Weekly Report

Reporting Period: July 13th, 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. Formally investigated window size using powers of 2 for size of window.

2. Investigated further geometric ergodicity and coupling. Was unable to find any generic Matlab or R tools to implement.

3. Documented Kim’s extra parameters.

4. Got Component working on windows, but was unable to get it to compile on Linux.

5. Installed R with apTreeshape package.

6. Continued work on generating tree space anchor points.

7. Created outline for presentation.

Work Planned For This Week
1. Rerun experiments that investigate how far apart randomly generated trees are.

2. Develop fully a system for generating tree space anchor points.
3. Modify Mr. Bayes to take samples from the hot chains.

4. Research Markov Chain coupling more. Come up with a coupling algorithm that the coupling lemma applies to.

5. Prepare for presentation on Thursday.

6. Continue to think about runtime profiles.

7. Learn how to use apTreeshape and discover if tree topology metrics are useful.

**Reflections**

This past week, I took detailed data on the window size parameter. The data below was generated using the Kim data set (27 taxa) without extra parameters. 2 million generations were run, sampling and diagnosing every thousand generations. Two runs were used, each using four chains. The results can be found in the next section.

I also looked further into the concept of geometric ergodicity. I discovered that our distribution is ergodic because it is positive and continuous. However, geometric ergodicity is more strict, and I found that is unlikely that our target distribution has this property. I also found on Dr. Bordewich's website the following quote:

> Due to the difficulty of proving rapid mixing, and stringent conditions required for other non-quantitative convergence tests (for example geometric ergodicity), many convergence diagnostics have become widely used to test whether an observed simulation has reached stationarity.

[http://www.dur.ac.uk/m.j.r.bordewich/Fellowship/fellowship.html](http://www.dur.ac.uk/m.j.r.bordewich/Fellowship/fellowship.html)

Which further confirms its unlikeliness.

I discovered that according to the coupling lemma, coupling times bound the total variation distance between the distribution discovered so far and the true distribution. Thus, if a coupling algorithm was implemented in MrBayes, it would be possible to determine the rate of convergence. However, if this rate is not geometric, there are not metrics to take advantage of this knowledge. However, I believe this knowledge would still be useful, and plan to learn more about chain coupling so that I can calculate the rate of convergence. This will take lower priority than investigating the breadth of search.

A final note on ergodicity, I was unable to find any Matlab or R packages or scripts that seem to investigate this. This is unfortunate, because code that actually implements coupling would be very helpful to me.
This past week, I looked into why the Kim data takes so long to run. It turns out in the .nex file there are extra parameters set up to partition the genes. When these parameters are removed, it runs much faster.

I was able to get Component to run on windows, but am still unable to get it to compile on Linux. The appropriate Pascal compiler has been installed, but the make file does not work. This week, I'll look closer at the documentation and see if I can get it to compile.

R and apTreeshape are now installed on my machine. I'll need to learn how to use apTreeshape so that I can investigate it's topology features, as well as generate random trees for investigating the breadth of search.

This past week I took important steps towards investigating the breadth of tree space. My mathematical analysis and experiments have shown that it is possible to map tree space using fully resolved trees. However, this contradicts earlier findings, so I'll need to rerun past experiments to confirm. If the past experiments confirm my results, I will continue by developing an algorithm to generate trees that cover the breadth of tree space. I will also investigate how many trees are required so that the anchor trees are at most n bipartitions apart, to see how finely treespace can be resolved. If the experiments contradict my results, I will have to resort to using multifuricating trees to map out tree space.

This week, something new I can investigate is testing the convergence between chains. When sampling, it may be possible to have the hot chain decide whether to accept without taking the temperature into account. If it does accept, this result would be stored off to file, and then the hot chain will return to sampling with heat. This data will also converge to the actual distribution, but how does its consensus tree compare with the cold run? If I have time this will be an interesting path to explore. If it begins to match the cold chain quickly, it might replace the use of multiple runs.

This Thursday I will be giving a ten minute presentation on my research. I have created an outline (see section below) and have begun preparing my power point presentation.
Window Results

Difference Between Window and Full Consensus Tree (Window Size=2)

Difference Between Window and Full Consensus Tree (Window Size=4)
Presentation Outline

1. What is computational phylogeny?
   a. Building evolutionary trees
   b. Characteristics->Alignment->Tree Search->Summary

2. My Area: Tree Search
   a. Looking for most likely tree
   b. Use of Bayesian and MCMC. Relation to hill climbing algorithms.
   c. Problem: size of tree space.

3. My Focus: When to Stop Search
   a. Well investigated problem in general, but we hope to use problem specifics to make stronger conclusions.
   b. Things looked into so far: consensus tree metric. Have found a good metric for local maximum.
   c. Things currently looking into: how quickly is estimate converging? How good of a breadth of search? Run time profiling.

Annotated Bibliography


   Thoughts: Talks about recent developments in Markov Chain Monte Carlo, and discusses ergodicity.


   Thoughts: Provides thorough mathematical bases for ergodicity, yet also applies to practical applications. Is optimistic about most data sets being geometrically ergodic because it must repeat some small set.

Thoughts: Talks about coupling lemma in context of hypergraphs. It's important to note that Bordewich does phylogeny research, and hasn't applied coupling lemma to phylogeny.


Thoughts: This is the real mathematical basis for Markov Chains, coupling, and ergodicity. Good for knowing background, but because all chains are matrices, it's hard to make practical connection on my own.

5. Generating Random Spamming Trees with Markov Chains (Part 1) by Ronitt Rubinfeld.

Thoughts: Goes over coupling, coupling lemma, and total variation distance.


Thoughts: This looks at coupling very closely, and gives another perspective on how to accomplish it.