

# MPW: the Metabolic Pathways Database

Evgeni Selkov, Jr<sup>1</sup>, Yuri Grechkin<sup>1</sup>, Natalia Mikhailova<sup>1</sup> and Evgeni Selkov<sup>1,2,\*</sup>

<sup>1</sup>Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142292 Pushchino, Moscow region, Russia and <sup>2</sup>Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, MCS-221, Argonne, IL 60439-4844, USA

Received October 7, 1997; Accepted October 24, 1997

## ABSTRACT

The Metabolic Pathways Database (MPW) ([www.biobase.com/emphome.html/homepage.html/pages/pathways.html](http://www.biobase.com/emphome.html/homepage.html/pages/pathways.html)) a derivative of EMP ([www.biobase.com/EMP](http://www.biobase.com/EMP)) plays a fundamental role in the technology of metabolic reconstructions from sequenced genomes under the PUMA ([www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html](http://www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html)), WIT ([www.mcs.anl.gov/home/compbio/WIT/wit.html](http://www.mcs.anl.gov/home/compbio/WIT/wit.html)) and WIT2 ([beauty.isdn.mcs.anl.gov/WIT2.pub/CGI/user.cgi](http://beauty.isdn.mcs.anl.gov/WIT2.pub/CGI/user.cgi)) systems. In October 1997, it included some 2800 pathway diagrams covering primary and secondary metabolism, membrane transport, signal transduction pathways, intracellular traffic, translation and transcription. In the current public release of MPW ([beauty.isdn.mcs.anl.gov/MPW](http://beauty.isdn.mcs.anl.gov/MPW)), the encoding is based on the logical structure of the pathways and is represented by the objects commonly used in electronic circuit design. This facilitates drawing and editing the diagrams and makes possible automation of the basic simulation operations such as deriving stoichiometric matrices, rate laws, and, ultimately, dynamic models of metabolic pathways. Individual pathway diagrams, automatically derived from the original ASCII records, are stored as SGML instances supplemented by relational indices. An auxiliary database of compound names and structures, encoded in the SMILES format, is maintained to unambiguously connect the pathways to the chemical structures of their intermediates.

## INTRODUCTION

The integrating role of metabolic pathway databases in bioinformatics is well recognized nowadays. A number of metabolic databases are accessible via the Web. Four of them are general-purpose databases: the Metabolic Pathways Database, MPW (1), accessible via PUMA ([www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html](http://www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html)), WIT ([www.mcs.anl.gov/home/compbio/WIT/wit.html](http://www.mcs.anl.gov/home/compbio/WIT/wit.html)), WIT2 ([beauty.isdn.mcs.anl.gov/WIT2.pub/CGI/user.cgi](http://beauty.isdn.mcs.anl.gov/WIT2.pub/CGI/user.cgi)), the NIH GenoBase ([http://specter.dcrf.nih.gov:8004/Pathway/pathway\\_toc\\_by\\_name.html](http://specter.dcrf.nih.gov:8004/Pathway/pathway_toc_by_name.html)), the Kyoto Encyclopedia of Genes and Genomes, KEGG ([www.genome.ad.jp/kegg/](http://www.genome.ad.jp/kegg/)) and the Gerhard's Michal Biochemical Pathways accessible from Swiss-Prot ([expasy.hcuge.ch/cgi-bin/search-biochem-index](http://expasy.hcuge.ch/cgi-bin/search-biochem-index)). The others represent the metabolisms of *Escherichia coli*, EcoCyc (2), *Haemophilus influenzae*, HinCyc (3), and soybean, SoyBase ([cgsc.biology.yale.edu/metab.html](http://cgsc.biology.yale.edu/metab.html)), and some specific processes like signal transduction ([www.genome.ad.jp/brite/CellCycleMaps.html](http://www.genome.ad.jp/brite/CellCycleMaps.html), [www.nihs.go.jp/taka/csndb.html](http://www.nihs.go.jp/taka/csndb.html)) or microbial biodegradation of pollutants (4). MPW is a derivative of EMP (5–7). Its pathway diagrams display primary and secondary metabolism, membrane transport, signal transduction, intracellular traffic, translation and transcription. The database is being updated with an increment of ~400–600 new maps a year (Fig. 1). Along with the stoichiometric skeletons of metabolic pathways, the diagrams represent the substrate and coenzyme specificity of enzymes, their sub-cellular locations, required prosthetic groups and cofactors, as well as the taxonomic occurrence of pathways.

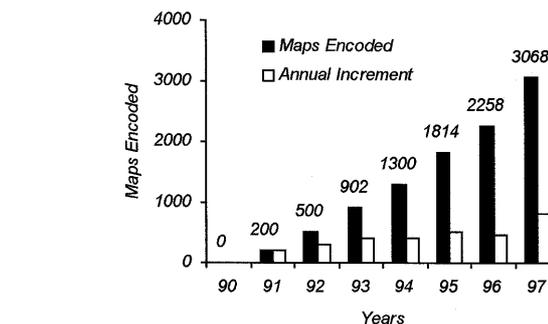


Figure 1. Dynamics of MPW growth. The figure for 1997 is a projection from October to the end of December 1997.

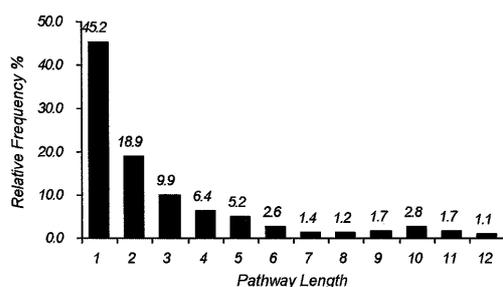
The MPW diagrams were originally encoded and distributed as formatted ASCII text. Later ([www.cme.msu.edu/MPW/](http://www.cme.msu.edu/MPW/)), we have found a way of converting these diagrams into the SGML format ([www.sgmlopen.org/sgml/docs/sgmldesc.htm](http://www.sgmlopen.org/sgml/docs/sgmldesc.htm)) to produce a new database ([beauty.isdn.mcs.anl.gov/MPW](http://beauty.isdn.mcs.anl.gov/MPW)) with computable objects and relations.

The MPW diagrams were originally encoded and distributed as formatted ASCII text. Later ([www.cme.msu.edu/MPW/](http://www.cme.msu.edu/MPW/)), we have found a way of converting these diagrams into the SGML format ([www.sgmlopen.org/sgml/docs/sgmldesc.htm](http://www.sgmlopen.org/sgml/docs/sgmldesc.htm)) to produce a new database ([beauty.isdn.mcs.anl.gov/MPW](http://beauty.isdn.mcs.anl.gov/MPW)) with computable objects and relations.

## WHAT IS A METABOLIC PATHWAY?

The definition of a metabolic pathway lies in the basis of metabolic bioinformatics. In accordance with the EMP nomenclature ([www.biobase.com/EMP](http://www.biobase.com/EMP)), a metabolic pathway is a set

\*To whom correspondence should be addressed at present address: 5020 South Lake Shore Drive, Apartment 2009, Chicago, IL 60615, USA.  
Tel: +1 773 493 4156; Fax: +1 773 288 5985; Email: [evgeni@mcs.anl.gov](mailto:evgeni@mcs.anl.gov)



**Figure 2.** The truncated pathway length distribution computed for the MPW collection. The actual distribution extends up to the length of thirty reaction steps.

of oriented reactions interacting under given physiological conditions via simple or apparently simple intermediates. This definition is based on the definitions of metabolic intermediates. A compound is considered as a *simple intermediate* if there is a unique couple *source flux*–*sink flux* controlling its concentration in a given intra- or extracellular compartment. If the number of fluxes balancing the compound concentration is greater than two, such a compound is considered as a *crossroad intermediate*.

It is tempting to define pathways on a purely topological basis. One could define a pathway as a set of reactions connecting two adjacent crossroads. In reality, however, such a definition splits the vast majority of known classical pathways into a large number of much shorter subpathways, since many of them run via crossroad intermediates. For instance, practically all the intermediates of the glycolytic system or of the TCA cycle are crossroad ones. Fortunately, under specific physiological conditions, leakage of intermediates into side reactions and pathways may be insignificant. This allows us to introduce an asymptotic definition of an *apparently simple intermediate*, as a crossroad compound whose concentration is controlled by the dominating unique couple, source flux–sink flux. Of course, the quantitative measure of this domination depends on the accuracy of experimental data and the desired theoretical approximation.

Thus, by practical necessity, the above general definition of metabolic pathways is asymptotic and conditional. This definition generates the pathway length distribution for MPW as shown in Figure 2. The distribution shows that one-step reaction pathways dominate the cellular metabolism. The relative frequency of occurrence of multistep pathways decreases with the pathway length in a nearly geometric progression.

## EMP NOMENCLATURE OF METABOLIC PATHWAYS

According to the EMP nomenclature ([www.biobase.com/EMP](http://www.biobase.com/EMP)), all systematic pathway names are generated in the form:

**substrates**–**products**\_function\_(**coenzymes**)\_(**locations**)\_[**comment**],

where **substrates**, **products** and **coenzymes** are the lists of substrates entering the first reaction of the pathway, of products leaving the last reaction, and of coenzymes consumed along the pathway, respectively; the term **locations** lists the cellular loci of the pathway enzymes. The substrates and products are listed in a decreasing order of their molecular masses. The coenzymes and locations are listed in order of their first appearance along the path. The nomenclature specifies the following categories of pathway **function**:

**anabolism,**  
**catabolism,**  
**electron transport,**  
**membrane transport,**  
**signal transduction.**

The distinction between **anabolism** and **catabolism** is based on the relative masses of the heaviest initial substrate and the heaviest end product. A pathway is classified as catabolic if the molecular mass of its end product is less than the molecular mass of the initial substrate. In the opposite case, it is classified as anabolic. As a rule, these definitions are in good agreement with the currently used ones. However, in some cases they may be different. In particular, the degradation of complex organic molecules may result in formation of acetyl-CoA or succinyl-CoA, both of which are considered, according to the nomenclature, as crossroad intermediates and hence as the end products of these pathways. Such a product is usually heavier than the initial substrate; therefore, the pathway must be classified as an anabolic one. This contradicts the apparently obvious catabolic function of the pathway.

A systematic pathway name can be confusingly long. Here are some examples:

pyruvate oxidation via bacterial TCA cycle:

**pyruvate**–**CO<sub>2</sub>**\_catabolism\_(**lipoamide**,**NADP<sup>+</sup>**,**CoA**,**ADP**,**FAD**,**NAD<sup>+</sup>**)\_(**cytosol**,**plasma membrane**),

a symport of **Fe<sup>3+</sup>**<sub>extracellular</sub> and pyochelin<sub>extracellular</sub> into cytosol of a Gram-negative bacterium:

**pyochelin**<sub>extracellular</sub>–**Fe<sup>3+</sup>**<sub>extracellular</sub>–**pyochelin**,**Fe<sup>2+</sup>**\_membrane\_transport\_(**ATP**,**NADPH**)\_(**outer membrane**,**periplasma**,**plasma membrane**),

electron transport from **NADH** to **O<sub>2</sub>** via micrococcal respiratory chain coupled with formation of transmembrane proton gradient:

**NADH**,**H<sup>+</sup>**–**O<sub>2</sub>**,**H<sup>+</sup>**<sub>extracellular</sub>\_electron\_transport\_(**plasma membrane**)\_[**Micrococcaceae**],

signalling from **cAMP** and **Ca<sup>2+</sup>** to myosin light chain:

**3',5'-cyclic AMP**,**Ca<sup>2+</sup>**–**myosin light chain signal**\_transduction\_(**ATP**)\_(**cytosol**).

(For sake of simplicity, the details of super- and subscripts encoding are dropped out here.) However, the nomenclature was not designed to substitute the existing common pathway names. Its purpose was purely computational. With the help of these precise names, one can easily compute many types of listings and automatically generate outlines of different depth and profiles chosen by the user.

Along with the systematic names, the database uses much shorter metabolic pathway names for listings and alternative names collected from the literature. The metabolic pathway names are formed from systematic ones by omitting coenzymes and default cytosolic locations, if the result remains unambiguous.

## CONVERSION OF THE ASCII DRAWINGS INTO A STRUCTURED FORM

The benefits of the ASCII text-based encoding became quite obvious when Ross Overbeek used the diagrams to connect them to sequence data and to create PUMA ([www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html](http://www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html)) and we later made our first attempt to automate detailed parsing of the pathway diagrams ([www.cme.msu.edu/MPW/](http://www.cme.msu.edu/MPW/)). Since then, we have further improved the recognition algorithm so that it can recognize

flawlessly more than three quarters of the original MPW records, leaving only minor gaps in the remaining ones. It provided us with a bulk of computer-readable metabolic data to perform thorough analysis and to develop a general model of a metabolic pathway.

According to this model, networks of intermediates are organized as sets locked within cellular compartments. The networks, each representing the traffic of a single intermediate within a single compartment, communicate to each other through directed 'parts', similar to parts used in electronic circuit diagrams. The parts represent chemical reactions, catalyzed or spontaneous, or membrane transport mechanisms. Each part has a number of assigned 'pins', representing chemical compounds that take part in the reaction. A pin possesses a type attribute indicating the role of the corresponding compound in the reaction: substrate, product, catalyst, coenzyme, cofactor or regulator. Connecting a pin to a network automatically allocates the corresponding compound inside or outside the cell compartment (e.g., cytosol or membrane), and affects the stoichiometry of the whole pathway by adding or subtracting a number of terms to the balance equation for the network.

This content-oriented model allows the same set of data to be used for various purposes, including generation of high quality graphics, typesetting and layout. However, its primary use is computing: integrating the data and deriving stoichiometric matrices, rate laws, and, ultimately, dynamic models of metabolic pathways.

The new MPW release ([beauty.isdn.mcs.anl.gov/MPW](http://beauty.isdn.mcs.anl.gov/MPW)) presents individual pathway diagrams, captured from the original ASCII records and available for further processing in the form of SGML instances supplemented with relational indices and an auxiliary database of compound names and their structures. The latter is based on the SMILES format ([daylight.com/dayhtml/smiles/index.html](http://daylight.com/dayhtml/smiles/index.html)) and is maintained to unambiguously resolve common compound names used to encode reactions in different pathway instances.

## MPW IN METABOLIC RECONSTRUCTIONS FROM SEQUENCED GENOMES

With the beginning of the era of large-scale genome sequencing, MPW started to play a key role in the metabolic reconstructions. Over 20 reconstructions based on complete or partial genomes are available from the PUMA/WIT/WIT2 systems. The reconstructions tend to be consistent models of cellular organizations. Each such model is an electronic album of metabolic maps.

Each reconstruction goes through two main phases. In the first phase, ORFs are assigned to the asserted functions; in the second phase the functions are assigned to fit the pathways best. Fitting pathways and functions, like fitting mathematical models, proceeds iteratively to attain the best match of the whole reconstruction to the available sequence data, biochemistry and phenotype knowledge. The fitting may require updating the MPW collection with new pathways or new versions of the old ones.

The biochemistry and phenotypic knowledge accumulated in MPW helps to identify the 'missing genes' owing to the functional interdependence among the proteins of a pathway and to the analogous interdependence among the pathways in the whole cell metabolism.

Figure 2 shows the distribution of pathway length plotted for MPW collection. The pathways of length more than one step constitute nearly half of the collection. It is clear that identifying a part of the proteins of a multistep pathway provides a strong clue for a focused search of the missing genes. The longer pathway, the stronger is the clue and the higher the probability of identifying the missing gene.

Of course, the pathway length is a weak measure of the functional interdependence among proteins, since many one-step pathways, like membrane transporting mechanisms, may be dependent on the concerted action of many proteins. Therefore, in a general case, the predicting power of a functional cluster depends on the number of proteins of the cluster.

One can easily see that the functional clustering of proteins extends itself far beyond the pathway limits. The whole cellular metabolism must be more or less balanced. Hence, any compound of the cell must have a source and a sink. This requirement reveals large functional clusters composed of many single-step pathways.

## PROJECTED DEVELOPMENT

Data modeling for a database of such complexity is an iterative process and is still far from its completion. Our immediate goal is to integrate MPW with EMP and WIT2, to install specialized authoring tools for the direct submission of pathways and enzymology data, and to develop software allowing analysis and simulation of the metabolism.

## ACKNOWLEDGEMENT

This work was supported by US Department of Energy under Contract W-31-109-Eng-38, and award OR00033-97CIS001.

## REFERENCES

- Selkov, E., Galimova, M., Goryanin, I., Gretchkin, Y., Ivanova, N., Komarov, Y., Maltsev, N., Mikhailova, N., Nenashev, V., Overbeek, R., *et al.* (1997) *Nucleic Acids Res.*, **25**, 37–38.
- Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. and Krummenacker, M. (1997) *Nucleic Acids Res.*, **25**, 43–50 [see also this issue (1998) *Nucleic Acids Res.* **26**, 50–53].
- Karp, P., Ouzounis, C. and Paley, S. (1996) In *Intelligent Systems for Molecular Biology 1996*, in press.
- Ellis, L.B.M. and Wackett, L.P. (1995) *Soc. Ind. Microb. News* **45**, 167–173.
- Selkov, E.E., Goryanin, I.I., Kaimachnikov, N.P., Shevelev, E.L. and Yunus, I.A. (1989) *Studia Biophys.* **129**, 155–164
- Sel'kov, E.E., Goryanin, I.I., Kaimachnikov, N.P., Shevelev, E.L. and Yunus, Y.A. (1990) In P.S. Glaeser, P.S. (ed.), *Scientific and Technical Data in a New Era*. Hemisphere Publishing Corp., pp. 22–27.
- Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., *et al.* (1996) *Nucleic Acids Res.*, **24**, 26–28.