

Pathway Bioinformatics

Peter Karp

Bioinformatics Research Group

SRI International

pkarp@ai.sri.com

Abstract

Pathway Bioinformatics is a subfield of Bioinformatics that is concerned with pathway algorithms, ontologies, visualizations, and databases [4]. This talk will provide an overview of the pathway databases and software under development in the Bioinformatics Research Group at SRI International, and will then discuss three pathway algorithms in detail.

A Pathway/Genome Database (PGDB) is a DB that couples pathway and genome information. SRI's BioCyc is a collection of PGDBs that includes organism-specific DBs such as EcoCyc [3] (for the bacterium *E. coli*) and HumanCyc. BioCyc also includes a general metabolic pathway DB called MetaCyc [2] that includes more than 450 pathways from 170 organisms. SRI's Pathway Tools software [1] has been licensed by more than 30 groups to create new PGDBs through a pathway inference process, to perform distributed refinement and updating of PGDBs, and to publish PGDBs on the Web. Pathway Tools also provides an API to facilitate global computations across PGDBs.

The first algorithmic problem we will discuss is graph layout of the complete metabolic network of an organism. Our long-term goal is to design an algorithm that will automatically create a graph layout whose structure emphasizes semantically meaningful relationships within the full metabolic network. Thus far, we have created a semi-automatic algorithm that segments the metabolic network into subregions based on the type of pathway (biosynthetic, catabolic, signaling, etc). Each subregion consists of a set of members consisting of individual pathways, and super-pathways (which are hierarchical clusters of pathways that share common metabolites). The layout system determines the topology of each pathway and super-pathway, and applies a node-positioning algorithm appropriate for that topology. Although each member is laid out automatically, the relative positions of members must be determined manually. The resulting graphs are intuitive to biologists and resemble manually constructed metabolic charts.

The second algorithm assists a scientist in determining whether the metabolic network of an organism stored within a PGDB is consistent with experimentally determined growth media for the organism [5]. Our approach is to map the metabolic network to a production system, and to map the growth media to a set of starting axioms for the production system. We then ask whether the propositions inferred by the production system include all metabolites that the cell requires for growth. This method has been used to validate the model of the *E. coli* metabolic network within EcoCyc.

The third algorithm assesses global properties of the *E. coli* metabolic network to provide insights about the system [6]. Properties measured by the algorithm include average number of substrates for each reaction, average number of reactions in which each substrate occurs, number of reactions with multiple isozymes, number of multifunctional enzymes, and number of reactions occurring in multiple pathways.

References

- [1] Karp, P.D., Paley, S. and Romero, P., (2002) "The Pathway Tools Software," *Bioinformatics* 18:S225-32.
- [2] Karp, P.D., Riley, M., Paley, S., and Pellegrini-Toole, A., (2002) "The MetaCyc Database," *Nucleic Acids Research* 30(1):59-61.
- [3] Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S., and Pellegrini-Toole, A., (2002) "The EcoCyc Database," *Nucleic Acids Research* 30(1):56-8.
- [4] Karp, P.D., (2001) "Pathway Databases: A Case Study in Computational Symbolic Theories," *Science* 293:2040-4.
- [5] Romero, P.R., and Karp, P.D., (2001) "Nutrient-Related Analysis of Pathway/Genome Databases," *Pacific Symposium on Biocomputing* pp471-82.
- [6] Ouzounis, C.O., and Karp, P.D., (2000) "Global properties of the metabolic map of *Escherichia coli*," *Genome Research* 10:568.