The Metabolic Pathway Collection: An Update

Evgeni Selkov,[1,2] Miliusha Galimova,[1] Igor Goryanin, Yuri Gretchkin,[1] Natalia Ivanova,[1] Yuri Komarov,[1] Natalia Maltsev,[2] Natalia Mikhailova,[1] Valeri Nenashev,[1] Ross Overbeek,[2] Elena Panyushkina,[1] Lyudmila Pronevitch,[1] and Evgeni Selkov Jr.[1]

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

Abstract. The Metabolic Pathway Collection from EMP is an extraction of data from the larger Enzymes and Metabolic Pathways database (EMP). This extraction has been made publicly available in hopes that others will find it useful for a variety of purposes. The original release in October 1995 contained 1814 distinct pathways. The current collection contains 2180. Metabolic reconstructions for the first completely sequenced organisms—*Haemophilus influenzae, Mycoplasma genitalium, Saccharomyces cerevisiae,* and *Methanococcus janaschii*—are all included in the current release. All of the pathways in the collections are available as ascii files in the form generated by the main curator, Evgeni Selkov. In addition, we are offering a more structured encoding of a subset of the collection; our initial release of this subcollection includes all of the pathways in *Mycoplasma genitalium,* and we ultimately intend to offer the entire collection in this form as well.

1 Introduction

As we noted in our original release of the Metabolic Pathway Collection, the diagrams in the collection are extracted from the Enzymes and Metabolic Pathways database (EMP) database [1], which covers the biochemistry of some 1400 different organisms. EMP includes encodings of over 10,000 journal articles and attempts to provide an increasingly comprehensive coverage of relevant enzymatic data.

The Metabolic Pathway Collection has been released [2] and made available on the Web to support efforts to encode metabolism, reconstruct the metabolism of sequenced organisms, and eventually to achieve a comprehensive encoding of the known metabolism of sequenced organisms. We have added several hundred new diagrams covering pathways required to characterize the sequenced microbial genomes.

2 Metabolic Reconstructions

During the past year, the collection has been used in efforts to develop metabolic reconstructions for a number of the initial complete genomes that have appeared; in particular, metabolic reconstructions have been made available for *Haemophilus influenzae, Mycoplasma genitalium, Saccharomyces cerevisiae,* and *Methanococcus janaschii*—the first four completely sequenced genomes. The metabolic

reconstructions are available through the WIT system, which was created to support the reconstruction of metabolism from sequence data [3]. These reconstructions are models attempting to reconcile known phenotypic and biochemical data with the released sequence data. They are presented as collections of metabolic pathways in which functional roles in the diagrams (normally enzymes, but occasionally noncatalytic roles) are directly linked to the sequences in the genome that are believed to encode the correponding proteins.

Since a primary goal of gathering and maintaining the collection is to support the creation of metabolic reconstructions, special attention has focused on the pathways in these initial genomes. It is clear that the collection will need to be expanded rapidly to account for the metabolic diversity represented by organisms for which genomes will appear during the coming year. Indeed, the recent release of the *Synechocystis sp.* [4] genome has already motivated a number of recent additions.

3 Improvements

The collection of pathways originated as the "working notes" of Evgeni Selkov. These diagrams can by themselves play a very useful role in constructing integrations of numerous form of biological data. However, in the current form they have some notable shortcomings. The most serious are as follows:

1. Standardized compound names are not consistently employed. There was a concerted effort to include synonyms, but without consistent adherence to a dictionary of standard compound names, alternative representation of the same compound inevitably arose.

2. The identification of functional roles in each diagram is usually achieved using EC numbers for enzymes. In cases where an EC number identifies only a class of roles and in cases in which no EC number has yet been assigned, an alternative mechanism was needed to remove ambiguities. Common enzyme names, genes that encode proteins corresponding to the role, and common functional designations for noncatalytic proteins all have been used. It is now felt that catalytic roles should be identified by EC numbers, and where more precision is required the descriptions employed in the Swiss Protein Data Bank [5] should be used. Similarly, any noncatalytic role included in a diagram should be identified by the description employed in the Swiss Protein Data Bank; the researchers constructing that database are making a laudable effort to standardize their descriptions, and we will attempt to conform to their decisions.

3. Users who wish to write software to manipulate metabolic pathways require a more structured encoding of the pathways. In many cases, they will need access only to the network of reactions described by the pathways (ignoring issues like locations, membranes, isosymes, and so forth). What is needed is a structured encoding of the rich representation informally used in the drawings, along with tools that produce extractions of subsets of the data in suitable formats.

Our efforts to address these issues, while constantly extending the set of drawings in the collection, are focused on parsing the drawings and producing structured documents in SGML. These structured documents are based on an underlying data model that captures the abstract notions of pathway, reaction, compound, compartment, functional role, membrane, compartment, prosthetic group, and so forth. The model is under development, and the goal will be to produce an extensible encoding

that will ultimately allow capturing all of the significant contents of the drawings as well as allowing an eventual imposition of significant additions (most notably regulatory mechanisms).

Such a structured encoding has as a primary purpose to capture the essential information from the drawings. The encoding combines the ability to reproduce the original drawing, render the pathway using a variety of output mechanisms, and precisely capture the information from the drawing. The point worth emphasizing is that capturing the contents in such a structured representation allows straightforward extractions into convenient representations of subsets of the complete representation. As a demonstration of these capabilities, we are releasing a tool that produces a simple, ascii relational representation of the pathways as a set of reactions, ignoring many of the complexities of the original drawings. The output of this tool is a set of tables that can be used in systems wishing to perform rudimentary computations on pathways. We provide the tables in both a simple ASCII format and encoded as Prolog facts.

4 Translation to More Convenient Subsets of the Data

When the encoding of all the drawings into SGML is complete, we will publish a precise description. At this point, we have committed to release an encoding of a significant subset of the drawings (including all of the pathways used in the reconstruction of *Mycoplasma genitalium,* as well as a significant number of other pathways in the collection) by early October 1996. We believe that most researchers building tools to utilize metabolic pathways would, at least initially, wish access to a relational representation which we also provide. The tables represented the extracted data can be summarized as follows (using the notation of Prolog):

Tables Representing Pathways, Reactions, Compounds, and Functional Roles

```
pathway(PathwayID,PathwayName).
reaction(ReactionID).
compound(CompoundID,Representation).
functional_role(RoldID,AdditionalDescription).
```

These relations correspond to the entities included in this simple model.

Tables Representing Relationships between the Entities

```
pw_substrate(PathwayID,CompoundID,Stoichiometry,Location,ReactionID).
pw_product(PathwayID,CompoundID,Stoichiometry,Location,ReactionID).
pw_to_reaction(PathwayID,ReactionID,Location).
reaction_to_substrate(ReactionID,CompoundID,Stoichiometry,Location,Connection).
reaction_to_product(ReactionID,CompoundID,Stoichiometry,Location,Connection).
reaction_to_role(ReactionID,FunctionalRole,Location).
```

The inclusion of Location is needed to meaningfully include a number of pathways in which transport occurs. The Connection argument in the reaction_to_substrate/5 and reaction_to_product/5

relations, as well as the ReactionID in the pw_substrate/5 and pw_product/5 relations, convey infomation that is useful when rendering the pathways (they establish the main inputs and outputs, as well as the significant couplings via intermediates).

5 Availability

As mentioned above, readers can access the Metabolic Pathway Collection from either 1] or through the WIT system [3]. We ask that those who find the collection useful in their work cite this article when giving credit.

Acknowledgments

This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Computational and Technology Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

References

[1] http://www.biobase.com/emphome.html/

[2] http://www.biobase.com/emphome.html/pags/pathways.html

[3] http://www.cme.msu.edu/WIT/

[4] http://www.kazusa.or.jp/cyano/cyano.html

[5] http://expasy.hcuge.ch/sprot/sprot-top.html