Weekly Report

Reporting Period: August 3rd, 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. Created poster.
2. Gathered results for breadth of search.
3. Created list of non-essential work to complete.
4. Solidified pipeline for convergence analysis.

Work Planned For This Week
1. Write final paper (due Thursday at noon).
2. Find a justifiable number of anchor points to use.
3. Give poster presentation.
4. Provide documentation for how to modify MrBayes for further work.

Reflections
Last week, I performed breadth of search analysis with both RF and Quartet Distance metrics. I was surprised to find that the RF anchors indicated data had searched the breadth of tree space, whereas the Quartet Distance did not. Furthermore, the RF distance provided practical feedback on convergence. I believe I have discovered what is causing these results. As you might remember, when tracking the changes in the consensus tree there are actually not that many changes that take place. This must mean that the starting trees are very good choices, and that the rest of the search is focusing on the little details. This means that even the hot chain never explores the breadth of tree space, because only the space around the starting tree provides promising results. The RF distance tracks the breadth of search within this narrow area. It makes sure a good number of couplet and triplet combinations are tried before the search is terminated.
I began searching for data sets that might refute the anchor point metric. I was able to find a data set where the consensus tree continues to change after the breadth metric indicates convergence. However, all of the changes made afterwards are just switching back and forth between resolving one triplet, and the consensus tree of rejected trees indicates that the search needs to continue at this point. It appears that using this metrics in combination provides an accurate diagnosis of convergence. I really wish I had time to fully modify MrBayes and HashRF so that these tests could be run faster, right now it can take several hours to run the breadth of search diagnostic (though this is still better than running multiple runs). Right now the last big hurdle is to justify the number of anchor points chosen.

This week, I’ll have to write a final paper summarizing all of this. I’ll want to run more taxa sets that are published, and make sure the trees I build from those sets match what are in the published papers. I’ll also be giving my poster presentation.

Before I leave, I’d like to create some documentation of my modifications of MrBayes. It could be that other lab workers will want to use similar modifications for collecting other kinds of data, and that would be easier if one knew where to start.