

GCB 535 / CIS 535
Fall 2004

Homework 3

Due in class, Wednesday, October 13, 2004

Blast theory

1. (8 pts) Blast plug-and-chug

a. Use a BLOSUM50 matrix to calculate the score for the second of the three MANSE GLUTATHIONE matches (p. 11 on the handout). You can get a BLOSUM50 matrix by going to <http://eta.embl-heidelberg.de:8000/misc/mat/blosum50.html> or doing a google search for "blosum50." (Note that the BLOSUM matrix given in the notes is a BLOSUM62 matrix, and moreover it has a couple of typos in it). Show your work if you want partial credit if your answer turns out not to be correct.

b. Use the formula given in class for the "bits" score, together with the fact that for the BLOSUM62 matrix, $\lambda = 0.320$, to check the bit score for the MAIZE GLUTATHIONE match given in the class handout (p. 11). Again, show your work.

c. Use the approximate relation E-score (= "Expect") = $-\log(1-P)$, where P is the P-value, to check the relation between "Expect" and P for the GTH_SILCO GLUTATHIONE S-TRANSFERASE match given in the handout (p. 12).

2. (6 pts) The "true" significance of any alignment depends on how well your problem fits with the assumptions that you make when choosing to use a particular alignment algorithm with particular parameters. List two different assumptions that you make when using a particular flavor of BLAST (doesn't have to be the same flavor for each assumption), and for each assumption, come up with a specific alignment task where BLAST leads to a misleading significance prediction, given the assumption you made, and explain why the significance prediction would be misleading. Your assumptions can be about the core mechanism of BLAST, the way significance scores are calculated, the values for the various parameters (gap, match, etc.), etc. Your tasks should be as specific as possible, for example "finding significant alignments to an intron sequence that has a high concentration of Cs and Gs." Be precise, but please don't write an essay.

3. (6 pts) The Blosom and PAM substitution matrices are all symmetric -- that is, for all x, y , $\text{score}(x, y) = \text{score}(y, x)$ where x and y are amino acids.

What does this fact mean about doing any alignment with two sequences given these symmetric substitution matrices? What if we decided that for some amino acid pair, let's say W and F, we specify a scoring matrix such that $\text{score}(W, F) > \text{score}(F, W)$. How would using such a scoring matrix affect doing your alignment? Give one biological justification for why you might want to use such a matrix. (Hint: think in evolutionary terms. What you come up with might not be "correct" with respect to how things actually work in biology, but that's ok).

Motif finding

4. (9 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 "AAAA."

- All four-nucleotide sequences containing an A in the 2nd position and either a G or a T in the 4th position (with equal probability).
- CCGA and any four-nucleotide sequence ending with a T.
- The sequences ACGT, TGCA, CTAG, and GATC.

5. (9 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.

- Consensus (Hertz and Stormo)
- Gibbs sampling
- Expectation Maximization

6. (12 pts) In the yeast high-affinity copper and iron transport system, **Mac1p** (also called **CuRE**), a copper-sensing TF, controls the transcription of *Fre1*, *Fre7*, *Ctr1* and *Ctr3*. In this exercise, we'll test the ability of a couple Motif finding tools to detect **Mac1** binding site for **Fre1** and **Ctr1**.

a) From SCPD, find and retrieve the sequence of the 500 basepairs upstream region of the following gene: *Fre1*, *Ctr1*. Save them in FASTA format. This can be done using "Retrieve promoter sequences" at SCPD. Then use predefined consensus to search for potential regulatory elements on the set of sequences you saved. What is the result you got? Does the TF binding sites returned contain Mac1(CuRE) sites?

b) Use AlignAce to search for conserved TF sites for the same set of sequences. Set the "Number of columns to align" to be 7; "Number of sites to expect" to be 10. What is the result you get? Which one(s) of the returned motifs do you think is Mac1 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" =10; "Minimum motif width" =6; "Maximum motif width" = 20"; "Distribution of motif occurrences" = Any number of repetitions. What is the result you got? Which one(s) of the returned motifs do you think is Mac1 binding site?

d) Binding sites for Mac1 have been identified experimentally for *Fre1* and *Ctr1*, you can find them here on SCPD: <http://cgsigma.cshl.org/cgi-bin/jz/getgene2?FRE1> (*Fre1*); <http://cgsigma.cshl.org/cgi-bin/jz/getgene2?CTR1> (*Ctr1*). Both contain 2 Mac1 sites. Based on these, which ones of the Mac1 sites you found in a), b) and c) are the real ones? (Beware of the difference in the indexing) What are the respective sensitivity and specificity values for the prediction for each gene using the two different methods?

Tools:

1. Genome sequence search and retrieval
 - <http://www.yeastgenome.org/> (SGD: *Saccharomyces* genome database)
 - <http://cgsigma.cshl.org/jian/> (SCPD: Promoter database of yeast)
2. Motif finding tools
 - <http://atlas.med.harvard.edu/cgi-bin/alignace.pl> (AlignAce)
 - <http://cgsigma.cshl.org/jian/HTML/searchputative.html>
 - <http://meme.sdsc.edu/meme/website/meme.html> (MEME)