Using Term Lists and Inverted Files to Improve Search Speed for Metabolic Pathway Databases

Greeshma Neglur[1], Robert L. Grossman[2], Natalia Maltsev[3], and Clement Yu[4]

[1] Laboratory for Advanced Computing,
University of Illinois at Chicago, Chicago, IL 60607, USA
neglur@lac.uic.edu
[2] Laboratory for Advanced Computing,
University of Illinois at Chicago, Chicago, IL 60607, USA
grossman@uic.edu
[3] Math and Computer Science Division,
Argonne National Laboratory, Argonne, IL 60439, USA
maltsev@mcs.anl.gov
[4] Department of Computer Science,
University of Illinois at Chicago, Chicago, IL 60607, USA
yu@cs.uic.edu

Abstract. This paper describes a technique for efficiently searching metabolic pathways similar to a given query pathway, from a pathway database. Metabolic pathways can be converted into labeled directed graphs where the nodes represent chemical compounds. Similarity between two graphs can be computed using a metric based on Maximal Common Subgraph (MCS). By maintaining an inverted file that indexes all pathways in a database on their edges, our algorithm finds and ranks all pathways similar to the user input query pathway in time, which is linear in the total number of occurrences of the edges in common with the query in the entire database.

1 Introduction

Understanding of the complex architecture of metabolic networks provides insights into the fundamental design principles underlying the structure and function of living organisms. Common ancestry leads to the similarity of many molecular functions observed in all domains of life (Eukaryotes, Prokaryotes and Archaeabacteria). However, differences in organisms' physiology and lifestyle result in divergent evolution and emergence of variants of metabolic organization and phenotypic features. Large amount of metabolic and proteomic data available in public databases now allows for systematic exploration of adaptive mechanisms that led to the diversification of biological systems and the emergence of metabolic pathways characteristic of particular taxonomic or phenotypic groups of organisms. Such evolutionary and comparative analysis of metabolic pathways represents one of the essential problems in life sciences and is essential for progress in medicine, biotechnology and bioremediation. Metabolic pathways corresponding to various metabolic processes in an organism may be represented as a labeled directed graph.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 168 – 184, 2006.

The basic elements of metabolic pathways are chemical reactions that include compounds (e.g., substrates, products) and enzymes. Hence, computing similarity between pathways involves matching the constituent reactions (that include substrates and enzymes) and the connectivity between them. Brute force methods of matching the structures of substrates, enzymes and the pathway itself involve graph isomorphism tests at three levels and turn out to be computationally very expensive. The problem is further complicated if we are trying to query a database consisting of thousands of pathways with an average pathway size of 20+ nodes. Hence, efficient techniques for querying pathway databases are essential.

Our technique is able to search and retrieve pathways from a database similar to a query pathway in *time linear in the total number of occurrences of the pathway edges that are in common with the query* in the entire database. We do this by employing a simple indexing technique that uses terms defined from pathway edges and inverted files containing these terms.

2 Related Work

Many metabolic pathway similarity computing algorithms [8, 9] are based on abstracting pathways as enzyme graphs, i.e., directed labeled graphs where nodes are labeled with enzyme EC [12] (Enzyme Commission) numbers and a directed edge from one node to another implies that the product of the former node is the substrate of the latter. EC numbers provide a hierarchical classification of enzymes based on the reactions they catalyze. The classification tree consists of 4 levels with a root. Each enzyme is assigned a string consisting of 4 numbers, each of which corresponds to a level, for example: 1.2.3.4. In [8] the algorithm takes as input a sequence of EC numbers representing the enzyme graphs, aligns one sequence to another and attempts to find all EC numbers with the same 4-level hierarchical numbers, scores the similarities and cuts the sequences by removing the identical EC numbers and each pair of sub-sequences is initialized to begin a new round of 3-level hierarchical EC number match and so on. Another polynomial time algorithm described in [9] uses Approximate Labeled Subgraph Homeomorphism. A disadvantage of the enzyme graph representation is that it does not incorporate the similarity verification (i.e., in terms of structural similarity or chemical formula/sequence similarity) of substrates and products[1] in the pathway graphs.

Another technique [11] overcomes this disadvantage by combining sequence information of substrates and enzymes with graph topology of the underlying pathway. Several algorithms that efficiently perform pairwise pathway comparison are known [8, 9, 11, 17, 19]. One of the popular techniques outlined in [18] is called PathBLAST, which performs pairwise protein-protein interaction network alignment to detect linear paths and clusters [19] that are conserved between different species. This approach incorporates a refined probabilistic model for protein interaction data and also includes an automatic system for laying out and visualizing the resulting conserved subnetworks. This method is useful in evolutionary analysis by comparing the same pathway from different organisms, but may not scale efficiently to search

[1] Products in a pathway are chemical compounds.