ARGONNE NATIONAL LABORATORY

9700 South Cass Avenue

Argonne, Illinois 60439

PUMA2 – An Environment for Comparative Analysis of Metabolic Subsystems and Automated Reconstruction of Metabolism of Microbial Consortia and Individual Organisms from Sequence Data*

by

Mark G. D'Souza, Jun Huan,† Samantha Sutton,‡ Margie Romine,§ and Natalia Maltsev

Mathematics and Computer Science Division

Technical Memorandum ANL/MCS-TM-240

December 1999

*This work was supported by the Office of Biological and Environmental Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

†Present affiliation: Wireless Solutions, Nortell Networks, 2201 Lakeside Blvd., Richardson, TX 75082

‡Present address: University of Illinois at Urbana-Champaign, Electrical Engineering Dept., Urbana, IL 61801

§Pacific Northwest National Laboratory, 902 Battelle Blvd., MSIN: P7-50, Richland, WA 99352

PUMA2—An Environment for Comparative Analysis of Metabolic Subsystems and Automated Reconstruction of Metabolism of Microbial Consortia and Individual Organisms from Sequence Data

by

Mark G. D'Souza, Jun Huan, Samantha Sutton, Margie Romine, and Natalia Maltsev

Abstract

We have created a prototype system called PUMA2—an interactive environment with the following goals:

- enable comparative analysis of the metabolic subsystems in different organisms,
- provide a framework for the automated reconstruction of the metabolism of microbial consortia and individual species, and
- provide a framework for the representation of expression data.

Analyses in PUMA2 are based on a collection of metabolic modules connected to sequence data. The results of such analyses are presented in graphical form based on hierarchical representation of the functional subsystems and annotated with sequence data and literature information. A set of new tools has also been developed for evaluation of the results and for efficient annotation.

Use of this environment will permit the following applications:

1. identification of signature genes for characterization of biological communities or single isolates,
2. development of a toolkit for visualization and annotation of user-defined pathways,
3. identification of unculturable organisms, and
4. prediction of properties of ecological niches based on microbial physiology.

Analysis of the genomes in PUMA2 will lead to formulation of conjectures about the functions of the hypothetical genes and modes of function of biological subsystems to be tested in the wet lab. We believe that using these conjectures can reduce dramatically the cost of the experimental work for the sequenced genomes and provide better understanding of the physiology of the organisms and microbial consortia.

PUMA2 could also be used for development of educational materials for courses in biochemistry and microbial physiology.

A working prototype of the PUMA2 system can be viewed at http://puma.mcs.anl.gov/PUMA2/.

1 Introduction

Recent years have witnessed a rapid accumulation of sequence data and data related to the physiology and biochemistry of organisms. Providing a functional context for this data is vital to understanding biological systems. Adequate representation of metabolism is one of the keys to the design of such a context.

Indeed, recent advances in biochemistry and elucidation of the metabolic pathways by which organisms produce energy and building blocks for biosynthesis led to the surprising conclusion that "virtually all biosynthetic pathways, polymerizations, and assembly processes are fundamentally similar in all bacteria ... [and] possess a high degree of similarity in all living cells" [1].

Arguably, exceptions to this conclusion exist: in some cases building blocks are synthesized by alternative routes; some modifications of the universal pathways are used by organisms to better adjust to current conditions and environments; and some organisms employ pathways and enzymes that are unique to them, usually reflecting specific conditions and restrictions imposed by an ecological niche. Moreover, catabolic processes show extraordinary metabolic diversity: almost any existing compound could be used as a source of energy or as a final electron acceptor. However, the turnover reactions—breakdown and resynthesis of macromolecules—show substantial similarities in various organisms.

2 PUMA2 and the Modular Approach

Motivated by these observations about metabolic similarity, we have developed an approach for representing metabolism as a network of interconnected modules. We define a "metabolic module" as a set of enzymes involved in particular physiological process (e.g., glycolysis, methanogenesis, nitrogen assimilation) with all known variations in different organisms, presented in the form of a metabolic diagram. Every module is connected to the other modules through common intermediates and components, thus constituting the architecture of a metabolic network.

The modular approach offers the advantage that it can be readily modified to accommodate the growing amount of metabolic information. Indeed, as was pointed out by Emmerling et al. [2], "to a large extent, molecular biology has been successful because it has dissected the complex cellular systems into smaller parts, which could be then analyzed and understood." Understanding of isolated cellular processes and subsystems is a prerequisite to comprehension of in vivo systems in which reactions and pathways are embedded in the whole metabolism.

We have developed a prototype system, called PUMA2 (http://puma.mcs.anl.gov/PUMA2), that has used the modular approach to analyze metabolic systems manually. The results have been extremely promising. We will further develop PUMA2 from a prototype to a production system. Metabolic modules will be thoroughly annotated with sequence information; relevant literature information; information on the "signature" enzymes or combination of the enzymes whose presence gives unambiguous evidence of occurrence of this module in a given organism or microbial community; and information on the alternative substrates used in the pathways constituting a module and resulting products (this data could aid identification of the potential microbial partners in

the microbial communities). The process will be done semi-automatically, dramatically reducing the time involved and the chance for error, while affording expert researchers an opportunity to revise the material based on their individual insight and intuition. New tools will be developed for analysis, representation, and annotation of the metabolic data. The major features of PUMA2 architecture are illustrated in Figure 1.

The organization of metabolic information in PUMA2 will provide the following features:

- a framework for comparative analysis of metabolic pathways across multiple genomes,
- automation of metabolic reconstruction from the sequence data for individual organisms and microbial consortia,
- a library of multiple sequence and reverse translated alignments for the signature enzymes to assist development of the primers for polymerase chain reaction (PCR), and
- a toolkit for entering and maintaining a set of user-defined pathways and its annotation with the literature and sequence information, allowing researchers to develop and maintain customized collections of metabolic pathways and annotations.

Comparison of PUMA2 with Other Metabolic Systems and Databases. Several metabolic information sources have been developed during the past decade. They contain a wealth of information about individual pathways (e.g., EMP, EcoSys, MetaCyc, SoyBase) or more generalized representation of the metabolic networks (e.g., KEGG, SwissProt). In Table 1 we list the most successful of these projects and briefly summarize the differences from PUMA2 project. As Table 1 indicates, no existing system provides the three key features needed for effective comparative analysis. PUMA2 thus fills an important gap left by the current metabolic systems and databases.

Successful development of the PUMA2 system will offer the following benefits:

- **Comparative analysis of metabolic systems.** Comparative analysis empowered by connection to genetic sequence information could have a substantial impact on diverse fields including medicine (assisting in finding metabolic pathways and enzymes that are unique to bacterial or eukaryotic pathogens and could be used as potential targets for antibiotics) and biotechnology and bioremediation (aiding in metabolic engineering, which requires deep understanding of metabolism and genetics of the organism to be used for modification).
- **Automation of the metabolic reconstructions from the sequence data.** Automation of the major steps of the metabolic reconstruction from the sequence data could aid high throughput analysis of bacterial genomes.
- **Characterization of unknown or unculturable organisms.** A library of signature enzymes could facilitate identification of the members of microbial consortia and individual organisms, and identification of pathogens in patients and environmental samples.

Current development effort in our work on the prototype PUMA2 focuses on the following four areas.

2.1 Metabolic Modules

Representation of metabolic and sequence data in the PUMA2 prototype is based on metabolic modules. As Figure 2 shows, each module is represented in two forms:

- A dynamically generated module diagram representing all similarities and differences that occur in different versions of a pathway
- A computer-readable encoding of a pathway, which could be used for comparative analysis of a pathway in different organisms

Currently, we have assembled (manually) preliminary versions of several modules (*de novo* pyrimidine biosynthesis, *de novo* purine biosynthesis, methanogenesis, glycolysis, etc.) (see http://puma.mcs.anl.gov/PUMA2/HTML/modules.html). Having demonstrated the feasibility of the approach, we now need to verify its accuracy, automate the process, and expand the number of modules significantly to reflect major physiological processes. A detailed explanation of the development of such a module for 'de novo' pyrimidine biosynthesis can be found at http://puma.mcs.anl.gov/PUMA2/HTML/pyrimidine_demo.html. All such examples of the modules represent only an initial stage of development and will be thoroughly updated and annotated.

2.2 A Tool for Drawing User-defined Metabolic Pathways and Summary Diagrams

Publicly available databases (EMP, KEGG, EcoCyc, etc.) contain a wealth of information about the metabolic pathways that exist in different organisms. However, in many cases, pathways are missing from these collections. This situation complicates representation of the metabolic data and development of the metabolic reconstructions. We intend to develop a convenient framework that will allow a user to easily generate standardized pathway diagrams and annotate them with biochemical, sequence, regulatory, and other types of relevant information.

As a first step in this direction the prototype contains a Web-based tool for visualization of the metabolic pathways (http://puma.mcs.anl.gov/PUMA2/HTML/tool.html). The user is provided with a form for entering data to be displayed. Entering the data requires following simple rules, described in a help file. The tool allows one to enter information in the form of EC numbers or text. The output of this tool is a dynamically generated pathway diagram (see Figure 3). Both the input form with the input data and the output diagram can be saved in a file, which will be useful for developing customized collections of metabolic data. The prototype tool enables the user to automatically link objects on a diagram to any user-defined URL. This feature allows extensive annotation of the pathway data with the sequence data, literature information, and so forth. Furthermore, the tool can be used for development of the overview diagrams representing a

generalized view of the parts of the metabolic networks. These summary diagrams can be linked to more detailed representations of the pathways. Such cascading representation of the metabolic networks will simplify navigation and representation of complex systems.

2.3 Automation of the Metabolic Reconstructions from the Sequence Data

We have explored several of the issues involved in automated reconstruction of metabolic modules in given organisms (see http://puma.mcs.anl.gov/PUMA2/HTML/met_recon.html). Currently, we present reconstruction by dynamically generated pathway diagrams connected to the sequence data for a particular organism. Different color codes are used to mark the enzymes that are present or absent in a genome under consideration, as well as to emphasize the presence or absence of the signature enzymes (see Figure 3).

We intend to investigate two different modes of automated metabolic reconstruction in PUMA2, according to the source of sequence data:

- *Automated metabolic reconstruction with the WIT2 system as a source of sequence data*: WIT2 currently contains data for 38 complete and almost complete genomes. WIT allows *interactive* genetic sequence analysis and *customized* gene annotations. It contains data on the conserved chromosomal gene clusters, For the diagrams generated from WIT2 data, different shapes are used to signify the enzymes that are a part of a conserved chromosomal gene cluster.
- *SwissProt database as a source of sequence data*: PUMA2 allows use of data from 189 prokaryotic and eukaryotic organisms from SwissProt, for which the number of entries is more than 100. The SwissProt database provides links to a number of sequence analysis tools and databases. However, metabolic reconstruction in PUMA2 using SwissProt data does not allow interactive genetic sequence analysis.

2.4 Automatic Evaluation of the Metabolic Model

Besides metabolic diagrams, the prototype PUMA2 system produces an evaluation of the probability of occurrence of certain alternative versions of a pathway in a particular organism. We have developed an algorithm for such analysis. All of the alternatives in the pathway are assigned a numerical score of the probability of its occurrence in an organism, and evidence based on the results of genetic sequence analysis for every possible version is presented. Table 2 represents an example of such an evaluation.

3 Scientific Applications

Once we have designed and developed the PUMA2 system, we plan to use it for several important scientific applications.

3.1 Reconstruction of Metabolism of Microbial Consortia

Prokaryotes may form associations of various complexity either with other prokaryotes or with eukaryotic cells and tissues. These "heterologous associations" include symbiotic and parasitic systems, commensalism, and neutralism. In most cases a high degree of metabolic interdependence exists between the members of such consortia. Associations could have nutritional advantages (fixation of nitrogen, supply of basic nutrients and accessory factors) or be parasitic (*Chlamydia* species in their "reticulate body" state are examples of intracellular parasitic bacteria).

In some cases the major metabolic relationships in such microbial associations are understood from biochemical experiments. Figure 4 represents an example of a carbon cycle. Our intent is to enable researchers (based on the library of metabolic modules) to predict a potential gene pool from a diagram such as the one depicted in Figure 4 and to assist development of a set of potential PCR primers for further detailing the metabolism of this complex system. Analysis of the sequence information in PUMA2 could also help estimate the impact of a particular organism in metabolism of a consortium (Figure 5).

In other cases, however, significant evolutionary lineages have been identified through molecular-phylogenetic methods rather than through traditional wet lab experiment. As N. Pace notes [14], "Because the molecular-phylogenetic identifications are based on sequence, as opposed to metabolic properties, microbes can be identified without being cultivated. Analysis of microbial ecosystems in this way is more than a taxonomic exercise because the sequences provide experimental tools—for instance, molecular hybridization probes—that can be used to identify, monitor, and study the microbial inhabitants of natural ecosystems." Our general approach to the analysis of metabolism of microbial consortia is represented in Figure 6.

3.2 Comparative Analysis of Organisms

The utility and power of comparative analysis in biological sciences are well understood and widely accepted. Insights that are gained from detailed comparisons of metabolic routes used by the organisms at varying lifestyles and phylogenetic distances will offer a framework for characterizing many mechanisms that still remain poorly understood (thermostability, biochemistry of high saline environments, etc.). The PUMA2 environment will allow comparative analysis of metabolic networks in different organisms. Each module will emphasize similarities and differences in executing particular functions that exist in different organisms and systems. This way of representing of metabolic process, we believe, allows a systematic analysis of the combinatorics of metabolic networks: connected to data on regulation, chromosomal gene clustering [15], and sequence data, this information will provide valuable insight into evolution of metabolic functions and genomes.

Representation of metabolic data both in human-readable and computer-readable forms will aid further detailed studies of diversity and conservation of function in biological systems. For example, it is known that bacteria synthesize unsaturated fatty acids by an anaerobic mechanism as opposed to an aerobic in mammals. The double bond is introduced into the growing acyl chain by beta-hydroxydecanoyl-ACP dehydrase—unique to bacteria—which could be viewed as a potential target for antibacterial agents [16]. Comparative analysis of the processes like DNA replication

protein secretion, biosynthesis of sterols, and biosynthesis of cellular walls also reveals potential drug targets. Elucidation of the new metabolic pathways and enzymes that are unique to bacterial or eukaryotic pathogens could provide invaluable insight into potential targets for antibiotics.

The *Riftia* symbiont and a number of other sulfur-oxidizing symbionts associated with invertebrate animals represent another example of the utility of comparative studies. These organisms proved to be fairly closely related to one another and also to the intensively studied organisms *E. coli* and *Pseudomonas aeruginosa*. Because of their phylogenetic proximity, many of the properties of the symbionts can be inferred from those of the well-studied organisms. For instance, one can predict with good confidence the nature of the ribosome and antibiotic-susceptibility patterns, the nature of the DNA-replicative machinery, the character of the RNA polymerase complex, the character of biosynthetic pathways and their regulatory mechanisms, the nature of the cell envelope and energy transduction schemes, and many other cellular properties of the symbionts [14].

3.3 Prediction of Properties of Ecological Nichess Based on Microbial Physiology

Very little is known about the ecology of many organisms. We believe that information about the functions of the genes in the genome and further metabolic reconstruction can suggest an organism's natural habitat. Understanding of catabolic and respiratory pathways of the microbial community inhabiting a particular ecological niche can provide important insights in mineral composition and occurrence of certain types of organic compounds in the environment. For example, the presence of dissimilatory nitrate reductase in an organism's genome suggests the presence of nitrates in the surrounding media. Similarly, the presence of the enzymes for catabolism of aromatic compounds points to contamination of the environment with these class of chemical substances.

3.4 Formulation of Conjectures about the Functions of Hypothetical Genes

Sequence data provides valuable resources for understanding cellular metabolism. Analysis of three well-studied biochemical pathways (the tricarboxylic acid cycle, pentose phosphate pathway, and glycolysis) from the 17 publicly available microbial genomes [17] has shown that these pathways may rarely occur as previously defined. Therefore, following whole-genome sequencing it has become necessary to redefine the "classical" biochemical steps leading from substrate to end-product for each pathway. Often, unique or alternative reactions appear to be required in order to maintain pathway functionality where expected enzyme reactions (as defined by the presence or absence of the corresponding genes) are "missing".

Various explanations have been suggested:

- the presence of low sequence similarity or novel genes that encode enzymes performing the same or similar functions,
- the presence of multifunctional enzymes,
- incorrectly assigned gene functions in genome databases, and

- known enzyme functions that have yet to be correlated with a gene sequence.

Modifications at the gene and/or functional levels, as well as the possible use of alternative enzymes, should be taken into consideration when reconstructing biochemical pathways for fully sequenced microbial genomes.

Analysis of the genetic sequence in PUMA2 performed within a framework of metabolic reconstruction could provide an important tool for discovering modifications in "classical" metabolic pathways and will lead to formulation of conjectures about the functions of the hypothetical genes and variations in organization of biological subsystems. We intend to produce a set of such conjectures, based on the results of our analysis, to be tested in the wet lab. We believe that using these conjectures can reduce dramatically the cost of the experimental work for the sequenced genomes and provide better understanding of physiology of the organisms and microbial associations.

3.5 Insight into Environmental Adaptation

Some organisms have more than one pathway to perform the same metabolic function. This property represents one of the major mechanisms of adaptation to changing environments. Visualization of expression data using metabolic modules in PUMA2 will allow display of the rates of functioning of the alternative metabolic routes in the same organisms under various experimental conditions.

Furthermore, comparative analysis of the metabolic pathways and enzymes employed by the organisms that reside in extreme environments (hypothermophiles, psychrophiles, halophiles, etc.) could provide important data for understanding the physiology of these extraordinary life forms.

3.6 Representation of Expression Data

The tools for automated metabolic reconstruction described above can also be modified for analysis and presentation of expression data. In collaboration with Dr. C. Giometti (Argonne), we have started to develop such tools, which exploit different coloring schemes to represent different levels of expression of the genes in the framework of metabolic reconstruction. Annotation of the expression data with information on potential gene clusters and usage of PUMA2 tools for comparative analysis of the metabolic pathways will also aid analysis of proteomics data.

Additionally, we note that some organisms have more than one pathway to perform the same metabolic function. This property represents one of the major mechanisms of adaptation to changing environments. Visualization of expression data using metabolic modules will allow display of the rates of functioning of the alternative metabolic routes in the same organisms under various experimental conditions.

We believe that PUMA2 will provide a useful environment for the studies of metabolism and physiology of the organisms and will help to bridge a gap between bioinformatics and functional genomics.

References

[1] Neidhardt, F., Ingraham, J., and Schaecheter, M., eds. *Physiology of the Bacterial Cell.* Sinauer Associates, Inc., Sunderland, MA, 1995.

[2] Emmerling, M., Bailey, J., and Sauer, T. Glucose Catabolism of *E. coli* Straints with Increased Activity. *Metab. Engineering* **1**, pp. 117-127, 1999.

[3] Bairoch, A., and Apweiler, R. The SWISS-PROT Protein Sequence Data Bank and Its Supplement TrEMBL in 1999. *Nucleic Acids Res.* **27**, pp. 49–54, 1999.

[4] Boehringer Mannheim Corp. Biochemical Pathways. Boehringer Mannheim GmbH - Biochemica, 1993.

[5] SWISS-PROT Protein Sequence Data Bank: http://www.expasy.ch.sprot.

[6] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**(1), pp. 29–34, 1999.

[7] KEGG: http://www.genome.ad.jp/kegg/kegg3.html.

[8] Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A., and Krummenacker, M. EcoCyc: Electronic Encyclopedia of *E. coli* Genes and Metabolism. *Nucleic Acids Res.* **27**(1), p. 55, 1999.

[9] EcoCyc: http://ecocyc.PangeaSystems.com/ecocyc/ecocyc.html.

[10] MetaCyc: http://ecocyc.pangeasystems.com:1555/server.html.

[11] Biocatalysis Database: http://dragon.labmed.umn.edu/ lynda/index.html.

[12] SoyBase: http://129.186.26.94/soybase/aboutsoybase.html.

[13] Maize Database: http://www.agron.missouri.edu/.

[14] Pace, N. R. A Molecular View of Microbial Diversity and the Biosphere. *Science* 276(5313), pp. 734–40, 1997.

[15] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G., and Maltsev, N. The Use of Gene Clusters to Infer Functional Coupling. *Proceedings of the National Academy of Science U.S.A.* **96**(6), p. 2896, 1999.

[16] Sutcliffe, J. A., ed. *Emerging Targets in Antibacterial and Antifungal Chemotherapy.* Routledge, Chapman and Hall, Inc., 1992.

[17] Minker, J. Informative and Cooperative Answers in Databases Using Integrity Constraints. In V. Dahl, et al., eds., *Natural Language and Understanding and Logic Programming*, North Holland, pp. 277–300, 1988.