

140.778 Homework 3

November 30, 2010

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 Θ and the background nucleotides are generated according to a background probability vector θ_0 . The length of the motif is W . W and θ_0 are known, but Θ is unknown. Assume that there are N DNA sequences and each sequence has exactly one motif site. Let \mathbf{A} be the location indicators of the motif sites. Derive a Gibbs sampler to find the motif sites after collapsing Θ (i.e., provide an algorithm that samples \mathbf{A} after integrating out Θ analytically).

(2) Now assume that the motif length W is unknown, but it follows a truncated Poisson prior. Derive a Markov Chain Monte Carlo algorithm to sample A and W conditional on the observed sequences.

(3) Implement your motif sampler. Download the data hwdata3-1.txt from the course website. The data contains 30 DNA sequences, each starting with a ">" and a sequence name. First, run your motif sampler in (1) by assuming that the motif length is 18 bp. You can use the empirical frequencies of A, C, G and T in the homework data set as your θ_0 . Based on the \mathbf{A} obtained from the last iteration of your chain, estimate the motif probability matrix Θ . Collect the sequences covered by the motif site based on the last sample of \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.

(4) Run your motif sampler in (2) by randomly choosing a W to start.

(5) (For fun if you have time. This part will not be graded): Implement a Gibbs motif sampler without collapsing the Θ . Now you need to iteratively sample Θ given \mathbf{A} , and \mathbf{A} given Θ . Compare the convergence

rate of the sampler with collapsing and the one without collapsing using the homework data.