Assignment #3 (Programming Assignment)

Comparison of Two Greedy Strategies for Multiple Sequence Alignment

Due Date: April 21, 2008, 11:59PM

Programming Assignment about Multiple Sequence Alignment

In this assignment, you are going to implement two greedy multiple sequence alignment strategies for aligning multiple protein sequences. You are going to evaluate the accuracy of the implemented techniques by comparing the alignment results on a subset of the BAliBASE benchmark 1.0. Below are short descriptions of the algorithms and the test procedure.

Standard Greedy Alignment

We have covered the greedy alignment technique in class. The algorithm is given below:

**Algorithm_Greedy_MSA**

Input: \( k \) protein sequences of different length

Output: A multiple alignment of the input sequences. All amino acids should be assigned either to a gap or amino acids of other proteins. (See the reference alignment dataset for sample alignments.)

1. Let \( S \) be the input set of sequences to be aligned
2. while (\( S \) contains more than one sequence/profile)
3. choose the two most similar sequences/profiles in \( S \) using Needleman-Wunsch pairwise alignment algorithm (use semi-global alignment)
4. let the most similar sequences/profiles be \( A \) and \( B \)
5. remove \( A \) and \( B \) from \( S \)
6. insert the profile \( P \) represented by the alignment of \( A \) and \( B \) into \( S \)
7. endwhile
8. report the alignment
Note that profiles do not contain sequence order (i.e., alignment) information, just the percentages of each amino acid at each position. Therefore, you need to also store the alignment along with the profile in order to be able to report the alignment at step (8).

**Triplet based Greedy Alignment**

The triplet based greedy alignment is a variation of the standard greedy strategy. At each iteration, instead of choosing the most similar pair of sequences/profiles, we choose the most similar **triplet** using the multiple sequence alignment of all possible triplets using dynamic programming (for 3 sequences). The score of a multiple alignment of three sequences A, B, and C using dynamic programming will be given by the value computed for the last cell of the dynamic programming partial scores table (which is a 3D table). Similar to the standard greedy technique you will use Needleman-Wunsch algorithm but will not penalize the terminal gaps (semi-global alignment). Below is a pseudo-code of the algorithm:

**Algorithm_Triplet_Greedy_MSA**

Input: $k$ protein sequences of different length

Output: A multiple alignment of the input sequences.

1. Let $S$ be the input set of sequences to be aligned
2. while $(S$ contains more than one sequence/profile)
3. if $(S$ contains less than three sequences/profiles)
4. let the two sequences/profiles be A and B
5. align A and B using Needleman-Wunsch pairwise alignment algorithm (use semi-global alignment)
6. remove A and B from S
7. insert the profile P represented by the alignment of A and B into S
8. else
9. choose the most similar triplet of sequences/profiles in S using the 3D Needleman-Wunsch alignment algorithm (use semi-global alignment)
10. let the most similar sequences/profiles be A, B and C
11. remove A, B, and C from S
12. insert the profile P represented by the alignment of A, B, and C into S
13. endif
14. endwhile
15. report the alignment

Again, note that profiles do not contain sequence order (i.e., alignment) information, just the percentages of each amino acid at each position. Therefore, you need to also store the alignment along with the profile in order to be able to report the alignment at step (15).
Parameters

You are going to use the BLOSUM62 matrix as the amino acid scoring matrix. Use linear gap penalty model with a gap penalty of -2. Do not penalize terminal gaps. In other words you will implement the semi-global alignment by dynamic programming (both 2D and 3D).

Comparison of the Two Techniques

You are going to test the two alignment algorithms on 12 test cases which are obtained from the BAliBASE benchmark 1.0. The test cases contain both input sequences and a reference alignment of the given input sequences. The reference alignment can be considered as the correct alignment of the input sequences. The test sequences and the reference alignments can be downloaded as a zip bundle from http://www.ceng.metu.edu.tr/~tcan/ceng465_s0708/Assignments/hw3_test_data.zip. The input sequences are given as .seq files in the zip bundle. Each line in a .seq file contains a protein sequence preceded by its protein name (separated by a tab). These .seq files will be processed by the techniques that you will implement and you will be generating your output alignments similar to the ones given in reference alignment .aln files. The format of the .aln file is similar to a .seq file. The only difference is the addition of gaps indicated by dots.

You are going to use two measures to compare the techniques, the Entropy Score and the Sum of Pairs Score. Compute the Entropy Scores and Sum of Pairs scores of the alignments generated by your programs and compare them with each other and with the scores of the reference alignments. To compute the sum of pairs score use the same scoring scheme (i.e., BLOSUM62 matrix with a linear gap penalty of -2). Which technique is better? Can they get close to the reference alignment scores? You will report your test results in a small report that you will provide along with the source code. Try to interpret your results and the reasons behind the results, do not just provide numbers.

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

Deliverables:
- The source code of your program(s). You may use any programming language.
- A short report containing your test results and your interpretation of the results.

Submission:
Send the deliverables as a zip bundle or as a tarball to your assistant Çelebi Kocair at celebi@ceng.metu.edu.tr by the due date. Use “[CENG465] HW#3 <Student_ID>” as the subject of your e-mail submission.

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.