CENG 465 – Introduction to Bioinformatics
Spring 2008-2009

Assignment #3 (for Biology or Genetics majors)
Comparison of BAliBASE2 Reference Alignments with Clustal-W

Due Date: April 29, 2009, 11:59PM

Assignment about Multiple Sequence Alignment

In this assignment, you are going to compare the entropy score of the "core blocks" in a given reference multiple sequence alignment to the entropy scores of conserved columns of Clustal-W alignments. Below are step by step details of the assignment.

1) You are going to retrieve the reference multiple sequence alignments from the BAliBASE 2.0 benchmark. You will use the "short" subset of the Reference 2 alignments at: http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE2/align_index.html#ref2 . There are 9 multiple sequence alignments in that set. For example the first alignment is the multiple alignment of the SH3 family and identified by the first sequence in the alignment: 1aboA. There are 15 sequences in the first alignment. The alignment can be viewed at: http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE2/ref2/test/1aboA_ref2.html or downloaded in RSF and MSF formats (if you want to use these formats you need to learn about them by yourself).
2) In the alignments, there are certain columns which are displayed as underlined capital letters. These are "core blocks" of the alignment. In other words, these are regions that are conserved well in that protein family. You are going to compute the "entropy" score for all the columns in each of these core blocks. In other words, if the total number of underlined columns is *n*, you will compute *n* entropy values for that reference alignment. For example, for the 1aboA alignment, you will compute 36 entropy scores for 36 columns. (Note that the column which mostly contains gaps in the final core block is also part of the core block and you need to compute the entropy score of that column by considering the gap symbol as a regular amino acid symbol). You may compute the entropy score manually (which will take time) or use any automation that you can manage (for example: by using Microsoft Excel)
3) Use Clustal-W (http://www.ebi.ac.uk/Tools/clustalw2/index.html) to generate multiple alignments for the same sets of sequences. You may get the sequences by either downloading the whole BAliBASE2 benchmark from the ftp site provided in their homepage or searching for the protein ids in the ASTRAL database at: http://astral.berkeley.edu/getseqs/ . (For example you may just enter 1aboA to retrieve the sequence of the first protein in the first alignment).

4) The Reference 2 alignment set of BAliBASE2 contains protein families aligned with a highly divergent "orphan" sequence. In other words, one of the sequences in the alignment is not so similar to the rest of the sequences. After computing the Clustal-W alignment for each set of reference sequences, use the "Scores Table" in the Clustal-W output to find the "orphan" sequence. The sequence with the lowest total score to the rest of the sequences will be the "orphan" sequence. Also, compute the entropy scores for the columns that have one of "*", ":", or "." symbols under the column. Compare these entropy scores to the scores you have computed for the BAliBASE reference alignments. Which one is better in general?
5) Use the WebLogo tool (http://weblogo.berkeley.edu/logo.cgi) to generate logo representations of the BAliBASE and Clustal-W alignments. Are the respective logos from the two different alignments similar? Which parts are different? Provide a comparison for the 9 reference alignments.

In your report you should give the orphan sequences you have found for each of the 9 protein families, give a comparison of the entropy scores you have computed for the reference alignments and the Clustal-W alignments, and provide a comparison of the logo representations of these alignments.

You may use your own judgment for any issue that is not specified clearly in this text.

Deliverables:

- A short report which contains a step by step description of the tasks that you have performed, your results, and your interpretation of the results.

Submission:
Submit your report as a plain text file (or if you want to use some formatting you may also submit a Word or PDF document) using the COW system.

Late Submission Policy:
Your final assignment grade will be penalized 20 points per late day.

CHECK THE NEWSGROUP REGULARLY FOR POSSIBLE UPDATES ON THE ASSIGNMENT.