Briefings in Bioinformatics, 17(5), 2016, 877–890

doi: 10.1093/bib/bbv079
Advance Access Publication Date: 10 October 2015
Paper

Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology

Peter D. Karp, Mario Latendresse, Suzanne M. Paley, Markus Krummenacker, Quang D. Ong, Richard Billington, Anamika Kothari, Daniel Weaver, Thomas Lee, Pallavi Subhraveti, Aaron Spaulding, Carol Fulcher, Ingrid M. Keseler and Ron Caspi

Corresponding author: Peter D. Karp, Bioinformatics Research Group, SRI International 333 Ravenswood Ave, Menlo Park, CA 94025, USA. E-mail: pkarp@ai.sri.com

Abstract

Pathway Tools is a bioinformatics software environment with a broad set of capabilities. The software provides genome-informatics tools such as a genome browser, sequence alignments, a genome-variant analyzer and comparative-genomics operations. It offers metabolic-informatics tools, such as metabolic reconstruction, quantitative metabolic modeling, prediction of reaction atom mappings and metabolic route search. Pathway Tools also provides regulatory-informatics tools, such as the ability to represent and visualize a wide range of regulatory interactions. This article outlines the advances in Pathway Tools in the past 5 years. Major additions include components for metabolic modeling, metabolic route search, computation of atom mappings and estimation of compound Gibbs free energies of formation; addition of editors for signaling pathways, for genome sequences and for cellular architecture; storage of gene essentiality data and phenotype data; display of multiple alignments, and of signaling and electron-transport pathways; and development of Python and web-

Peter D. Karp is the Director of the Bioinformatics Research Group at SRI International. He received the PhD degree in Computer Science from Stanford University. SRI International is a multi-disciplinary non-profit research institute headquartered in the San Francisco area.

Mario Latendresse received a Ph.D. in Computer Science and has published many papers in bioinformatics, programming language design and compilation, malware detection, functional languages, and high performance computing.

Suzanne M. Paley is a Computer Scientist in the Bioinformatics Research Group at SRI, with interests in representation and visualization of biological knowledge. She has M.S. degrees in computer science and chemistry, and has been a developer of Pathway Tools since the project's inception.

Markus Krummenacker, a Scientific Programmer, has worked Pathway Tools/BioCyc for 15 years. His interests reside in the intersection and synergy between (bio-)chemistry and computer science, ultimately culminating in atomically precise manufacturing.

Quang D. Ong received a BS in Industrial Management; he works as a Scientific Programmer and System Administrator.

Richard Billington received a MS from the University of Pennsylvania in Computer and Information Science. He is a Senior Software Developer at SRI.

Anamika Kothari is a Scientific Database Curator at SRI International. She holds a graduate Degree in Chemistry and Biochemistry, from Mumbai University, India.

Daniel Weaver is a systems biologist and holds a PhD in Biophysics from the University of Michigan. His research interests focus on cell metabolism and immune activity in the human immune system, and more generally on systems biology.

Thomas Lee is a Senior Research Engineer at SRI International. Thomas has a BS in Computer Science from the University of Nebraska and a MS in Computer Science from the University of Wisconsin; his research interests include machine learning, intelligent control, planning and scheduling.

Pallavi Subhraveti is a Scientific Programmer/Release Manager at SRI International. She received a BS in Biology from UC Riverside and a MS in Computer Science from Cal State Northridge.

Aaron Spaulding is a Senior Computer Scientist and Interaction Designer at SRI s Artificial Intelligence Center, working in the intersection of design, user experience, and artificial intelligence. Mr. Spaulding holds a Masters of Science Degree in Human-Computer Interaction from Carnegie Mellon University.

Carol Fulcher, Ph.D. is a Scientific Database Curator at SRI International.

Ingrid M. Keseler is a Senior Scientific Database Curator in the Bioinformatics Research Group at SRI International. She curates the EcoCyc database for *E. coli* and has created the BsubCyc and CdiffCyc databases for *Bacillus subtilis* and *Peptoclostridium difficile*.

Ron Caspi, curator of the MetaCyc database, received his PhD in biology from the Scripps Institution of Oceanography, La Jolla, CA. He is secretary of the IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN) and a member of the Enzyme Commission (EC).

Submitted: 7 May 2015; **Received (in revised form):** 14 August 2015

services application programming interfaces. Scientists around the world have created more than 9800 Pathway/Genome Databases by using Pathway Tools, many of which are curated databases for important model organisms.

Key words: metabolic pathways; metabolic models; systems biology; computational genomics

Introduction

Pathway Tools [1–4] is a software environment for management, analysis, simulation and visualization of integrated collections of genome, pathway and regulatory data (This article incorporates some text from [1] by permission of the publisher.). Pathway Tools handles many types of information beyond pathways and offers extensive capabilities. The software has been under continuous development within the Bioinformatics Research Group (BRG) within SRI International since the early 1990s. Pathway Tools addresses several different use cases in bioinformatics and systems biology:

- It supports development of organism-specific databases (DBs) (also called model-organism DBs) that integrate many bioinformatics data types.
- It supports scientific visualization, web publishing and dissemination of those organism-specific DBs.
- It performs computational inferences from sequenced genomes including prediction of an organism's metabolic network, prediction of metabolic pathway hole fillers and prediction of operons.
- It enables creation of steady-state quantitative metabolic flux models for individual organisms and for organism communities.
- It provides tools for graph-based analysis of biological networks, such as for identification of metabolic choke points, dead-end metabolites and blocked reactions.
- It provides tools for analysis of gene expression, metabolomics, proteomics and multi-omics data sets.
- It provides comparative analyses of organism-specific DBs.
- It supports metabolic engineering.

This article describes aspects of the software that are new since [1], which are summarized in Table 1. An online article that provides a complete description of Pathway Tools is available [5].

Pathway Tools is focused around a type of model-organism DB called a Pathway/Genome Database (PGDB). A PGDB integrates information about an organism's genes, proteins, metabolic network and regulatory network. Pathway Tools has several components. The 'PathoLogic' component enables users to create a new PGDB from the annotated genome of an organism, containing the genes, proteins, biochemical reactions and predicted metabolic pathways and operons of the organism.

The 'Pathway/Genome Editors' let PGDB developers interactively refine the contents of a PGDB, such as editing a metabolic pathway or an operon, or defining the function of a newly characterized gene.

The 'Pathway/Genome Navigator' supports querying, visualization and analysis of PGDBs. Whereas all other Pathway Tools components run as desktop applications only, the Navigator can run as both a desktop application and as a web server. The Navigator enables scientists to quickly find information, to display that information in familiar graphical forms, and to publish a PGDB to the scientific community via the web. The Navigator provides a platform for systems-level analysis of high-throughput data by providing tools for painting combinations of gene expression, protein expression and metabolomics data onto a full metabolic map of the cell, onto the full genome and onto a diagram of the regulatory network of the cell.

The new 'MetaFlux' component enables construction and execution of steady-state metabolic flux models from PGDBs. MetaFlux has modes to accelerate development of metabolic models, and to use metabolic models to simulate both gene and reaction knockouts. Pathway Tools provides a unique environment for metabolic flux modeling: by combining a tool for reconstructing metabolic networks from genome annotations with metabolic-model debugging tools such as a reaction

Table 1. New Pathway Tools capabilities introduced since 2010 [1], and the section of this article in which they are described.

New capability	Section
MetaFlux metabolic modeling component	Metabolic modeling with MetaFlux
Metabolic route search	Metabolic route searching using the RouteSearch tool
Calculation of pathway abundance for metagenomics analysis	Calculation of pathway abundance for metagenomics
Computation of reaction atom mappings	Atom mappings
Computation of Gibbs free energies of metabolites	Computation of metabolite Gibbs free energies
Signaling pathway editor	New editing tools
Cellular architecture editor	New editing tools
Sequence editor	New editing tools
Storage of conditions of cellular growth	Conditions of cellular growth
Storage of gene essentiality data	Gene essentiality
Storage of organism phenotype data and genome metadata	Organism phenotype data and genome metadata
Cross-organism search	Cross-organism search
Sequence pattern searching	Sequence-based query and visualization tools
Display of signaling pathways	Visualization tools for individual biological entities
Display of electron transfer pathways	Visualization tools for individual biological entities
Display of multiple sequence alignments	Sequence-based query and visualization tools
SmartTables	SmartTables: large-scale manipulation of PGDB object groups
Graphing of omics data on pathway diagrams	Figure 2
Python API	Computational access to PGDB data
Import of UniProt sequence features	Computational access to PGDB data
Import of Gene Ontology annotations	Computational access to PGDB data
Export of pathways to Cytoscape	Computational access to PGDB data
Expansion of web services	Computational access to PGDB data

gap-filler, the software enables rapid development of metabolic models from sequenced genomes. And by tightly coupling the metabolic model with other enriching information such as the sequenced genome, chemical structures and regulatory information, Pathway Tools-based metabolic models are easier to understand, validate, reuse, extend and learn from.

Pathway Tools includes a sophisticated ontology and DB application programming interface (API) that enables programs to perform complex queries, symbolic computations and data mining on the contents of a PGDB. For example, the software has been used for global studies of the *Escherichia coli* metabolic network [6] and genetic network [7].

Pathway Tools is seeing widespread use across the bioinformatics community to create PGDBs in all domains of life. More than 5600 groups to date have licensed the software. As well as supporting the development of the EcoCyc [8] and MetaCyc [9] DBs at SRI, and SRI's BioCyc collection of 5700 PGDBs [9], the software is used by genome centers, experimental biologists and groups that are creating curated DBs for a number of different organisms (see the Supplementary Material for a more detailed listing of available PGDBs).

New metabolic informatics capabilities

Atom mappings

The atom mapping of a reaction specifies for each reactant non-hydrogen atom its corresponding atom in a product compound. Pathway Tools contains an algorithm for computing atom mappings, described in [10]. Essentially, this approach computes atom mappings that minimize the overall cost of bonds broken and made in the reaction, given assigned propensities for bond creation and breakage. This algorithm has been applied to compute atom mappings for almost all of the reactions in the MetaCyc DB.

Atom mappings are used in two other parts of Pathway Tools. Atom mappings are used in the rendering of Pathway Tools reaction pages, to depict the conserved chemical moieties in a reaction. Conserved moieties are depicted by using the same color on the reactant and product sides. The bonds made or broken by a reaction are identified from the atom mapping for the reaction, and are colored black. Atom mappings are also used in the RouteSearch module of Pathway Tools described in section 'Metabolic route searching using the RouteSearch tool'.

Computation of metabolite Gibbs free energies

The MetaCyc DB provides the standard Gibbs free energy of formation for its compounds, and the change in Gibbs free energy for its reactions. These data were calculated by an algorithm within Pathway Tools. The algorithm first calculates the free energy of formation at pH 0 and ionic strength 0 ($\Delta_f G^0$) by using a technique based on the decomposition of the compounds into chemical groups with known free-energy contributions to the overall energy, based on the method of [11]. Then, the standard Gibbs free energy at pH 7.3 and ionic strength 0.25 ($\Delta_f G^{'0}$) is computed based on a technique developed by Robert A. Alberty [12]. In his technique, Alberty proposes to use several protonation states for some compounds, but we simplified the technique by always using only one protonation state, the state stored in MetaCyc.

The change in standard Gibbs free energy of reactions, $\Delta_r G^{'0}$, is computed based on the $\Delta_f G^{'0}$ values of the compounds involved in the reaction. The $\Delta_f G^{'0}$ could not be computed for some of the compounds in MetaCyc owing to the impossibility of decomposing them into the groups provided by the technique of [11]. Consequently, the $\Delta_r G^{'0}$ is not computed for any reaction that has a substrate for which its $\Delta_f G^{'0}$ is not stored in MetaCyc.

Calculation of pathway abundance for metagenomics

PathoLogic computes abundances of metabolic pathways based on gene abundances, which is useful for comparing the metabolic profiles of different microbial communities. Gene abundances are specified in the annotated genome file [PathoLogic format (PF) only].

No preprocessing of the gene abundances (such as outlier removal) is done by PathoLogic. The abundance of a pathway is computed based on the gene abundances involved in the pathway. More precisely, assume that R is the set of reactions in pathway P for which gene abundances are specified, $|R|$ is the size of R and g_a is the given abundance of gene g. The abundance of a pathway P is

$$\sum_{r \in P} r_a/|R| \text{ where } r_a = \sum_{g \text{catalyzes} r} g_a$$

That is, the abundance of a pathway is the sum of the abundances of the genes catalyzing the reactions of the pathway, divided by the number of reactions of the pathway for which gene abundances are given. Notice that this formula does take into account all the known isozymes catalyzing a reaction and spontaneous reactions do not take part in the computation.

New supported datatypes

Conditions of cellular growth

The Pathway Tools schema can now store conditions of cellular growth that include the chemical composition of the growth medium, pH, temperature and aerobicity. This representation enables us to capture low-throughput information about conditions of cellular growth, and high-throughput information such as Phenotype Microarray (PM) [13, 14] data sets.

Gene essentiality

The schema now supports representation of gene-essentiality experiments. Our representation links growth phenotype (no growth, limited growth or growth) under a given gene knockout with the conditions of cellular growth expressed as per section 'Conditions of cellular growth'. Multiple phenotypic observations can be recorded for a given gene knockout and growth condition to express conflicting experimental outcomes.

Organism phenotype data and genome metadata

The schema has been extended to support representation of microbial phenotypic data to enable users to query among the many genomes stored within a Pathway Tools Web site to find organisms pertinent to their research. Our representation adapts the minimum information about a genome sequence [15] standard to incorporate metadata about the sample from which the organism was derived (e.g. geographic location, depth, health-or-disease state of host, human microbiome site), plus phenotypic information about the organism itself (e.g. relationship to oxygen, temperature range and pathogenicity).

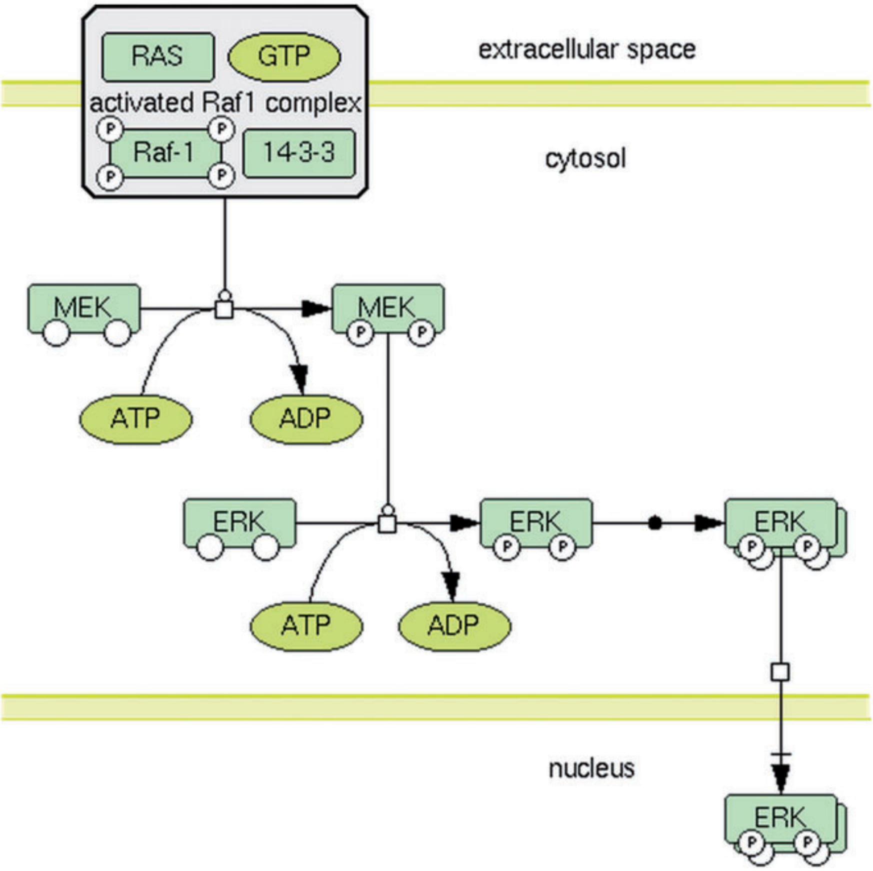


Figure 1. The HumanCyc MAP kinase cascade, a signaling pathway diagram.

New editing tools

A new signaling pathway editor enables users to interactively construct and edit a signaling-pathway diagram by using a toolkit of icons and operations inspired by CellDesigner [16] (see Figure 1). Updates to the visual representation are automatically translated back to changes to component reactions and proteins.

Because Pathway Tools now supports storage and display of glycan structures (using an icon-based style that follows the conventions of Consortium for Functional Glycomics), the Compound Editor now interfaces to the GlycanBuilder java applet [17, 18].

The Organism Editor now supports editing of phenotypic information, such as pathogenicity and relationship to oxygen, and sample collection data such as date, geographic location, host, body site.

A new Cellular Architecture Editor enables users to specify exactly which set of cellular components are present in an organism or cell type, with appropriate defaults derived from the organism's taxonomy.

A new Sequence Editor supports interactive, visual editing of the nucleotide sequence for a replicon, allowing insertion, deletion and replacement of arbitrary sections of sequence. Coordinates of all objects affected by the edits are updated automatically.

New visualization and query capabilities

Visualization tools for individual biological entities

Genes/proteins/RNAs: Since our last publication [1], we merged the individual display pages for a gene and its product (protein or RNA) into a single page that combines all information in one central place. These pages are quite extensive, listing information such as the map position of the gene on the chromosome, a graphical depiction of the chromosomal region containing the gene and available gene-essentiality information. A new diagram, the regulation-summary diagram, integrates all known regulatory influences on the gene and gene product into a single figure. Common to all protein types is the ability to graphically display information about protein regions (such as phosphorylation sites and active sites) using a protein-feature ontology that we developed.

Pathways: We recently added the ability to display gene expression and metabolomics data on pathway diagrams (see Figure 2). Support for display of signaling pathways (see Figure 1) was added recently; signaling pathway layout is performed by the user.

Electron transfer reactions and pathways: A crucial role in cellular metabolism is played by electron transfer reactions (ETRs), which are of key importance in the energy household of a cell. In a series of redox steps, the high-energy electrons from some compounds drive the pumping of protons across a cell membrane, to maintain the proton motive force needed for ATP synthesis. We designed and implemented drawing code for a special ETR diagram, which shows the enzyme complex embedded in a membrane, and which schematically depicts the flow of electrons from one redox half reaction to another. Inside the membrane, the quinone/quinol cofactor is shown together with an indication of the cell compartments that are sources or sinks of the protons. An additional

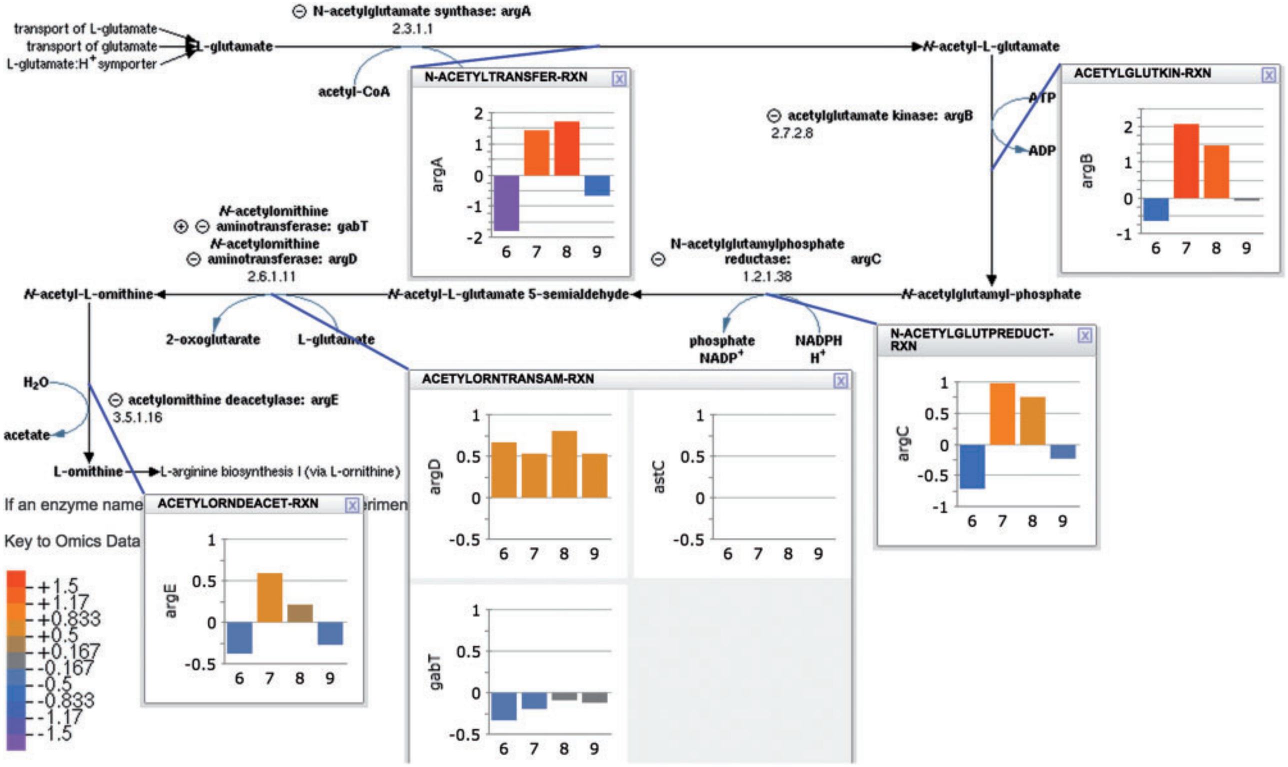


Figure 2. The EcoCyc L-ornithine biosynthesis pathway shown with omics pop-ups containing time-series data from a gene-expression experiment.

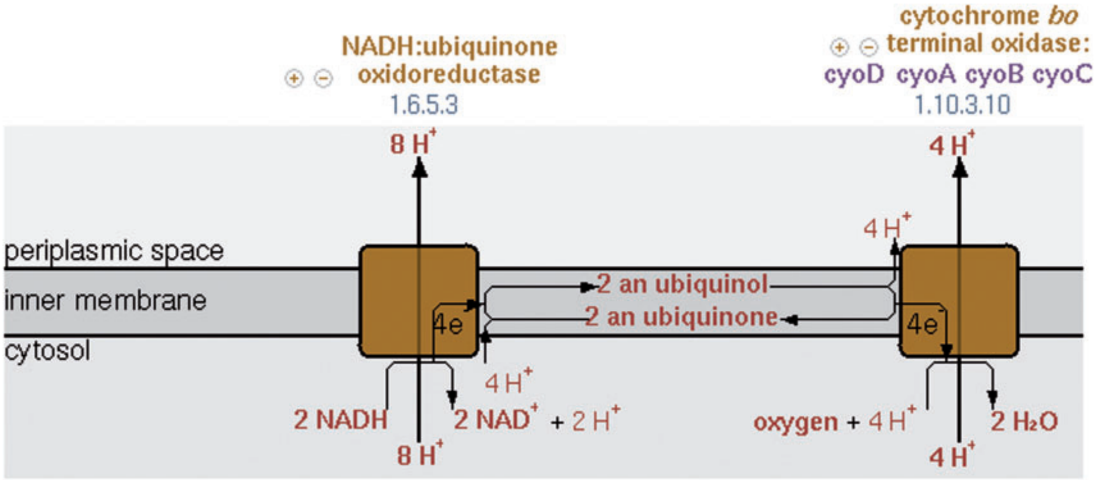


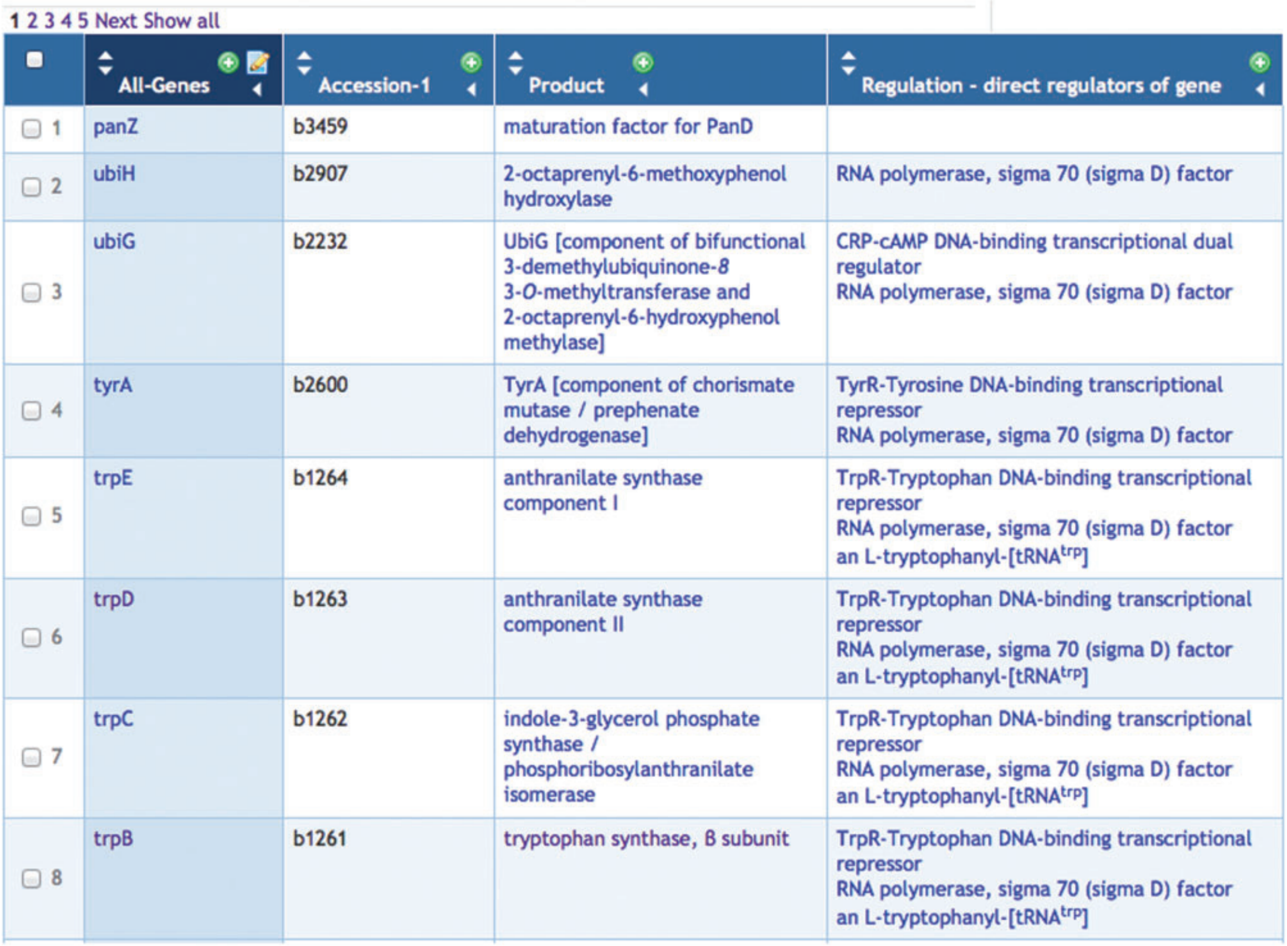
Figure 3. An electron transfer pathway diagram.

vectorial proton transport reaction can be added to the diagram. This results in displaying the flow of all substrates and products relative to the cellular compartments, in a similar way to what is customary in the biomedical literature. Pathways consisting of several ETRs joined together can also be depicted (see Figure 3).

SmartTables: large-scale manipulation of PGDB object groups

SmartTables is a recent addition to Pathway Tools, which enables users to construct and manipulate groups of PGDB objects through a spreadsheet-like user interface [19] (SmartTables were previously called Web Groups). SmartTables provide many powerful operations to biologist end users that previously would have required assistance from a programmer, and our user surveys indicated that SmartTables are reasonably easy for biologists to use [19]. Altogether, 2700 users of BioCyc.org have created more than 31 000 SmartTables.

A typical SmartTables use case is for a user to define a SmartTable by importing a list of PGDB objects from a file. For example, a user could define a metabolite SmartTable by importing a list of metabolites from a metabolomics experiment, where the metabolites are specified by metabolite name, BioCyc identifier, PubChem identifier or KEGG identifier. (The set of objects in a SmartTable can also be defined from a query result, from any column of an existing SmartTable, or from the set of, say, all genes in a PGDB.)



1 2 3 4 5 Next Show all

□	All-Genes	Accession-1	Product	Regulation - direct regulators of gene
□ 1	panZ	b3459	maturation factor for PanD	
□ 2	ubiH	b2907	2-octaprenyl-6-methoxyphenol hydroxylase	RNA polymerase, sigma 70 (sigma D) factor
□ 3	ubiG	b2232	UbiG [component of bifunctional 3-demethylubiquinone-8 3-*O*-methyltransferase and 2-octaprenyl-6-hydroxyphenol methylase]	CRP-cAMP DNA-binding transcriptional dual regulator RNA polymerase, sigma 70 (sigma D) factor
□ 4	tyrA	b2600	TyrA [component of chorismate mutase / prephenate dehydrogenase]	TyrR-Tyrosine DNA-binding transcriptional repressor RNA polymerase, sigma 70 (sigma D) factor
□ 5	trpE	b1264	anthranilate synthase component I	TrpR-Tryptophan DNA-binding transcriptional repressor RNA polymerase, sigma 70 (sigma D) factor an L-tryptophanyl-[tRNA^{trp}]
□ 6	trpD	b1263	anthranilate synthase component II	TrpR-Tryptophan DNA-binding transcriptional repressor RNA polymerase, sigma 70 (sigma D) factor an L-tryptophanyl-[tRNA^{trp}]
□ 7	trpC	b1262	indole-3-glycerol phosphate synthase / phosphoribosylanthranilate isomerase	TrpR-Tryptophan DNA-binding transcriptional repressor RNA polymerase, sigma 70 (sigma D) factor an L-tryptophanyl-[tRNA^{trp}]
□ 8	trpB	b1261	tryptophan synthase, β subunit	TrpR-Tryptophan DNA-binding transcriptional repressor RNA polymerase, sigma 70 (sigma D) factor an L-tryptophanyl-[tRNA^{trp}]

Figure 4. A gene SmartTable. Column 1 shows the gene name, column 2 shows the *E. coli* genome 'b-number' accession number for the gene (a property); column 3 shows the gene product name (a property). Column 4 shows the result of a transformation in which the regulator(s) of each gene were computed.

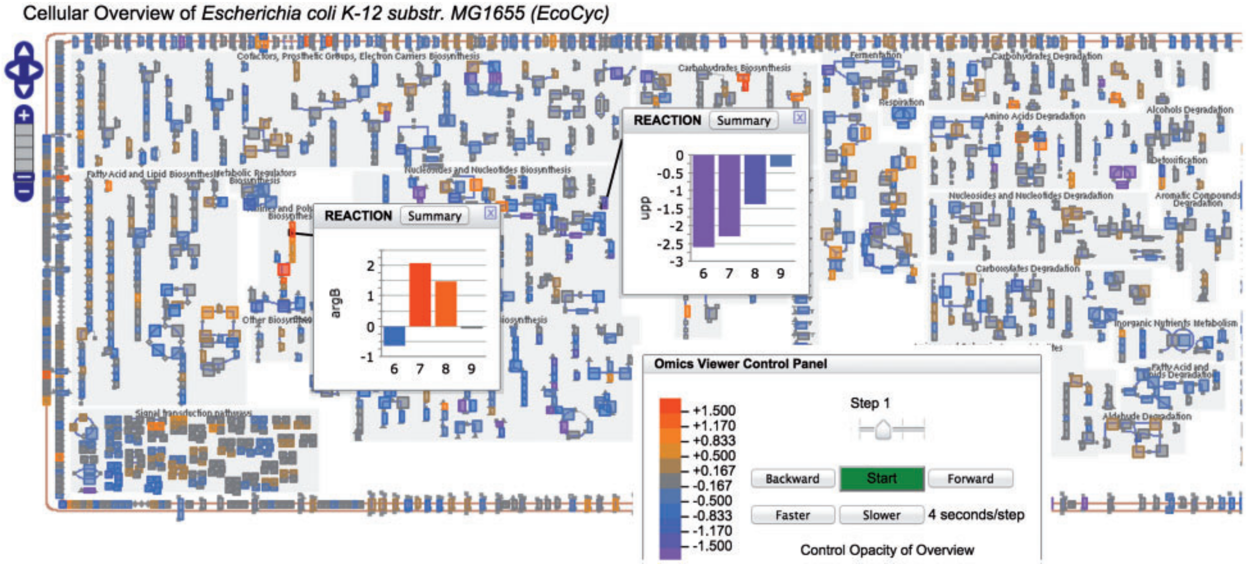

Cellular Overview of *Escherichia coli K-12 substr. MG1655 (EcoCyc)*
Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis
Carbohydrates Biosynthesis
Fermentation
Carbohydrates Degradation
Respiration
Amino Acids Degradation
Alcohols Degradation
Detoxification
Fatty Acid and Lipid Biosynthesis
Metabolic Regulators Biosynthesis
Nucleosides and Nucleotides Biosynthesis
Nucleosides and Nucleotides Degradation
Aromatic Compounds Degradation
Carbohydrates Degradation
Inorganic Nutrients Metabolism
Fatty Acid and Lipids Degradation
Aldehyde Degradation
Other Biosynthesis
Signal transduction pathways
REACTION
Summary
argB
2
1
0
-1
6
7
8
9
REACTION
Summary
upp
0
-0.5
-1
-1.5
-2
-2.5
-3
6
7
8
9
Omics Viewer Control Panel
+1.500
+1.170
+0.833
+0.500
+0.167
-0.167
-0.500
-0.833
-1.170
-1.500
Step 1
Backward
Start
Forward
Faster
Slower
4 seconds/step
Control Opacity of Overview

Figure 5. The Pathway Tools Cellular Overview diagram for EcoCyc, painted with gene-expression data. Omics pop-ups are shown for two genes.

The user can browse the set of objects in a SmartTable by paging through the table, and can modify the information displayed about each object by specifying which table columns to include (see Figure 4). SmartTable columns are derived from the PGDB attributes available for each object, and can include information such as chemical structures, molecular weights, links to other DBs and nucleotide and protein sequence. A variety of filters and set manipulations are provided for SmartTables, such as removing or retaining all rows that match a user query, and computing the union, intersection and set difference of two SmartTables. SmartTables are stored in the user's online web account, and a desktop version of SmartTables is also provided. SmartTables are private by default, but the user can make them public, share SmartTables with selected other users or archive them in a frozen form in conjunction with a publication.

Several more advanced SmartTable operations are provided. 'Transformations' compute new columns from relationships in a PGDB. For example, column 4 in Figure 4 is a transformation column that shows one or more regulators for each gene in column 1 that has been computed from PGDB relationships. Other gene transformations available include computing the metabolic pathways in which a gene's product occurs and computing the amino-acid changes caused by sequence variants. Different transformations are available for different datatypes. For example, the transformations available for a metabolite SmartTable include computing the reactions in which a metabolite occurs, the pathways in which a metabolite occurs, the proteins for which the metabolite is a ligand and mapping the compounds to their equivalents in another PGDB.

A user can perform a statistical enrichment analysis on a gene or metabolite SmartTable to detect overrepresented metabolic pathways or Gene Ontology (GO) terms, or overrepresented metabolic pathways, respectively. In addition, a SmartTable of genes or metabolites can be visualized on the cellular overview.

System-level visualization of metabolic networks

Pathway Tools can automatically generate organism-specific metabolic charts that we call Cellular Overview diagrams [2]. The diagram can be interrogated interactively and used to analyze omics data sets. Recently, the diagram was reengineered for the web mode of Pathway Tools. The diagram can be generated as a PDF file for printing as a large-format poster. Example posters can be downloaded from [20].

Figure 5 depicts the Web Cellular Overview at low resolution painted with gene-expression data. It contains all known metabolic pathways and transporters of an organism (online example: [21]; example with animated display of omics data: [22]). Each node in the diagram represents a single metabolite, and each line represents a single bioreaction. Biosynthetic pathways are in the left half of the diagram; catabolic pathways are in the right half.

Omics data (e.g. gene-expression or metabolomics measurements) for a given organism can be painted onto the cellular overview to place this data in a pathway context and to enable the user to discern the coordinated expression of entire pathways (such as the TCA cycle) or of important steps within a pathway. The user can click to create omics pop-ups that graph all available time points for particular reactions or metabolites of interest. Omics data may be loaded from a data file, from a GEO data set retrieved via web services or from a SmartTable. The data may be generated from a variety of experimental technologies, including microarrays and next-generation sequencing. Lower-level data processing must be performed external to Pathway Tools and must produce an expression value (or series of values) for each gene, protein, metabolite or reaction. The input data files can mix values for multiple types of entities such as genes and metabolites. Each entity can be specified using one or more names or identifiers to maximize the chances that Pathway Tools will recognize each entity. For more information on omics data file formats, see [23]. In web mode, the user has a choice of several color schemes—in desktop mode the color scheme is fully customizable.

Cellular Overview diagrams are generated automatically using an advanced layout algorithm [2]. Automated layout is essential to enable the diagram to accurately depict the underlying DB content as that content evolves, without requiring time-consuming manual updates by curators that are bound to overlook some updates. In addition, automated layout enables generation of organism-specific cellular overviews that reflect the exact pathway content of each organism-specific PGDB in large PGDB collections such as BioCyc.

The Cellular Overview has many capabilities (described in more detail in [2]), including semantic zooming of the diagram (where the highest magnification corresponds to the detail shown in the poster version); highlighting of user-requested elements of the diagram (such as metabolites or pathways); highlighting large, biologically relevant subnetworks (such as all reactions regulated by a given transcription factor); and highlighting comparative analysis results, such as comparison of the metabolic networks of two or more PGDBs.

System-level visualization of genome maps

The Pathway Tools genome browser displays a selected replicon (chromosome or plasmid), and enables the user to zoom into a region of the replicon by gene name or by coordinates. The browser supports semantic zooming: as the user moves deeper into the genome, additional features are displayed, such as promoters and terminators. The browser has been extended so that at high magnification, the genome sequence and the amino-acid sequence of coding regions become visible, and intron and exon boundaries are shown.

Sequence-based query and visualization tools

The following new tools support query and visualization of sequence data from a Pathway Tools web server.

Nucleotide Sequence Viewer: Gene pages include links to view or download the nucleotide or RNA/protein sequence for the gene. When viewing the nucleotide sequence, an option is provided to include an additional upstream and/or downstream flanking region of any desired length. This option makes it easy to, for example, view the sequence of a regulatory region surrounding a gene of interest. Alternatively, the user can enter specific start and end coordinates and the desired strand to view the sequence of any arbitrary portion of the chromosome.

Sequence Pattern Search (PatMatch): The PatMatch facility enables searching within a single genome for all occurrences of

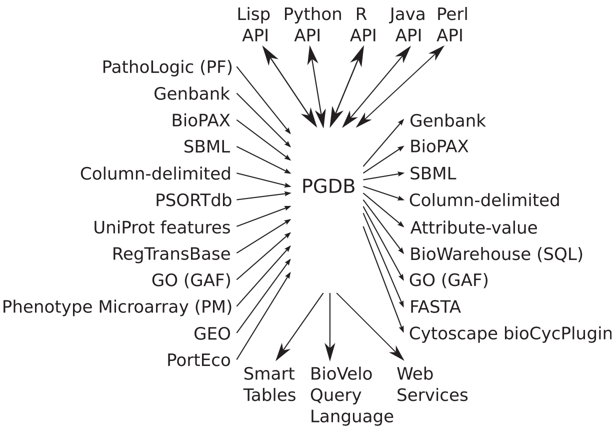


Figure 6. Pathway Tools supported formats and APIs for data import and export.

a specified short nucleotide or peptide sequence (less than about 20 residues), with the ability to specify degenerate positions. The user can specify the kind and number of allowable mismatches, and whether to search coding regions only, intergenic regions or the entire genome. Examples of situations in which this facility might be useful include searching for all occurrences of a particular regulatory motif upstream of any gene, or all occurrences of a known cofactor binding motif within proteins.

Multiple Sequence Alignment: From a gene page, users can request a multiple sequence alignment between the nucleotide or amino-acid sequences of that gene and its orthologs in a user-specified set of organisms. Alignments are displayed using MUSCLE [24].

Cross-organism search

We recently added a tool for searching across all organisms within a Pathway Tools web server, such as for the 5700 organisms at BioCyc.org. The cross-organism search tool [25] searches for user-specified combinations of words in the Common-Name/Synonyms attributes, and/or the Summary attribute. It can search all types of objects in a given PGDB, or in user-specified object types, such as genes and/or pathways. It can search all organism DBs present in the Pathway Tools web server, or it can search user-specified sets of organisms, such as all organisms within a selected taxonomic group. Indexing and searching is implemented using SOLR [26].

Computational access to PGDB data

In addition to the user-friendly graphical interfaces to PGDBs provided through the web and desktop versions of Pathway Tools, the software supports multiple methods for importing and exporting data from files and via programmatic interactions, which are summarized in Figure 6.

Programmatic access through APIs. Programmers can access and update PGDB data directly [27, 28] by writing programs in the Python, R, Java, Perl and Common Lisp languages. R, Java and Perl queries are executed using systems called RCyc [29], JavaCyc [30] and PerlCyc [31]. New programmatic access tools will now be summarized.

PythonCyc is a Python package that enables programmatic access to Pathway Tools. The package provides the basic functions to access and modify PGDB data. It also exposes more than 150 functions of Pathway Tools, among them, the MetaFlux module. The PythonCyc package is hosted on GitHub and is a separate installation from Pathway Tools. Full API documentation and a tutorial is available online. Please consult the URL [32] for access to the package, the API documentation and the tutorial.

Data import formats. Pathway Tools can import data from many sources into a PGDB. First of all, PathoLogic can create or update a PGDB based on a PF [33] or GenBank file [34]. A PGDB can also be populated from an SBML [35, 36] or BioPAX [37, 38] file.

Several specialized data import operations are also supported. The following sources can be imported for proteins. UniProt sequence annotation features [39] can be fetched from a Biowarehouse [40] server that was loaded with SwissProt and TrEMBL. GO annotations can be loaded from a GAF [41, 42] file. PSORTdb [43] cellular localization data can be imported from tab-delimited files.

Gene regulatory data can be imported from RegTransBase [44], which is a SQL DB. Growth conditions versus growth media can be imported from PM [13, 14] files. High-throughput expression data can be obtained from NCBI GEO [45] via web services.

File export formats. Pathway Tools can export PGDBs into several file formats that we have developed, which include tab-delimited tables and an attribute-value format (see [46]). Pathway Tools can also export subsets of PGDB data to other common formats including SBML [35, 36], BioPAX [37, 38], GO (in the GAF format) [41, 42], GenBank [34] and FASTA [47]. The bioCycPlugin for Cytoscape [48] makes use of web services in conjunction with BioPAX, to select and export pathways into the Cytoscape environment.

Pathway Tools web services [49] enable programmatic retrieval of numerous data types, based on submitted HTTP GET or POST commands, and have expanded substantially in recent years. Users can access BioVelo queries, a Metabolite Translation Service, and can also invoke omics visualization services and SmartTable manipulations via web services. Services for SmartTables include creation, retrieval, copying and deletion; applying many transformations; and changes like adding and deleting rows, columns and cells.

Graph-based metabolic network analyses

Metabolic route searching using the RouteSearch tool

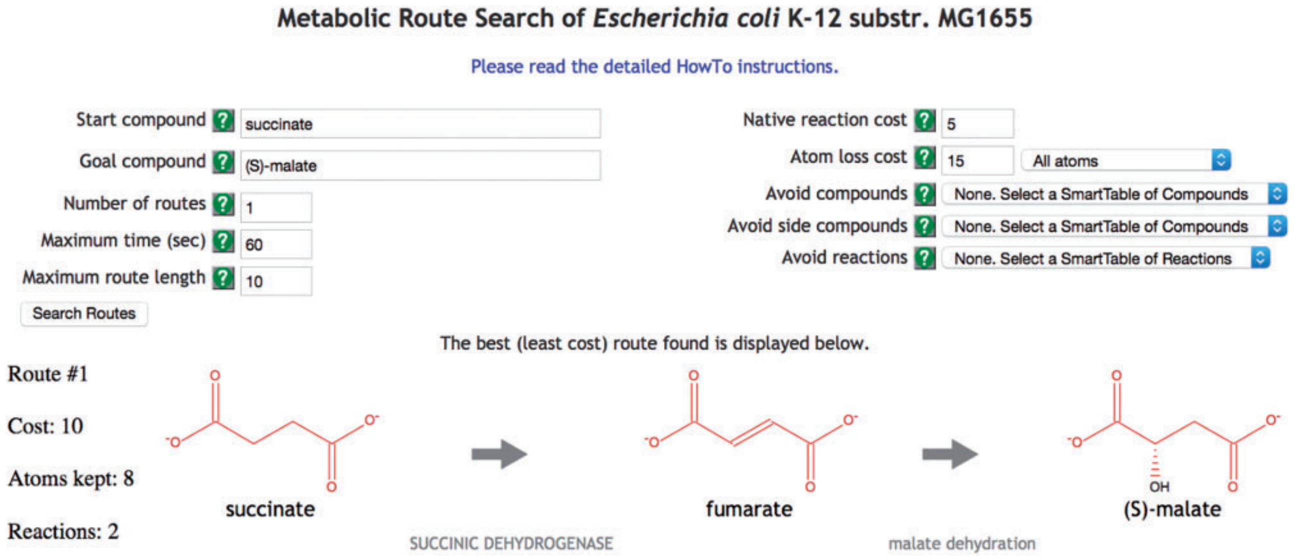
RouteSearch [50] is a Pathway Tools component that enables the exploration of the reaction network of a PGDB, and engineering of new metabolic pathways. RouteSearch computes optimal metabolic routes (that is, an optimal series of biochemical reactions that connect start and goal compounds), given various cost parameters to control the optimality of the routes found. RouteSearch can display several of the best routes it finds using an interactive graphical web page (see Figure 7).

When RouteSearch is used for engineering new pathways, it uses the MetaCyc DB as its external reaction DB for new reactions to include in an organism. The cost for adding one reaction from MetaCyc to a route is selected by the user. Typically this cost, an integer, is larger than the cost of adding one reaction from the organism to a route. Thus, a MetaCyc reaction would be added to a route only if it causes the route to conserve more atoms from the start compound to the target compound. The cost of losing one atom from the start compound is also selected by the user, and it is typically larger than the cost of adding one reaction to a route, from either the organism or the external library of new reactions.

In computing optimality, RouteSearch takes into account the conservation of non-hydrogen atoms from the start compound to the goal compound. The more atoms that are conserved, the more efficient the transformation from start to goal. To compute the number of conserved atoms, RouteSearch uses precomputed atom mappings of reactions that are available in MetaCyc. An atom mapping of a reaction gives a one-to-one correspondence of each non-hydrogen atom from reactants to products. RouteSearch is available only in web mode in Pathway Tools.

Computation of blocked reactions

The MetaFlux component of Pathway Tools computes the blocked reactions in a reaction network (see section ‘Reports generated by MetaFlux’)—reactions that can never carry flux because of blockages in the network.



Metabolic Route Search of *Escherichia coli* K-12 substr. MG1655

Please read the detailed HowTo instructions.

Start compound: succinate

Goal compound: (S)-malate

Number of routes: 1

Maximum time (sec): 60

Maximum route length: 10

Native reaction cost: 5

Atom loss cost: 15 All atoms

Avoid compounds: None. Select a SmartTable of Compounds

Avoid side compounds: None. Select a SmartTable of Compounds

Avoid reactions: None. Select a SmartTable of Reactions

The best (least cost) route found is displayed below.

Route #1

Cost: 10

Atoms kept: 8

Reactions: 2

Figure 7. The RouteSearch web interface is shown with the result of one short pathway found. The arrows represent reactions and are tagged with the protein names catalyzing them. All atoms are conserved from the start compound (succinate) to the target compound ((S)-Malate) in this simplified example.

Metabolic modeling with MetaFlux

The MetaFlux component of Pathway Tools is used to develop and execute quantitative metabolic flux models for individual organisms and for organism communities. MetaFlux uses the steady-state modeling technique of flux-balance analysis (FBA), which can be used to predict the phenotypes of an organism, or a community of organisms, based on a specification of available nutrients in the growth environment. MetaFlux can be controlled via a graphical user interface and via a Python API.

MetaFlux offers several modes of operation:

1. Solving mode: Execute a metabolic model for a single organism or for a community of organisms.
2. Development mode: Generate hypotheses on how to fill gaps in a developing metabolic model.
3. Knockout mode: Run metabolic models under gene-knockout scenarios.

The MetaFlux model development process

Pathway Tools provides a unique environment for the development of metabolic flux models for several reasons. First, it includes a range of tools that support fast and accurate development of metabolic models from annotated genomes. Second, metabolic models developed with MetaFlux are highly accessible to the user, and are coupled with extensive enriching information, resulting in models that are easier to understand and reuse.

The high-level steps for developing a metabolic model from an annotated genome using Pathway Tools are as follows. For more information on the genome-scale metabolic reconstruction process, we suggest the comprehensive COBRA-based [51] overview published by Thiele and Palsson [52].

1. The PathoLogic tool computes a qualitative metabolic reconstruction in two phases; in Phase I it infers the reactome of an organism from its annotated genome. By combining the enzyme-name matching tool with the extensive reaction information in the MetaCyc DB, we obtain an extensive mapping of annotated enzymes to metabolic reactions.
2. In Phase II, PathoLogic infers metabolic pathways. This step fills a significant number of missing (gap) reactions for the following reason. Imagine that inferred pathway *P* contains three reactions, *A*, *B* and *C*. Imagine that enzymes were identified in the genome for *A* and *B* (thus causing *P* to be inferred as present), but that no enzyme was identified for *C*—*C* is a 'pathway hole'. Because all reactions in a pathway must be present in a PGDB for the pathway to be present, *C* will be created in the PGDB as part of creating *P*. Thus, pathway inference also infers the presence of pathway reactions that were not initially identified.
3. The pathway hole filler attempts to identify enzymes that catalyze pathway–hole reactions such as *C*. (Note this step is optional and is informative in nature, because it does not modify the set of reactions in the model.)
4. The user can request that MetaFlux compute an initial set of biomass metabolites for the organism. MetaFlux will do so if the organism falls within the 12 taxonomic groups for which MetaFlux has defined biomass compositions, obtained from the experimental literature.
5. The user supplies an objective function and constraints on metabolite uptake and secretion. These can be based on experimental observations of the organism under study, or can be set to arbitrary values to explore the theoretical behavior of metabolism.
6. The MetaFlux gap filler identifies missing reactions, nutrients and secretions that enable a model to be solved. It can be run on one compartment at a time to simplify the gap-filling process for eukaryotic organisms. MetaFlux tools for computing blocked reactions and dead-end metabolites identify potential errors and omissions in the metabolic network definition.
7. The reaction network, objective function and constraints of the metabolic model are adjusted by the user until its predictions match experimental results.

When developing a metabolic model with MetaFlux, the reactions and metabolites within the model are derived from (and stored in) the PGDB. MetaFlux automatically generates the system of linear equations for the model from the PGDB. Thus, to modify the reactions within a model, the user edits the PGDB; to inspect the reactions and metabolites within a model, the user can query the PGDB using the plethora of Pathway Tools query and visualization operations. The entire PGDB/model can be published on a Web site using Pathway Tools, where all reactions and pathways that use a given metabolite are listed on the Pathway Tools metabolite page for that compound; all reactions within a given cellular location or using a given set of reactants and products can be found using reaction searches.

Furthermore, compared with other modeling environments, a metabolic model stored within a PGDB contains extensive additional enriching information. Chemical structures within a PGDB enable reaction mass and charge balancing. Chemical structures and reaction atom mappings aid users in understanding the chemistry of reaction transformations. Pathways arrange reactions into biologically meaningful groupings. Couplings between reactions, enzymes and genes enable reasoning about the roles of multi-subunit complexes, isozymes and gene knockouts. Regulatory information supports inferences about metabolic regulation. Model testing and validation are facilitated by PGDB storage of growth media and growth experiment results, and of gene-knockout experiment results.

Taken together, the Pathway Tools modeling environment—with its extensive tools for inspecting metabolic-model content and its enriching information—renders MetaFlux models significantly easier to understand, learn from, validate through inspection, reuse and extend than models produced with other metabolic-modeling software environments.

Description of an FBA model

The description of a FBA model is provided to MetaFlux partly via a text file called an FBA input file and partly via a PGDB. Typically, the FBA file specifies many parameters, but we will describe only the most important ones. Some parameters are only relevant for specific modes of MetaFlux, so that we will present these parameters when describing that mode.

Solving mode

Solving mode computes flux values for the reactions in the metabolic model given four inputs: a set of nutrient compounds, a set of secreted compounds, a set of biomass metabolites that are synthesized by the cell and a set of metabolic reactions. The first three inputs are supplied by the FBA file. The set of metabolic reactions are provided by the PGDB, but may be altered by the FBA file. For example, the following file describes a (simple) FBA model for *E. coli*:

```
pgdb: ecoli

reactions:
metab-all           # Include all metabolic reactions of
                     PGDB ecoli.
mal -> fum + water  # Example of including a reaction by
                     reaction equation.

biomass:
CYS[CCO-CYTOSOL] 0.0054
GLN[CCO-CYTOSOL] 0.2987
GLT[CCO-CYTOSOL] 0.2987
GLY[CCO-CYTOSOL] 0.3431

nutrients:
GLC[CCO-PERI-BAC]                    :upper-bound 10.0
OXYGEN-MOLECULE[CCO-PERI-BAC]        :upper-bound 1.0

secretions:
CARBON-DIOXIDE[CCO-PERI-BAC]
WATER[CCO-PERI-BAC]
```

Note that this example is meant to show the syntax of the file describing a model; it is not meant to show a working model. The `pgdb` parameter specifies the PGDB to be used in this model. The set of model reactions is specified by using the `reactions` parameter. The keyword `metab-all` specifies all metabolic reactions from the PGDB. A reaction equation can be provided to describe a reaction that is to be present in the model, but is not present in the PGDB.

The set of biomass metabolites are specified by parameter `biomass`; each metabolite is specified either by metabolite name or unique identifier, plus a compartment identifier in square brackets, and an optional coefficient. Nutrients and secretions are provided in the same manner, but no coefficients are allowed. Upper and lower bounds can be provided to constrain nutrient uptake rates or secretion production rates.

Modeling a community of organisms

Almost all organisms live in a larger ecosystem. MetaFlux can solve a model describing an organism community by combining individual FBA input files for each organism in the community, plus one COM input file describing the organism interactions. A typical COM description lists the FBA models to be combined, optionally with the abundance of each organism in the community. It lists the compartments shared by the organisms—the compartments in which the secretions of some organisms can be used as nutrients by other organisms. For example, the following COM file specifies two models that exchange metabolites in the periplasmic space, and three nutrients that are provided to the community.

```
community-name: ecoli-ile

fba-files:
ecoli-strain-A.fba :abundance 2
ecoli-strain-B.fba :abundance 3

exchange-compartments: [CCO-PERI-BAC]

community-nutrients:
GLC[CCO-PERI-BAC]           :upper-bound 20
OXYGEN-MOLECULE[CCO-PERI-BAC]       :upper-bound 10.0
AMMONIUM[CCO-CYTOSOL]
```

A MetaFlux community model can be executed in a steady-state mode using FBA, and in a time-dependent mode using dynamic FBA. In steady-state mode, the objective function defined by MetaFlux for the FBA model of a community of organisms is the sum of the growth of the organisms multiplied by their abundances. A solution file will be produced that describes the growth of the community, and the growth rate for each organism.

In MetaFlux dynamic community mode, inspired by COMETS [53], each organism has an independent objective function, and dynamic FBA is used to create a temporal simulation.

Development mode

Development mode can be used to create an FBA model or to discover what is wrong with a model that does not grow when growth is expected. The main parameters used in development mode are the four parameters `try-biomass`, `try-reactions`, `try-secretions` and `try-nutrients`.

When MetaFlux is provided with a list of metabolites for the `try-biomass` in development mode, the software tries adding these metabolites to the biomass reaction. That is, in development mode, MetaFlux will output as a solution the largest subset of the try-biomass metabolites that it can produce as biomass. In the early phases of developing a model, typically the entire biomass reaction is specified in try-biomass and no metabolites are specified for the biomass parameter.

The try-biomass parameter cannot specify any metabolite with a negative coefficient not included in a group. This is required because any such metabolite could be used as a 'free' nutrient for the organism. On the other hand, inside a group, negative coefficients are allowed because MetaFlux tries to produce the entire group of metabolites, not any one of them independently. A group of metabolites is supposed to form a cohesive unit where any metabolite specified with a negative coefficient is used to produce some other metabolites of that group. Therefore, declaring all metabolites that have negative coefficients in some groups enables the entire biomass reaction to be used as a try-biomass set, indicating for all metabolites which ones can or cannot be produced.

The try-reactions parameter can be used to try to add candidate reactions to the model to increase the number of try-biomass metabolites produced by the model. Therefore, the try-reactions parameter is typically used when at least one biomass metabolite is not produced. The candidate reactions are selected from MetaCyc. A single keyword, metacyc-metab-all, instructs MetaFlux to try all the metabolic reactions of MetaCyc. Alternatively, a list of candidate reactions can be specified by their unique identifiers. MetaFlux tries to produce as many biomass metabolites as possible from the try-biomass section, by adding as few reactions as possible from the try-reactions set. This computation is performed using optimization as a Mixed-Integer Linear Program, which can be computationally expensive. MetaFlux has also a fast development mode [54] that can be used with the try-reactions parameter only. It may run much faster than the general development mode and may provide different solutions. It uses an heuristic that does not necessarily provide an optimal solution.

Knockout mode

Knockout mode is used to computationally evaluate the impact of removing genes or reactions from an FBA model. This mode is used to predict essential genes of an organism for a given growth environment, and can also be used to evaluate the accuracy of a model if experimental gene-knockout data are available. MetaFlux can compute single, double or higher numbers of simultaneous knockouts.

When run, MetaFlux solves the model without any knockouts, and then solves for each reaction (or gene) to knockout by deactivating that reaction (or the reaction(s) associated with that gene). Note that a given gene may deactivate one, none or several reactions, because some genes may have isozymes, or catalyze several reactions. The user can request a summary of the results of modeling each knockout, and can request that a complete solution file be produced for each gene knockout.

Outputs generated by MetaFlux

Whichever mode is used to execute MetaFlux, the following output files are generated: the solution file that contains a description of the active reactions and metabolites used and produced, the log file to describe issues that may exist in the model and a data file for the computed fluxes that can be displayed with the Cellular Overview map of Pathway Tools.

In solving mode, the main output is the computed optimal assignments of reaction fluxes. For an organism-community model, solution, log and data files are produced for each organism.

In development mode, the outputs are the set of biomass metabolites that can be produced; the try nutrients used (if any); the try secretions produced (if any); the reactions that actively carry flux; a minimal list of suggested reactions to be added to the model; and reactions whose directions are reversed (if any) to produce otherwise unproducible biomass metabolites. Development mode also identifies which biomass metabolites could not be produced by the model (if any), despite the additions from the try sets.

Reports generated by MetaFlux

The set of reactions specified by the PGDB and the FBA input file may contain reactions that cannot, or that should not, be used in a model. MetaFlux checks each reaction to ensure that its inclusion in the model will not invalidate the model. In particular, all reactions are verified to be mass and charge balanced. Reactions that are unbalanced, or cannot be shown to be balanced, are excluded from the model, but are listed in the log file.

Another step in model execution is to instantiate the generic reactions of a model according to the compounds available in the PGDB. A generic reaction has some compound classes as products or reactants (e.g. 'a sugar'). Each computed reaction instantiation is added to the model.

Blocked reactions. MetaFlux analyzes the network of reactions specified by the PGDB plus the FBA input file to detect if some reactions are blocked. A blocked reaction is a reaction that can never have a positive flux, given the nutrients, secretions, biomass metabolites and reactions specified for a particular model execution. That is, blocked reactions are a function not only of the network, but of the cellular growth conditions. A 'basic blocked reaction' *R* has at least one reactant *M* that is not produced by any reaction in the model, and that is not provided as a nutrient; or *R* has a product *M* that is not used as a reactant by any other reaction and that is not secreted, and is not specified as a biomass metabolite with a positive coefficient. A metabolite that caused a basic blocked reaction to be blocked is called a 'basic blocking metabolite'.

Blocked reactions can never carry flux because in steady-state modeling, because the producing and consuming fluxes for every metabolite must be balanced. But the preceding metabolites *M* could not have balanced fluxes if a blocked reaction *R* carried flux, because according to the preceding definition, *M* must have either zero reactions that produce it or zero reactions that consume it.

Additional blocked reactions can be identified by eliminating basic blocked reactions from the model, causing more reactions to become basic blocked reactions. This process of removing basic blocked reactions from the model is repeated until no more reactions become blocked. The detection of blocked reactions is done before the linear solver is called (that is, this detection does not depend on the fluxes of reactions, but is a static evaluation of the model).

The set of blocked reactions is listed in the log file, grouped by basic blocking metabolite (the reactions are also grouped by pathways). The basic blocking metabolites are the root causes of blocked reactions, thus their identification is valuable for model debugging.

Painting MetaFlux fluxes on the cellular overview

MetaFlux also generates a data file that can be used with the Cellular Overview of Pathway Tools (see section 'System-level visualization of metabolic networks'). The graphical user interface of MetaFlux enables the user to click one button to invoke the Cellular Overview with the data from that file. The Cellular Overview displays all the reactions, grouped by pathways, of the organisms, and highlights, with colors indicative of flux values, the reactions that have positive fluxes. This overview enables the user to visually assess the metabolic activities of a model solution.

For a community of organisms, a data file is generated for each individual organism, and the graphical user interface gives direct access to each individual Cellular Overview of the organisms involved in the community.

Software and DB architecture

Pathway Tools consists of 640 000 lines of Common Lisp code, organized into 20 subsystems. In addition, 24 000 lines of JavaScript code are used within the Pathway Tools web interface. Pathway Tools runs on the Macintosh, Linux and Microsoft Windows platforms. Pathway Tools was ported to 64-bit architectures in recent years.

The main bioinformatics modules of Pathway Tools are the Navigator, Editors and PathoLogic, plus a chemoinformatics subsystem that includes tools such as SMILES [55] generation and parsing, a chemical substructure matcher, plus a large set of shared utilities that we call the Pathway Tools core. Pathway Tools uses an object-oriented DB system called Ocelot [56]. The Pathway Tools user interface relies on a graph layout and display package called Grasper [57], and web and desktop graphics packages called CWEST and CLIM (the Common Lisp Interface Manager). Other software used by and included with Pathway Tools are (bioinformatics) textpresso; MUSCLE [24]; PatMatch [58]; BLAST; libSBML [59]; (chemoinformatics) Marvin [60]; GlycanBuilder [18]; InChI [61]; (lisp) ARNESI; 5am; cl-json; cl-store; (other) SCIP [62]; ghostscript; SKIPPY [63]; Yahoo User Interface library (YUI) [64]; SOLR [26]; and MySQL.

Summary

Pathway Tools treats a genome as far more than a sequence and a set of annotations. Instead, it links the molecular parts list of the cell both to the genome and to a carefully constructed web of functional interactions. The Pathway Tools ontology defines an extensive set of object attributes and object relations that enables representing a rich conceptualization of biology within a PGDB, along with enabling querying and manipulation by the user. Furthermore, a PGDB can be transformed into a quantitative metabolic model for the organism.

Pathway Tools provides a broad range of functionality. It can manipulate genome data, metabolic networks and regulatory networks. For each datatype, it provides query, visualization, editing and analysis functions. It provides model-organism DB development capabilities, including computational inferences that support fast generation of comprehensive DBs, editors that enable refinement of a PGDB, web publishing and comparative analysis. A family of curated PGDBs has been developed using these tools for important model organisms.

The software also provides visual tools for analysis of omics data sets, and tools for the analysis of biological networks.

Software availability

Pathway Tools runs on Macintosh, Windows and Linux. It is freely available to academic and government researchers; a license fee applies to commercial use. See http://BioCyc.org/download.shtml.

Key Points

- The Pathway Tools software is a comprehensive environment for creating model organism databases that span genome information, metabolic pathways and regulatory networks.
- Pathway Tools inference capabilities include prediction of metabolic pathways, prediction of metabolic pathway hole fillers, inference of transport reactions from transporter functions and prediction of operons.
- Its metabolic modeling capabilities include flux-balance analysis modeling for individual organisms and organism communities, with model gap filling and the ability to model gene knockouts.
- Pathway Tools provides interactive editing tools for use by database curators.
- Omics data analysis tools paint genome-scale data sets onto a complete genome diagram, complete metabolic network diagram and complete regulatory network diagram.
- Other tools include comparative analysis operations, dead-end metabolite and blocked-reaction analysis of metabolic networks and metabolic route searching.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgments

Pathway Tools has benefited from advice, input and contributions from many scientists during its lifetime. We thank Tomer Altman for developing RCyc.

Funding

This work was supported by grants GM75742, GM080746 and GM077678 from the National Institutes of Health. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

References

1. Karp P, Paley S, Krummenacker M, *et al*. Pathway Tools version 13.0: Integrated software for pathway/genome informatics and systems biology. *Brief Bioinform* 2010;**11**:40–79. http://bib.oxfordjournals.org/cgi/content/abstract/bbp043, doi: 10.1093/bib/bbp043.
2. Paley S, Karp P. The pathway tools cellular overview diagram and omics viewer. *Nucleic Acids Res* 2006;**34**:3771–8. http://nar.oxfordjournals.org/cgi/content/full/34/13/3771
3. Karp P, Paley S, Romero P. The pathway tools software. *Bioinformatics* 2002;**18**:S225–32.
4. Karp P, Paley S. Integrated access to metabolic and genomic data. *J Comput Biol* 1996;**3**:191–212. http://www.ai.sri.com/pubs/papers/Karp96:Integrated/document.ps

5. Karp P, Latendresse M, Paley S, *et al*. *Pathway Tools version 19.0 Overview: Software for Pathway/Genome Informatics and Systems Biology, 2015*. Http://www.ai.sri.com/pkarp/misc/ptools15.pdf
6. Ouzounis C and Karp P. Global properties of the metabolic map of *Escherichia coli*. *Genome Res* 2000;**10**:568.
7. Karp P. Pathway databases: a case study in computational symbolic theories. *Science* 2001;**293**:2040–4.
8. Keseler IM, Mackie A, Peralta-Gil M, *et al*. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res* 2013;**41**:D605–12.
9. Caspi R, Altman T, Billington R, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 2014;**42**:D459–71. http://nar.oxfordjournals.org/content/42/D1/D459
10. Latendresse M, Malerich J, Travers M, *et al*. Accurate atom-mapping computation for biochemical reactions. *J Chem Inf Model* 2012;**52**:2970–82.
11. Jankowski MD, Henry CS, Broadbelt LJ, *et al*. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys J* 2008;**95**(3):1487–99.
12. Alberty RA. Thermodynamics of biochemical reactions. *Wiley InterScience*, John Wiley & Sons, Hoboken, New Jersey2003.
13. Bochner BR, Gadzinski P, Panomitros E. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res* 2001;**11**(7):1246–55.
14. Bochner B. Global phenotypic characterization of bacteria. *FEMS Microbiol Rev* 2009;**33**(1):191–205.
15. Field D, Farrity G, Gray T, *et al*. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 2008;**26**:541–7.
16. Funahashi A, Morohashi M, Kitano H. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico* 2003;**1**(5):159–62.
17. Ceroni A, Dell A, Haslam SM. The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med* 2007;**2**:3.
18. Damerell D, Ceroni A, Maass K, *et al*. The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments. *Biol Chem* 2012;**393**(11):1357–62.
19. Travers M, Paley S, Shrager J, et al. Groups: knowledge spreadsheets for symbolic biocomputing. Database 2013. http://database.oxfordjournals.org/content/2013/bat061.abstract
20. BioCyc genome and metabolic map posters. http://biocyc.org/posters.shtml
21. EcoCyc Cellular Overview. http://ecocyc.org/overviewsWeb/celOv.shtml
22. Cellular Overview with Animated Display of Gene Expression Data. http://biocyc.org/ov-expr.shtml
23. How to Use a Pathway Tools Website. http://biocyc.org/PToolsWebsiteHowto. shtml#OmicsDataAnalysis
24. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**(5):1792–7.
25. Pathway Tools Data-File Formats. http://bioinformatics.ai.sri.com/ptools/flatfile-format.html
26. SOLR Website. http://lucene.apache.org/solr/resources.html
27. Krummenacker M, Paley S, Mueller L, *et al*. Querying and computing with BioCyc databases. *Bioinformatics* 2005;**21**:3454–5. http://bioinformatics.oxfordjournals.org/cgi/reprint/21/16/3454
28. Querying Pathway/Genome Databases. http://brg.ai.sri.com/ptools/ptools-resources.html
29. RCyc API to Pathway Tools. https://github.com/taltman/RCyc
30. JavaCyc API to Pathway Tools. http://solgenomics.net/downloads/index.pl
31. PerlCyc API to Pathway Tools. http://solgenomics.net/downloads/perlcyc.pl
32. PythonCyc API to Pathway Tools. http://bioinformatics.ai.sri.com/ptools/pythoncyc.html
33. Billington R, Caspi R, Kothari A, *et al*. Pathway Tools User's Guide version 19.0, 2015. SRI International, Menlo Park CA.
34. Genbank Format. http://www.ncbi.nlm.nih.gov/collab/FT/#7.1.2
35. Hucka M, Finney A, Bornstein BJ, *et al*. Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. *Syst Biol (Stevenage)* 2004;**1**(1):41–53.
36. SBML. http://www.sbml.org/
37. BioPAX. http://www.biopax.org/
38. Demir E, Carry MP, Paley S, *et al*. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 2010;**28**(12):935–42.
39. UniProt Sequence annotation (Features). http://www.uniprot.org/help/sequence_annotation
40. Lee T, Pouliot Y, Wagner V, *et al*. BioWarehouse: A bioinformatics database warehouse toolkit. *BMC Bioinformatics* 2006;**7**:170. http://www.biomedcentral.com/1471-2105/7/170
41. Ashburner M, Ball C, Blake J, *et al*. Gene Ontology: Tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.
42. GO Annotation File Formats. http://geneontology.org/page/go-annotation-file-formats
43. Yu NY, Laird MR, Spencer C, *et al*. PSORTdb–an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic Acids Res* 2011;**39**:D241–4.
44. Cipriano MJ, Novichkov PN, Kazakov AE, *et al*. RegTransBase–a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics* 2013;**14**:213.
45. Barrett T, Wilhite S E, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* 2013;**41**:D991–5.
46. Pathway Tools Data-File Formats. http://bioinformatics.ai.sri.com/ptools/flatfile-format.html
47. FASTA Format. http://www.ncbi.nlm.nih.gov/blast/fasta.shtml
48. bioCycPlugin for Cytoscape. http://www.cgl.ucsf.edu/cytoscape/bioCycPlugin/
49. Pathway Tools Web Services. http://biocyc.org/web-services.shtml
50. Latendresse M, Krummenacker M, Karp PD. Optimal metabolic route search based on atom mappings. *Bioinformatics* 2014;**30**:2043–50.
51. Schellenberger J, Que R, Fleming R, *et al*. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 2011;**6**:1290–307.
52. Thiele I and Palsson B O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 2011;**5**(1):93–121.
53. Harcombe WR, Riehl WJ, Dukovski I, *et al*. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep* 2014;**7**(4):1104–15.
54. Latendresse M. Efficiently gap-filling reaction networks. *BMC Bioinformatics* 2014;**15**:225.
55. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31–6.

56. Karp P, Chaudhri VK, Paley SM. A collaborative environment for authoring large knowledge bases. *J Intell Inform Syst* 1999;**13**:155–94. http://www.ai.sri.com/pkarp/pubs/99jiis.pdf
57. Karp P, Lowrance J, Strat T, *et al*. The grasper-CL graph management system. *LISP Symb Comput* **7**:245–282, 1994.
58. Yan T, Yoo D, Berardini TZ, *et al*. PatMatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic Acids Res* 2005;**33**:W262–6.
59. Bornstein BJ, Keating SM, Jouraku A, *et al*. LibSBML: an API library for SBML. *Bioinformatics* 2008;**24**(6):880–1.
60. Marvin Chemical Editor. http://www.chemaxon.com/products/marvin/
61. Stein SE, Heller SR, Tchekhovskoi D. An open standard for chemical structure representation: the IUPAC chemical identifier. In *Proc. 2003 International Chemical Information Conference (Nimes)*, 2003, pp. 131–43.
62. SCIP Home Page. http://scip.zib.de/
63. SKIPPY – Read and Write GIF Files with Common Lisp. http://www.xach.com/lisp/skippy/
64. Yahoo User Interface Library. http://developer.yahoo.com/yui/