Weekly Report

Reporting Period: June 15th, 2009

Submitted By: William Dawson (dawsonwd1@gcc.edu)

Advisor: Dr. Tiffani L. Williams

Summer Goals
1. Develop, assess, and implement a set of rubrics for detecting convergence of a Bayesian phylogenetic analysis.

2. Gain a strong knowledge of computational phylogeny and the current challenges it faces.

3. Experience research firsthand.

4. Improve knowledge and skills in the areas of science and computer science.

Work Completed
1. Performed follow up investigation into Average Standard Deviation of Split Frequency and Potential Scale Reduction Factor.

2. Wrote code that can calculate consensus tree for each chain, halt program on a condition, only calculate data for cold chain, periodically stores consensus tree, and compares RF of consensus tree across runs.

3. Completed research plan.

4. Began read through Felsenstein book and other papers.

5. Finished running Kim data.

Work Planned For This Week
1. Experiment with a burn-in parameter when tracking consensus tree.

2. Create a sliding window for building consensus tree. Use it to analyze whether we should keep running because current set is hurting overall consensus tree.
3. Determine when search is in a minimum, store that consensus tree.
4. Run Kim data with different heat parameters.
5. Determine whether Johnson paper is useful.
6. Read more papers and more of Felsenstein book.

Reflections

This week I sought to better understand the ASDoSF and PSRF consensus metrics. Every time a partition is added to a node in the sample tree, a count variable is incremented. When a partition is removed, that count decrements. ASDoSF keeps track of these between runs and calculates a standard deviation between them.

PSRF keeps track of the tree length, the substitution rate between any two values (such as A->C), the stationary state frequencies, the shape of gamma across sites, and the proportion of invariable sites. PSRF is calculated across runs, not chains, because when you use heated chains then the posterior distribution is incorrect. Thus chains can’t be used for convergence detection, and can be only used to prevent getting trapped at a local maximum.

I also spent this week doing further reading to gain better foundational knowledge as well as to explore new ideas. This included Felsenstein’s book, Nylander 2007, Mengersen 2008, and Aldouni 2005. This week I plan on reading Johnson 1998, Hastings 1970, and Huelsenbeck 2002. I believe the Johnson paper might be particularly useful and will require close reading.

This week the Kim data which was running for 10 million generations completed running. This was the data that had a high ASDoSF yet a stable consensus tree. It was found that after the 7000th generation, the consensus tree began to change wildly. This data will need to be further examined. First, it will be run again now with the ability to periodically save the consensus tree. Second, it will be run with high heating parameters (doubling to 0.4 seems like a good start).
This week I have a few leads to follow. First, I should implement a hard coded burn-in parameter. Second, I should create a sliding window for calculating the consensus tree. This will let me know how much the consensus tree is being changed from the recent data, and will help predict when the program should run further. I should also explore what sort of information the strict consensus tree can provide me. Finally, it should be possible to use RF values to determine when the search has likely ended up on a maximum. It might be useful to record the number of unique maximum, and build a strict consensus tree between them.