ORGANIZING AND COMPUTING METABOLIC PATHWAY DATA IN TERMS OF BINARY RELATIONS

S. GOTO, H. BONO, H. OGATA, W. FUJIBUCHI, T. NISHIOKA,[a] K. SATO,[b] M. KANEHISA

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

A new database system named KEGG is being organized to computerize functional aspects of genes and genomes in terms of the binary relations of interacting molecules or genes. We are currently working on the metabolic pathway database that is composed of three interconnected sections: genes, molecules, and pathways, which are also linked to a number of existing databases through our DBGET retrieval system. Here we present the basic concept of binary relations and hierarchical classifications to represent the metabolic pathway data. The database operations are then defined as an extension of the relational operations, and the path computation problem is considered as a deduction from binary relations. An example of using KEGG for the functional prediction of genomic sequences is presented.

1 Introduction

The first complete genome of an organism, $\phi x174$, was determined in 1977,[1] which was followed by the explosion of DNA sequence data for specific genes, as well as for small genomes of viruses and organelles. The first complete genome of a free-living organism, *Haemophilus influenzae*, was determined in 1995,[2] which would certainly be followed by the explosion of complete genomic sequences and complete gene catalogs of a number of organisms from bacteria to eukaryotes. Whereas the experimental technologies have been refined to systematically analyze a whole genome, the computational methods for deciphering functional implications are still based on the analysis of each gene or gene product at a time. A systematic computational analysis is required for functional prediction of a whole genome.

At present, the functional data is computerized as an attribute of each sequence, for example, in the features tables of the sequence databases and in the motif libraries such as PROSITE.[3] Thus, the basic idea for functional prediction is to search similarities of each sequence in the sequence databases or in the motif libraries and then to extend sequence similarities to functional

[a]Present address: Graduate school of Agriculture, Kyoto University, Sakyo-ku, Kyoto 606-01, Japan

[b]Permanent address: Cray Research Japan LTD., 12-25 Hiroshiba-cho, Suita-shi, Osaka 564, Japan

similarities. The problem here is the lack of a suitable measure for the correctness of functional assignment, especially for the similarities in the so-called twilight zone. The sequence-function relationships of single molecules represent how individual components of a biological system work, and they do not contain higher level information of how components are connected to form a functional unit, such as a metabolic pathway, a signal transduction and an operon. As long as the assignment is made for each component separately, it will be difficult to check if such a functional unit is properly formed.

We have initiated a project named KEGG,[c] Kyoto Encyclopedia of Genes and Genomes, to computerize current knowledge of the information pathways of genes and gene products, which may be considered as wiring diagrams of biological systems. KEGG consists of three types of data, pathways, genes, and molecules, that are linked with each other and with the existing databases through our DBGET [d] integrated database system.[4] Currently, we focus our attention to the metabolic pathways and enzyme genes. The database project LIGAND [5] for enzyme reactions, enzymes, and metabolic compounds is also tightly coupled with the KEGG project. In this paper, we present the basic concept of the binary relation between two interacting molecules or genes as a fundamental element to represent the pathway, the manipulation of binary relations based on relational operations, and the path computation as a deduction from binary relations. We also briefly mention how the pathway data can be utilized to make a functional assignment from a gene catalog.

2 Data Representation

2.1 Basic Data Item

The basic data item in KEGG is a gene, a gene product, a metabolic compound, or any other molecule in a cell, which may be identified in the form:

database:entry

or

organism:gene

where 'database' is the database name such as genbank and swissprot, 'entry' is the entry name or the accession number, 'organism' is the organism name, and 'gene' is the gene name. This is the uniform identifier both in KEGG and DBGET. For example,

EC:6.3.2.3

and

[c] http://www.genome.ad.jp/kegg/kegg.html
[d] http://www.genome.ad.jp/dbget/dbget.links.html

Table 1: Examples of binary relations

Data item 1	Data item 2	Type of relation
genbank:DROALPC	medline:93273796	factual data and reference
embl:DMALPC	sp:CTNA_DROME	translation from nucleotides to amino acids
genbank:DROALPC	embl:DMALPC	same accession
cpd:C00118	cpd:C00111	substrate and product
EC:4.2.1.1	EC:5.3.1.1	two consecutive enzymes in the metabolic pathway
D.melanogaster:dpp	D.melanogaster:sax	two interacting genes (ligand and receptor)
E.coli:tpiA	sp:TRIS_ECOLI	gene product as stored in public database
E.coli:tpiA	EC:5.3.1.1	gene and function
B.subtilis:tpi	EC:5.3.1.1	gene and function
E.coli:tpiA	B.subtilis:tpi	orthologous genes

cpd:C00051

represent glutathione synthase and glutathione, respectively, in the LIGAND database, and

E.coli:tpiA

represents triosephosphate isomerase gene in *E. coli*.

2.2 Binary Relations

To represent the data of interacting molecules or genes, we use the simplest form of representation: the binary relation that corresponds to the pairwise interaction. Any higher level interactions involving more than two components at a time will be approximated by the collection of pairwise interactions. The concept of binary relations has already been utilized in the LinkDB database[6] of the DBGET system. For example, the link between two database entries:

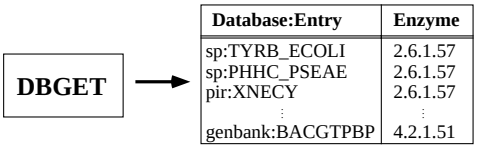
database1:entry1 $\Rightarrow$ database2:entry2

and the link between a database entry and a gene name:

database:entry $\Rightarrow$ organism:gene

are usually provided by each database and implemented in LinkDB/DBGET. Examples of these and other types of binary relations are shown in Table 1.

KEGG incorporates the links that exist in LinkDB/DBGET (Fig. 1(a)) and, in addition, contains more biological links of two interacting molecules



Database:Entry	Enzyme
sp:TYRB_ECOLI	2.6.1.57
sp:PHHC_PSEAE	2.6.1.57
pir:XNECY	2.6.1.57
⋮	⋮
genbank:BACGTPBP	4.2.1.51

(a) Relation between a database entry and an enzyme. This relation is extracted from LinkDB [6] in the DBGET system.[4]

Enzyme	Substrate	Product
2.6.1.57	Prephenate	Pretyrosine
4.2.1.51	Prephenate	Phenylpyruvate
4.2.1.51	Chorismate	Prephenate
⋮	⋮	⋮

(b) Binary relation between chemical compounds (substrate and product) that participate in a chemical reaction. This relation is extracted from LIGAND Chemical Database for Enzyme Reaction.[5]

(c-1)

Map	Enzyme
Phe, try and trp biosynthesis	2.6.1.57
Phe, try and trp biosynthesis	4.2.1.51
⋮	⋮
Terpenoid biosynthesis	2.5.1.10

(c-2)

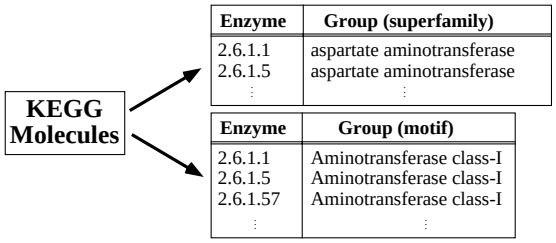
Map	Enzyme1	Enzyme2
Phe, try and trp biosynthesis	4.6.1.4	4.2.1.51
Phe, try and trp biosynthesis	4.2.1.51	2.6.1.57
⋮	⋮	⋮
Terpenoid biosynthesis	2.5.1.1	2.5.1.10

(c) Relationship between an enzyme and its location (map) in the metabolic pathways (c-1). Binary relation between two enzymes that appear consecutively in the pathway (c-2). These relations are extracted from the pathway diagrams of KEGG.

Organism:Gene	Enzyme
H. influenzae:HI1145	4.2.1.51
H. influenzae:HI0196	4.6.1.3
⋮	⋮
H. sapiens:ADCY1	4.6.1.1

(d) Relation between a gene and an enzyme. It is extracted from the KEGG hierarchical text data in the genes section.

Figure 1: Relational tables used in the computation of pathways



(e) Relation between an enzyme and the group to which it belongs. This relation is derived from the KEGG hierarchical text data in the molecules section. Currently, there are four classifications based on EC numbers,[7] PIR superfamilies,[8] SCOP 3D-folds [9] and PROSITE motifs.[3]

Figure 1: Relational tables used in the computation of pathways (contd.)

or genes. For the metabolic pathways, the two types of binary relations are identified. One is the substrate-product relation in the form of

$$\text{cpd:C00118} \Rightarrow \text{cpd:C00111}$$

which is extracted from the LIGAND database (Fig. 1(b)). The other is the relation of two consecutive enzymes appearing in the known pathways, such as

$$\text{EC:4.2.1.1} \Rightarrow \text{EC:5.3.1.1}$$

This relation is extracted from the pathway diagrams of KEGG (Fig. 1(c)).

2.3 Hierarchical Classifications

A hierarchy is often used to represent functional and structural similarities of genes and molecules. The hierarchical classification of EC numbers is based on the nature of chemical reactions catalyzed by enzymes. The degree of similarity in sequences and 3D structures of proteins is used for classifying superfamilies and folds. The taxonomy is the classification of organisms, which is important for extending sequence and 3D structural similarities to functional similarities. In the current version of KEGG, we construct enzyme classification tables according to EC numbers,[7] PIR superfamilies,[8] SCOP 3D-folds,[9] and PROSITE motifs [3] (Fig. 1(e)).

Once the entire genome is sequenced and the complete catalog of genes is obtained, it is natural to attempt to classify all genes according to their functions. Riley's hierarchical classification and categorization of *E. coli* genes [10] is an example of such an attempt. In KEGG, the functional classification is based on the actual pathway data which we consider as an objective criterion for categorization of functional units. In the WWW implementation of KEGG,

the classification tables can be viewed and manipulated as what we call hierarchical text data, where branches of the tree may be expanded or collapsed by simply clicking on the headings and subheadings.

2.4 Pathway Diagrams

The metabolic pathway diagrams such as by Boehringer[11] and by the Japanese Biochemical Society [12] represent a consensus view of known metabolic pathways for humans to understand. Starting from these two compilations, we have computerized the metabolic pathway data in the form of about 80 graphical diagrams. Each diagram contains enzyme objects that can be manipulated together with the relational operations. The diagram is intended as a drawing of all chemically feasible pathways, rather than a consensus of known pathways. Thus, although the reference diagram has to be drawn and continuously updated manually, the organism-specific pathways are automatically generated by matching the enzymes in the gene table with the enzymes on the pathway diagrams. For example, Figure 2 shows the phenylalanine, tyrosine and tryptophan biosynthesis pathway for *H. influenzae*, where each box represents an enzyme with the EC number inside and those enzymes found in *H. influenzae* are marked for easy recognition. Thus, the organism-specific pathway can be seen by following the connection of marked enzymes.

In the WWW implementation of KEGG, the subsets of enzymes stored in the public databases including PDB,[13] PIR,[8] SWISS-PROT,[14] GenBank,[15] and OMIM,[16] can also be viewed as marked boxes on each pathway diagram. The process of marking these enzyme subsets and the organism-specific pathways is automatic and daily updated in KEGG. In all the pathway diagrams, marked or unmarked, an enzyme in the box is a clickable object (see Fig. 2) to retrieve the corresponding entry in the LIGAND database, and then related entries can be obtained from a number of databases through DBGET. The boxed enzyme is also a searchable object; the enzymes as they appear in the known pathways can be searched and viewed graphically.

3 Computation

3.1 Partial Join Operation

The process of marking organism-specific pathways is based on the matching of the gene catalog (Fig. 1(d))

$$\text{organism:gene} \Rightarrow \text{EC:number}$$

and the list of enzymes on each pathway diagram (Fig. 1(c))

$$\text{map:accession} \Rightarrow \text{EC:number}$$

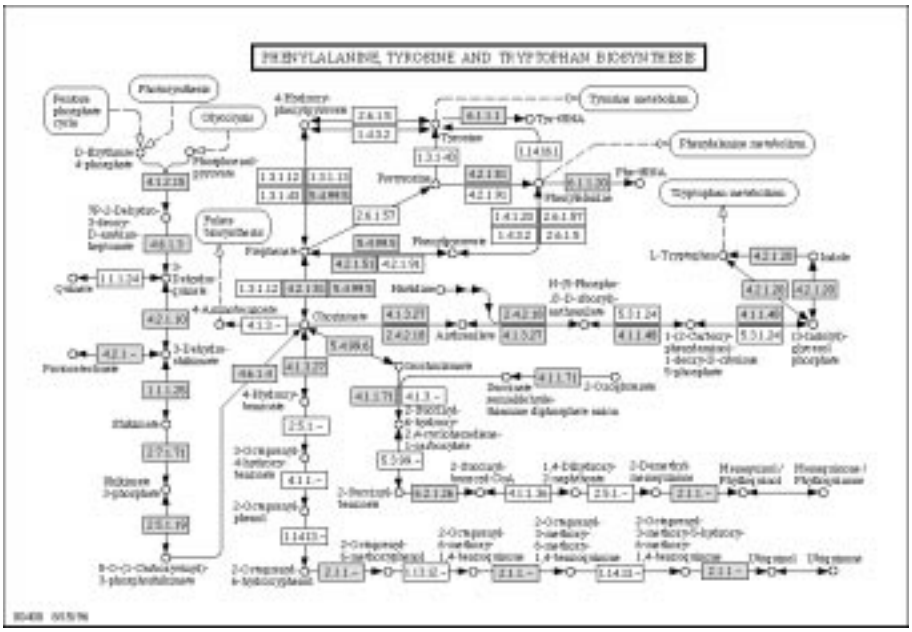


Figure 2: Phenylalanine, tyrosine and tryptophan biosynthesis pathway for *H. influenzae*

where 'map' is the database name for the pathway diagrams and 'accession' is the accession number of each diagram. The matching process is made by a join operation on EC:number. In the standard relational database the join is made by exactly matching the values in two columns, but the join here contains an interpretation of the EC number classification where a wild card is permitted for the lowest levels of the numbering scheme.

By applying this partial join operation in a more organized way, the process of query relaxation [17] in the deductive database can be implemented. Suppose an enzyme is found to be missing after examining the connectivity of the pathway. Because the EC number assignment is usually made by sequence similarity, it is necessary to check if other enzymes belong to the same superfamily of similar sequences. This can be done by going up the hierarchy of the superfamily (Fig. 1(e)):

$$\text{EC:number} \Rightarrow \text{SF:number}$$

and then examining all EC numbers in the superfamily (going down the hierarchy).

3.2 Path Computation

When no appropriate enzymes are found even by going up the hierarchy, the next step is to examine if alternative reaction paths can be found, given two compounds as the initial substrate and the final product. Suppose enzyme E catalyzes a chemical reaction with substrate X and product Y. Then the reaction is represented by the following form in the deductive database.

$$reaction(E, X, Y).$$

When the conversion of compound X to compound Y is a multistep process consisting of a number of enzymes, the enzymatic pathway is represented by:

$$path(X, Y, [E]) \leftarrow reaction(E, X, Y).$$
$$path(X, Y, [E \mid EL]) \leftarrow reaction(E, X, Z), path(Z, Y, EL).$$

This is a simple deduction from a collection of binary relations of substrates and products or, equivalently, a list of enzymes. We have been experimenting this path computation by the deductive database CORAL. [18] However, when we actually provide the path computation capability in KEGG we will use the C++ library that we are currently developing for efficient manipulation of binary relations and hierarchies.

Another type of problem in path computation is to compute from a given list of enzymes all possible pathways starting and ending at all possible compounds. This will become necessary for analyzing a collection of enzymes that have not matched on any of the known metabolic pathways. Apparently, there are still a number of unknown pathways, especially for secondary metabolisms and metabolisms that are turned on under stressful conditions. This type of computation may help understand such additional pathways.

4 An Example

4.1 Identifying Missing Enzymes

As shown in Fig. 2, the connectivity of marked enzymes can be used as a criterion to assess the accuracy of gene finding and functional prediction from genomic sequences. Especially, missing enzymes in each organism can be found immediately by just looking at the connectivity. In the case of the phenylalanine biosynthesis for *H. influenzae*, the enzymes missing are, among others, tyrosine transaminase (EC 2.6.1.5), aromatic amino acid transaminase (EC 2.6.1.57), phenylalanine dehydrogenase (EC 1.4.1.20), and cyclohexadienyl dehydrogenase (EC 1.3.1.43).

According to the current knowledge in the map, the lack of these enzymes results in the incompletion of the biosynthesis, which may be fatal to the organism. There are basically two possibilities that we need to examine to resolve this problem. The first case is that the genes coding for the missing enzymes have not been identified in the gene finding process, even though the organism actually has them. If the complete genome is already sequenced, the gene finding and functional assignment have to be repeated more carefully.

The second case is that the organism does not have the genes coding for those enzymes, and it uses other chemical reactions (enzymes) to produce phenylalanine. In this case, we need to search for alternative pathways from the source chemical compound to the target chemical compound, starting from the list of all enzymes identified or predicted in the genome.

4.2 Searching Alternative Assignments Using Hierarchies

As a way to re-examine the functional assignment of genes, the superfamily, motif and 3D-fold classifications and partial join operations can be used. Using the CORAL system we searched alternative assignments of enzyme genes for the three possible pathways from prephenate to phenylalanine in Fig. 2. The EC numbers of the missing (unmarked) enzymes were given to the system, and it returned the result shown in Table 2. If the enzyme 2.6.1.57 had been assigned to 2.6.1.1, or the enzyme 1.4.1.20 had been assigned to 1.4.1.4, the pathway from prephenate to phenylalanine would have been formed. In fact, two genes HI0286 and HI1617 are given the same EC numbers 2.6.1.1 according to the result of the sequence similarity search. They may require more careful analysis. The enzyme 1.4.1.4 was, however, correctly assigned and utilized in the other sections of our pathway database. We will probably need to search hypothetical proteins and unassigned genes in the *H. influenzae* genome for this missing enzyme.

4.3 Computing Alternative Paths

The path computation becomes necessary when we cannot find appropriate enzymes. Unfortunately, the current version of our database is not adequate to systematically perform the computation. For the example of Fig. 2, the LIGAND database describes the chemical reaction of enzyme 2.6.1.57 as the conversion between an aromatic amino acid and an aromatic oxo acid. It does not specify pretyrosine and prephenate, and we do not currently have a hierarchy of compound names. The reaction between prephenate and phenylpyruvate by enzyme 5.4.99.5 is not included at all in the current version of LIGAND.

Table 2: Result of the search of alternative enzymes for those participate in the phenylalanine biosynthesis for *H. influenzae*

Missing	Alternative	Classification used
2.6.1.57	2.6.1.1	motif (Aminotransferases class-I pyridoxal-phosphate attachment site)
2.6.1.5	2.6.1.1	motif (Aminotransferases class-I pyridoxal-phosphate attachment site) superfamily (aspartate aminotransferase)
1.4.1.20	1.4.1.4	motif (Glu / Leu / Phe / Val dehydrogenases active site)
1.4.3.2	1.4.3.5	EC number classification

In any case, we tried to compute alternative paths from phenylpyruvate to phenylalanine, but no paths were found. We believe, however, as we add more reactions and more compounds to LIGAND, which are then converted to relational tables in KEGG, and work out a hierarchy of compound names, the path computation will become a practical tool for biologists.

5 Discussion

The complete genome sequences of at least five free-living organisms, *H. influenzae*,[2] *M. genitalium*,[19] *M. jannashii*,[20] *Synechocystis sp.* [21] and *S. cerevisiae*, are already available via WWW or FTP servers of the original sequencing groups. They also provide the list of predicted coding regions and predicted gene functions. The major purpose of the KEGG metabolic pathway database is two-fold. First, we wish to establish an integrated view of higher functional aspects of gene products, namely, how they are interacting, for an increasing number of organisms. Second, we will provide a practical tool for making assignments of enzyme genes from genomic sequences.

For the first purpose, there are also other metabolic pathway databases, such as EcoCyc,[22] HinCyc,[23] EMP, [24,25] which often contain more detailed information of specific pathways and specific organisms than KEGG. Perhaps, the major feature of KEGG is its link capabilities, both in terms of the linking to the existing databases and different organisms and in terms of the linking to compute a pathway from binary relations. Because of these links we expect KEGG can be used to extract different information from the other metabolic pathway databases.

In KEGG the binary relations in the relational tables are effectively uti-

lized for path and other computations using logic. Of course, we are well aware of the complications of biological problems, which may sometimes be better treated by object-oriented technologies. However, a complex data representation may result in an ineffective computation. It remains to be seen whether the simple representation and efficient computation in KEGG can result in biologically significant observations. Toward that end, we are in the process of manually extracting main substrates, products, and cofactors for each of the enzymatic reactions of all known metabolic pathways. By representing each reaction as a collection of binary relations, we will try to reproduce the known pathways. At the same time, we need to computerize the synonyms and hierarchies of compound names. We are working on this in conjunction with the COMPOUND section of the LIGAND database.

The DBGET/LinkDB system is based on the model of loosely-coupled integration, where different databases with different schemas are integrated at the level of data entries. Thus, entries in different databases can be retrieved uniformly and links are made between related entries in different databases. This approach has been successful in view of the proliferation of WWW and the rapid progress of genome projects. The text-based DBGET system has been extended to the multimedia environment, and DBGET has been and will continue to be able to cope with the ever increasing number and volume of daily updated databases. Since KEGG inherits and shares DBGET/LinkDB capabilities, it will also be able to provide the most up-to-date metabolic pathways for a number of organisms including all organisms with complete genomes known.

The algorithm used in computing pathways between two compounds is based on Mavrovouniotis's algorithm.[26] Here it is implemented in a simpler way by representing reactions as binary relations, which enables us to treat the path computation as a logical derivation. Using various hierarchies and the framework of query relaxations,[17] the computation of alternative pathways can be accomplished. For example, Gaasterland and Selkov [25] adopted the taxonomy of *Mycoplasma* as a measure of proximity between organisms for use in the query relaxation. As a next step it will be necessary to combine different types of hierarchies, such as taxonomy, sequence similarity, and compound similarity, to compute alternative pathways.

Acknowledgement

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Science' from the Ministry of Education, Science, Sports and Culture in Japan. The computation time was provided by the

Supercomputer Laboratory, Institute of Chemical Research, Kyoto University.

References

1. F. Sanger *et al*, *Nature* **265**, 678 (1977).
2. R.D. Fleischmann *et al*, *Science* **269**, 496 (1995).
3. A. Bairoch *et al*, *Nucleic Acids Res.* **24**, 189 (1996).
4. Y. Akiyama *et al*, in *Proceedings of the 1995 Meeting on the Interconnection of Melecular Biology Databases* (P. Karp *et al* eds) ⟨http://www.ai.sri.com/~pkarp/mimbd/95/abstracts.html⟩.
5. M. Suyama *et al*, *Comput. Appl. Biosci* **9**, 9 (1993).
6. S. Goto *et al*, in *Proceedings of the 1995 Meeting on the Interconnection of Melecular Biology Databases* (P. Karp *et al* eds) ⟨http://www.ai.sri.com/~pkarp/mimbd/95/abstracts.html⟩.
7. Enzyme Nomenclature, Academic Press Inc. (1992).
8. D.G. George *et al*, *Nucleic Acids Res.* **24**, 17 (1996).
9. A.G. Murzin *et al*, *J. Mol. Biol.* **247**, 536 (1995).
10. M. Riley *et al*, *Escherichia coli and Salmonella: Cellular and Molecular Biology 2nd Ed.*, 2118–2202, ASM Press (1996).
11. M. Gerhard ed., Biological Pathways, Third Edition. *Boehringer Mannheim* (1992).
12. T. Nishizuka ed., Metabolic Maps. *Biochemical Society of Japan* (1980) (in Japanese).
13. F.C. Bernstein *et al*, *J. Mol. Biol.* **112**, 535 (1977).
14. A. Bairoch and R. Apweiler, *Nucleic Acids Res.* **24**, 21 (1996).
15. D.A. Benson *et al*, *Nucleic Acids Res.* **24**, 1 (1996).
16. P. Pearson *et al*, *Nucleic Acids Res.* **22**, 3470 (1994).
17. T. Gaasterland *et al*, *J. Intelligent Information Systems* **1**, 293 (1992).
18. R. Ramakrishnan *et al*, The CORAL Manual: A tutorial introduction to CORAL (1993).
19. C.M. Fraser *et al*, *Science* **270**, 397 (1995).
20. C.J. Bult *et al*, *Science* **273**, 1058 (1996).
21. T. Kaneko *et al*, *DNA Res.* **3**, 109 (1996).
22. P.D. Karp *et al*, *Nucleic Acids Res.* **24**, 32 (1996).
23. P.D. Karp *et al*, *Ismb* 4, 116 (1996).
24. E. Selkov *et al*, *Nucleic Acids Res.* **24**, 26 (1996).
25. T. Gaasterland and E. Selkov, *Ismb* **3**, 127 (1995).
26. M.L. Mavrovouniotis, in *Artificial Intelligence and Molecular Biology* (L. Hunter ed.), 325 (1993).