Coupling Hidden Markov Models for the Discovery of *Cis*-Regulatory Modules in Multiple Species (Supplemental Notes)

Qing Zhou Wing Hung Wong

February 2007

Pilot study on the use of multiple chains in MultiModule

We compared two strategies for running MultiModule on the muscle-specific data set. In the first strategy, we ran two long chains, each for 5000 iterations. In the second strategy, we ran 10 shorter chains, each for 1000 iterations. Please note that the computational cost for these two strategies is exactly the same. We monitored the log-likelihood of the current parameters conditional on ortholog alignments along the iteration, i.e. $\log P(S|\Psi, A)$ as given in equation 14 in the main text. In order to make this statistic comparable, we fixed the alignment for each ortholog group as the same initial alignment built from ordinary multiple alignment methods. From Figure 1A, we see that the log-likelihood of the two long chains increases dramatically (> 300 for both chains) within the first 200 iterations, and then reaches a more stable phase in which it moves quite locally with mild changes in likelihood. The fact that these two chains never merge demonstrates the multimodality of our target distribution. From the last 500 samples generated from the 10 short chains, one sees that these combined samples form a mixture of at least five distinct local modes in the likelihood space (Figure 1B), whereas each of the long chains was trapped in one of them for a long time. In this sense, running multiple short chains enhances the chance of exploring more local modes with comparable computing cost.

Now the question is how to combine the samples and estimates from multiple chains. We find that the estimation of P_m, the posterior module probability, is quite consistent across different runs (see Figure 4A in the main text). This is because module sampling is based on the summation of the probabilities of all possible motif combinations, which effectively smoothes the likelihood function in a great degree. Thus we developed the combined prediction in section 3.2 in the main text. With a population of short chains, each of them may capture some local structure of the joint posterior distribution quickly, in this case different motifs and their binding sites. Then major local structures are utilized in conjunction with average P_m to generate combined predictions.

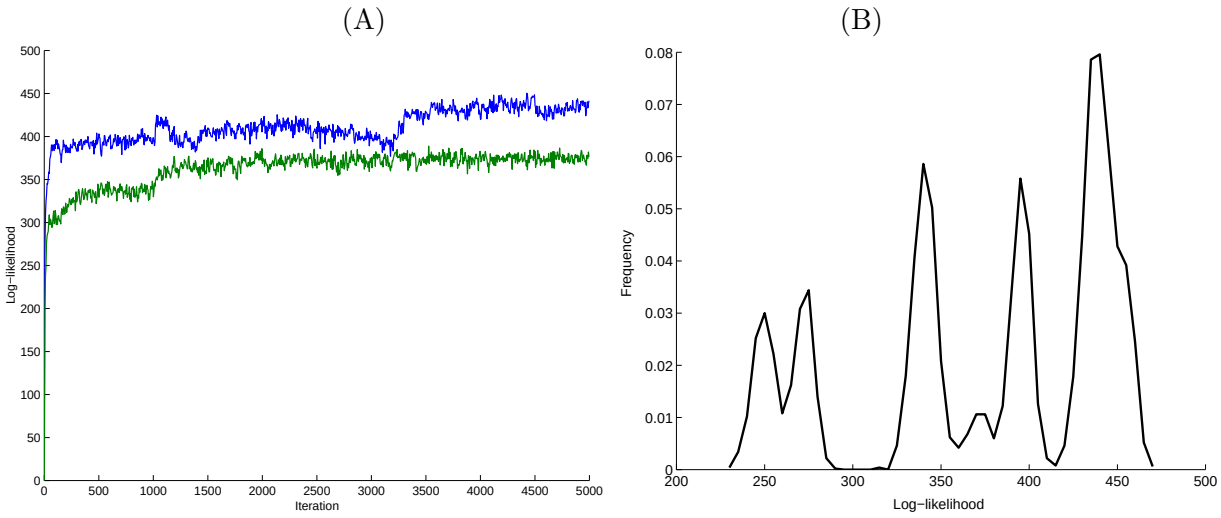

Figure 1: Comparison between running MultiModule with long and short chains. (A) The log-likelihood along 5000 iterations of two independent chains. (B) The histogram of the log-likelihood from 10 short chains. The values of the log-likelihood are all relative to a uniform baseline.

Programs and data sets

Executable programs in Linux and OSX systems of MultiModule are available freely for academic use at "http://www.stat.ucla.edu/∼zhou/MultiModule/". Test data sets used in this article can be downloaded at the same website.