3 WIT/WIT2: METABOLIC RECONSTRUCTION SYSTEMS

Ross Overbeek, Niels Larsen, Natalia Maltsev, Gordon D. Pusch, and Evgeni Selkov

Argonne National Laboratory, Argonne, IL 60439

Introduction: What Is Metabolic Reconstruction?

For the past few years, we have been developing metabolic reconstructions for organisms that have been sequenced, and we have made a number of these working models available. By the term *metabolic reconstruction* we mean the process of inferring the metabolism of an organism from its genetic sequence data supplemented by known biochemical and phenotypic data. Our initial software system to support metabolic reconstruction was called WIT (for "What Is There?") and has been in use since mid-1995 (http://www.cme.msu.edu/ WIT/) [7]. Recently, a second system, which we have called WIT2, has been made available (http://www/mcs.anl.gov/home/overbeek/WIT2/CGI/ user.cgi). In this chapter we discuss the central design issues in constructing such systems, along with the basic steps that must be supported by any such system.

Representation of Metabolism

The most basic decisions center on how to represent the metabolism of an organism. Clearly, a topic of such complexity might well warrant an extremely abstruse computational representation. Indeed, the efforts that have been spent in representing chemical compounds give some indication of the potential magnitude of the problem.

In considering this problem, we have found it useful to draw an analogy to the representation of an automobile as it appears in any auto parts store. In this context, the auto overview and parts catalog give an accurate, high-level abstraction that does not include any real discussion of the "intermediates". It is an effective representation, but it does not convey the details of how energy is generated and

distributed, how control mechanisms function, or how dynamic behavior is constrained.

We have developed an approach for representing the metabolism of an organism that is based on similar simplifications:

- We begin with a set of metabolic pathway diagrams. For our purposes, these diagrams are an arbitrarily structured and complex representation of a functional subsystem. Hence, we call them *function diagrams*. Just as an abstract drawing in an auto parts catalog attempts to convey the essential relationship of a set of functionally related parts, the function diagrams that we use attempt to convey the functional grouping of a set of proteins (normally, the set of enzymes that catalyze the reactions depicted in a metabolic pathway).

- Function diagrams themselves can (and should be) a well-structured representation of the functional interactions of proteins. This will be critical to support systems that base computations on the details of the interactions. For the purpose of metabolic reconstruction, however, none of this is necessary. A minimal function diagram composed simply of a list of protein identifiers would work just as well (i.e., we could use a set of minimal function diagrams in which each diagram was nothing but a list of enzymes along with any additional noncatalytic proteins).

- The central issue in this highly simplified framework now becomes how to assign unique identifiers to the *functional roles* in the diagrams. For function diagrams describing metabolic pathways, the enzyme number is usually adequate. However, some enzyme numbers are imprecise (i.e., they describe a class of enzymes), and there is the issue of what identifier to use for noncatalytic functional roles. As a (not completely adequate) solution, we use a slightly distilled version of the Swiss Protein Data Bank descriptions. (The people maintaining the Swiss Protein Data Bank have been making heroic efforts to standardize descriptions of protein functional roles, and whenever possible we simply exploit their efforts.)

The initial set of function diagrams that we now include in WIT/WIT2 comes from the *Metabolic PathwayDdatabase* built by Evgeni Selkov [8]. It now contains well over 2500 pathways and variants of pathways. These have been supplemented by a much smaller set of additional function diagrams from other sources.

Another way to summarize our representation of functional groupings is to say that we begin with two relational tables: (1) the *diagram-role table*, which contains two columns: the diagram identifier and the functional role identifier; and (2) the *protein-role table*, which also contains two columns: the protein sequence identifier and the functional role identifier .

Swiss Protein Data Bank entries are one class of protein sequences, and for them the "protein sequence identifier" is just the accession number. When other classes of

protein sequence are used (e.g., ORFs from a newly sequenced genome), appropriate identifiers are used.

A metabolic reconstruction for a genome amounts to the entries in the protein--role table corresponding to ORFs from the genome, and a third table, the *asserted-diagrams table*, which is a list of the diagrams that have been asserted for the genome.

We stress that our approach of using arbitrary function diagrams and treating them as no more than collections of functional roles is a critical simplification. Such a simplification makes it possible to proceed with our goal of creating metabolic reconstructions without facing the detailed issues required to make inferences about the metabolic network. At the same time, if the actual function diagrams are a well-structured representation of the functions, such inferences will become commonplace (and useful in supporting the derivation and analysis of metabolic reconstructions).

How Is Metabolic Reconstruction Done?

Once the ORFs for a newly sequenced genome have been determined [1,2], we must carry out four steps: (1) assign functional roles, (2) assert the functional diagrams, (3) determine missing functions, and (4) balance the model.

Initial Assignments of Function

Our first step is to make initial assignments of their functional roles. This is done in two substeps: first, assignments are automatically generated for cases in which there appears to be relatively little ambiguity, and second, a manual pass through the ORFs with strong similarities but no assigned function is made.

Techniques for automatically assigning functional roles are advancing rapidly. We currently use the following approach for a translated ORF *x* from genome *g1*:

1. Compute similarities between the ORF and all sequences in the nonredundant protein sequence database. Save those above some designated threshold.
2. Consider similarities against ORFs from a completely sequenced genome *g2*. If x is similar to *y* from *g2*, and *y* is the protein in *g2* closest to *x*, we say that *y* is a *best hit* (BH) against *x*. If *x* is also the best hit in - *g1* against *y*, then we say that *y* is a *bidirectional best hit* (BBH).
3. Collect the set of BBHs for *x*. If the functional roles already assigned to those BBHs are all identical, assign the same functional role to *x*.

This is a quite conservative approach, although it can still lead to errors. Following the automated assignment of function, we recommend that the user of WIT/WIT2 make a pass through the set of ORFs that have strong similarities to other proteins with known functional roles but for which no automated assignment could be made. WIT2 allows the user to peruse the BBHs for each protein, to align the protein against other proteins of known function, to analyze regions of similarity, and so forth. At this point, assignment of function is still a process of thoughtfully

considering a wide range of alternatives, and the background of the user determines the quality of the assignments. We believe that the rapid addition of new genomes and the accumulation of a growing body of probable assignments of function, together with consistency checks based on clustering protein sequences, will lead to a situation in which most of the currently required judgment can be eliminated. However, we are not yet close to that point.

An Initial Set of Pathways

Once the initial assignment of functional roles has been completed (i.e., once the initial version of the entries in the protein-role table for the newly sequenced genome has been generated), one normally proceeds to the assertion of function diagrams (i.e., to the addition of entries to the asserted-diagrams table for the genome). As the collection of analyzed genomes increases, it becomes ever more likely that each new genome will contain a substantial similarity to a genome that has already been analyzed. If a fairly similar (biochemically and phenotypically) organism has already been analyzed, it is useful to begin the analysis of the new organism by asserting the diagrams that are believed to exist from the already analyzed organism. Some of the asserted pathways are likely to be wrong, but their removal can be deferred until after the initial assignment of pathways.

In any event, the user should move through the major areas of metabolism and ask the system to propose diagrams that might correspond to functionality present in the organism. A system supporting metabolic reconstruction should be able to support such requests. As we learn more about the reasoning required to accurately assert the presence of pathways, the proposal of pathways by the system can become increasingly precise. For now, we employ a very straightforward approach.

First, we take the entire collection of pathways and assign a score to each pathway. The score for a pathway is

$$(I + 0.5U) / (I + U + M),$$

where I is the number of functional roles in the diagram that have been connected to specific sequences in the genome, M is the number that have not been connected and for which known examples from other genomes exist, and U is the number of unconnected roles for which no exemplar exists from other genomes. This is a crude measure of the fraction of the functional roles that have been identified, considering that there are U roles for which reasoning by homology is impossible at this point.

Then, we sort the pathways by score and present to the user those that exceed some specified threshold. The user is expected to go through each proposed pathway and either assert it to the asserted-diagrams table or simply ignore the proposal.

Locating Missing Functions

After we have accumulated an initial set of asserted diagrams, a pass through this asserted set must be made, focusing on the functional roles that remain unconnected to specific ORFs in the genome. Here, the system can provide a very useful function by collecting all known sequences that have been assigned the functional role, tabulating all similarities between ORFs in the new genome and these existing exemplars, and summarizing which of the existing ORFs is most likely to perform the designated functional role. Without a tool like WIT/WIT2, this process would be extremely time-consuming (and, in fact, would almost never be done systematically). In WIT2, we made the design decision to precompute similarities between all ORFs from the analyzed genomes and between these ORFs and entries in the nonredundant protein database maintained by NCBI. This allows an immediate response to requests to locate candidates for unconnected functional roles, summarizing BHs, BBHs, and all other similarities. The disadvantage of such a design commitment is that the collection of similarities is out of date almost immediately. Such a trade-off is commonly faced in developing bioinformatics servers. In our case, the severity of the problem is inevitably reduced by the addition of more genomes – that is, while the system may well not have access to all relevant similarities, the chances of establishing a solid connection between a new sequence and a previously analyzed sequence with an established function improve dramatically as the set of completely sequenced (and increasingly analyzed) genomes grow.

Once the system has located candidates for an unconnected functional role, the process of actually coming to a conclusion about whether a given sequence should be connected to the functional role is arbitrarily complex and corresponds to the types of decisions made while doing the initial assignments. In this case, however, the user of the system has the additional knowledge that assignments based on weak similarities may be strongly supported by the presence of assignments to other functional roles from the same diagram. This represents one of the pragmatic motivations for developing metabolic reconstructions: they offer a means of developing strong support for assignments based on relatively weak similarities.

We emphasize that the assertion of specific diagrams (i.e., pathways) should be considered in the context of known biochemical and phenotypic data. A variety of assignments cannot be made solely based on sequence similarities. For example, one might consider the choice between malate dehydrogenase and lactate dehydrogenase. Although examples of sequences that play these roles are extremely similar (exhibiting almost arbitrarily strong similarity scores), the choice between these functional roles often can be made only by using biochemical evidence or a more detailed sequence analysis based on either the construction of trees or the analysis of "signatures" (i.e., positions in the sequence that correlate with the functional role). Similarly, the choice between assigning a functional role of aspartate oxidase, fumarate reductase, or succinate dehydrogenase will require establishing an overview of the lifestyle of the organism, followed by a detailed analysis of all related sequences present in the genome. These examples are unusually difficult; in most cases the determination of function is much more straightforward. Even in these cases, however, the accumulation of more data will dramatically simplify things.

Balancing the Model

We turn now to the more difficult and critical step of balancing the model. By *balancing*, we mean considering questions of the following form:

> "Since we know this compound is present (because we have asserted a given pathway for which it is a substrate), where does it come from? Is it synthesized, or is it imported?"

This consideration holds for all substrates to pathways, coenzymes, prosthetic groups, and so forth. In addition, we need to consider the issue of whether products of pathways are consumed by other cellular processes or are excreted.

To begin this process, the user must first make tables including all substrates of asserted pathways and all products of asserted pathways. As we stated above, our simplified notion of function diagram does not require that substrates and products be included. However, if one wishes to automate this aspect of metabolic reconstruction (which we have not yet done), the data must be accurately encoded. Once such tables exist, we can remove all compounds that occur as both substrates and products. Two lists remain:

1. A list of substrates that are not synthesized by any process depicted in any of the asserted function diagrams, and
2. A list of products that are not consumed by an processes depicted by asserted diagrams.

The user must go through these lists carefully and assess how best to reconcile the situation. This task may require searching for a protein that might be a potential transporter, asserting a new pathway for which a limited amount of evidence exists, or formulating some other hypothesis about what is going on.

Once the user has analyzed the situation as it relates to substrates and products of pathways, a similar analysis must be applied to known cofactors, coenzymes, and prosthetic groups. In this case, the logical issue of potential producers and consumers of specific compounds must be analyzed, but additional issues relating to volumes of flows can be analyzed. At this point, most of this type of analysis requires a substantial amount of expertise, and many of the decisions are necessarily impossible to make with any certainty. The situation is exacerbated by the difficulty of determining the precise function of a wide class of transport proteins, as well as by the potential for broad specificity for many enzymes. In this regard, while the situation is currently tractable only for those with substantial biochemical backgrounds (and not always by them), it is clearly possible that rapid advances in our ability to perform more careful comparative analysis and to acquire biochemical confirmation of conjectures will gradually simplify this aspect of metabolic reconstruction, as well.

Coordinating the Development of Metabolic Reconstructions

A metabolic reconstruction can be done by a number of individuals, often sharing a single model that is developed jointly. WIT2 includes the capability for multiple users either to work jointly on a single metabolic reconstruction or to develop such reconstructions in isolation. This is achieved as follows:

- For each organism, a list of *master users* is installed. When these users alter a model, the change is visible by all users of the system.
- When a user logs into a version of WIT2, he chooses a "user ID". Any set of users sharing the same ID will be working on the same model.
- When any non-master user alters a model (asserts the existence of a diagram or makes an entry to the protein-role table), the change is visible only to the group of users sharing the same user ID. The model constructed within a given user ID should be viewed as an extension to the "standard" model generated by the master users.
- A metabolic reconstruction for an organism (corresponding to a designated user ID) can be exported (i.e., converted to an external format), which can later be imported to any other version of WIT2 that includes the data for the organism.

Our intent is that users develop metabolic reconstructions on many distinct Web servers, but that they be able to conveniently import the efforts of others working on the same genome.

Where Do We Stand?

At this point we are attempting to develop and maintain metabolic models for well over twenty organisms representing a remarkable amount of phylogenetic diversity (http://wit.at.msu). The development of these initial models will be, we believe, far more difficult than the efforts required to add new models for more organisms that are similar to these initially analyzed organisms. On the other hand, unicellular life exhibits an enormous amount of diversity; and when the task of analyzing multicellular organisms is contemplated, it is clear that an enormous amount of work is required to attain even approximate metabolic reconstructions.

As we develop these initial models, we have noted a clear core of functionality that is shared by a surprisingly varied set of organisms. Techniques for developing clusters of proteins that are clearly homologous and that perform identical functions in distinct organisms are now beginning to simplify efforts to develop metabolic reconstructions. Such techniques are also leading to a clear hypothesis about the historical origins of specific functions.

The task of constructing a detailed overview of the functional subsystems in specific organisms is closely related to the issue of characterizing the functions or genes in the gene pool. While specific organisms often have been analyzed in isolation, it is rapidly becoming clear that comparative analysis is the key to

understanding even specific genomes and that characterization of the complete gene pool for unicellular life is far more tractable than previously imagined. Our goal is to develop accurate, although somewhat imprecise, functional overviews for unicellular organisms and to use these as a foundation for the analysis of multicellular eukaryotes. Just as protein families derived from unicellular organisms are beginning to form the basis for assigning function to many eukaryotic proteins, an understanding of the central metabolism of eukaryotes will be built on our rapidly expanding understanding of the evolution of functional systems within unicellular organisms.

A Growing Interest in Connecting Metabolic and Sequence Data

The growing perception that the metabolic structure must be encoded and used to interpret the emerging body of sequence data has resulted in a number of projects. Here we summarize the most successful of these projects at this time. With interest expanding so rapidly, the reader is encouraged to do a network search for other sites, which we believe will continue to appear at a growing rate.

- KEGG (http://www.genome.ad.jp/kegg/kegg3.html) [4]: This outstanding effort, based at Kyoto University in Japan, represents an attempt to maintain metabolic overviews for sequenced genomes. It has connected the genes from specific organisms to metabolic functions with excellent visual depictions of metabolic maps.

- Boehringer Manheim Biochemical Pathways (http://expasy.hcuge.ch/cgi-bin/search-biochem-index): This excellent collection of metabolic pathways has been recently integrated into the SwissProt effort, allowing one to move between pathways, enzymes, and sequence data.

- EcoCyc (http://www.ai.sri.com/ecocyc/ecocyc.html - Overview) [5]: This database is a detailed encoding of the metabolism of Escherichia coli and Haemophilus influenzae. Besides just the metabolic network, this collection includes some of the kinetic and thermodynamic parameters (when they are known).

- Biocatalysis/Biodegradation Database (http://dragon.labmed.umn.edu/~lynda/index.html) [3]: This database covers a small, but significant, set of pathways that are of special interest in the area of xenobiotic degradation.

- SoyBase (http://probe.nal.usda.gov:8000/plant/aboutsoybase.html): This databases captures genetic and metabolic data for soybeans.

- Maize DB (http://teosinte.agron.missouri.edu/) [6]: This database is a comprehensive collection of maize genetic and biochemical data.

Availability of the Pathways, Software, and Models

The PUMA (http://www.mcs.anl.gov/home/compbio/PUMA/Production/puma.html), WIT (http://www.cme.msu.edu/WIT/) [7], and WIT2 (http://www.mcs.anl.gov/home/overbeek/WIT2/CGI/user.cgi) systems were developed at Argonne National Laboratory in close cooperation with the team of Evgeni Selkov in Russia. The beta release for WIT2 has been sent to four sites and is currently available. The first actual release of WIT2 is scheduled for October 1997. It will include all of the software required to install WIT2 and develop a local Web server, all of our metabolic reconstructions for organisms with genomes in the publicly available archives, and detailed instructions for adding any new genomes to the existing system (perhaps, for local use only). Just as widespread availability of the Metabolic Pathway Database has stimulated a number of projects relating to the analysis of metabolic networks, we hope that the availability of WIT2 will foster the development and open exchange of detailed metabolic reconstructions.

Acknowledgments

R.O. was supported by the U.S. Department of Energy, under Contract W-31-109-Eng-38. N.L. was supported by the Center for Microbial Ecology at Michigan State University (DEB 9120006). We also thank the Free Software Foundation and Larry Wall for their excellent software.

References

1. Badger, J. H., and Olsen, G. J. CRITICA: Coding Region Identification Tool Invoking Comparative Analysis, Molec. Bil. Evol., 1977, in press.
2. Borodovsky M., and Peresetsky, A. Deriving Non-Homogeneous DNA Markov Chain Models by Cluster Analysis Algorithm Minimizing Multiple Alignment Entropy. Comput Chemistry, 18, no. 3, 1994, pp. 259-267.
3. Ellis, L.B.M., and Wackett, L. P. A Microbial Biocatalysis Database, Soc. Ind. Microb. News. 45, no. 4, 1995, pp. 167-173.
4. Kanehisa, M., Toward Pathway Engineering: A New Database of Genetic and Molecular Pathways, Science and Technology Japan, 59, 1996, pp. 34-38.
5. Karp, P, Riley, M., Paley, S., and Pellegrini-Toole, A. EcoCyc: Electronic Encyclopedia of E. coli Genes and Metabolism, Nucleic Acids Research, 25, no. 1, 1997
6. Nelson, O., Coe, E., and Langdale, J., Genetic Nomenclature Guide. Maize. Trends Genet., March, 1995, 20-21
7. Overbeek O., Larsen, N., Smith, W., Maltsev, N., and Selkov, E.. Representation of Function: The Next Step. Gene-COMBIS (on-line): 31 January 1997; Gene 191, no. 1,: GC1-9
8. Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., .Panushkina, E., Pronevitch, I., Selkov Jr., E.., and Yunus, I. The Methabolic Pathway Collection from EMP: The Enzymes and Metabolic Pathways Database. Nucleic Acids Research, 24, no. 1 (database issue), 1996, pp. 26-29.