such as reactions, compounds, and publications are exported. Enzymes and genes can also be exported. The dialog lists some situations in which you may or may not want to include enzymes and genes in your export file.

To import from a file created using the above commands, select **File -> Import -> Pathways from File…** and supply the file name. Any frames in the export file that do not exist in the current PGDB are created. Frames that already exist in the current PGDB are generally not overwritten or modified, even if they are different in the export file (thus, this facility is not useful for exchanging updates to existing frames between PGDBs). Check any imported pathways to make sure that they look correct.

4.2 SBML EXPORT

The **File -> Export -> Selected Reactions to SBML File…** command brings up a dialog panel that allows selection of a set of reactions to be written to an SBML file. SBML is an XML-based format for capturing models of biochemical reaction networks. See http://www.sbml.org/

4.3 GENBANK EXPORT

The **File -> Export -> Selected Chromosome to Genbank File…** command allows exporting all the gene annotations of a chromosome to a Genbank file. If the PGDB has more than one chromosome/replicon, a small popup menu allows selection of the chromosome to export. Thereafter, a dialog panel allows specifying the name of the output file. A default filename is suggested, which consists of the frame ID of the selected chromosome, together with the customary ".gbk" file suffix. Pathway Tools will be busy for about a minute (for a few thousand genes), while writing out the file.

Not all of the rich data in a PGDB can be represented adequately, so there may be some information loss. E.g. there is no obvious and standardized way to export gene synonyms. The comments for genes are written to the /note feature qualifier, but comments for the corresponding gene products (usually a protein) are not written out.

The resulting Genbank file can be read in by other, external gene annotation tools. It is know to be readable by **Artemis**, version 7.1.

Genbank files can also be read by Pathway Tools. For this, please see vol.II of the User Guide for the description of Pathologic. The combined read and write capability allows e.g. using the Pathway Hole Filler to improve the annotation of existing Genbank files, by running them

through Pathologic and the Pathway Hole Filler, and then exporting the resulting data again in the Genbank format.

4.4 LINKING TABLE EXPORT

The **File -> Export -> Generate Link Tables…** command allows generating a set of Tab-delimited files containing IDs and names of various objects to help with creating Database links to and from external sources. A dialog panel allows selection of a directory for these files. The suggested default location is the "data/**"** subdirectory. For more information about these tables, please consult the section "Creating Links between a PGDB and External Databases" in User Guide vol.II.

4.5 FULL FLAT FILE DUMP

The **File -> Export -> Entire DB to Flat Files** command allows writing out the full set of flat files for the currently selected PGDB that we also make available for download from http://biocyc.org , and so it takes some time until this command completes. The various objects in our BioCyc schema are written both in column-delimited and attribute-value formats. Also, a BioPAX file is written as well. For a detailed list of the files, their contents, and their formats, please see the online description at http://bioinformatics.ai.sri.com/ptools/flatfile-format.html .

4.6 FRAME IMPORT/EXPORT

The **File -> Import -> Frames from File…** and **File -> Export -> Selected Frames to File…** commands can be used to import and export collections of frames to character-delimited files, as well as to an attribute-value format that resembles MEDLINE export format. Frames to be exported can be selected by means of query commands (i.e., the Answer List), or by browsing through the hierarchy of object classes.

In general, this system is intended to allow users to export a set of data, edit it, and re-import the edited data. In addition, this mechanism can be used to create new frames.

4.6.1 Frame Export

Begin by opening the Export dialog, by selecting **File -> Export -> Selected Frames to File**. This brings up the dialog shown in Figure 4-1. Because of the complexity of the export process, the dialog displays only items that are relevant to your current set of choices. In general, items further down on the window are dependent on those higher up, so as you change your choices for items near the top, items lower down may appear or disappear. The complete list of settings for export is as follows:

- **Source of frames to be exported**: This item grayed out unless there are frames on the Answer List. It allows you to choose whether exported frames will come from the Answer List, or will be specified by selecting a class from the Pathway Tools class hierarchy.
- **Choose Slots**: This item appears only if " Answer List" is selected as the source of frames

for export. It displays a menu showing the complete list of slots for the frames on the Answer List. You must choose at least one slot from this menu in order to export from the Answer List.

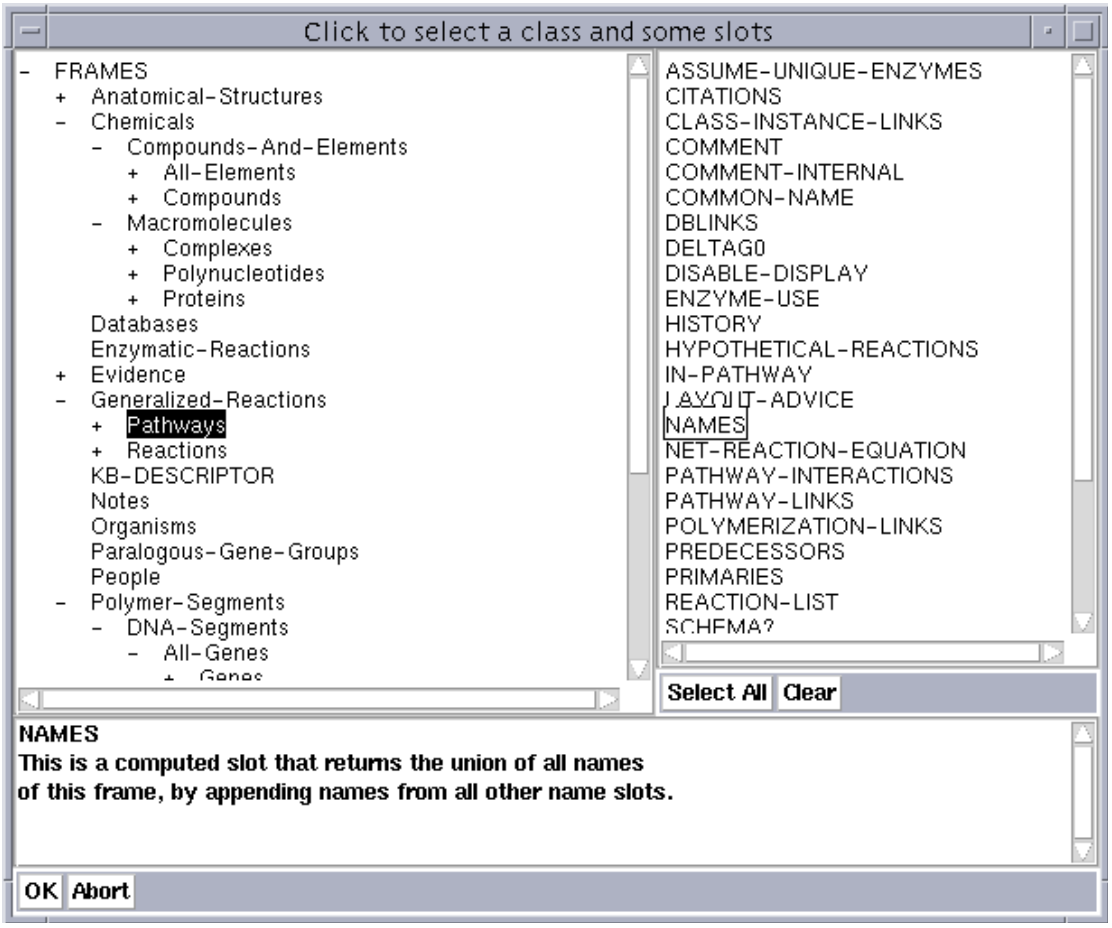


Figure 4-2 Pathway Tools class and slot browser

- **Browse Frames**: This item appears if "Browse" is selected as the source of frames for export. It brings up a browser window for classes and slots, shown in Figure 4-2. The left-hand pane of this window displays a tree of classes, similar to the file folder browsers found in most operating systems. Classes that have a "+" to their left have subclasses that are not yet displayed. Clicking on the "+" opens up another level of the hierarchy, displaying a list of subclasses that may themselves have subclasses. Once a level of hierarchy is displayed, the "+" symbol changes to a "-", and clicking on the "-" closes the sublevel of the hierarchy. Rolling the mouse over a class displays its name, and in some cases additional documentation, in the documentation pane immediately below the "OK" and "Abort" buttons. You can select only one class at a time. Once a class has been selected, the set of slots for that class is displayed in the right pane of the browser. Roll the mouse over a slot to display documentation for the slot. Clicking on a slot toggles whether or not it is selected. You must select at least one slot in order for export to proceed. Clicking "Select All" selects all slots, and clicking "Clear" deselects all slots. Once you have selected a class and a set of slots, click "OK" to return to the main dialog. Click "Abort" to return to the main dialog.

- **Include**: This item appears only if "Browse" is chosen as the source of exported frames. Selecting "Direct only" limits export to frames that are direct instances of the selected class, whereas "Direct and Indirect" exports all frames that are instances of classes that are descendants of the selected class, down to the lowest level of the class hierarchy. For example, if "Reactions" is the selected class, "Direct only" selects a very small number of frames, whereas "Direct and Indirect" selects a very large number of frames, because in addition to the small number of frames that are direct instances of the "Reactions" class, this also selects all Binding Reactions, Transport Reactions, and so on. In many cases, selecting a higher-level class and "Direct only" results in no frames being selected at all, because most frames are instances of classes at the bottom level of the class hierarchy.
- **File format**: This item allows you to select between two export file formats, "Delimited Columns" and "Attribute-Value". See the description of Delimited Column and Attribute-Value formats below for more detail on these formats.
- **Column delimiter**: This item appears only if "Delimited Column" is chosen as the file format. It offers three choices: "TAB", "Comma", and "Other". Choosing "TAB" or "Comma" produces TAB-separated or comma-separated output files, respectively, which are the most common formats used for import into spreadsheet programs such as Excel. If you choose TABs or commas as column separators, avoid exporting slots that may include these characters in their values. Choosing "Other" allows you to choose any single ASCII character as a column separator. For text slots such as comments, it is generally best to choose a little-used character, such as "^", as a column separator.
- **Multiple slot value delimiter**: This item appears only if "Delimited Column" is chosen as the file format. When using Delimited Column format, slots with multiple values are handled by concatenating the values, with a user-specified delimiter character between values. If the same character is specified when the file is re-imported into Pathway Tools, these fields are split back into multiple slot values on import. We recommend that you use a character even more uncommon than the one used for a column delimiter for this purpose, because many slots have multiple values. The initial default for this item is '$'.
- **Output file**: This item allows you to choose the name of the file in which to place exported data. If the file name is not specified as a fully qualified file path, the export file is created in the directory you were in when you started Pathway Tools. A "Browse" button allows you to choose a file name by means of a standard file browser.
- **Include file header**: If this item is checked, the file begins with a documentation header specifying the name of the file, the name of the DB from which the data was exported, the date and time that the export operation was begun, the username of the user who performed the export, the names of the classes of the exported data (or top-level class only, if frames were chosen by browsing), and the names and documentation strings of all slots exported, whether or not those slots actually had any values in the frames exported.

Once all fields have been filled in, click "OK" to begin the export operation. A progress bar is displayed during the export process, and a window showing the number of frames exported is displayed when exporting is complete.

Exported files can be edited with a standard text editor, or any program capable of decoding the export format. In particular, most spreadsheet programs should be able to import column-delimited files easily. In a column-delimited file, the first line following the documentation header contains the names of the slots for exported frames, and is interpreted as a header on import.

4.6.2 Frame Import

The **File -> Import -> Frames from File** brings up a dialog similar to the one used for specifying export parameters, shown in Figure 4-3. As in the case of the export dialog, items that are not relevant in the context of the choices specified by items higher up in the window are hidden.

- **File format**: With this item, you can choose between "Delimited Columns" and "Attribute-Value" formats. See below for details of these formats.
- **Column delimiter**: This item appears only if "Delimited Columns" is chosen as the import format. As on the export dialog, it allows you to choose, TAB, Comma, or some other ASCII character as the column delimiter. The value of this item must match the contents of the import file.
- **Multiple slot value delimiter**: This item appears only if "Delimited Columns" is chosen as the import format. Within columns values in the imported data, this character is used to split single column values in the file into multiple slot values in the DB. As is the case for the column delimiter, the character specified in this field must match the contents of the import file.
- **If object exists**: The Pathway Tools import facility allows very flexible handling of frames that match existing frames, and of slot values for slots that already contain values. This item offers the following choices:
    - **Replace entire object:** Import erases all existing slot values for the frame, including those for slots not found in the import fil**e**, and repopulates the frame's slots with values from the file. You should use this option with caution.
    - **Update slots:** Import augments the existing slot values for the frame with values from the import file. This item causes the "If slot value exists" item, described below, to appear.
    - **Log, do not import:** The name of the frame in the import file is noted in a log file, but the frame is not imported into the DB.
    - **Ignore completely (no logging):** The frame from the import file is silently discarded, without being logged. This option can be used to verify that an import file can be parsed without errors, without affecting the DB,
    - **Ask user each time:** This option allows different frames from the import file to be dealt with individually, at the user's choice. When a duplicate frame is encountered, a dialog appears, giving the name of the frame, and a menu of choices. See the description of the "If slot value exists" item for details of the choices provided. The dialog also includes a checkbox, which allows any subsequent duplicate objects to be dealt with based on the current choice, with no

further user input. If "Update slots" is chosen, you may still be asked about the disposition of individual slot values.

- **If slot value exists**: This item appears only if "Update objects" is selected as the value of "If object exists", and it controls what happens when the import process encounters slots that already have values. The possible choices for this item are

    - **Replace existing value**: any existing value(s) for the slot is erased, and the value(s) in the import file replaces them. We expect this to be the most commonly used option, because it ensures that edits to the export file are incorporated into the PGDB.

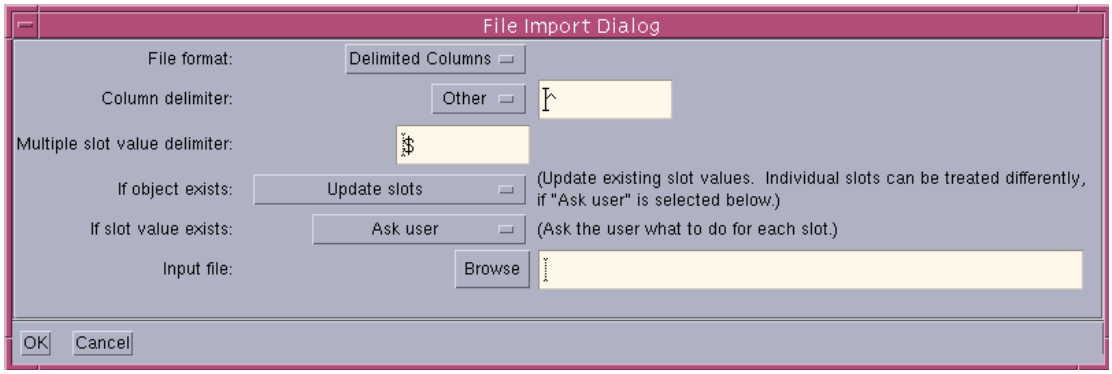


Figure 4-3 File Import dialogue

    - **Add to existing values:** The existing value for the slot is augmented with values from the import file.
    - **Log and continue:** The existing slot values are left undisturbed, and the values in the file are noted in a log file.
    - **Ignore completely:** The existing slot values are left undisturbed, and the values in the file are discarded.
    - **Ask user:** If a slot with existing values is encountered, a dialog appears, showing the existing slot values and the values found in the import file. Available choices are the same as those listed above for the "If slot value exists" item.

The dialog also includes a checkbox, which allows any subsequent duplicate values for the slot in question to be dealt with based on the current choice, with no further user interaction. This applies only to the current slot: any other slots that have "ask user" selected will still cause the dialog to appear.

- **Input file**: You can enter the name of the import file in the text field for this item, or use the "Browse" button to pop up a graphical file browser. The file name must include the file's extension, and unless the file is in the directory you were in when you started Pathway Tools, it must also include the full directory path for the file.

Once all fields have been filled in, click "OK" to begin the import. A progress bar appears, and starts moving once the import process begins actually updating the DB. If "ask user" was specified for dealing with either existing frames or existing slot values, you may be presented

with a dialog for dealing with the frame or slot in question. In each case, the available choices are the same ones listed above for dealing with these cases. Once import has completed, a window appears, summarizing the data imported. Any values in the import file that are an exact match for the existing values in the DB are ignored, and do not change the DB in any way. If you suspect that the import has corrupted the PGDB, or if import aborts before completion, you should use **File -> Revert Current DB** to restore the PGDB to its previous state. Correct any errors in the import file, and try the import again.

- **About** the log file: If "log" is chosen as the disposition of an import value that clashes with an existing frame or slot value, an entry is written to the log file. The log file is written in the same directory as the import file, and its name is the same as that of the import file, with the extension ".log" appended.

4.6.3 Supported file formats for frame import and export:

Detailed documentation for Pathway Tools' import and export file formats can be found at **`http://bioinformatics.ai.sri.com/ptools/import-export-file-formats.txt`**.

This documentation is reproduced below for convenience.

Column delimited format:

In this format, each line in the file represents a single frame in the DB. Columns are separated by a single character, of the user's choice. Multiple values for single slots are put into single columns. All the values of the slots are concatenated into a single string, with individual values delimited by a single user-specified character.

Following the documentation header, which is the same for both column-delimited and attribute-values formats, the first line in a column delimited file is a header line, with columns separated by the same character that delimits columns in the rest of the file. If all the frames in the file are of the same class, the first column of the header is the name of the class; otherwise, it is the string "FRAME". The last column of the header is always the string "CLASSES".

Within the portion of the file that represents individual frames, the first column is always the name of the frame, and the second column is the names of the classes from which the frame inherits, with multiple class names separated by the same character used to delimit multiple slot values. Subsequent columns are in alphabetical order of the slot name.

Attribute-value format:

This format is similar to MEDLINE format. Frames are separate by blank lines. Within each record, attribute names, which may be class, frame, or slot names, are at the beginning of each line, followed by the string "- ", followed by the value of the attribute. Attribute values can span multiple lines, in which case the second and subsequent lines of the value begin with spaces. Multiline values are reassembled into single strings containing newlines on import. Multiple slot values are handled by putting each value into a single attribute-value pair. On import, the user

interface for dealing with multiple slot values is identical to that for column-delimited files: the user is queried about all of the values for a single slot with a single dialog.

5 DATABASE SHARING

5.1 PUBLISHING YOUR DATABASES

The process of publishing a database involves three steps:

1. The directories and files that make up the database are packaged and compressed into a single archive file.
2. The packaged file is moved to an FTP file server that allows anonymous access for retrieval of stored files.
3. The location of the packaged database, and essential information about its contents, is sent to the central registry server, from which it is visible to Pathway Tools users across the entire Internet.

Fetching of shared databases is done through a standard FTP server. If you publish one of your databases, you do not need to have Pathway Tools running in order for others to retrieve and install the database.

5.1.1 Details of What Happens during Each Step:

1. Packaging the database

   The “tar” program is used to package the entire directory tree for the database into a single file, which is then compressed using the “gzip” program. The compressed file is then moved to the standard directory for temporary files (this is different for Windows and for Unix/ Linux—see below for details).

2. Moving the archive to an FTP server

   If your site configuration allows it, FTP (File Transfer Protocol) is used to copy the packaged database from the standard temporary directory on your system, to the server that you specified in the Database Sharing preferences dialog. Note: this step does not depend on an external “ftp” program, but runs entirely within Pathway Tools. Once the file has been copied to the FTP server, the temporary file is deleted.

3. Registering

   Pathway Tools contacts the central registry server over the Internet, and transfers information about your database that will allow others to browse and download it. The information stored includes:

   Database authors

   The species name and strain

   Genome

   Version

   Copyright

   A comment for the entire database

Contact email address

The URL of a license agreement for the database, if there is one

We have made every effort to automate the process of sharing databases, but for security reasons, anonymous FTP servers are usually organized in a way that requires files that are copied to the server to be manually moved to a different directory for subsequent public retrieval. Our user interface allows you to perform each of the three required steps separately, in case a manual step is required between the packaging/storing and registering steps. Because of the wide range of platforms and network configurations in general use, we can provide only general instructions for copying files to your FTP server manually, if necessary. The details of this process are dependent on the setup of your site—please consult your local system administrator for more information on how to do this at your location.

5.1.2 Preliminary Step: Setting Preferences

Before actually beginning the processing of your databases for publishing, you need to fill in the Database Sharing preferences dialog. This is available from the main command menu via *Preferences -> Database Sharing*. Note that these preferences are needed *only* for publishing your own databases, not for fetching shared databases from other sites.

The first time that you invoke the preference dialog for database sharing, an initial dialog pops up, allowing you to enter some basic properties of the way your site is set up. The user interface for PGDB publishing is customized to provide the simplest possible user interface, based on the constraints of your site's configuration.

Unless there are major changes to the way your site is administered, you should not need to run this initial setup more than once. Nevertheless, the main preferences dialog has a button that displays this dialog again, so that it is always possible to go back and change your answers to the questions in the initial dialog.

As shown in the figure, the initial dialog contains a good deal of explanatory text, and is intended to be self-documenting.

Once the two questions on the initial setup dialog have been answered, the main preferences dialog is displayed. It contains the following items:

- **Enable PGDB sharing functionality**: This checkbox can be used to completely disable PGDB sharing, if desired. In general, enabling PGDB sharing should have little effect on the operation of Pathway Tools. The one exception to this is that displaying the summary page for all organisms queries the server for databases that were installed via the database sharing facility, to see if a newer version is available. If the server does not respond within a few seconds, the query is simply skipped, so the summary page should display fairly quickly even if the servers for these databases are completely unavailable. If you notice a delay in displaying the summary page, temporarily disabling PGDB sharing may eliminate the delay, and will have no effect on the functioning of any of your databases.
- **Rerun initial setup**: Pushing this button will cause the initial setup dialog to appear, and will allow you to change your answers to the basic setup questions.

The rest of the preferences dialog is divided into two parts. The first set of fields contains

information needed for storing files on an FTP server. If you answered "no" to the question about using FTP for storing files in the initial settings dialog, this section will not appear at all.

FTP server to which files will be uploaded: Enter the Internet name or address for the FTP server to which you will connect for storing your archived databases, just as you would enter it for manually starting an FTP session.

Username for storing to the FTP server above: This is the username that will be used to log in to the FTP server. For security reasons, we do not store passwords between sessions. A pop-up dialog for entering the password will appear the first time that you store a database in each Pathway Tools session.

Directory to which archive files will be copied: Enter the full pathname for the directory. This is what you would give for a `cd` command, when using FTP to transfer files by hand. In most installations, this will *not* be the same directory from which remote users will retrieve archived PGDBs.

The second set of fields contain information that will be sent to the central registry server, in order for users at other sites to connect and retrieve your databases. These values *must* be filled in before contacting the registry server for the final step in publishing your PGDB, regardless of how you copy files to the FTP server prior to making them available for retrieval.

FTP server from which archived databases will be retrieved: This should be the Internet name or address of the server, as it would be entered by a user at another site on the Internet. Note that this may not be the same as the server name or address that you entered above for storing files, even though it may be the same computer.

Directory from which archived database will be retrieved: This should be a full pathname, as described above. In most instances, this will simply be `pub`, since by convention that is the standard directory for anonymous FTP access.

Once you have filled in all of the fields in the preferences dialog and saved your preferences, you can continue with the process of publishing local databases.

A NOTE ON LICENSING: The first time that you use any portion of the database sharing facility, whether for publishing your own databases or for retrieving databases from other sites, the click-through license agreement for the Pathway ToolsRegistry will appear. Please read carefully the text of the license agreement. If you are in agreement with all terms of the license agreement, click "I ACCEPT" to continue. If any of the terms of the license agreement are not acceptable, click "I DO NOT ACCEPT". The database sharing facility will not function until the license agreement has been accepted.

The main user interface for publishing your PGDBs will differ, depending on how you answered the basic setup questions. There are three possible cases, each documented separately below. Please refer to the section that matches the setup of your site.

Case 1: If you answered "Yes" to "use FTP for uploading", and "yes" to "is it necessary to move uploaded files before they can be retrieved".

> In this case, the publishing process will have two steps: "package & upload" and "register".

Step 1: Package and Upload

Select the corresponding checkbox for the databases that you intend to publish, and when you are satisfied with your choices, click the button labeled "Package and upload selected databases". The packaging and uploading process will probably take several minutes, during which a series of messages will detail the steps being performed. When the process is complete, a "Done" message will appear in the message window.

Before continuing with the "Register" step, you must now move your packaged databases to their final location on your FTP server, from which they can be retrieved using anonymous FTP. If you are not sure how to do this at your site, please consult your local system administrator.

Step 2: Register

As noted above, you *must* have moved your files to their proper locations for retrieval by outside users before performing this step. Select the databases that you want to publish by left-clicking the corresponding checkbox in the column labeled "Select for publishing". If you require users to execute a click-through license agreement for the database, select that checkbox as well.

Before completing the registration process, we strongly recommend that you click the "Refresh (check FTP server)" button. This command connects to the FTP server and directory that you have designated in the Database Sharing preferences dialog, and tries to obtain the file size (not the contents) of the archive file for each of the databases listed. If the file is missing, or for some reason Pathway Tools is unable to connect to the server, any error messages encountered are displayed in the message window, and the database is grayed out and will not be processed by the "Register" command. **NOTE**: When you click "Register", the same check is done for each database to be registered, and processing of all databases will be stopped. Running the "Refresh" command allows you to fix any problems *before* you actually try to register your databases.

If you plan to use a click-through license for any of your databases, the "Test a click-through license" button prompts for a URL, then displays the contents of that URL just as it will be displayed to users who want to install your databases. This gives you a chance to fine-tune the content and presentation of your license agreement files, before they are presented to actual users.

When you are satisfied with your choices, click the "Register selected files" button. If you have specified that a license agreement will be required for any of your database, you will be prompted for the URL of the license agreement. Assuming that there are no problems with your FTP server, the registration process should complete very quickly.

Case 2: "No" to "use FTP for uploading".

In this case, Pathway Tools automates only the packaging of your databases, and all of

the work of moving the archived databases to an FTP server is done outside of Pathway Tools.

Click on the corresponding checkbox for the databases that you intend to publish. When you are satisfied with your choices, click the "Package selected databases" button, and the packaging process will begin. Packaging normally takes a few minutes to complete, during which time a series of messages will keep you informed of the process's progress.

When packaging has completed, a message will show the location of the files containing the packaged databases. Before running the "Register" step, you must move these files into the locations that you entered in the Database Sharing preferences dialog, from which they can be retrieved by users at other locations.

Once the files have been moved to your FTP server and are ready to be retrieved, the "Register" step is identical to that for Case 1, above. Follow the instructions for Step 2 above to complete the process of publishing your databases.

Case 3: "Yes" to "use FTP for uploading", "no" to "is it necessary to move uploaded files before they can be retrieved".

In this case, the entire publishing operation can be accomplished with a single click.

Select the databases that you want to publish by left-clicking the corresponding checkbox in the column labeled **"Select for publishing"**. If you require users to execute a click-through license agreement for the database, select that checkbox as well. When you are satisfied with your choices, click the **"Publish selected databases"** button.

The process of packaging, storing, and registering your databases will take several minutes to complete. A sub-window at the bottom of the publishing dialog will display information about the various steps involved as they are performed.

If no errors are encountered during the publishing process, a **"Done"** message will appear in the message window. Your databases will be available for immediate retrieval by other Pathway Tools users. The **"Browse Downloadable PGDBs"** command should show your databases along with those contributed by other users.

5.1.3 About click-through licenses

If desired, you can require users at other sites to execute what is commonly referred to as a "click-through" license, before they are allowed to retrieve and install your databases on their system. If the user clicks on "**I DO NOT ACCEPT**", the database will not be retrieved from the FTP server on which it is stored.

In order to create a click-through license, you need to put the text of your license into a file, which can be accessed by means of a standard Web URL. Pathway Tools supports a limited set of HTML tags, to allow you to add boldface, italics, and a few other formatting options. Specifically, the **B, I, H1, H2, H3,** and **H4** tags are supported. Other HTML formatting, such as table directives, will be ignored. HTML formatting cannot be extended across line breaks: if you have boldface text that extends across multiple lines, you need to add <**B**> at the

beginning of each line, and **</B>** at the end of each line. Lines separated by only a single new-line will be filled, to allow the text to fit neatly into the window that Pathway Tools puts up. Multiple new-lines separate paragraphs. Note that the filling process does not preserve word breaks across lines, so it is necessary to put a space at the beginning of each line in a paragraph. For an example of a file with this type of formatting, see the SRI click-through license at:

http://bioinformatics.ai.sri.com/ptools/downloadable-database-license.html

If you view this file using the **"View -> Source"** command available in most Web browsers, you can see the formatting tags.

6 TROUBLESHOOTING

Pathway Tools detects the occurrence of an internal error. A dialog window pops up that contains an error message and asks you if you want to exit back to Unix, or to reset the software. You can select either choice with the mouse.

6.1 FREQUENTLY ASKED QUESTIONS

A list of Pathway Tools Frequently Asked Questions is available on the SRI Web site at **http:// www.ai.sri.com/pkarp/ptools/faq.html.**

6.2 REPORTING PROBLEMS

If you encounter problems with the software, or if you see errors in the scientific information in a data set, or if you have suggestions about the program, contact us by sending electronic mail to **biocyc-support@ai.sri.com**.

When the Pathway Tools software detects an internal error, it writes a descriptive file called **error.tmp** in your home directory. That file helps SRI Technical Support track down the problem, so you will need to mail the file to **biocyc-support@ai.sri.com**. Mail the file only once for every distinct bug that you encounter; that is, if you repeatedly encounter what appears to be the same bug, tell us about it only once.

Include a description of the problem, and be as thorough as possible. Include the names of relevant database objects, or a description of the operations you had performed using Pathway Tools before an error occurred. Also tell us what version of Pathway Tools you are running. The version number appears in the title bar of the Navigator main window. More complete version information can be obtained by typing this command at the Unix command line: **ptools -id**.

A GUIDE TO THE PATHWAY TOOLS SCHEMA

All Pathway/Genome Databases (PGDBs) used by the Pathway Tools software -- including the EcoCyc and MetaCyc PGDBs -- must conform to the schema (ontology) described herein. The objects and the relationships between these objects are utilized in this computerized description of metabolic and genomic information. Understanding the schema is essential for both users and developers of Pathway/Genome Databases who are using the Pathway Tools software.

In defining a conceptualization of knowledge for computer use, it is essential to employ precise definitions and distinctions. The fidelity of a computer representation determines the degree to which meaningful computations and analyses can be performed with the information in computer form. Unfortunately, many concepts in biology are not defined with the required precision. For example, a half dozen biologists could easily supply a half dozen conflicting definitions for the terms "gene," or "metabolic pathway." You may discover that our definitions of the class names and attribute names employed herein do not match the definitions that you prefer. We ask you to acknowledge that (a) biology is not yet well enough formalized that every biologist can expect to employ the same definitions, and (b) the definitions used in this document are much more thorough and precise (and therefore useful) than those offered in most biological databases.

Much of the discussion in this document refers to the EcoCyc database (DB), but the same schema is used for all other DBs managed by Pathway Tools. This schema may change in future versions of the software.

PGDBs are stored within a *frame knowledge representation system* (FRS). An FRS is a kind of object-oriented database system. The DB consists of a collection of *frames*, where each frame encodes information about a single object, such as an enzyme, a gene, or a biochemical pathway. For a more precise discussion of FRSs, see [5].

Instance frames describe specific biological objects, such as a specific gene or a specific metabolic pathway. *Class frames* describe general types of biological objects, such as the class of all genes. Each frame contains one or more *slots*. A slot describes an attribute or a property of the object that the frame represents. Each slot makes sense for (is valid in) a particular set of classes. For example, the slot **`EC-Number`** makes sense only for frames in the Reactions class, whereas the slot **`Synonyms`** is valid in all classes.

The current Pathway Tools ontology contains several hundred classes arranged in a taxonomic hierarchy. Figure 6-1 shows some of the major classes, and their relationships. An arrow that points from class A to class B indicates that B is a child of A, and therefore that A is a more general class that subsumes B. For example, the class **`Proteins`** can be subdivided into the subclasses **`Polypeptides`** (monomers) and **`Protein-Complexes`** (multimers). Subclasses inherit slots from their parents, for example, **`Polypeptides`** inherits all slots defined in **`Proteins`**, and some additional slots are also defined in **`Polypeptides`**. The classes in this figure whose names are shown in bold are described in more detail in the remainder of this appendix.

The top-level classes in this figure describe physical entities and processes. More specifically, **`Chemicals`** describes atoms and complete chemical compounds, **`Polymer-Segments`** describes regions within polymers such as proteins and DNA, and **`Organisms`** describes the

biological organism modeled within a PGDB. The class **Chemicals** is subdivided into small-molecular weight compounds (class **Compounds**) and atoms (not shown), and into macromolecules (**Macromolecules**). **Macromolecules** includes subclasses such as DNA and **RNA; DNA** includes subclasses that describe different types of replicons such as chromosomes and plasmids. The different subclasses of **Polymer-Segments** include different types of DNA sites such as transcription start sites and terminators, and longer regions such as genes. On the process side, **Generalized-Reactions** describes both individual biochemical reactions, and biochemical pathways. The class **Enzymatic-Reactions** describes information specific to the pairing of an enzyme with a reaction that the enzyme catalyzes, such as its activators, inhibitors, and cofactors.

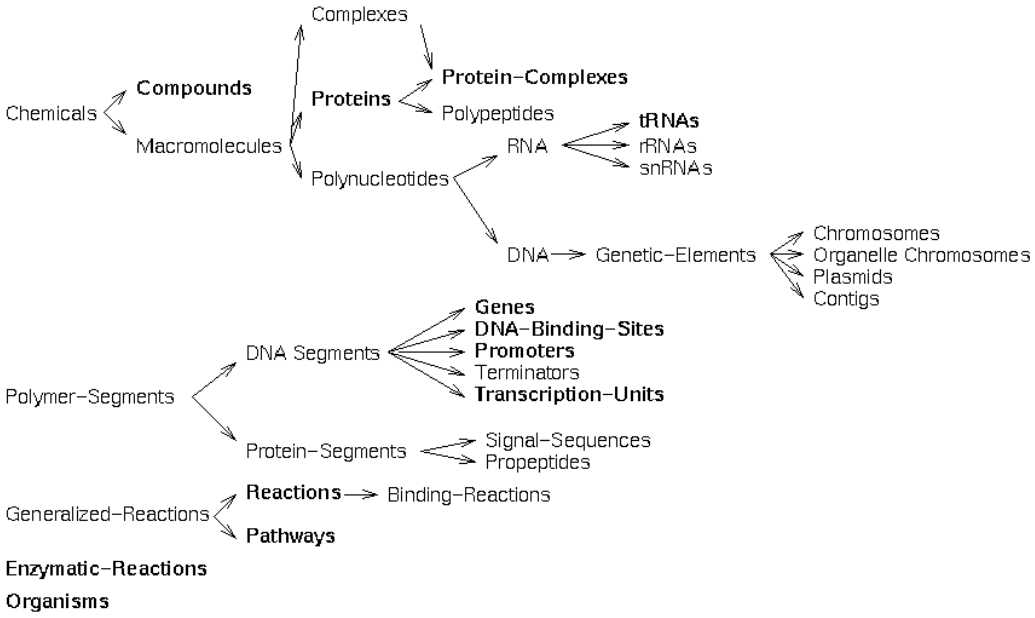


Figure 6-1 Some of the main classes defined in the schemas of Pathway/Genome Databases. The arrows denote the specialization–generalization relationship; for example, this figure indicates that all polypeptides are proteins, because class Proteins is the parent (generalization) of Polypeptides. A class inherits slots from its parents; therefore, the slots that are applicable to a given class are those slots inherited from its parent, plus those slots defined locally in that class.

The sections that follow describe the major classes and their slots in more detail. For additional discussions of the representations employed in the Pathway Tools ontology, see [10, 7].

A.1 SLOTS VALID IN MULTIPLE CLASSES

We begin with a discussion of slots that are used in several different EcoCyc classes.

This guide sometimes displays slot names capitalized, and sometimes in lowercase. In fact, all slot names are all uppercase in the DB itself. However, for the purpose of writing Lisp queries to Pathway/Genome Databases, all slot names can be written as lowercase in those queriesbecause the Lisp interpreter translates all symbol names to uppercase (except for symbol names written between vertical bars). Hyphens separate multiple words in a slot name.

A.1.1 Common-Name

This slot defines the primary name by which an object is known to scientists — a widely used and familiar name (in some cases arbitrary choices must be made). This field can have only one value; that value must be a string.

A.1.2 Synonyms

This field defines one or more secondary names for an object — names that a scientist might attempt to use to retrieve the object. These names may be out of date or ambiguous, but are used to facilitate retrieval — the Synonyms should include any name that you might use to try to retrieve an object. In a sense, “Synonyms” is misleading because the names listed in this slot may not be exactly synonymous with the preferred name of the object.

A.1.3 Names

Values of this slot are computed by combining the values of all other name-related slots for this frame: slots Common-Name, Systematic Name, Synonyms, N–Name, N–1–Name, and N+1–Name.

A.1.4 Comment

The Comment slot stores a general comment about the object that contains the slot. The comment should always be enclosed in double quotes.

A.1.5 Citations

This slot lists general citations pertaining to the object containing the slot. Each value of the slot is a citation of the form [reference-ID], where reference-ID is a Medline unique identifier.

A.2 CLASS BINDING REACTIONS

In a binding reaction, a set of reactants bind together to form a single product that is a complex of the reactants.

A.2.1 Reactants

The entities that are the reactants that bind together to form a complex.

A.2.2 Activators

This slot lists entities that activate the current binding reaction. To indicate an entity that is a complex of two other entities, use a list notation. For example, if a binding reaction involving RNA polymerase and a transcription start site is inhibited by a complex consisting of a DNA binding site and a protein, that complex could be indicated as `(TRP-OP. TRP-REP)` where `TRP-OP` is the identifier of an operator region, and `TRP-REP` is the identifier of a repressor protein.

A.2.3 Inhibitors

This slot lists entities that inhibit the current binding reaction. Use the same notation to indicate complexes as is used for the `activators` slot.

A.3 CLASS COMPOUNDS

The class Compounds describes small-molecular-weight chemical compounds — typically compounds that are substrates of metabolic reactions, or compounds that activate or inhibit metabolic enzymes.

A.3.1 Appears-In-Left-Side-Of, Appears-In-Right-Side-Of

Lists the one or more reactions in which this compound occurs as a reactant or product, respectively.

A.3.2 Aromatic-Rings

Each value in this slot is a list of atom numbers; that list of atoms constitutes a single aromatic ring. For example, the list might specify that atoms 1, 2, 5, 6, 10, 20 are in one aromatic ring (see slot Structure-Atoms).

A.3.3 Atom-Charges

This slot lists the charges of specific atoms within the compound. Each value of the slot is a list of the form (A C) where A is the index of an atom in slot Structure-Atoms, and C is the charge of that atom.

A.3.4 Charge

Lists the chemical charge for this compound.

A.3.5 Chemical-Formula

Lists the empirical formula for this compound. Each value of this slot is a list of the form (ATOM COUNT) where ATOM is the ID of a frame for the corresponding chemical element, and COUNT is the number of occurrences of that atom in this compound. For example, molecular oxygen, O_2, would be represented as (O 2) with a space in between the letter O and the number 2.

A.3.6 Display-Coords-2D

This slot lists coordinates for the display of the chemical structure of this compound in two dimensions. The values of this slot correspond one-to-one to the values of slot Structure-Atoms. Each value of this slot is a list of the form (X Y) and consists of the X-Y display coordinate of the corresponding atom in Structure-Atoms. The coordinates are real numbers with no specified minimum or maximum values they are re-scaled at display time.

A.3.7 Gibbs-0

Provides the standard Gibbs free energy of formation of the compound. The values are in units of kilocalories/mol, assuming the common state in aqueous solution at pH=7 and T=25C.

A.3.8 Molecular-Weight

Provides the molecular weight of this compound in Daltons.

A.3.9 N-Name, N-1-Name, N+1-Name

These slots are used when displaying the names of polymeric compounds in pathways that increase or decrease the lengths of the polymers. The names indicate a polymer of length N, length N-1, and length N+1. As an example, see the compound at **`http://biocyc.org/META/NEW-IMAGE?type=COMPOUND&object=CPD-1301`**.

A.3.10 Smiles

Provides a representation of the chemical structure of this compound using the SMILES chemical encoding system. Note that the value of this slot is computed using an attached procedure; do not attempt to store a value into this slot.

A.3.11 Structure-Atoms

This slot is one of several that are used to encode the chemical structure of a compound. This slot lists all the distinct atoms in the compound, with multiple entries for atoms of the same element that occur more than once. For example, water could be described as the list (H H O). The atoms are listed in no special order. However, other slots refer to the atoms in the compound according to their position in this list; for example, the first hydrogen is atom 0, and the oxygen is atom 2.

A.3.12 Structure-Bonds

This slot describes the chemical bonds within a compound. Each bond is encoded as a list of the form (A1 A2 B-TYPE) where A1 is the index in slot Structure-Atoms of the first atom in the bond, A1 is the index of the second atom in the bond, and B-TYPE encodes the type of the chemical bond. Valid bond types are the numbers 1, 2, and 3 for single, double, and triple bounds, and the symbol :AROMATIC for aromatic bonds. For example, to specify that a double bond exists between the first and fifth atoms, use the list (0 4 2).

A.4 CLASS DNA-BINDING-SITES

This class describes DNA regions that are binding sites for transcription factors.

A.4.1 Relative-Center-Distance

This slot defines the distance from the center of this binding site to another DNA region of interest.

A.5 CLASS ENZYMATIC REACTIONS

Frames in the class Enzymatic-Reactions describe attributes of an enzyme with respect to a particular reaction. For reactions that are catalyzed by more than one enzyme, or for enzymes that catalyze more than one reaction, multiple Enzymatic-Reactions frames are created one for each enzyme/reaction pair. For example, Enzymatic-Reactions frames can represent the fact that two enzymes that catalyze the same reaction may be controlled by different activators and inhibitors. See [10] for more details.

A.5.1 Enzyme

This slot lists the enzyme whose activity is described in this frame. More specifically, the value of this slot is the key of a frame from the class Protein-Complex or Polypeptide.

A.5.2 Required-Protein-Complex

Some enzymes catalyze only a particular reaction when they are components of a larger protein complex. For such an enzyme, this slot identifies the particular protein complex of which the enzyme must be a component.

A.5.3 Reaction

The value of this slot is the key of a frame from the Reaction class — the second half of the enzyme/reaction pair that the current frame describes. In fact, this slot can have multiple values, which encode the multiple reactions that one catalytic site of an enzyme catalyzes.

A.5.4 Activators, Inhibitors

This collection of slots lists inhibitor and activator compounds for this enzymatic reaction. Each compound is listed under the slot whose name best describes the mechanism of action of the compound, if known. For competitive activators or inhibitors, the competing compound is currently listed in a comment. The full list of activator and inhibitor slot names is

ACTIVATORS-ALLOSTERIC

ACTIVATORS-NONALLOSTERIC

ACTIVATORS-UNKMECH

INHIBITORS-COMPETITIVE

INHIBITORS-NONCOMPETITIVE

INHIBITORS-UNCOMPETITIVE

INHIBITORS-ALLOSTERIC

INHIBITORS-IRREVERSIBLE

INHIBITORS-NEITHER

INHIBITORS-UNKMECH

The fact that a compound is listed in this slot makes no commitment as to whether the effects of the compound on the enzyme are of physiological relevance, as opposed to whether the effects are known purely because of in vitro studies.

A.5.5 Physiologically-Relevant

This slot is a companion to the preceding slots for activators and inhibitors. It contains a list of those activators and inhibitors whose effects are known to be of physiological relevance. That is, this slot indicates which of the compounds named in the preceding slots are physiologically relevant. The values of this slot are therefore a (possibly empty) subset of all values of the activators and inhibitors slots.

A.5.6 Cofactors, Prosthetic-Groups

The literature uses terms such as coenzyme, cofactor, and prosthetic group in an extremely inconsistent fashion. In version 2.8 of EcoCyc (March 1996) we adopted the usage of these terms that was developed by Eugeni Selkov (Gene Selkov) for use in the Enzymes and Metabolic Pathways (EMP) database.

"Class Reactions" in section A.13 defined the substrates of a reaction as the union of its reactants and its products. After Selkov, we define a coenzyme to be a specialization of substrates, namely, substrates with a relatively stable, conserved moiety, whose main function is group transfer among different enzymes and pathways. Example: NAD. EcoCyc does not define a special slot for coenzymes.

Also after Selkov, we define cofactors and prosthetic groups to be compounds that are required for an enzyme to catalyze a reaction, but that are unchanged by the reaction. Thus, cofactors and prosthetic groups are (loosely speaking) activators of an enzyme in the sense that the enzyme is not active when these compounds are absent. However, cofactors and prosthetic groups have an infinite "activation degree," thus distinguishing them from those compounds that will be listed in the **`Activators`** slots (whose definitions were given earlier in this section); when **`Activators`** are missing, the enzyme still functions, but at a lower rate.

The distinction between cofactors and prosthetic groups is that prosthetic groups are covalently or tightly bound to an enzyme, whereas cofactors are not. The corresponding slot names are **`Cofactors`** and **`Prosthetic-Groups`**.

A slot called **`Cofactors-Or-Prosthetic-Groups`** identifies compounds whose binding affinity to the enzyme is unclear.

A.5.7 Alternative-Substrates, Alternative-Cofactors

These slots record variability in the substrates and cofactors that have been observed for this enzymatic reaction. If, for example, the literature indicates that Mn+2 can substitute for Mg+2 as a cofactor in this reaction, we would list the following as a value for the Alternative-Cofactors slot: (Mg+2 Mn+2).

The Alternative-Substrates slot describes the substrate specificity of an enzymatic reaction. We use the Alternative-Substrates when the complete equation is not known for an alternative reaction, or when the alternative reaction is not physiologically important, or is not a member of a known pathway. Each value of the **`Alternative-Substrates`** slot is a list whose first member is a compound that was specified as a substrate; the remaining elements of the list are compounds that can serve as alternative substrates for the first compound.

Each value of the **`Alternative-Cofactors`** slot is a list whose first member is a compound that was specified as a cofactor or prosthetic group; the remaining elements of the list are compounds that can serve as alternatives for the first compound.

An annotation on a value for either of these slots is assumed to apply to each alternative substrate/cofactor listed in the value. If an annotation is intended to apply to only one such compound (or other subset), two (or more) values should be used instead, where the substrate is repeated as the first element of each value, and the alternative compounds are divided among the values according to the applicability of the annotations.

A.5.8 Reaction-Direction

This slot specifies the directionality of a reaction. The slot is particularly important to fill for reactions that are not part of a pathway, because for such reactions, the direction cannot be determined automatically, whereas for reactions within a pathway, the direction can be inferred from the pathway context. This slot aids the user and software in inferring the direction in which the reaction typically occurs in physiological settings, relative to the direction in which the reaction is stored in the database. Possible values of this slot are

:REVERSIBLE: The reaction occurs in both directions in physiological settings.

:PHYSIOL-LEFT-TO-RIGHT, :PHYSIOL-RIGHT-TO-LEFT: The reaction occurs in the specified direction in physiological settings, because of several possible factors including the energetics of the reaction, local concentrations of reactants and products, and the regulation of the enzyme or its expression.

:IRREVERSIBLE-LEFT-TO-RIGHT, :IRREVERSIBLE-RIGHT-TO-LEFT: For all practical purposes, the reaction occurs only in the specified direction in physiological settings, because of chemical properties of the reaction.

A.6 CLASS GENES

Each frame in the class **`Genes`** describes a single gene, meaning a region of DNA that defines a coding region for one or more gene products. Multiple gene products may be produced due to modification of an RNA or protein.

A.6.1 Left-End-Position, Right-End-Position

These slots encode the position of the left and right ends of the gene on the chromosome or plasmid on which the gene resides. "Left" means the end of the gene toward the coordinate-system origin (0). Therefore, the **`Left-End-Position`** is always less than the **`Right-End-Position`**.

In the EcoCyc DB, the values of this slot were taken directly from Genbank entry U00096 submitted by the Blattner laboratory.

A.6.2 Centisome-Position

This slot lists the map position of this gene on the chromosome in centisome units (percentage length of the chromosome). The centisome-position values are computed automatically by Pathway Tools from the **`Left-End-Position`** slot. The value is a number between 0 and 100, inclusive.

A.6.3 Transcription-Direction

This slot specifies the direction along the chromosome in which this gene is transcribed; allowable values are "+" and "-".

A.6.4 Product

This slot holds the ID of a polypeptide or tRNA frame, which is the product of this gene. This slot may contain multiple values for two possible reasons: a given gene might be translated from more than one start codon, giving rise to products of different lengths; the product of the gene may undergo chemical modification. In the latter case, the gene lists all modified forms of the protein in its **`Product`** slot.

A.6.5 Evidence

Describes evidence for the defined function of this gene. Currently, we distinguish between function that is determined experimentally (value :**`Experiment`**), and function that is determined through computational sequence analysis (value :**`Sequence-Analysis`**).

A.6.6 Interrupted?

If True, indicates that the specified gene is interrupted, that is, has a premature stop codon.

A.7 CLASS ORGANISMS

Each PGDB contains a single frame that is an instance of class Organisms, and whose unique ID is the same as the unique ID that was assigned to the PGDB itself when the PGDB was created. For example, the BsubCyc PGDB contains a frame whose unique ID is BSUB. This frame encodes information about the organism described by the PGDB, and encodes information about the PGDB itself. Much of the latter information is displayed by the Pathway/Genome Navigator, such as in the organism summary page for the organism.

A.7.1 PGDB-Authors

A list of the names of the authors of this DB. The names are displayed on a summary page for this organism. It is appropriate to suffix each name with the author's institution, for example, "John Doe, University of New Jersey". Use one slot value per author.

A.7.2 PGDB-Copyright

The contents of this slot should be a copyright notice for this database, if one is desired. The copyright notice should preferably fit in one line because it will be printed at the bottom of every Web page served for this organism database by the Pathway Tools Web server. Example: “Copyright 1999 University of New Jersey.”

A.7.3 PGDB-Footer-Citation

The value of this slot should be a single literature citation, in the form of a string, such as "Bioinformatics 12:155 2002". This citation, if present, is printed at the bottom of each Web

page served for this organism, within the following text: "Please cite XXXCyc as **CITATION** in publications resulting from its use."

A.7.4 PGDB-Home-Page

The URL of a Web page describing this PGDB. Authors can use this page to provide more background information about the PGDB.

A.7.5 PGDB-Name

The name of the database for this organism, when the database name is to be printed somewhere by Pathway Tools. Examples: "EcoCyc," "PlasmoCyc." The suffix "Cyc" is not required.

A.7.6 PGDB-Unique-ID

An integer unique ID for this PGDB that differentiates it from other PGDBs. This ID is used to build unique IDs for newly created frames that are created in this PGDB so that (a) when frames are copied among PGDBs, we know what PGDB the frame originated in, and (b) we can ensure that two frames in two different PGDBs that have the same ID do in fact refer to the same biological entity. The MetaCyc DB has a PGDB-Unique-ID of NIL; all other DBs should have a non-NIL value for this slot.

A.7.7 Strain-Name

Specifies the strain name for the organism.

A.7.8 Taxonomic-Domain

The taxonomic domain to which this organism belongs. Possible values are :EUBACTERIA, :EUKARYOTES, :ARCHAEA, :VIRUSES.

A.7.9 Contact-Email

The email address of a person who serves at the primary contact for this PGDB, such as to receive questions or bug reports from users of the PGDB.

A.7.10 Genome

A list of all replicons (chromosomes and plasmids) in the genome of the organism.

A.8 CLASS PATHWAYS

Frames in class Pathways encode metabolic and signaling pathways.

A.8.1 Net-Reaction-Equation

This slot specifies the net chemical transformation accomplished by a pathway, including the stoichiometry, and is written in the same form as the Reaction Equation slot.

A.8.2 Pathway-Interactions

This slot holds a comment that describes interactions between this pathway and other biochemical pathways, such as those pathways that supply an important precursor.

A.8.3 Predecessors

This slot describes the linked reactions that compose the current pathway. Since pathways have a variety of topologies — from linear to circular to tree structured — pathways cannot be represented as simple sequences of reactions. A pathway is a list of reaction/predecessor pairs. That is, each value of this slot is of the form (reaction-ID pred-ID*) where reaction-ID is the key of a reaction in the pathway, and each pred-ID is the key of a reaction in the pathway that directly precedes the reaction-ID reaction. For example, to represent the combined pathway for tyrosine and phenylalanine synthesis (see Figure 6-2), this predecessor list might be used:

predecessors:(chorismatemut-rxn),

(prephenatedehydrat-rxn chorismatemut-rxn),

(pheaminotrans-rxn prephenatedehydrat-rxn),

(prephenatedehydrog-rxn chorismatemut-rxn),

(tyraminotrans-rxn prephenatedehydrog-rxn)

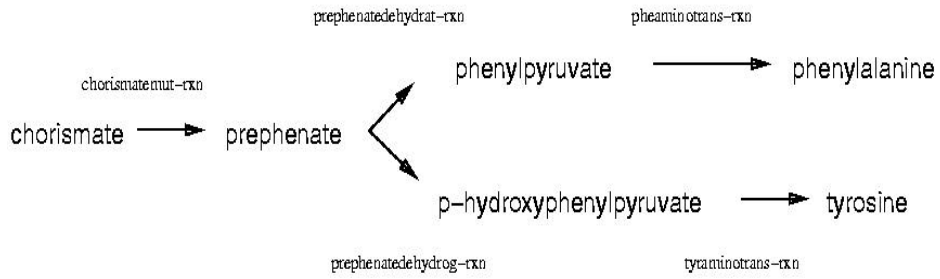


Figure 6-2 This pathway shows the combined synthesis of the amino acids tyrosine and phenylalanine from chorismate. Each reaction is labeled by its key in the EcoCyc DB. For example, the key for the reaction that converts prephenate to phenylpyruvate is prephenatedehydrat-rxn.

The first reaction in the pathway has no predecessor, so there is only one key within the first value. Since prephenate is a branch point in the pathway, two reactions in the pathway list the reaction that synthesizes prephenate as a predecessor.

Alternatively, any value for this slot can be another pathway key, which means that the current pathway inherits all the predecessor values of the indicated pathway. In other words, the current pathway is a superpathway of the indicated pathway. Thus, a more compact way of representing the combined pathway for tyrosine and phenylalanine synthesis would be to use the following predecessor list:

predecessors: tyrsyn, phesyn

In actuality, this latter representation is the preferred one and is required in order for the combined pathway to be determined to be a superpathway of either the tyrosine or phenylalanine (which of course should be the case). The advantage of specifying the predecessor list in this way (aside from being more compact and easy to read) is that if the subpathway is ever modified, the changes will automatically propagate to the superpathway.

A.8.4 Reaction-List

This slot lists all reactions in the current pathway, in no particular order. It is computed from the predecessors slot.

A.8.5 Hypothetical-Reactions

A list of reactions in this pathway that are considered hypothetical, probably because presence of the enzyme has not been demonstrated.

A.8.6 Assume-Unique-Enzymes

By default it is assumed that all enzymes that can catalyze a reaction will do so in each pathway in which the reaction occurs. That default assumption is encoded by the default value of **FALSE** for this slot; when you want to assume that only one enzyme exists in the DB to catalyze every reaction in this pathway, this slot should be given the value **TRUE**.

This slot can be used for consistency-checking purposes, that is, in a pathway for which this slot is **TRUE**, there should not exist any reactions that are catalyzed by more than one reaction.

A.8.7 Enzyme-Use

By default it is assumed that all enzymes that can catalyze a reaction will do so in each pathway in which the reaction occurs. This slot is used in the case that this assumption does not hold, that is, if a reaction is catalyzed in a particular pathway by only a subset (or none) of the possible enzymes that are known to catalyze that reaction. Therefore, this slot can be used only when the value of the **assume-unique-enzymes** slot is **FALSE** (because multiple enzymes catalyze some step in the pathway).

The form of a value for the slot is **(reaction-ID enzymatic-reaction-ID-1... enzymatic-reaction-ID-n)**. That is, each value specifies a reaction, and specifies the one or more enzymatic reactions that catalyze that reaction in this pathway. If no enzymatic reactions are specified, then none of the enzymes that are known to catalyze the reaction do so in this pathway.

For example, under aerobic conditions the oxidation of succinate to fumarate is catalyzed by succinate dehydrogenase in the forward direction, and, under anaerobic conditions, by fumarate reductase in the reverse direction. The TCA cycle is active only in aerobic conditions, so only succinate dehydrogenase is used in this pathway. This fact would be recorded as follows:

enzyme-use: (succ-fum-oxred-rxn succinate-oxn-enzrxn)

A.8.8 Primaries

When drawing a pathway, the Navigator software usually computes automatically which compounds are primaries (mains) and which compounds are secondaries (sides). Occasionally, the heuristics used are not sufficient to make the correct distinction, in which case you can specify primary compounds explicitly. This slot can contain the list of primary reactants, primary products, or both for a particular reaction in the pathway. Each value for this slot is of the form **(reaction-ID (primary-reactant-ID-1 ... primary-reactant-ID-n) (primary-product-ID-1 ... primary-product-ID-n))**, where an empty list in either the reactant or product position means that that information is not supplied and should be computed.

For example, in the purine synthesis pathway, we want to specify that the primary product for the final reaction in the pathway should be AMP and not fumarate. The primary reactants are still computed. The corresponding slot value would be

primaries: (ampsyn-rxn () (amp))

A.8.9 Species

This slot is used only in pathway frames in the MetaCyc DB, in which case the slot identifies the one or more species in which this pathway is known to occur experimentally.

A.8.10 Disable Display

When the value is true, this slot disables display of the pathway drawing for a pathway.

A.8.11 Super-Pathways

This slot lists direct super-pathways of a pathway. Its value is computed by the Pathway Tools consistency-checking code by comparing the predecessors slots of each pathway in the DB, and therefore should not be set by the user. A pathway key does not have to be explicitly specified in a predecessor slot in order for the superpathway relationship to be detected.

A.8.12 Sub-Pathways

This slot is the inverse of the Super-Pathways slot. It lists all the direct subpathways of a pathway. The values of this slot are computed automatically; see slot **Super-Pathways**.

A.8.13 Pathway-Links

This slot indicates linkages among pathways in pathway drawings. Each value of this slot is a list of the form **(cpd other-pwy*)**. The Navigator draws an arrow from the specified compound pointing to the names of the specified pathways, to note that the compound is also a substrate in those other pathways. If no other pathways are specified, then links are drawn to and from all other pathways that the compound is in (i.e., if the compound is produced by the current pathway, then links are drawn to all other pathways that consume it, and vice versa).

A.8.14 Polymerization-Links

This slot controls drawing of polymerization relationships within a pathway. Each value of this slot is of the form **(cpd-class product-rxn reactant-rxn)**. When both reactions are non-nil, an identity link is created between the polymer compound class cpd-class, a product of product-rxn, and the same compound class as a reactant of reactant-rxn. The **PRODUCT-NAME-SLOT** and **REACTANT-NAME-SLOT** annotations specify which slot should be used to derive the compound label in product-rxn and reactant-rxn above, respectively if one or both are omitted, **COMMON-NAME** is assumed. Either reaction above may be nil; in this case, no identity link is created this form is used solely in conjunction with one of the name-slot annotations to specify a name-slot other than **COMMON-NAME** for a polymer compound class in a reaction of the pathway.

A.8.15 Class-Instance-Links

Each value of this slot is a reaction in the pathway. Two annotations (in addition to the usual possibilities) are available on this slot: **REACTANT-INSTANCES** and **PRODUCT-INSTANCES**, whose values are compounds. If one of the reactants of the slot-value reaction is a class *C* and the **REACTANT-INSTANCES** are instances of *C*, then the instances are drawn as part of the pathway, with identity links to the class. The **PRODUCT-INSTANCES** are treated similarly.

A.8.16 Layout-Advice

Each value of this slot is a dotted pair of the form **(advice-keyword . advice**, and represents some piece of advice to the automatic pathway layout code. Currently supported advice keywords are

- **:CYCLE-TOP-CPD**: The advice is a compound key. In pathways containing a cycle, the cycle will be rotated so that the specified compound is positioned at twelve o'clock.
- **:REVERSIBLE-RXNS**: The advice is a list of reactions that should be drawn as reversible, even when the pathway is being drawn to show pathway flow (rather than true

reversibility).

:CASCADE-RXN-ORDERING: The advice is a list of reactions that form a partial order for reactions in a cascade pathway (i.e. the 2-component signalling pathways).

A.9 CLASS POLYPEPTIDES

Frames of class Polypeptides are monomers consisting of a single polypeptide chain.

A.9.1 Gene

This slot contains a value that identifies the gene that encodes the polypeptide.

When a polypeptide exists in two forms, modified and unmodified, both forms contain the same value in their **Gene** slots.

A.9.2 Features

This slot links the polypeptide to any protein features that have been defined for it. When a polypeptide exists in multiple forms, each form will link to the same set of features.

A.9.3 Splice-Form-Introns

This slot lists any introns that were spliced out of the gene in order to generate this polypeptide. Values of this slot are of the form (start-bp end-bp).

A.10 CLASS PROMOTERS

Frames in this class define transcription start sites.

A.10.1 Absolute-Plus-1-Pos

The absolute base pair position of the transcription start site on the DNA strand.

A.11 CLASS PROTEIN-COMPLEXES

Frames of class Protein-Complexes are multimeric proteins composed of multiple subunits. The subunits of a protein complex may themselves be protein complexes, although eventually the subunits must bottom out as polypeptides.

A.11.1 Components

This slot lists the subunits of a protein complex. Each subunit is either a polypeptide or a protein complex; therefore, each slot value is the key of a polypeptide frame or a protein-complex frame.

The coefficient of each component of the protein complex is listed as an annotation of the component value under the label **`Coefficient`** [5].

A.12 CLASS PROTEINS

The class of all proteins is divided into two subclasses: protein complexes and polypeptides. A polypeptide is a single amino acid chain produced from a single gene. A protein complex is a multimeric aggregation of more than one polypeptide subunit. A protein complex may in some cases have another protein complex as a component.

A.12.1 Component–Of

This slot lists the protein complex(es) that this protein is a component of, if any. Protein complexes may contain other protein complexes as components.

A.12.2 DNA-Footprint-Size

For proteins that bind to DNA, the number of base pairs on the DNA strand that the binding protein covers.

A.12.3 Locations

This slot describes the one or more cellular locations in which this protein is found, one of

`:Cytoplasm`: Protein is floating in the cytoplasm

`:Membrane`: Associated with either the inner or the outer membrane

`:Inner-Membrane`: Somehow associated with, or a component of, the inner membrane

`:Outer-Membrane`: Somehow associated with, or a component of, the outer membrane

`:Periplasm`: Space between the inner and outer membrane

`:Extracellular`: Protein is secreted to the external environment of the cell

A.12.4 Modified-Form

This slot points from the unmodified form of a protein, to one or more chemically modified forms of that protein. For example, the slot might point from the unmodified form of a polypeptide (or a protein complex) to a phosphorylated form of that polypeptide (or protein complex).

A.12.5 Molecular-Weight-KD

This computed slot lists the known molecular weight(s) of a macromolecule in kilodaltons by taking the union of the slots Molecular-Weight-Seq and Molecular-Weight-Exp.

A.12.6 Molecular-Weight-Seq

This slot lists the molecular weight of the protein complex or polypeptide, as derived from sequence data. Units: kilodaltons.

A.12.7 Molecular-Weight-Exp

This slot lists the molecular weight of the protein complex or polypeptide, derived experimentally. Multiple values of this slot correspond to multiple experimental observations. Units: kilodaltons.

A.12.8 Neidhardt-Spot-Number

This slot lists the Neidhardt-Spot-Number of the protein, reflecting its electrophoretic behavior in two-dimensional electrophoresis [15].

A.12.9 pI

This slot lists the pI of the polypeptide.

A.12.10 Species

This slot is used in proteins only in the MetaCyc DB, in which case it identifies the species in which the current protein is found.

A.12.11 Unmodified-Form

This slots points from a chemically modified form of some protein, to the native unmodified form of that protein (e.g., from a phosphorylated form to the unphosphorylated form).

A.13 CLASS REACTIONS

Frames within the Reaction class describe properties of a biochemical reaction independent of any enzyme or enzymes that catalyze that reaction. A reaction is a biochemical transformation that interconverts two sets of chemical compounds (which includes small metabolites, proteins, and DNA regions), and may translocate compounds from one cellular compartment to another. Most reactions are written in a conventional direction that has been assigned by the Enzyme Nomenclature Commission, but that direction may or may not be the predominate physiological direction of the reaction. Reaction substrates can include small-molecular-weight compounds (for metabolic reactions), proteins (such as in signaling pathways), and DNA sites (such as for reactions involving binding of transcription factors to DNA).

Two novel features of our conceptualization with respect to previous metabolic databases are to separate reactions from the enzymes that catalyze them, and to use the EC numbers defined by the International Union of Biochemistry and Molecular Biology (IUBMB) to uniquely identify reactions, not enzymes. (In database terms, the EC number is a key for the Reaction class.) The reason for this separation is that the *catalyzes* relationship between reactions and enzymes is many-to-many: a given enzyme might catalyze more than one reaction, and the same reaction might be catalyzed by more than one enzyme. Frames in the class Enzymatic-Reaction describe the association between an enzyme and a reaction.

You should always write transport reactions in the predominate direction in which the reaction occurs. Transport reactions are encoded by labeling substrates with their cellular compartment. For example, if a given substrate is transported from the periplasm to the cytoplasm, it would be labeled with "periplasm" as its compartment as a reactant, and with "cytoplasm" as its compartment as a product. The default compartment is the cytoplasm, so the cytoplasm label may be omitted. These labels are implemented as annotations in Ocelot.

A.13.1 EC-Number

This slot holds the EC (Enzyme Commission) number associated with the current reaction, if such a number has been assigned by the IUBMB. This slot is single valued.

A.13.2 Official-EC?

The value of this slot is NO if the current reaction either was not defined at all by the Enzyme Commission, or if the current equation stored for that reaction is not the equation assigned by the EC (e.g., we have corrected the EC equation). Otherwise, the value is YES, which is the default inherited value.

A.13.3 Left, Right

These slots hold the compounds from the left and right sides, respectively, of the reaction equation. Each value is either the key of a compound frame, or a string that names a compound (when the compound is not yet described within the DB as a frame). The terms *reactant* and *product* are not used because these terms may falsely imply the physiological direction of the reaction.

The coefficient of each substrate, when that coefficient is not equal to 1, is stored as an annotation on the substrate value. The annotation label is **`COEFFICIENT`**.

The substrates of transport reactions are also described using the **`Left`** and **`Right`** slots. However, the values of these slots are annotated to indicate their compartments. For example, a transporter that moves succinate from the periplasm to the cytoplasm, accompanied by hydrolysis of ATP in the cytoplasm, would be described with **`succinate`** and **`ATP`** as the values of the **`Left`** slot, and with **`succinate`**, **`ADP`**, and P*i* as the values of the **`Right`** slot. The **`succinate`** in the **`Left`** slot would be annotated with **`Periplasm`** under the label **`Compartment`**. The other substrates need not be annotated with a compartment because the default compartment is taken to be the cytoplasm.

A.13.4 Substrates

The value of this slot is computed automatically — its values may not be changed by the user. The values of the slot are computed as the union of the values of the **`Left`** and **`Right`** slots.

A.13.5 DeltaG0

This slot contains the change in Gibbs free energy for the reaction in the direction the reaction is written.

A.13.6 Spontaneous?

This slot is true in the case when this reaction occurs spontaneously, that is, it is not catalyzed by any enzyme.

A.13.7 Species

This slot is used to indicate that a reaction is known to occur in an organism in the case where the enzyme that catalyzes the reaction is unknown. In such cases, the value for this slot in a given reaction would be the symbolic identifier of the species for the organism for the current PGDB.

A.13.8 Balance-state

This slot describes what is known about the element balance state of this reaction. Its possible values are:

NIL -- The balance state is unknown

:BALANCED -- The last time the reaction-balancing code checked this reaction, it was balanced.

:UNBALANCED-UNKNOWN -- The reaction is unbalanced, and the curation team cannot determine how to balance it.

:UNBALANCED-INCOMPLETE -- The reaction is unbalanced because the complete reaction equation is not known.

:UNBALANCED-DIFFERING-R-GROUPS -- The reaction is unbalanced because it contains substrates with R-groups where the structures of the two R-groups differ.

A.14 CLASS TRANSCRIPTION-UNITS

Frames in this class encode transcription units, which are defined as a set of genes and associated control regions that produce a single transcript. Thus, there is a one-to-one correspondence between transcription start sites and transcription units. If a set of genes is controlled by multiple transcription start sites, then a PGDB should define multiple transcription-unit frames, one for each transcription start site.

A.14.1 Components

The **Components** slot of a transcription unit lists, in 5' to 3' order, the DNA segments within the transcription unit, including transcription start sites, DNA binding sites, and genes.

A.14.2 Extent-Unknown?

The value of this slot should be True when it is not known to how many genes the transcription unit extends; that is, it is not known which is the last gene in the transcription unit.

A.15 CLASS TRNAS

Frames of this class encode both charged and uncharged tRNAs.

A.15.1 Anticodon

This slot contains a string as a single value, which lists the three letters that make up the anticodon bases on the tRNA. The direction in which the letters are listed is 5' to 3' with respect to the tRNA. This is the reverse of, and complementary to, the sequence of the recognized codons.

A.15.2 Codons

This slot contains possibly multiple values as strings, which list the three letters that make up the base triplets recognized by the anticodon on the tRNA. The direction in which the letters are listed is 5' to 3' with respect to the coding strand of genes.

BIBLIOGRAPHY

1. A. Bairoch. The enzyme data bank in 1995. *Nuc. Acids R*es., 24:221--222, 1996.
2. F.R. Blattner, G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, and Y. Shao. The complete genome sequence of *Escherichia coli* k--12. *Science*, 277:1453--1462, 1997.
3. NCBI DDBJ, EMBL. *The DDBJ/EMBL/GenBank Feature Table Definition*, version 2.0 edition, December 1997. http://www.ncbi.nlm.nih.gov/collab/FT/index.html.
4. P. Karp. Database links are a foundation for interoperability. *Trends in Biotechnology*, 14:273--279, August 1996.
5. P. Karp and T. Gruber. The generic frame protocol. Available via World Wide Web URL `http://www.ai.sri.com/~gfp/doc/paper.html`, 1995.
6. P. Karp and M. Mavrovouniotis. Representing, analyzing, and synthesizing biochemical pathways. *IEEE Expert*, 9(2):11--21, 1994.
7. P. Karp and S. Paley. Representations of metabolic knowledge: Pathways. In R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls, editors, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 203--211, Menlo Park, CA, 1994. AAAI Press.
8. P. Karp and S. Paley. Automated drawing of metabolic pathways. In H. Lim, C. Cantor, and R. Robbins, editors, *Proceedings of the Third International Conference on Bioinformatics and Genome Research*, pages 225-238. World Scientific Publishing Co., 1995. See also WWW URL ftp://ftp.ai.sri.com/pub/papers/karp-bigr94.ps.Z.
9. P. Karp and S. Paley. Integrated access to metabolic and genomic data. *Journal of Computational Biology*, 3(1):191-212, 1996.
10. P. Karp and M. Riley. Representations of metabolic knowledge. In L. Hunter, D. Searls, and J. Shavlik, editors, *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, pages 207-215, Menlo Park, CA, 1993. AAAI Press.
11. P. Karp, M. Riley, S. Paley, A. Pellegrini-Toole, and M. Krummenacker. EcoCyc: Electronic encyclopedia of E. coli genes and metabolism. *Nuc. Acids Res.*, 26(1):50--53, 1998.
12. P.D. Karp. A knowledge base of the chemical compounds of intermediary metabolism. *Computer Applications in the Biosciences*, 8(4):347--357, 1992.
13. M. Riley. Functions of the gene products of Escherichia coli. *Microbiological Reviews*, 57:862--952, 1993.
14. J.-F. et al. Tomb. The complete genome sequence of the gastric pathogen helicobacter pylori. *Nature*, 388:539--547, 1997.
15. R.A. VanBogelen, P. Sankar, R.L. Clark, J.A. Bogan, and F.C. Neidhardt. The gene-protein database of *Escherichia coli:* Edition 5. *Electrophoresis*, 13:1014–1054, 1992.
16. Edwin C. Webb. Enzyme Nomenclature, 1992: Recommendations of the nomenclature

committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. Academic Press, 1992.

17. D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31--36, 1988.

INDEX

L

M

O

P

Q

R

S

T

U

V

W

X