Weekly Report

Reporting Period: June 22nd, 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. Read more papers and more of Felsenstein book.
2. Obtained paper for Kim data set.
3. Added features to track resolution rate, calculated a windowed consensus tree, build a consensus tree from rejected trees, and record when a chain swap is made.
4. Created a program that combined all features developed and gathered data using it.
5. Further investigated Average Standard Deviation of Split Frequency and Potential Scale Reduction Factor.
6. Constructed annotated bibliography.
7. Integrated version of HashRF that works with unresolved trees.
8. Attempted to get Qdist working but won't work with unresolved trees.
10. Read Kevin Liu et al. Paper
Work Planned For This Week

1. Discover why Average Standard Deviation of Split Frequency is so sensitive to local change, how it relates to consensus tree.

2. Develop a written plan of attack or evaluation strategy for using Johnson 1998 paper.

3. Run combined data collection program on Kim data.

4. Get good results from Kim and compare with paper.

5. Tweak some more with window size.

Reflections

This week I spent some time reading to gain a better background on the subject matter. This included spending lots of time in Felsenstein's book, as well as the papers listed in the annotated bibliography. Most of these simply built better foundational knowledge about Bayesian statistics, phylogeny, and markov chain monte carlo. I also read the Kevin Liu paper which presented the interesting idea of combining the tree calculation and the sequence alignment. I thought this paper sounded promising, but was concerned about whether bootstrapping would be even more time consuming using it.

I found that the Johnson paper may provide a practical strategy for how to detect convergence. It uses two independent runs, and uses their coupling times to map dependent draws from the posterior (as a result of us using MCMC) to independent draws. By using this information, one could use traditional statistical techniques on the data, which would give us a confidence interval for convergence. The downside might be that his methods are too conservative to be practical. In the coming week, I should either come up with a plan of attack for implementing his method, or a set of metrics to determine if his methods are too conservative.

This week I continued to add features to MrBayes for data collecting purposes. You can now track resolution rate, a windowed consensus tree, a consensus tree built from the rejected trees, and when chain swaps happen. These features, as well as past features, were all combined into one program. I then integrated this program with a version of HashRF that allows for unresolved trees as input, and collected data. I would have integrated Qdist, but unfortunately it does not work with unresolved trees. The graphs below show data collected over 1 million generations using the anolis.nex file. Samples and diagnostics were taken every 1000 generation. Next week, I hope to run this experiment on the Kim data set to see what sort of trends a more complex data set reveals.
One interesting finding from this data was that the consensus tree built from the rejected trees and the consensus tree built from the accepted trees actually converge to the same consensus tree. I discovered that this was a side effect of the tree topology rearrangement algorithm. By tweaking the parameters so that bolder trees were proposed, it was possible to make the consensus tree of rejected trees converge to the normal consensus tree at a much slower rate.

I was disappointed to discover that the windowed consensus tree appears too sensitive to be used. However, it does not appear to be any more sensitive then the average standard deviation of split frequency. It could be that by tweaking the window size a better metric can be established. The window metric still shows promise, as demonstrated by the graphs below.
This week I also explored the idea of calculating the PSRF across chains instead of across runs. I accomplished this by changing the heating parameters on only one of the runs, so that it would in effect work like a chain. Running data on primates.nex for 100000 generations sampling every 10 generations, I found that the PSRF by default averages to 1.002. When the second run is given the heat of the other three chains, it changes to 1.006, 1.013, and 1.030. Thus, by calculating the PSRF by comparing with the least heated of the hot chains, it appears that one can get an alright diagnostic.

Next week, I will continue to look into the ASDoSF metric, determining why it is so sensitive to local change, and how it relates to the consensus tree. This task is challenging because of the cryptic code used to calculate it, but by continuing to look into it I hope to fully understand how it is calculated.

A final thing to look into next week is the Kim data set. I am extremely skeptical of the results found in its associated paper, it seems to me that they didn't run it for nearly long enough to get any sort of convergence. My goal is to get this data to converge according to multiple metrics, and then compare the outputted consensus tree with the one in the paper.

Thoughts: Using markov chains means we are taking dependent, not independent samples from the sample domain, making each sample less useful. Johnson shows a way that uses coupled chains to put a lower bound on how likely our samples represent independent draws from the domain. This might be useful, you can then use traditional statistics on the data. But may be too conservative.

2. Potential Applications and Pitfalls of Bayesian Inference of Phylogeny by JOHN P. HUELSENBECK, BRET LARGET, RICHARD E. MILLER, AND FREDRIK RONQUIST.

Thoughts: A good explanation of using Bayesian techniques. It also talks a lot about the need for choosing a good prior and why their method is faster than Maximum Likelihood methods. It also overviews their convergence methods.


Thoughts: An early implementation of Bayesian analysis. Good for pitfalls of MCMC, talks a lot about rearranging the tree.


Thoughts: See visually what convergence means, overview of several methods with good detail on Gelman and Rubin.

Location:

5. Comparison of methodologoes to assess the convergence of Markov chain Monte Carlo methods by Salaheddine El Aldouni, Anne-Catherine Favre and Bernard Bobee.

Thoughts: Comparison of methods of convergence.


Thoughts: Excellent overview of Mr. Bayes and really talks about what is happening behind the scenes, how its convergence metrics work.
7. **AWTY:** a system for graphical exploration of MCMC convergence in Bayesian phylogenetics by John A. A. Nylander et al 2007.

Thoughts: Takes MrBayes output and makes it graphical, uses tree distance, log likelihood, and split frequencies.


Thoughts: Foundational work, but very difficult to understand, with lots of math. Has discussion of error sources at the end, including issue of using dependent samples with floating point error.

9. **Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees** by Kevin Liu et al.

Thoughts: Combines sequence alignment and max likelihood computations together. I thought their claim about speed is hindered by bootstrapping though.