

Integrated pathway/genome databases and their role in drug discovery

P.D. Karp (pkarp@PangeaSystems.com), M. Krummenacker, S. Paley and J. Wagg
Pangea Systems, Inc.
4040 Campbell Ave.
Menlo Park, CA 94025, USA.

Appears in *Trends in Biotechnology*, 17(7):275–281 1999.

Integrated pathway/genome databases describe the genes and genome of an organism, as well as its predicted pathways, reactions, enzymes and metabolites. In conjunction with visualization and analysis software, these databases provide a framework for improved understanding of microbial physiology and for antimicrobial drug discovery. We describe pathway-based analyses of the genomes of a number of medically relevant microorganisms, and a novel software tool that provides visualization of gene expression data on a diagram showing the whole metabolic network of the microorganism.

The method of inferring the function of a DNA or protein sequence by analogy to the functions of other similar sequences has had a profound impact on our ability to identify the functions of sequenced genes. Similar reasoning can be applied by analogy to biological pathways to identify the presence of known metabolic pathways in the annotated genome of an organism. Just as we can predict the function of an unknown sequence S by searching a reference database (DB) of sequences for those that are similar to S , we can also predict pathways from a sequenced genome by analogy to a reference DB of pathways.

A number of groups have developed techniques for predicting the metabolic pathways of an organism from its genome and for producing integrated pathway/genome databases that model the resulting predictions [1]. The techniques used range from manual analysis to automated computational analysis and the resulting databases vary according to both the types of information they contain and the software tools they make available for the querying, visualization and analysis of that information. Such projects include the KEGG project [2,3],* the WIT project [4],** and a project at Pangea Systems that has produced the collection of microbial pathway/genome DBs described in this article. We will describe the information contained in these DBs, the methods used to create them, and some of the ways they may be exploited to support antimicrobial drug discovery.

Microbial-pathway datasets

Pangea Systems has developed a collection of integrated metabolic pathway/genome DBs for a set of medically relevant microorganisms (Table 1). These DBs consist of a set of microorganism-specific pathway/genome datasets, and a software environment for querying, analyzing and visualizing these datasets. Each dataset combines information about the genome and the metabolic network of one microorganism. The metabolic network is described in terms of four biological object types: (1) the pathways that compose the network; (2) the reactions that compose each pathway; (3) the metabolic compounds (substrates, activators, inhibitors); and (4) the enzymes that catalyze these reactions (Fig. 1). Pathways include those of biosynthesis, degradation, energy production and intermediary metabolism, for compounds such as amino acids, carbohydrates, fatty acids, nucleotides and enzyme co-factors. A few macromolecule pathways are included, but the DBs currently do not contain transmembrane-transport, gene-regulation or signal-transduction pathways (with the excep-

* See URL <http://www.genome.ad.jp/kegg/kegg.html>.

**See URL <http://wit.mcs.anl.gov/WIT2/>.

tion of the EcoCycTM *Escherichia coli* DB [5], which does contain signaling pathways).

Genomes are described in terms of three biological object types: (1) genomic maps of the sequenced chromosome(s) and plasmid(s) of the microorganism; (2) the constituent genes; and (3) the corresponding gene products (Fig. 1). The primary link between the representation of a genome and that of a metabolic network is between the gene products that encode enzymes and the reactions those enzymes catalyze.

The EcoCyc DB contains a comprehensive collection of metabolic pathway/genome information drawn from the primary literature and from the full genomic sequence of *E. coli*, including all experimentally identified *E. coli* small-molecule metabolic pathways.*** The metabolic pathways for the other microorganisms were computationally predicted from corresponding publicly available annotated genomes using the PathoLogicTM software (Fig. 1). PathoLogic uses the information contained in EcoCyc to transform annotated microbial genomes into pathway/genome DBs. The resulting DBs are referred to as “EcoCyc clone DBs,” because they store information using the same schema employed for EcoCyc. The resulting set of microbial DBs facilitate comparative metabolic-pathway analyses. For example, Table 2 shows a pairwise comparison of the predicted pathways shared by all possible pairs of organisms in the collection. There is significant variation in both the total number of pathways in a given microorganism, and in the number of pathways that it shares with other organisms.

Methodology for creating the microbial datasets

The datasets for all microorganisms except *E. coli* were created from the corresponding annotated genome using PathoLogic. The EcoCyc DB served as the reference pathway DB. However, to provide broader coverage of microbial pathways, a reference pathway DB under development at Pangea that contains pathways from many microorganisms may be used as the reference DB in place of EcoCyc. PathoLogic transforms an annotated genome for a microorganism *M* into an integrated pathway/genome DB. In most cases, the publicly available Genbank record for *M* served as the input to PathoLogic. For each microorganism, the program assessed the evidence for the presence of specific metabolic pathways (those described in EcoCyc) in *M*. The genome and any pathways for which evidence was found (see below) are then incorporated into a pathway/genome DB for *M*. The program then generates a set of reports that outline the evidence for the presence of EcoCyc pathways in *M*. Pathways are grouped by both functional category and by the evidence scores assigned to them by the PathoLogic software. These reports are written in the HyperText Markup Language (HTML) and may be viewed using a World Wide Web browser.

The process by which PathoLogic transforms an annotated genome into a pathway/genome DB involves both automatic and interactive stages; unless stated otherwise, the operations outlined below should be assumed to be automatic. A parsing tool extracts information from the

***See URL <http://ecocyc.PangeaSystems.com/ecocyc/>.

inputted annotated genome. For example, gene names, start and end nucleotide positions, and corresponding polypeptide products. No sequence analysis is performed by the PathoLogic tool — it simply extracts relevant information from the annotated genome. An EcoCyc clone DB is then created. Subsequently, DB objects for each chromosome and their component genes and open reading frames (ORFs) are generated, along with entries for corresponding polypeptides. The polypeptides are assigned the product names given in the annotated genome. All EcoCyc DB entries for reactions, compounds and pathways are then imported from EcoCyc into the clone DB.

The next stage involves linking polypeptides with enzymatic activity to reaction objects. Enzyme Commission (EC) number assignments in the input annotated genome are used to automatically create links between polypeptides and the reactions they catalyze. However, sequence annotations do not always include EC numbers for every enzyme (some Genbank genomes contain no EC numbers at all). Furthermore, some known enzyme-catalyzed reactions do not have assigned EC numbers. For these reasons, a special purpose name matching tool is used to automatically match enzyme names to one or more imported reactions. This tool uses a lookup table (that maps enzyme names and/or synonyms to reactions) constructed from four distinct information sources: (i) the EcoCyc DB; (ii) the Enzyme DB [6]; (iii) a DB developed at Pangea and (iv) an optional user provided file that maps enzyme names/synonyms to reactions.

A graphical tool for defining protein complexes is executed next. This tool automatically identifies all reactions linked to more than one polypeptide, under the assumption that annotated genomes commonly assign the same enzymatic activity to each subunit of an enzyme complex. For each such reaction, a list of all polypeptides linked to the reaction is presented to the user. In some cases these polypeptides represent the subunits of a larger enzymatic complex, whereas in other cases they represent isozymes. When the user determines from inspection of the group of polypeptides that they do constitute enzyme subunits, they can direct the program to create a DB object for the enzymatic complex, and to link the complex to both the reaction it catalyzes and its component polypeptides.

A Pangea scientist next surveys the biomedical literature to identify pathways present in M but absent from EcoCyc. These pathways are added to the clone DB using Pangea's graphical pathway editor. *E. coli* pathways grouped as probably absent from M (see below) are removed from the DB, along with any entries for reactions and compounds that do not participate in pathways grouped as possibly or probably present in M (see definitions below). At this point, a metabolic overview of the predicted metabolic network of M is generated semi-automatically.

A key feature of PathoLogic involves analyzing a genome for an organism M to assess the evidence for the presence of known *E. coli* metabolic pathways in M . A score is assigned to a pathway P based on the number of reactions in P that are predicted to occur in M . We know that a reaction occurs in M if it is linked to an enzyme of M . The score consists of three numbers, X, Y, Z , that summarize the genomic evidence for P in M . X denotes the total

number of reactions in P , Y denotes the number of these reactions known to occur in M and Z denotes the number of reactions in Y known to be used in other pathways. For example, the score assigned to the *E. coli* pathway for valine biosynthesis in *Helicobacter pylori* is 4, 2, 1, meaning that 2 of the 4 reactions required for this pathway are catalyzed by *H. pylori*, and one of these two occurs in one or more other pathways (Fig. 2). The Z value is recorded because, if a reaction R occurs in two or more pathways ($P1, P2, \dots$), we should exercise caution in counting R as evidence for $P1$ in M because the enzyme that catalyzes R might in fact be present in the genome only because of its role in the other pathways. However, intuitively, the larger the ratio $Y : X$, calculated from a pathway score, the more probable it is that this pathway is present in M . Pathways are grouped by score into three major groups: those probably absent from M ($Y = 0$), those possibly present in M ($0 < Y/X < 0.5$) and those probably present in M ($0.5 \leq Y/X$).

For a given microorganism, whenever a predicted pathway is displayed, this display clearly depicts the degree of evidence for the pathway by indicating which enzymes within the pathway were found in the genome and which were not. For those enzymes that were not found, the enzyme name is not drawn in the diagram (Fig. 2).

The KEGG and WIT approaches to pathway prediction differ from the PathoLogic approach in several respects. The KEGG and the WIT groups do not take an annotated genome (such as a Genbank record) as their starting point. They begin with raw protein sequences, which they reanalyze using automated sequence-analysis techniques to assign EC numbers to each enzyme. Because automated function-prediction techniques are not likely to be as accurate as the painstaking manual analysis of sequences performed by experts on most genome-sequencing teams, this approach will probably result in less accurate predictions of protein function than those in the original genome annotations.

The KEGG approach to pathway prediction relies on a generic, multi-species conceptualization of pathways. A KEGG pathway such as methionine biosynthesis is defined as the union of all reactions related to methionine biosynthesis that have been observed across a number of organisms. Thus, that generic pathway may not occur in its entirety in any one organism. When performing a pathway prediction for a new organism, the KEGG group highlights on the generic pathway diagram those enzymes present in the genome of that organism. The KEGG group does not build distinct database objects for the different variants of a pathway (i.e., the different pieces of a generic pathway) that exist in different species. They therefore cannot encode the pathway exactly as it is postulated to occur in that species, nor do they take into account experimental information about what pathway variant is present in a given species. In contrast, the WIT and Pangea approaches use different pathway database objects for each species to encode exactly that pathway variant that is thought to be present, and integrate literature-research findings.

Neither the WIT nor the KEGG databases can distinguish multiple isozymes that catalyze a reaction from multiple subunits of a single enzyme complex. Those databases do not encode

enzyme complexes.

Identification of false positive/negative functional annotations

Genomics-based drug discovery relies strongly on the accuracy of the functional annotation of a genome. Although the frequency of incorrect functional annotations in the sequence databases has not been firmly established, a recent study by Brenner estimates the error rate to be 8% in full microbial genomes [7]. Incorrect annotations can cost a company significant time in pursuing incorrect targets. The EcoCyc annotation of the *E. coli* genome is probably the most reliable of all annotated microbial genomes because of the extensive experimental work on *E. coli*. It therefore provides a strong reference for the annotation of additional microbial genomes.

Pathway analysis of a genome can help to identify both false-positive (incorrect) functional annotations and false-negative annotations (unidentified genes) through an examination of the pathway distribution of gene annotations. The pathway analyses identify pathway holes, which are missing steps within a pathway that is largely present. Pathway holes may be due to enzymes that have not yet been identified within the genome and are hidden among the uncharacterized ORFs. Alternatively, holes may reflect pathway variation and occur at points where the microorganism catalyzes a different step that is not present in *E. coli*. Pathway analysis also identifies singleton steps, which are single steps in a pathway where the majority of steps in the pathway are predicted to be absent. Singleton steps may be due to incorrect functional annotations; their predicted functions should be carefully verified.

The pathway reports generated by PathoLogic identifies pathway holes. Literature research may then be undertaken to establish whether enzymatic activities corresponding to such holes have been isolated biochemically from the corresponding microorganism or from one or more closely related microorganisms. If biochemical evidence is found, this provides compelling evidence that corresponding genes are present within the genome. Subsequent gene finding efforts may be prioritized based on such biochemical evidence.

Operations of the metabolic overview

The EcoCyc graphical user interface (GUI) has been extended to support querying, navigating and analyzing the information contained within a set of microbial pathway/genome DBs. This GUI contains visualization tools for all of the data types within a pathway/genome DB, including genes, enzymes, small molecules, reactions, pathways and complete genomic maps. Because these tools have been discussed in detail elsewhere [5,8], this article focuses on a new component of the GUI software for visualizing and querying the full metabolic network of an organism, which is called the Metabolic Overview (Fig. 3). Each microbial DB contains a representation of the full metabolic network of the corresponding species.

The computational operations and queries supported by the Overview illustrate many of the applications of pathway/genome DBs to drug discovery. Pathway/genome DBs support a whole-metabolic-network approach to drug discovery, in which target identification and validation occurs at the level of the metabolic network and individual pathways, as well as at the level of individual gene products. The ability to consider individual genes and enzymes from the perspective of the metabolic network facilitates selection and evaluation of enzymatic targets likely to be essential to the overall function of microbial metabolic networks.

The WIT project does not provide an Overview-like visualization. The KEGG project does support such a diagram, however, it is static, meaning that it does not support any of the highlighting or comparative operations discussed here.

Identification queries

The simplest way to query the Overview is for the user to move the mouse over a compound or a reaction within the Overview, which will cause the program to identify, in the lower left corner of the screen, the compound or reaction that the user is pointing to, as well as the pathway within which that compound or reaction is found. In this way the user can identify any element of the diagram.

The user may also request that the program highlight objects on the Overview according to a number of different criteria. Compounds may be highlighted by whole name (for example, to highlight all occurrences of arginine), by name substring, by compound class (find all amino acids), or by substructure search (highlight all compounds containing the SMILES string CC(=O)C(=O)O — SMILES is a language for writing chemical structures in terms of character strings [9]). Reaction steps may be highlighted by EC number (highlight reaction 1.2.3.4, or highlight all reactions in class 1 (oxidoreductases)), by enzyme name (highlight the reaction catalyzed by anthranilate synthetase), or by gene name (highlight the reaction catalyzed by the product of *fumA*). Reactions may also be highlighted according to their substrates (find all reactions containing both pyruvate and acetyl-CoA). Pathways may be highlighted by several criteria, such as by name, by substring or by pathway class (highlight all biosynthetic pathways for amino acids). When the user performs a sequence of highlighting operations, each result is highlighted in a different color, with an additional color used to show overlapping results.

Querying properties of the metabolic network

To investigate the regulation of the metabolic network, the user may highlight on the Overview those reactions whose catalysis is modulated by specified compounds, or classes of compounds. EcoCyc contains detailed information about the small-molecule activators and inhibitors that each enzyme is known to interact with, broken down by the mechanism of modulation (competitive, allosteric, etc). For example, the user may highlight all steps whose catalyzing enzyme is activated by ATP, or all steps that are inhibited by an entire class of compounds, such as

the amino acids. The EcoCyc data on small molecules known to interact with an enzyme (substrates, inhibitors, activators and cofactors) may facilitate the selection of lead compounds (compounds expected to inhibit enzyme targets). Some of these lead compounds may be structural analogues of small molecules that participate in the microbial metabolic network and would, therefore, be subject to comparable metabolic transformations. Thus, pathway/genome DBs may help to identify pathways by which a given lead compound may be metabolized. In addition, EcoCyc contains citations to experimental assays of function for some enzymes that could be used as part of the lead-evaluation process.

One type of drug target probably to be avoided is a metabolic enzyme for which the organism has one or more isozymes. The user may highlight all such steps within the Overview diagram (currently, 79 such steps are identified in EcoCyc). By contrast, metabolic enzymes that are used in multiple pathways are attractive targets, because knocking out a single protein could disrupt multiple pathways. An enzyme could be used in multiple pathways because it is multifunctional and/or because the reaction that it catalyzes is used in multiple pathways. EcoCyc lists 101 multifunctional enzymes and 99 reactions that are used in multiple pathways for *E. coli*. Microbial enzymes that do not have human homologs may also be attractive targets [10], although if the human and microbial enzymes have diverged sufficiently, they may exhibit differential drug responses.

The Overview may also be used to visualize the pathway distribution of genes whose knock-out mutants produce a given phenotype, such as those genes essential for growth under certain experimental conditions. A set of such genes may be loaded from a data file and the corresponding reaction steps highlighted on the Overview to examine whether these essential genes cluster within a small number of pathways, which genes of these pathways are not essential, and why.

Species comparisons

The Overview provides species-comparison operations in order to support the design of antimicrobial agents with a desired spectrum. The user may query for all reactions that are either shared by a specified set of the organisms in the collection, or that are unique to one organism with respect to a specified set of organisms. For example, the user might highlight on the Overview for *Saccharomyces cerevisiae* all reaction steps that are unique to yeast with respect to all other organisms in the collection. When two consecutive highlighting operations are performed, the overlap of the resulting sets are shown in black and the steps unique to each query are shown in a unique color. For example, in Fig. 3 with the *H. pylori* metabolic overview as the reference point, the user first highlighted all steps shared with *Haemophilus influenzae* and then highlighted all steps not shared with *E. coli*.

One advantage to asking comparative questions at the level of pathways rather than genes results from the incompleteness of genome function analyses — many enzymes within the genomes have not yet been identified. However, the presence of pathways can be inferred from

a partial enzyme complement.

Pathway-based analysis of expression data

Each microbial pathway/genome DB provides an *in silico* model of a microorganism that may be used to predict microbial metabolic properties and thereby support subsequent design of experiments to test these predictions (e.g., gene-expression studies). To aid in the interpretation of the resulting data, the Pathway Tools provide a facility for pathway-based visualization and interpretation of protein and gene-expression data.

Expression data for a given organism may be loaded from a data file and superimposed on the Overview diagram for that organism. The data file might contain a single absolute expression level for each member of a set of genes, or it might contain two expression values for each gene if the user wishes to compare relative levels of expression. Fig. 4 shows the results of superimposing expression data from DeRisi et al. [11] on the *S. cerevisiae* Overview. The data shown are the ratios of the last to the first points in this dataset, which were measured during growth under low and high glucose, respectively. This visualization allows the user to discern the coordinated expression of entire pathways (such as the TCA cycle) and changes in the expression of individual enzymes within a pathway (such as the two red steps in the gluconeogenesis pathway, which is two pathways to the left of the TCA cycle).

Pathway-based visualization of gene-expression data allows the user to rapidly identify the pathways switched on or off under a given set of experimental conditions (e.g., to see which genes or pathways are active during growth on media that simulates the biochemical environments experienced by one or more microorganisms within an infected host). Furthermore, comparable data obtained from one or more microorganisms in the presence or absence of different concentrations of a lead compound will help to identify genes that are up- or down-regulated in the presence of this compound. Some of these changes may be compensatory. For example, down-regulated genes may encode proteins that antagonize the decreased function of the targeted enzyme. Alternatively, the genes being up-regulated may encode proteins that functionally compensate for the decreased function of the targeted enzyme, or that encode enzymes that metabolize the lead compound, or transport it or one of its metabolites from the microorganism.

This visualization tool can also aid in validating algorithms for inference of new pathways from expression data by asking whether those algorithms are able to rediscover existing pathways.

Conclusions

Microbial pathway/genome DBs and related software tools will become increasingly important resources for the development of antimicrobial agents. Indeed, many of the bioinformatics tools discussed in this context generalize to other pharmacologic therapeutics. Consequently, in the

future, pathway/genome DBs will become useful resources for drug development in general.

Acknowledgments

We thank Fidel Salas for valuable discussions. This work was supported by grant 1-R01-RR-07861-01 from the National Center for Research Resources, and by Pangea Systems.

Organisms	Pathways	Genes	Genome size (bp)
<i>Escherichia coli</i>	138	4668	4639221
<i>Mycobacterium tuberculosis</i>	103	3974	4411529
<i>Bacillus subtilis</i>	112	4221	4214814
<i>Haemophilus influenzae</i>	91	1746	1830140
<i>Saccharomyces cerevisiae</i>	84	6526	12147823
<i>Helicobacter pylori</i>	74	1590	1667867
<i>Mycoplasma pneumoniae</i>	38	706	816394

Table 1

A summary of the information content of each microbial pathway/genome DB. The number of pathways defined in the DB, the number of gene objects and the length of the genome in base pairs are listed for each organism. The EcoCyc DB was created through manual curation only; the other DBs were created computationally and extended with manual curation.

	<i>Escherichia coli</i>	<i>Mycobacterium tuberculosis</i>	<i>Bacillus subtilis</i>	<i>Haemophilus influenzae</i>	<i>Saccharomyces cerevisiae</i>	<i>Helicobacter pylori</i>	<i>Mycoplasma pneumoniae</i>
<i>Escherichia coli</i>	138	103	107	90	84	73	36
<i>Mycobacterium tuberculosis</i>		103	91	79	82	70	35
<i>Bacillus subtilis</i>			112	81	79	69	36
<i>Haemophilus influenzae</i>				91	67	62	32
<i>Saccharomyces cerevisiae</i>					84	64	34
<i>Helicobacter pylori</i>						74	29
<i>Mycoplasma pneumoniae</i>							38

Table 2

Pathways shared between all pairs of organisms. Each element gives the number of pathways in common, between a pair of organisms. Only one entry is given per organism pair (rather than two identical values) resulting in an upper triangular table. If two values had been given the table would have been symmetric with diagonally opposite elements equal.

References

- [1] P. D. Karp. Metabolic databases. *Trends in Biochemical Sciences*, 23:114–116, 1998.
- [2] S. Goto, H. Bono, H. Ogata, W. Fujibuchi, T. Nishioka, K. Sato, and M. Kanehisa. Organizing and computing metabolic pathway data in terms of binary relations. In *Pacific Symposium on*

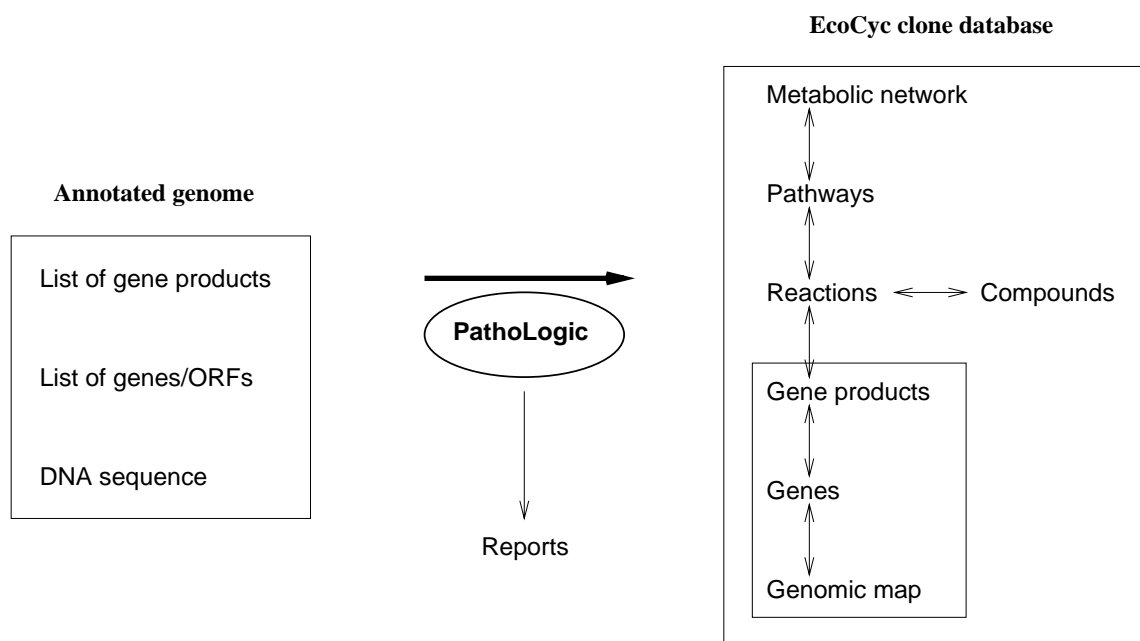


Fig. 1. A pathway/genome DB, denoted on the right, is one output of the PathoLogic program. Its input is an annotated genome, such as in the form of a Genbank flatfile. A second output of the program is a set of reports summarizing the evidence for the predicted pathways of the organism.

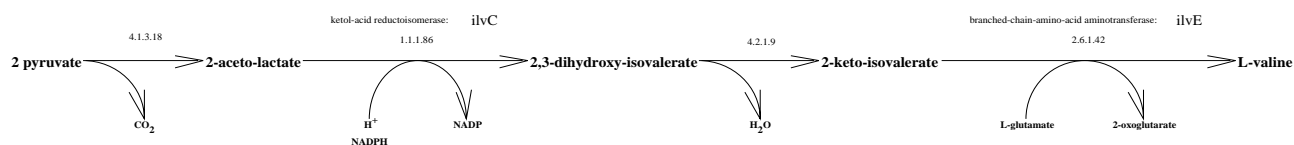


Fig. 2. The predicted *Helicobacter pylori* pathway for valine biosynthesis to which PathoLogic assigned a score of 4, 2, 1

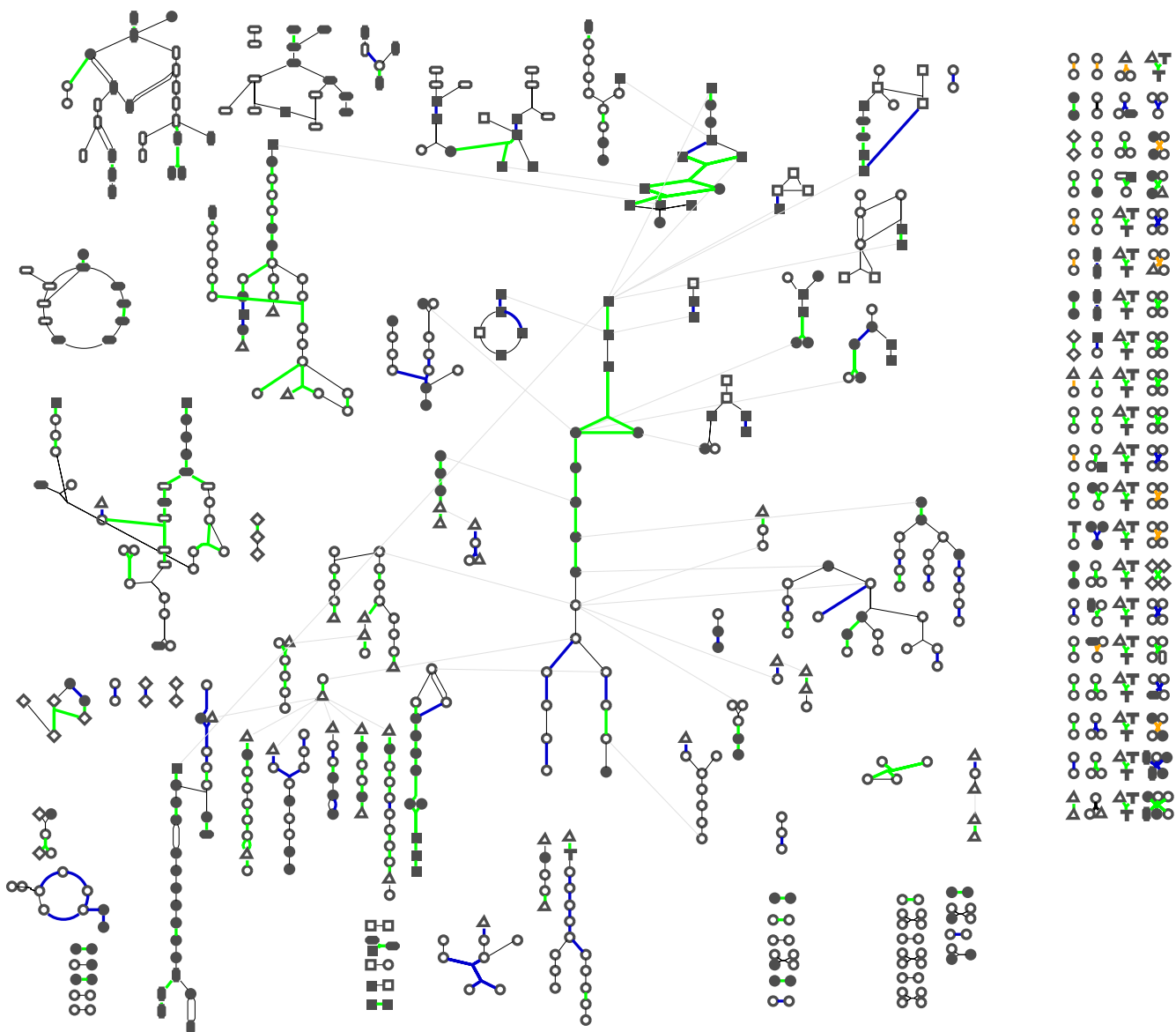


Fig. 3. The predicted metabolic network of *Helicobacter pylori*. Each line in this diagram represents a single enzyme-catalyzed Reaction and each node represents a single metabolite. Glycolysis is in the middle of the diagram, with biosynthetic pathways to their left, catabolic pathways to their right and reactions that have not been assigned to a pathway grouped along the far right hand side. The shape of each metabolite encodes its chemical class, e.g., amino acids are shown as triangles, and shaded nodes indicate phosphorylated compounds. The grey lines denote reaction steps possibly present in *Helicobacter pylori* but for which one or more enzymes were not identified in the annotated *Helicobacter pylori* genome analyzed by PathoLogic. The reactions highlighted in color show the results of a species comparison and encode the following information: blue, *Helicobacter pylori*; green, *Helicobacter pylori* and *Haemophilus influenzae*; orange, *Helicobacter pylori* and not *Escherichia coli*.

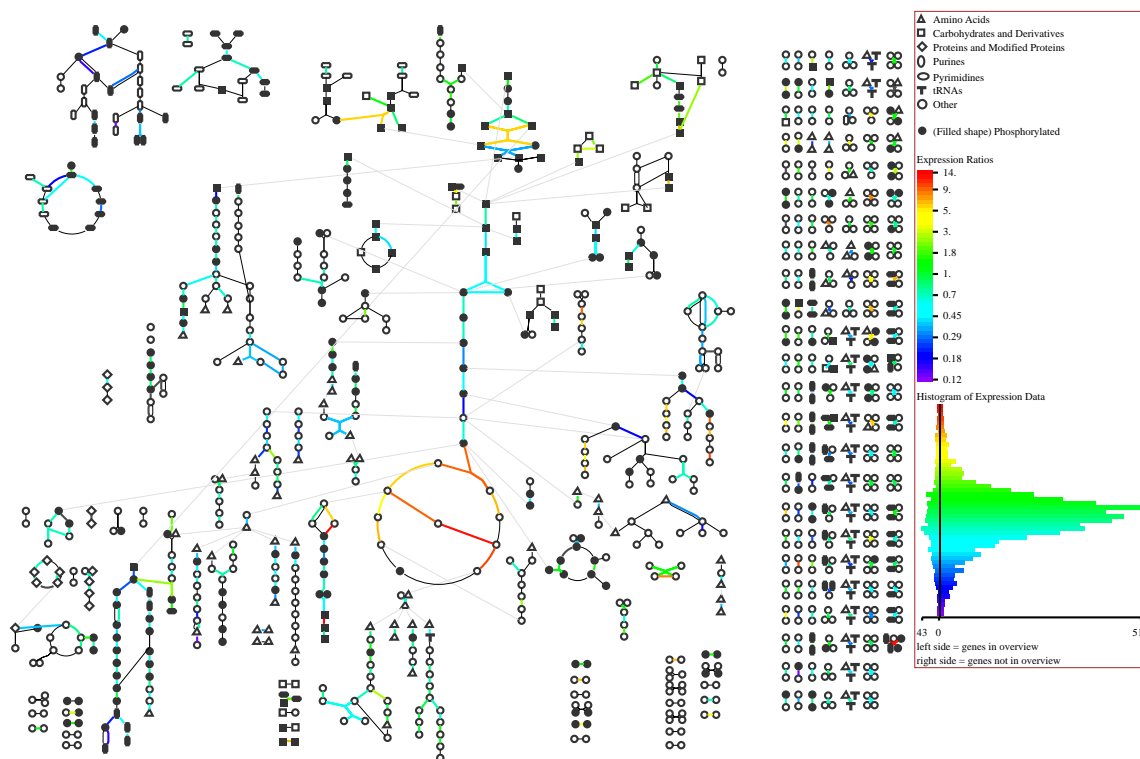


Fig. 4. Expression data superimposed on the predicted metabolic network of *Saccharomyces cerevisiae*. The color of each reaction step reflects the expression levels of the gene(s) that code for the enzyme(s) that catalyze each reaction. As indicated in the color bar in the key to the right, reactions colored yellow, orange and red represent genes the expression of which increased during the shift from high to low glucose; steps colored deep blue and purple represent genes the expression of which decreased during that shift. Steps are colored grey when the corresponding gene is unknown.

Biocomputing '97, pages 175–186, 1997.

- [3] H. Bono, H. Ogata, S. Goto, and M. Kanehisa. Reconstruction of amino acid biosynthesis pathways from the complete genome. *Genome Research*, 8:203–10, 1998.
- [4] Evgeni Selkov, Miliusha Galimova, Igor Goryanin, Yuri Gretchkin, Natalia Ivanova, Yuri Komarov, Natalia Maltsev, Natalia Mikhailova, Valeri Nenashev, Ross Overbeek, Elena Panyushkina, Lyudmila Pronevitch, and Evgeni Selkov Jr. The metabolic pathway collection: an update. *Nuc. Acids Res.*, 25(1):37–38, 1997.
- [5] 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.
- [6] A. Bairoch. The ENZYME databank in 1995. *Nucl Acids Res*, 24:221–222, 1996.
- [7] SE. Brenner. Errors in genome annotation. *Trends in Genetics*, 15(4), 1999.
- [8] P. Karp and S. Paley. Integrated access to metabolic and genomic data. *Journal of Computational Biology*, 3(1):191–212, 1996.
- [9] 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.

- [10] MY Galperin, DR Walker, and EV Koonin. Analogous enzymes: Independent inventions in enzyme evolution. *Genome Research*, 8:779–90, 1998.
- [11] J.L. DeRisi, V.R. Iyer, and P.O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.