Homework 2 (Part 2)

Due date: Dec 10 24:00, Monday

Gibbs Sampler for DNA motif discovery:

(1) Assume that DNA is a mixture of motif sites and background nucleotides. The motif sites are generated according to a probability matrix $\mathbf{\Theta}$ and the background nucleotides are generated according to a background probability vector θ_0. The length of the motif is W. W is known, but $\mathbf{\Theta}$ and θ_0 are unknown. Assume that there are N DNA sequences and each sequence has exactly one motif site. Let $\mathbf{A} = (A_1, A_2, \ldots, A_N)$ be the location indicators of the motif sites. Derive a Gibbs sampler to find the motif sites after collapsing $\mathbf{\Theta}$(i.e., provide an algorithm that samples $\mathbf{A}$ and θ_0 after integrating out $\mathbf{\Theta}$ analytically). Please notice a "*shift mode*" problem which is discussed in the paper "CollapsedGibbs_Liu_1994_JASA" and the solution is highlighted for you.

(2) Implement your motif-finding sampler using "*hw2_part2_data.txt*." The data contains 30 DNA sequences, each starting with a ">" and a sequence name. First, run your motif-finding sampler in (1) by assuming that the motif length is 18 bp ($W = 18$). After burn-in, collect posterior samples for $\mathbf{A}$ and θ_0, and then use samples' posterior mode $\widehat{\mathbf{A}}$ to estimate $\mathbf{A}$ and samples' posterior mean to estimate θ_0. Furthermore, according to $\widehat{\mathbf{A}}$, estimate the motif probability matrix $\mathbf{\Theta}$. Finally, record the sequences covered by the motif site based on the $\widehat{\mathbf{A}}$, save them. Then go to the website `http://weblogo.berkeley.edu/logo.cgi` to create a sequence logo (a way to visualize motif) for the motif you found.