Project 2: Multiple Genome Alignment by Clustering Pairwise Matches

Shravya Reddy Konda
Email:shravyak03@gmail.com

April 22,2008

Abstract
The goal is to implement a multiple genome alignment algorithm by using a sequence clustering algorithm to combine local pairwise genome sequence matches produced by pairwise genome alignments, e.g, BLASTZ. Sequence clustering algorithms often generate clusters of sequences such that there exists a common shared region among all sequences in each cluster. To use a sequence clustering algorithm for genome alignment, numerous local alignments between a pair of genomes must be handled. This algorithm does not need to make a guide tree to find out the order of the multiple alignment, thus performing quite well over existing algorithms. In this project, after implementing this algorithm, I will compare its performance with one of the existing tools for finding multiple alignments.

Motivation:

A heuristic alignment approach which is normally used for multiple sequence alignment problem is not suitable to be applied to multiple genome alignment problem due to the following reasons:

a) It is hard to utilize the guide tree since there are many local regions where their phylogenetic relationships are different from those of the “entire” genome. Even duplications of a gene within a genome have their own phylogenetic relationship.

b) The greedy progressive alignment works due to the guide tree. Given that the utilization of a guide tree is not trivial for genome alignment as discussed above, the greedy progressive alignment strategy should be avoided.

c) The iterative alignment requires generation of profiles – alignment of multiple sequences – while combining pairwise matches. Generating and aligning profiles for long genome sequences is not practical.

Project Description:

The authors of the paper who proposed this algorithm called it BAG which stands for Bi-connected components and Articulation point based Grouping of sequences. This can typically handle protein sequences which are 1000 amino-acid characters or less. This strategy was found to detect multiple homologous regions accurately.

In this project, I will explore the details of the BAG algorithm and implement it. Then, I will compare the performance of this algorithm to an already existing tool for multiple alignment and check which algorithm works better for multiple genome alignment.