26–28 Nucleic Acids Research, 1996, Vol. 24, No. 1

The metabolic pathway collection from EMP: the enzymes and metabolic pathways database

Evgeni Selkov, Svetlana Basmanova, Terry Gaasterland[1], Igor Goryanin, Yuri Gretchkin, Natalia Maltsev[1], Valeri Nenashev, Ross Overbeek[1,*], Elena Panyushkina, Lyudmila Pronevitch, Evgeni Selkov, Jr and Ilya Yunus

Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142292 Pushchino, Russia and [1]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

Received September 26, 1995; Revised and Accepted October 26, 1995

ABSTRACT

The Enzymes and Metabolic Pathways database (EMP) is an encoding of the contents of over 10 000 original publications on the topics of enzymology and metabolism. This large body of information has been transformed into a queryable database. An extraction of over 1800 pictorial representations of metabolic pathways from this collection is freely available on the World Wide Web. We believe that this collection will play an important role in the interpretation of genetic sequence data, as well as offering a meaningful framework for the integration of many other forms of biological data.

INTRODUCTION

The Enzymes and Metabolic Pathways database (EMP) is a database on the biochemistry of some 1800 different organisms (1). Over 10 000 journal articles have been encoded into quantitative data for almost all documented enzymes. The curation of the encoded data and of the pictorial representations of pathways is an ongoing project centered at the Laboratory of Mathematical Simulation of Multienzyme Systems at the Institute of Theoretical and Experimental Biophysics of the Russian Academy of Sciences, in Pushchino, Russia.

The effort to build EMP was initiated in 1984. The initial motivation was to support internal projects in the mathematical simulation of cell metabolism. The goal of the effort was to encode as much of the known data relating to enzymology as possible. An attempt was made to define a database format that allowed one to encode the types of factual assertions made by authors of papers in enzymology (2). Once an initial format was defined, the process of encoding research papers began.

When metabolic pathways were presented in papers, they were encoded into EMP. Often, it was necessary to add the EC (enzyme nomenclature) numbers. As the collection of pathways grew, it became possible for an increasing number of cross-checks to be performed, and numerous variations of common pathways were recognized. Each pathway has been labeled with EC numbers when known. In 1995, these pathways were extracted from EMP and made freely available to other researchers (see below for information on distribution). While this paper describes the collection of over 1800 metabolic pathway diagrams, these represent a small portion of the overall contents of EMP; the metabolic pathways do indeed offer a useful functional overview, but the vast majority of the content of EMP is detailed data relating to specific enzymes (encoded into slightly over 300 distinct field types).

Contents of the Metabolic Pathway Collection

Currently, the metabolic pathway collection available to users via the World Wide Web contains 1814 distinct pathway drawings. Each chart is associated with one or more taxonomic groups. Table 1 gives the most represented taxonomic groups, along with the number of associated diagrams.

There are over 250 other taxonomic groups for which fewer than 10 metabolic pathways are included in the collection.

Table 2 shows the number of enzymes covered by the pathways for the three most represented organisms (the reader should note that this number does not represent the number of sequenced enzymes, but rather the number of distinct EC numbers that occur in the metabolic pathway diagrams). The reader should be aware that the same enzyme may appear in a number of charts.

We have made the pathway collection available in two forms:

1. It can be browsed in the context of the World Wide Web application PUMA, which can be reached via the following URL:
   http://www.mcs.anl.gov/home/compbio/PUMA/Production/puma.html
2. It can be acquired via anonymous FTP from the BioBase server (details below).

The collection available from the FTP site is broken into the following broad categories for ease of distribution:

Amino Acid Metabolism
Aromatic Hydrocarbons
Carbohydrate Metabolism
Coenzymes and Vitamins
Electron Transport
Enzyme Metabolism

* To whom correspondence should be addressed

Halide Metabolism
Hydrogen Metabolism
Intracellular Transport
Lipid Metabolism
Nitrogen Metabolism
Nucleic Acid Metabolism
Membrane Transport
One-Carbon Metabolism
Oxygen Metabolism
Phosphate Metabolism
Protein Metabolism
Purine Metabolism
Pyrimidine Metabolism
Signal Transduction
Sulfur Metabolism

Users acquiring these distinct subsets will wish to impose a more refined structure and some appropriate browsing mechanism. We realize that there are a number of different ways in which the collection could be organized to support ease of access. Our intent is to provide access in two of these ways, via the EMP database and through the PUMA system on the WWW, and to provide mechanisms for other groups to restructure and display the collection for their own purposes.

Table 1. Number of charts for organism and taxonomic groups

Number of Charts	Organism/Taxonomic Group
667	*Escherichia coli*
531	*Haemophilus influenzae*
418	*Homo sapiens*
379	*Mammalia*
135	*Rattus norvegicus*
397	*Rodentia*
115	*Saccharomyces cerevisiae*
316	*Aves*
101	*Ascomycotina*
57	*Salmonella typhimurium*
56	*Oryctolagus cuniculus*
52	*Bos taurus*
51	*Embryophyta*
47	*Pseudomonas*
40	*Bacillaceae*
407	*Mycoplasma capricolum*
366	*Archaea*
29	*Sus scrofa*
26	*Mus musculus*
19	*Pseudomonas putida*
19	*Clostridium*
16	*Pseudomonas* sp.
14	*Tetrahymena pyriformis*
14	*Gallus gallus*
13	*Clostridium acetobutylicum*
12	*Neurospora crassa*
12	*Fungi imperfecti*
12	*Canis familiaris*
359	*Sulfolobus solfataricus*
10	*Pseudomonas fluorescens*
10	*Protozoa*
350	*Bacillus subtilis*

Table 2. Number of enzymes for distinct organisms

Organism	Distinct enzymes in charts
Escherichia coli	434 enzymes
Haemophilus influenzae	335 enzymes
Homo sapiens	204 enzymes

A nomenclature for pathways

Early in the EMP project, it became clear that it was necessary to develop a nomenclature of metabolic pathways, and rules for generating systematic names were formulated. This was necessary to manage such a large collection (which will grow to contain thousands of pathways and variations of pathways). This nomenclature allows those wishing to restructure the collection or to write a browser for the collection to have access to a compact representation of the function of the pathway.

The systematic name contains the initial substrates, final products, the function of the pathway, coenzymes, and cellular location of the pathway enzymes. Every metabolic pathway record includes this characteristic systematic pathway name. In addition, each record includes a shorter, but still unequivocal, recommended pathway name. Finally, a set of common names for the pathway are also encoded.

A brief description of the systematic name would appear as
Substrates–Products_Function_
(Coenzymes)_(Locations)_[Comment]

For example, one of the versions of the Entner–Doudoroff Pathway encoded in the database is characterized by the name
D-glucose–pyruvate_catabolism_
(ATP,_NADP('+),_NAD('+),_ADP)_(cytosol)
while the recommended name would be
Glucose–pyruvate_catabolism_
[via D–glucono–1 ,5–lactone_6–phosphate]
and the common name would be
Entner–Doudoroff pathway.

It should be noted that a number of pathways in the collection (those encoded early in the project) do not strictly conform to the above conventions, and we are attempting to correct omissions as quickly as possible.

PUMA, one means of access to the collection

We believe that this collection of metabolic pathways offers a powerful way to organize access to other important categories of biological data. As an illustration, we have added the pathways to the PUMA system developed at Argonne National Laboratory to offer integrated access to biological data to support interpretation of genomes (the URL for PUMA is given above). Within the context of PUMA, the user has access to a general functional overview of organisms in which the pathways have been broken into small functional categories. By simply clicking on a functional category, one has access to the set of pathways corresponding to the function. The enzymes in the pathways have been connected to alignments, protein sequences, and related enzymatic and sequence databases. The compounds connect to documents that summarize the sets of pathways that utilize the given compound as a substrate, product, or intermediate.

In addition to a general overview that attempts to organize the entire collection of metabolic pathways, PUMA also offers access

to over 200 functional overviews of specific organisms. Each of these overviews has been constructed to include the metabolic pathways that have been identified for the specific organism. We hasten to add that these overviews are far from complete; indeed, they do not contain many pathways present in the literature. However, we are making a concerted attempt to offer as complete a metabolic picture as possible for those organisms being sequenced within the growing number of genome initiatives.

As an example of what can be offered via the metabolic pathways, we encourage the reader to access PUMA and examine the collection for *Escherichia coli*. When a specific functional category is accessed, the user is shown which pathways implement the function. For each pathway, the enzymes for which sequence exists are shown (links to acquire the sequence are provided), alignments containing known sequences are indicated, and a list of the unsequenced enzymes is provided. In the case in which a sequence does not exist and for which the corresponding sequence does not exist for any other organism (i.e., the sequence for the enzyme does not exist for any organism), the fact is noted. The significance of including the EC numbers for each pathway becomes obvious: they allow the pathways to be easily connected to the rich and growing collection of databases containing enzymatic data (3–5).

Reconstruction of the metabolism of organisms from sequence data

One of the most significant applications of this collection of pathways will be to support the analysis of the metabolism for organisms for which a substantial amount of sequence data already exists. Indeed, now that several complete genomes have been sequenced, the utility of metabolic overviews becomes increasingly apparent. It is also worth noting as an aside that assignment of function to coding sequences in a genome is a difficult task, and that access to an accurate metabolic characterization of the organism or a close relative often sheds light on the function of specific genes.

We are now completing collections of pathways to represent what is known of the metabolism of *Haemophilus influenzae, Mycoplasma capricolum, Mycoplasma genitalium* and *Sulfolobus solfataricus*. As more genome sequences are completed, we will add their 'constructions of metabolism' to the collection. The process of reconstructing the metabolism of an organism from sequence data involves using the sequence data to establish enzymes that are known to be present, supplementing this list by using the biochemical literature (since many sequenced genes have an unknown function, while the literature often explicitly identifies the presence or absence of specific functions), using what is known about phylogenetically close organisms, and finally integrating these different sources of information. The process of weighing different forms of evidence often reveals inconsistencies and offers one of the important means for gradually addressing errors across the public databases.

The reconstruction of the metabolism of these first complete genomes is a challenging task that requires familiarity with a great deal of the biochemical literature. However, since these early organisms were selected to ensure phylogenetic and biochemical diversity, we believe that each step toward formulating a more complete picture of their metabolism should dramatically simplify the task for other related organisms. Indeed, we are working on tools to support and automate sections of the task (6).

Comments on the format of the pathways

This collection of metabolic pathways originated as the 'working notes' of Evgeni Selkov. As an experiment, we spent a fair amount of effort converting them to a set of relations that could be used in standard relational databases. This approach appeals to those with a background in databases, and it certainly simplifies a number of important issues relating to maintenance and organization of the collection. However, the real choice is between keeping the primary form of the data as charts (supplying software to produce the relational representation) or keeping the collection as a relational database (supplying software to produce the charts). In some sense the choices appear equivalent. Our experience suggests that operationally this might not be completely accurate. Certainly, the main curator of the collection finds it far more convenient to encode and work with the actual drawings. It also seems likely to us that, as other experts create and exchange encoded pathways, the basic exchangeable unit should probably be drawings and not encoded tables.

We clearly need to move toward a situation in which the drawings conform rigidly to a set of minimal standards that allow ease of conversion directly to a relational format, along with tools that will render drawings in a variety of forms for different purposes. To do this properly is more challenging than it appears at first glance. We will attempt to facilitate moves in this direction, but it must be emphasized that the charts often contain information well beyond that given by the reaction equations, and this information must not be lost (indeed, it seems likely to us that the relational encoding will naturally evolve to contain a subset of the information contained in the drawings).

Availability

Downloadable demo versions of EMP are available via anonymous FTP. The collection of metabolic pathways described in this article is publicly available (free of charge) via anonymous FTP. For further information, send an empty electronic mail message (or one containing the single word INFO) to emp_info@biobase.com.

Those who do not have access to electronic mail may write to Biological Databases Inc., 2004 South Wright Street, Urbana, IL 61801, USA.

ACKNOWLEDGEMENT

This work has been supported in part by the US Department of Energy.

REFERENCES

1 Selkov, E., Goryanin, I., Kaimachnikov, N., Shevelev, E. and Yunus, I. (1990) In Glaeser, P. (ed.), *Scientific and Technical Data in a New Era.* Hemisphere Publishing Corp., pp. 22–27.
2 EMP User Manual: A Guide to the Enzymes and Metabolic Pathways Database (1995) Release 1.0, September.
3 Bairoch, A. (1994) *Nucleic Acids Res.*, **22**, 3626–3627.
4 Bairoch, A. and Boeckmann B. (1994) *Nucleic Acids Res.* **22**, 3578–3580.
5 Suyama, M., Ogiwara, A., Nishioka, T. and Oda, J. (1993) *Comput. Appl. Biosci.*, **9**, 9–15.
6 Gaasterland, T., Maltsev, N. and Olrerbeek, R. (1995) In Proceedings of the Second International Meeting on Integration of Molecular Biological Databases, Cambridge, England, July.