Optimal Mapping of Multi Source Trees into DAG in Biological Network

Qiong Cheng, Alexander Zelikovsky

Department of Computer Science,
Georgia State University, Atlanta, GA 30303,
{cscqxcx,alex}@cs.gsu.edu

Abstract. The effective and efficient prediction and reconstruction of biological networks for a variety of inter-species and intra-species organisms have been a challenging research topic in computational system biology. They are becoming more and more important with the immense increase of good-quality data now available from high-throughput genomic, proteomic and metabolomic sources.
Our research focuses on the network alignment problem on a class of biological networks which can be modeled as directed acyclic graphs*(DAG)* such as phylogenetic networks, and metabolic pathways. Our objective is to obtain the minimal cost of mapping of multi source tree into DAG; and our approach is to employ dynamic programming to obtain the approximately optimal solution. We implemented the algorithm and are evaluating the preliminary results on metabolic network.
Further efforts are underway to analyze the computational complexity of the specified *DAG* mapping problem and to build a general tool for automatically mining feasible patterns for the reconstruction of multiple bimolecular networks at different levels and dimensions.

Key words: System Biology, Graph Theory, Graph Mapping

1 Introduction and Related Work

Graph theoretic formalism is adopted to study the subgraph mapping problem. Some papers in pathway mapping only consider the maximal similarity of sequences composing the network. Their authors assumed that the networks have the same fixed topology. However there are some others who consider not only sequence similarity but also network topological similarity. Pinter et al. modeled metabolic network of pathways as outgoing trees with different size and reduced the problem to the approximate labeled tree homeomorphism one[1]. Kelly et al. rebuilt a global alignment graph in which each vertex represents a pair of proteins and each edge represents a conserved interaction, gap, or mismatch; and then they try to find the highest-scoring path with limited length and no consecutive gaps or mismatches([4], [5] and [6]).

Both vertex similarity and network topology similarity are considered in network alignments. Chen et al. formulated the problem as an integer quadratic

problem to obtain the global similarity score based on the mapping of as many as node-node similarity and as many as edge-edge similarity[2]. Yang et al. employed an exhaustive searching approach to find the vertex-to-vertex and path-to-path mappings having the maximal mapping score with the limited length of gaps or mismatch[3].

To explore the general mapping problem, we focus on the solution to a class of problems of mapping of multi source tree to the networks which can be modeled as directed acyclic graphs. A solution based on dynamic programming is proposed in this paper so that a vertex-to-vertex and edge-to-path mappings with the minimal homomorphism cost is obtained. The flexible interface can be provided to support different vertex similarity calculation and different approaches to generate random graphs.

2 Network Alignment Problem Formulation

2.1 NOTATION AND DEFINITIONS

Definition 1. *A **directed** path is a sequence of distinct vertices $v_0, v_1, ...v_i, ...v_k$ and edges $e_0, e_1, ...e_i, ...e_{k-1}$ such that e_i is an edge directed from v_i to v_{i+1}, for all $i < k$. We denote the path with the first vertex v_0 and the last vertex vertex v_k by $P_{0,k}$. In this paper, **path** will always mean 'directed path'. Two or more disjoint Paths mean that they share at least no last vertex with each other.*

Definition 2. ***Pattern tree*** *$T =< V^T, E^T, L^T >$ is a multi-source tree in which V^T is the vertex set, E^T is the edge set, and L^T is the label set. Every vertex in V^T has unique label in L^T. The multi-source tree is a directed acyclic graph(DAG), whose underlying undirected graph is a tree [1].*

Definition 3. ***Text*** *$G =< V^G, E^G, L^G >$ is a directed acyclic graph(DAG) in which V^G is denoted as vertex set, E^G is denoted as edge set, and L^G is denoted as label set. Every vertex in V^G has its unique label in L^G.*

Definition 4. *A **mapping** from pattern $T =< V^T, E^T, L^T >$ to text $G =< V^G, E^G, L^G >(f : PatternT \rightarrow TextG)$ is called **homomorphic** if*

1. *every vertex in V^T is mapped to a vertex in V^G $(f_v : V^T \rightarrow V^G)$;*
2. *every edge e=(u, v) in E^T is mapped to a path in $P^G (f_e : E^T \rightarrow P^G)$. The path is a ordered sequence $u_0 = f(u), u_1, u_2, ..., u_k = f(v)$ of vertices in P^G;*

Definition 5. *Let $\Delta : L^T \times L^G \rightarrow \Re$ denote the 2-place relation on the set of all ordered pairs of a vertex in V^T and a vertex in V^G to a real number. δ represents the penalty score for deleting a vertex from a path. The **cost of a homomorphic mapping** $f : T \rightarrow G$ is*

$$cost(f) \sum_{v \in V_T} \Delta(v, f(v)) + \delta \times \sum_{e \in E_T} (|f(e)| - 1)$$

2.2 PROBLEM FORMULATION

Given pattern tree T and text graph G, find a homomorphic mapping that minimize the homomorphic cost. The homomorphic mapping allows vertices repeated occurrence.

An example of the mapping between metabolic pathways are as follows.Our objective in the example is to find a mapping of pattern graph into the pydrimidine ribonucleotide ribonucleoside metabolism pathway so that the mapping score is minimized.

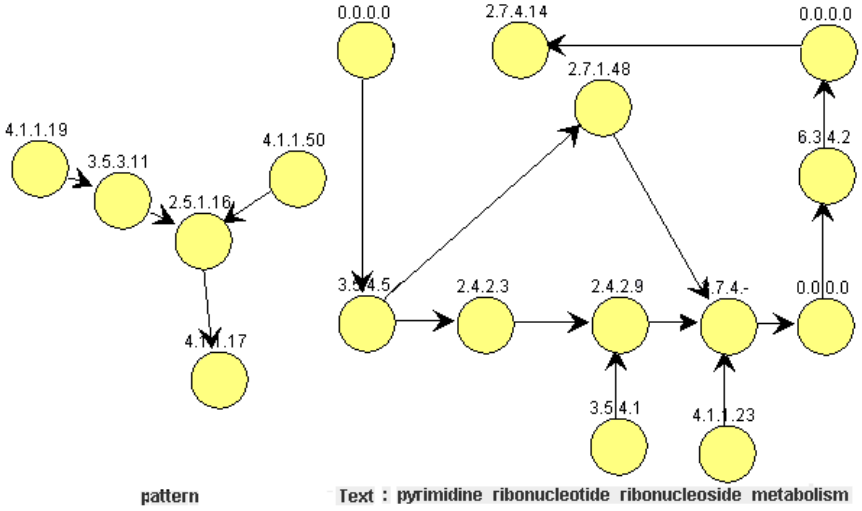

Fig. 1. Pattern graph is a multi source tree. Text graph is a pydrimidine ribonucleotide ribonucleoside metabolism pathway which is a DAG.

2.3 Dynamic Program Algorithm

We build a two dimension table DT for our dynamic program by the post order traversal. Every column is a homomorphic mapping from every vertex of pattern T to a vertex of text G. Every row is a mapping from a vertex of pattern T to every vertex of text G. The default value in every entry is initialized to infinity.

The recursive function for the dynamic program is as follows,

$$DT(u^T, v^G) = \min\{\Delta(u^T, v^G) + \sum_{u_i^T \in u^T.children} \min_{v_i^G \in v^G.children} DT(u_i^T, v_i^G), \quad (1)$$

$$PenalityOfDelete(v) + \min_{v_i^G \in v^G.children} DT(u, v_i^G)\}\} \quad (2)$$

3 Future Work

More meaningful experiments are in need. We will generate the pattern graph randomly by network motifs with small number of vertices of different and random labels such as with the size of 3, 4 and 5 vertices[7]. These generated patterns will be used to verify the reliability of our algorithms.

4 Conclusion

We devote ourselves to the studies of mining graph data and their application of studying biological evolution and reconstruction. We reformulated the problem proposed in [1], designed and implemented the dynamic programming. Further efforts are in need to evaluate our algorithm so that we could have a deeper understanding towards the studies of biological evolution and reconstruction.

Acknowledgments. We acknowledge Dr.Ron Y Pinter and his student Oleg Rokhlenko for sharing their executable code and graphic user interface with us.

References

1. Ron Y Pinter, Oleg Rokhlenko, Esti Yeger-Lotem, and Michal Ziv-Ukelson, *Alignment of metabolic pathways* Bioinformatics, **21(16)**: 3401-8, (Aug 2005)
2. Zhenping Li, Yong Wang, Shihua Zhang, Xiang-Sun Zhang, Luonan Chen, *Alignment of Protein Interaction Networks by Integer Quadratic Programming* EMBS '06. 28th Annual International Conference of the IEEE : 5527-5530, (Aug. 2006)
3. Qingwu Yang, Sing-Hoi Sze *Path Macting and Graph Matching in Biological Networks* Journal of Computational Biology Jan 2007, Vol. 14, No. 1: 56-67 : 5527-5530, (2007)
4. Brian P. Kelly, Roded Sharan, Richard M. Karp, Taylor Sittler, David E. Root, Brent R. Stockwell, *Conserved pathways within bacteria and yeast as revealed by global protein network alignment* PNAS : 11394-11399, (Sep. 30 2003)
5. Brian P. Kelly, Bingbing Yuan, Fran Lewitter, Roded Sharan, Brent R. Stockwell and Trey Ideker, *PathBLAST: a tool for alignment of protein interaction networks* Nucleic Acids Research, Vol.32 : W83-W88, (2004)
6. Roded Sharan, Silpa Suthram, Ryan M. Kelley, Tanja Kuhn, Scott McCuine, Peter Uetz, Taylor Sittler, Richard M. Karp, and Trey Ideker, *Conserved patterns of protein interaction in multiple species* PNAS, Vol.102 : 1974-1979, (2005)
7. R. Milo, S.Shen-Orr, S. Ltzkovitz, N. Kashtan, D. Chklovskii, U. Alon, *Network Motifs: Simple Building Blocks of Complex Networks* Science www.science.org : 824-827, (Oct. 25 2002)