RegAlign: An Algorithm for the Alignment of Gene Regulatory Networks

George Davidescu

Recently, information on gene regulatory networks is being amassed at an unprecedented level, resulting in a need for the detailed analysis of this data. Network alignments are useful measures of the similarity of two networks. RegAlign is a new tool whose function is to accurately produce a global alignment of two given regulatory networks and to identify their most similar components and motif structures.

In this project, regulatory networks were represented as static directed graphs. Nodes represent genes (transcription factors or otherwise) and contain information such as the name and nucleotide or amino acid sequence of their respective gene. Directed edges between nodes represent regulatory interactions and are of two types, activation or repression. The networks are represented in KGML (KEGG Markup Language) format, an XML format currently used by KEGG (Kyoto Encyclopedia of Genes and Genomes) to represent other types of biological networks.

The core algorithm of RegAlign is based on NetAlign[1], a network alignment application developed by Andrew McKim, with some modifications. Unlike most biological network alignment techniques the algorithm takes into account not only the similarity between the nodes of the two networks being aligned, but also the structure of the networks and the presence of motifs.

Although there are some options as to how the alignment is performed, the general steps are as follows. The algorithm first compares all nodes in one network to every node in the other network and vice versa, computing the nodes' pair-wise similarities. It then compares the local network structure (or neighbourhood) around every node in each network and the neighbourhood of every node in the other network, computing a neighbourhood similarity value for each pair of nodes. Munkres' algorithm is then applied to generate a maximum-weight bipartite graph (Fig 1.) from the two networks by matching the most similar nodes of the two networks, based on the nodes' pair-wise and neighbourhood similarities. Finally, the aligned nodes are reported and an overall similarity score of the two networks is computed.

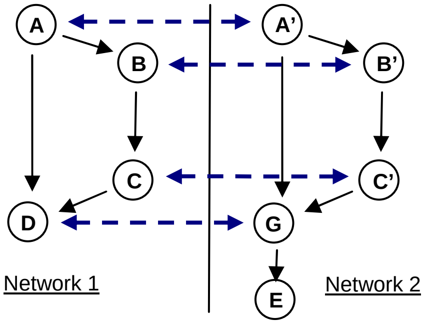

Fig 1. A Regulatory Network Alignment

Much insight can be gleaned from biological network alignment information. The comparison of two regulatory networks can be used to infer how closely related two organisms are, what their relation in a phylogenetic tree is or how homologous networks in related organisms evolved since they diverged. Regulatory network similarity measures could also reveal how well a model organism approximates another organism, for example for gauging the accuracy of animal trials for prospective commercial drugs.

An issue that surfaced during this project is the present lack of a standard format for publishing gene regulatory network data. It is hoped that collaboration between members of the bioinformatics community will lead to the standardization of regulatory data in the future. The project has some options in mind to this end, among them the adoption of KGML.

1. McKim, Andrew. (2007) *Aligning biological networks*. MCS Thesis. University of New Brunswick