Markov-chain Monte Carlo sampling

New characteristics

In addition to recording contact distance sums, root degree, height, and number of leaves, I updated the program to also save the average branching and the ladder distance of each sample. Then I ran the Markov chain twice to gather data about these new characteristics. Under the thermodynamic distribution, I used an initial mixing time of 10,000,000, collecting 10,000 samples with a gap size of 10,000 — this run took an hour and forty minutes on my laptop. And under the uniform distribution, I used an initial mixing time of 100,000, collecting 10,000 samples with a gap size of 1,000 — this run took under a minute on my laptop. These parameters are based on the results of the Gelman-Rubin convergence diagnostic on samples obtained under the uniform and thermodynamic distributions (this is the "shrink factor" row in the two tables I sent during the past three weeks).

Plots based on these runs are attached. For the characteristics where I have access to the expected distribution under the uniform distribution, these plots compare the experimental values obtained under the thermodynamic distribution to the expected values under the uniform distribution. This is the case for contact distance sums, root degree, height, and number of leaves. Otherwise, the experimental values under the thermodynamic distribution are compared to the experimental values under the uniform distribution — I did this for average branching and ladder distance.

Kolmogorov-Smirnov test

I used the one-sample discrete KS test from the R package dgof to compare the values for root degree, height, and number of leaves under the thermodynamic distribution (using the parameters given above) to the expected values under the uniform distribution, and got p-values of 0, indicating that for each characteristic, the distribution of values is definitely different between the thermodynamic and uniform distributions.

RNAdb

I generated some rows of the CSV that will be used to create the new database. However, I found that most of the sequence files contain the N nucleotide code, which GTfold can't handle. I asked Emily how to handle this code but I haven't heard back from her yet. In the meantime, I imported the CSV file as a new table for the RNAdb website. Emily said that once the data is in the table, she will write/modify the Java GUI to make a front end to correspond with the database, so right now I'm not making further progress on this project.