

GCB/CIS535 HW3 Motif Finding due 10/7/05

1. (10 pts) For each of the following motif descriptions, design a positional weight matrix that will assign the highest possible probability to that motif. Assume all nucleotides have an equal background probability (note that there is no need to normalize the matrix with the background probabilities, since they're all the same). For each case, also report the score obtained from applying the matrix on the string "GGGG."

a. All four-nucleotide sequences containing a G in the 2nd position and either an A or a C in the 4th position (with equal probability).

b. CACAT and any four-nucleotide sequence ending with a G.

c. The sequences CACATGT, CACATGG, CTGNCCY. (Y means C or T, N means any four-nucleotide)

2. (10 pts) Many genomes contain an approximately equal proportion of As, Cs, Gs, and Ts. *Plasmodium falciparum*, the parasite responsible for causing malaria, has a particularly (A+T)-rich genome – that is, there are significantly more As and Ts in the *Plasmodium* genome than Cs and Gs. How would this fact affect the ability of the following motif finding techniques to return significant hits? As a way to be specific about your descriptions, compare how motif finding would function over the *Plasmodium* genome as compared with motif finding over a genome with equal proportions of bases.

a. Gibbs sampling

b. Expectation Maximization

3. (15 pts) The mammalian circadian oscillator is an intracellular mechanism composed of a set of interlocking transcription/translation feedback loops that complete one cycle each day. In mouse, the *Period* genes (*Per1* and *Per2*) and *Cryptochrome* genes (*Cry1* and *Cry2*) are at the center of the core feedback loop, which is required for a functional clock. These genes are transcriptionally activated by the basic helix–loop–helix PAS transcription factors CLOCK and BMAL1, which heterodimerize and bind to E-box enhancer (CACGTG) elements in the promoters of these genes. We'll try different motif finding software in this exercise to detect E-box enhancer.

a) From ENSEMBLE and MGI, find and retrieve the sequence of the 1000 base pairs upstream region of the following gene: *Per1*, *Cry1* and *Cry2*. Save them in FASTA format.

b) Use AlignAce to search for conserved TF sites for the same set of sequences. Set the

“Number of columns to align” to be 6; “Number of sites to expect” to be 6. What is the result you get? Which one(s) of the returned motifs do you think is Clock binding site?

c) Use MEME to search for conserved TF sites for the same set of sequences. Set “Number of different motifs” = 5, “Minimum number of sites” =2; “Maximum number of sites” =7; “Minimum motif width” =6; “Maximum motif width” = 7”; “Distribution of motif occurrences” = any number of repetitions. What is the result you got?

Tools:

1. Genome sequence search and retrieval

<http://www.informatics.jax.org/> (MGI: *Mouse Genome Informatics*)

<http://www.ensembl.org/> (ENSEMBL)

2. Motif finding tools

<http://atlas.med.harvard.edu/cgi-bin/alignace.pl> (AlignAce)

<http://meme.sdsc.edu/meme/website/meme.html> (MEME)