Group: Rei Doko

Parallelized phylogenetic tree search

Abstract

In the angiosperms (flowering plants) there are monocots, eudicots, and magnoliids. There is an interesting pattern in monocots, which is that in many of their genomes, their G+C content is bimodally distributed, or just generally elevated, unlike in the other flowering plants. This indicates some kind of shift in the rate of substitution biased towards G and C. To test this hypothesis, and potentially locate that shift, I would have to use a nonhomogeneous substitution model which can vary across lineages, which is not common. Thankfully, Phylogenetic Analysis by Maximum Likelihood (PAML) [1] is a set of programs which implement wonderfully rich models for estimating phylogenetic trees, branch lengths, substitution rates, etc. Baseml is one such program in PAML that supports nonhomogeneous substitution models. For that, however, tree topology search is disabled. I've already written a wrapper for baseml for basic tree topology search, but it's single-threaded. I want to parallelize the search this wrapper is doing.

Strategies

Essentially, tree search boils down to the following:

1. Select a starting tree topology and evaluate its likelihood
2. Pick a neighboring tree using some rearrangement strategy (NNI, SPR, or TBR)
3. Evaluate likelihood of the neighboring tree
4. Repeat unless the whole neighborhood of trees has been visited and no higher likelihood tree remains

The easiest thing to do would be to run multiple searches with different initializations in parallel, without sharing any memory, and then picking the best result. Once that works, I can use MPI to coordinate the processes a bit more. The idea being that there is a main process which keeps track of which tree topologies have been tried, their likelihoods, etc., and coordinating to the processes what topologies should be tried next.

Benchmarking & Optimization

There's two things I care about: tree accuracy and overall time. I would consider this project a success if any combination of the following happens:

- Tree accuracy/likelihood improves
- Tree accuracy stays the same, but the results are faster

I've already put together a pipeline for generating simulating datasets under nonhomogeneous models, so that gives me an easy way to test performance. I recently found another method that supports non-homogeneous substitution models called BppML [2], but I have yet to use that. It seems to work similarly to PAML though, so it could be an interesting comparison.

Group: Rei Doko

References

[1] Z. Yang, “PAML 4: Phylogenetic Analysis by Maximum Likelihood,” *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, Aug. 2007, doi: 10.1093/molbev/msm088.

[2] J. Dutheil and B. Boussau, “Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs,” *BMC Evolutionary Biology*, vol. 8, no. 1, p. 255, Sep. 2008, doi: 10.1186/1471-2148-8-255.