NSERC research project
By Oana Sandu
Supervisor: Pr. Mathieu Blanchette
McGill Center for Bioinformatics

Enforcing ordering constraints in sets of DNA Motifs

Abstract:
The bioinformatics program Footprinter predicts functional DNA regions by comparing sections of the genome of related species. It reports a set of motifs along with the position(s) at which each motif occurs in each sequence. Although biologically, actual functional sequences tend to be in an order that is highly conserved across species, Footprinter doesn't take this into account. The purpose of this project is to process its output to eliminate false positives by reporting the maximal set of motifs that are in a consistent order.

In the simplest version of the problem, each motif appears exactly once in each sequence; we modeled the problem by a directed acyclic graph in which vertices are motifs and directed edges from a motif to another indicate that they are in the same order in all of the sequences. The largest consistent subset of the motifs is given by the longest path in the graph. It can be found by a dynamic programming algorithm in polynomial time.

In a slightly harder version of the problem, a motif can be reported as part of the solution and be present in only some of the sequences and not in others. In this case, there can be cycles in the graph used to model the problem, hence finding a longest path is harder. Also, for the solution to make sense, we enforce that the longest path has to be transitive.

The algorithm created to solve this version of the problem tries to guess an upper bound on the length of the longest transitive path. It then determines whether a subset of vertices of that size could exist, by analyzing the set of degrees of the vertices. The algorithm then reduces the upper bound until it converges to the length of the longest transitive path, then it reports a solution. A number of optimizations were done to this algorithm to reduce its running time.

In the hardest version of the problem, each motif can appear any number of times in any of the sequences. Since there no longer is a one-to-one correspondence between one occurrence of a motif in a sequence and an occurrence in another, the problem is harder to model. The algorithm tries to simplify this by making educated guesses about which occurrences of a motif across sequences correspond to each other.