Automated Metabolic Reconstruction at the Molecular Level

Masanori Arita
ari@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

1 Introduction

The analysis of metabolism in bacteria facilitates a new application area such as drug synthesis or toxin degradation, using genetically engineered bacteria. However, the understanding of bacterial metabolism is still premature, because in a newly sequenced bacterium, gene functions are determined for half its total ORFs. Only from sequence similarity, it is impossible to decide functions of the remaining ORFs.

It seems speculative to estimate metabolism from only half of its component, but the good news is basic metabolism is conserved throughout species. Therefore, the basic strategy for predicting pathways is to (1) map known functions to the pathways we already have[1]; and to (2) fill gaps by inserting a missing piece of a metabolic jigsaw puzzle. The latter step corresponds to the prediction of ORF functions, and not only sequence similarity but substrate motif becomes available in this analysis.

For this purpose, we designed and implemented Automated Metabolic Reconstruction (AMR) system, which includes the database of chemical molecules, the database of enzymatic functions, and the deduction engine for computing putative pathways.

2 Method

Information on molecular structures is essential for the pathway reconstruction, in order to check the validity of pathways and the similarity of molecules. AMR considers atomic structures only, and does not introduce any abstraction at the level of chemical groups or components. The system is written in C^{++} using LEDA library package [4].

Database for Compounds: Molecular structures are input in SMILES notation [6] with chiral information. Input data is internally converted and stored in graph format. Aromatic rings are automatically detected at the same time. Many algorithms have been proposed for the normalization of molecular graph structures[2], but AMR supports an original efficient algorithm.

Database for Enzymes: Given a reaction formula, AMR computes which atoms correspond to which, between both hand-sides of the reaction. This mapping information is pre-computed and stored in the database. The structural matching of graphs is performed by an original efficient algorithm.

Data of compounds and enzymes are input in an original language, which can describe a hierarchical structure of cellular compartments. Therefore, the system can treat the same compound differently according to the compartment where it is located.

Deduction Engine: The information in the above two databases is considered a graph, in which atoms in compounds and the mapping of enzymes correspond to nodes and edges, respectively. Such

[1] A well-maintained metabolic map is on-line. http://www.genome.ad.jp/kegg
[2] See [2, 3] for the recent results.

edges are called *enzyme edges*. A metabolic pathway from compound S to T corresponds to a sequence of edges from an atom in S to T. In addition to these nodes and edges, the graph is augmented with *hypothetical edges* and nodes, generated from about 20 basic reactions representing dehydrogenase, kinase, and other common enzymes. All edges are assigned weights, whose biological interpretation is a *likelihood* of the corresponding reaction. The computation of pathways in this graph is realized with k shortest paths algorithm by Eppstein [1].

3 Result and Discussion

Currently, over 700 compounds are stored in the database, and the search process can efficiently compute pathways. For example, it can reproduce glycolysis and amino acid syntheses in bacteria. The advantage of our system is the enumeration of all the logically possible pathways, including hypothetical reactions.

The overall strategy resembles that of an expert system for finding synthetic paths of chemical compounds [5], but ours has several advantages over previous systems. First, because of the graph matching procedure, the enzyme database is generated only from reaction formulas. Enzyme data is different for each bacterium, but there is no laborious data-input process. Second, the graph theoretic representation of metabolism enabled us the search of much longer pathways, compared to a few steps in previous systems. Finally, a chemical molecule is merely an instance in our framework, and the deduction engine is applicable to other datatypes. Therefore, other biological cascades can be written in our system, provided that appropriate hypothetical edges and their weights can be defined.

If a hypothetical edge(s) can recover an important metabolism, the corresponding enzyme(s) is likely to be coded in the ORFs of unknown function. The prediction of gene functions with substrate motifs is an immediate application of this system. However, the output of the system depends sorely on edge weights, and their adjustment is quite important. The computational learning of these weights from examples is our future work.

Acknowledgments

The author thanks Prof. Morris, Prof. Green (U. Washington), Prof. Akutsu (U. Tokyo), and Prof. Dandekar (EMBL Heidelberg) for their continuous help and encouragement. This work is supported by the Research Fellowship of the Japan Society for the Promotion of Science for Young Scientists.

References

[1] Eppstein, D., Finding the k Shortest Paths, *Proceedings FOCS '94*, 154–165, 1994.

[2] Fan, B.T., Barbu, A., Panaye, A. and Doucet, J.P., Detection of Constitutionally Equivalent Sites from a Connection Table, *J. Chem. Inf. Comput. Sci.*, 36:654–659, 1996.

[3] Faulon, J.L., Isomorphism, Automorphism Partitioning, and Canonical Labeling Can Be Solved in Polynomial-Time for Molecular Graphs, *J. Chem. Inf. Comput. Sci.*, 38:432–444, 1998.

[4] LEDA is a library of data types and algorithms of combinatorial computing. `http://www.mpi-sb.mpg.de/leda/leda.html`

[5] Suzuki, E., Akutsu, T., and Ohsuga, S., Knowledge-based system for computer-aided drug design, *Knowledge-based Sys.*, 6:114–126, 1993.

[6] Weininger, D., Weininger, A. and Weininger, J. L., SMILES 2. Algorithm for Generation of Unique SMILES Notation, *J. Chem. Inf. Comput. Sci.*, 29:97–101, 1989.