Biological Data Analysis (CSE 182) : Assignment 4

Logistics

For the written part, please submit the answer either electronically, or with a hard copy.

Pattern matching

1. Build an aho-corasick automaton for a dictionary containing 3 words. Show all failure links, and transition links. Submit a sheet of paper with the automaton hand-drawn. The words are: CAMPERS, AMPERE, and AMINO (18pts.).

2. You are given the following: A database D (represented as a single sequence), a family F of 20 sequences, and a scoring matrix M (40pts.).

   (a) Design a profile based algorithm to find homologs of F in D. Submit a written (pseudo-code) description of the algorithm, and your reasoning on why it is appropriate.

   (b) Implement the algorithm, and apply it to finding novel homologs of F in the database D. Report the homologs you found in the output file.

   (c) Compute an empirical P-value for the homologs, by first computing a distribution of scores on a random database.

   (d) Repeat the experiment from the previous step with the new family F_2, containing 1000 sequences. Are the homologs of F_2 different from F? Explain your results.

3. One way to get rid of the discrepancy in the results of searching with results is to remove redundant sequences from the family. Design an algorithm *non-red* which takes a set F of sequences, a threshold $30 \leq T \leq 100$ and outputs a set $F' \subseteq F$ such that no pair of sequences in F' has percent sequence identity greater than T. There are more than one ways to do this, none of which are very good. Therefore, describe your algorithm in pseudo-code, and say if your approach is perfect, or describe the shortcomings of your approach (30).

4. (10pts) The PROSITE database describes a pattern (PS00342) for "Microbodies C-terminal targeting signal". The pattern is:[STAGCN] - [RKH] -[LIVMAFY]

   (a) Assuming that the Swissprot database is randomly generated with all residues equally likely (size: 78M), what is the number of hits you would expect to find simply by chance. Is this a useful pattern?

   (b) Download swissprot (http://www.uniprot.org/downloads) in Fasta format. Record the version number, and empirically compute the frequencies of each amino-acid. Use these frequencies to refine your estimate of the expected number of hits.

5. (2pts.) What language did you use? How much time did you take to do the assignment? Who did you discuss your homework with?