Project: implementing a Gibbs sampler

November 7, 2014

1. Implement a Gibbs sampler for finding a motif that occurs once in each of N input sequences, as follows:

   - The input sequences are read from a file, one line per sequence
   - The length ℓ of the expected motif is an input parameter
   - The motif is expected to be described by a position weight matrix of dimensions $4 \times \ell$
   - Given N putative occurrences of the motif, the likelihood of their being sampled from a PWM is given by integrating over the space of PWMs, as discussed in class. Use this likelihood as the probability (ie assume a uniform prior on all possible alignments of sequences).

   The Gibbs sampler works as follows:

   - Initialise a random configuration of one site per sequence (eg, all sites at position 0)
   - At each step, take one sequence at random, and sample a new position for it (keeping others fixed) from all allowed choices, weighted by the likelihood for each choice, as described above. Eg, if the length of that sequence is L then you have $L - \ell + 1$ possible choices, for each of which you have a likelihood. Normalise the likelihoods (so that they sum to 1) and pick from that distribution.
   - Repeat this at least 10 times per sequence. Then perform "simulated annealing", ie introduce a fictitious inverse temperature β which is increased slowly, so you are sampling from probabilities P^β (normalised!) instead of P.

   Test on the attached file of sequences, using $\ell = 10$.