Implementing Parallel Processing for Genetic Algorithm Based Phylogenetic Tree Estimation From Aligned Sequences

Project Proposal for CMSE 822

Group Member: Md Alamin (alaminmd@msu.edu)

1 Abstract

Phylogenetic tree construction is a widely researched topic in the field of bioinformatics or genomics. It shows the evolutionary relation among the different species being descended from a common ancestor. The inputs for phylogeny estimation algorithms are genomic sequences and the output is a phylogenetic tree showing the relationship between those sequences. Constructing phylogenetic tree is a difficult computational problem. For, example for only 14 taxa, there are more than seven trillion possible unrooted phylogenetic trees. Genetic algorithm can play a vital role here to reduce the computational effort to a substantial amount compared to a conventional heuristic search methods. The intend of this project is to apply parallel processing while performing the optimization through out the population. This project will refer to [2] which works with the similar problem.

2 Project Description

This project will present a genetic algorithm based phylogenetic tree construction from aligned sequences.

The initial population will be a number of randomly generated trees. So, each chromosome for the GA will be a phylogenetic tree along with its branch lengths and the values of other parameters comprising the substitution model

used. For fitness function natural log likelihood of the tree for the given dataset will be used. Tress with higher likelihood values will produce more offspring to the next generation.

The substitution model intended to use for the tree is HSY nucleotide substitution model. The parameters that will be embedded with the trees are k (a value for the transition/transversion rate ratio), the branch lengths and equilibrium nucleotide frequencies.

Rank based selection operator will be used with elitism. Rank is based on the natural log likelihood of that individual. Mutations will be implemented on branch level and topological level. A topological mutation involves removing a randomly chosen sub tree and reattaching it at a randomly chosen site on the remaining tree. For crossover, swapping a sub tree under a randomly selected branch will be used between the parents. The algorithm will be terminated either after a fixed number of generations or if the likelihood score for the best individual of a population is not improved for a fixed number of generations.

3 Parallelization Strategies

Calculation of the likelihood value for a single phylogenetic tree is a computational demanding task. If the size of population of the GA is 100, it takes a substantial amount of time for calculating fitness of each individual sequentially. I intend to calculate the fitness for each individual of the population parallel. I plan to use Cython with OpenMP support for the parallel processing part.

4 Benchmark and Optimization

For the datasets it is intended to use the simulated SATe-I [1] datasets. There are 37 model conditions. Each model condition has 20 replicate datasets. In each of replicate datasets, there remain the input sequence file which is in FASTA format and its model tree which is in Newick format. After selecting

a model condition I may need to customize the dataset to reduce the number of taxa because I want to test the algorithm in a small dataset first.

Normalized Robinson Foulds distances will be used for the performance measure. It compares the topological differences between the GA generated final trees and the true trees.

From this project, I expect to the get a clear and concise idea about the implementation of parallel computing in Bioinformatics and also about the accuracy and limitations of GA in phylogenetic tree reconstruction.

References

[1] S. Nelesen C. R. Linder T. Warnow K. Liu, S. Raghavan. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees,. *Science*, 324(5934), 2009.

[2] A. Skourikhine. Phylogenetic tree reconstruction using self-adaptive genetic algorithm. *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering, Arlington, VA, USA*, pages 129–134, 2000.