

EXAM I

L519: Bioinformatics: Theory and Applications
School of Informatics
Indiana University – Bloomington
March 24, 2003

due date: BEFORE lab on April 3, 2003

Working Directory Use `/gpfs/L519/exam1` on IBM SP. `/gpfs` on IBM SP and `/bigscr/bioinfo/` on Solar are temporary files systems and files on them may be lost in case of the system failure. *Be sure to make copies of your work* somewhere else.

Preliminary on clustering Clustering techniques apply *when there is no class to be predicted* but rather when the instances are to be divided into natural groups. Clustering divides instances into *groups* or *clusters* which may be exclusive, overlapping, probabilistic, or hierarchical. Correct clustering is a quite difficult task to achieve since we do not know how many clusters exist in the data. In some cases, the number of clusters k is fixed and we are asked to group instances into k clusters. Examples of such clustering techniques includes k -means clustering and Expectation-Maximization with k different distributions.

We will perform a clustering analysis using a set S of protein sequences. To make our life easier, the number of clusters is fixed to 3. Most clustering techniques assumes that there are features of an instance and clustering is done on the feature space. One difficulty in clustering sequences is that we do *not* know features – not to be confused with the sequence feature (like annotations) – of protein sequences. Thus, unfortunately, it is almost impossible to use existing clustering techniques directly for sequence clustering. This exam will attempt to cluster proteins sequences when the number of clusters are given using various sequence analysis techniques including BLAST, CLUSTALW, PSI-BLAST, PHI-BLAST, HMMER, MEME, etc.

Data : `/gpfs/L519/exam1/allthree`

PROBLEM 1: Seed set generation (25pt): A set of sequences is in the file `allthree` and the number of clusters is fixed to 3. We will tackle this problem as *classification* of sequences after selecting three seed sets (A, B and C) of sequences, each of which corresponds to a cluster.

1. Perform all pairwise comparisons using `blastp`.
2. Write an algorithm to select three seed sets of sequences (each set of n sequences), $A = \{a_1, a_2, \dots, a_n\}$, $B = \{b_1, b_2, \dots, b_n\}$, and $C = \{c_1, c_2, \dots, c_n\}$.
3. (A possible method) There is no guarantee that your algorithm will pick up the three seed sets correctly. One advice it to try several values for n and see if the following conditions are met for any value n .
 - (a) All pair of sequences in A shows a very high sequence similarity. Test this for B and C .
 - (b) There is no observable sequence similarity (by BLAST) between two sequences that belong to different seed sets.

PROBLEM 2: Classification using BLAST (25pt) For each unassigned sequence s , assign s to a cluster in a greedy fashion based on sequence similarity to sequences in the seed sets. An ideal case would be that s is matched to *all* seed sequences in a cluster, say A , while not matched to any seed sequence in two remaining clusters, say B and C . However, you will encounter ambiguity in assigning s to a cluster. *Define your own criteria.*

PROBLEM 3: Classification using HMMER (25pt) Build three HMMs, M_A , M_B and M_C , using `hmmbuild`. Perform the same greedy cluster assignment for each sequence s as in Problem 1. Refinement of the multiple sequence alignment using any tool including MEME is encouraged.

PROBLEM 4: PSI-BLAST search with PSSM (25pt: Bioinformatics major only) The `blastpgp` program allows us to build a position specific scoring matrix (PSSM) with `-C` option and restart search with a previous PSSM with `-R` option. In addition, users can control construction of PSSM. One obvious way is to use a higher cutoff score for matches and another is to input a carefully calibrated multiple sequence alignment for PSSM construction. You may want to try MEME to control the multiple sequence alignment for each seed cluster. See the exam web page for additional information.

1. Devise a method to build the best performing PSSMs and
2. Assign s to a cluster by performing three PSI-BLAST searches with the three PSSMs against the set of all remaining sequences.

Discussion Forum `exam1` on `oncourse.iu.edu`

Restriction on Discussion Students may help others to understand the exam better or to show how to run particular commands. However, personal communication related to the solution to this exam is prohibited. Use the discussion forum `exam1` on `oncourse.iu.edu`.

Submission A written document and all related programs should be turned in BEFORE lab on April 3, 2003. The exam will be accepted until 5pm on April 4 with 10% late penalty deduction. After that time, no exam will be accepted.

Directory and submission :

- Create a directory `/gpfs/L519/exam1/yournetwork-id`
- Put all scripts on `/gpfs/L519/exam1/yournetwork-id` with names `p#` where `#` denotes the problem number.
- For *each* problem, write a document that clearly describes your method and the result. Turn in the documents and put an electronic version on `/gpfs/L519/exam1/yournetwork-id` with a name `exam1.your-networkid.pdf` (suffix `pdf` can be changed to an appropriate type).
- Create the `README#` file for each problem where a line under `RUN` will execute your program without any modification.