

# 6.095/6.895 Fall 2005 - Problem Set 1

Due: September 26, 2005 at 8pm

If you feel that any question is ambiguous, state any assumptions you need to make and consult them with the TA during office hours (to be announced soon) or via email.

1. **Global/Local Alignments.** Consider the sequences: AGGTGAT and AGTAA, and the alignment parameters:
  - Match: 3
  - A/G or C/T Mismatch: -1 (transition)
  - Other Mismatch: -2 (transversion)
  - Gap: -4

Calculate the dynamic programming matrix and the best alignment when using:

- (a) The Needleman-Wunsch algorithm (global alignment)
- (b) The Smith-Waterman algorithm (local alignment)

If there is more than one best alignment, give all such alignments.

2. **Dynamic Programming.** Give a dynamic programming recurrence for computing the optimal global alignment of three sequences. You do not need to describe how to fill in the dynamic programming table. Assume that you have a function,  $s(i, j, k)$ , that will provide the score of aligning three nucleotides and/or gaps.
3. **Motif Finding.** Here, we will perform the exhaustive motif finding described in lecture and recitation. For this problem, we will define a motif as a sequence of six characters from  $\{A, T, G, C\}$ . You will need to use the two data files available on the website for this assignment. `allinter` contains all the intergenic regions of DNA for one species of yeast. `allintercons` lines up with `allinter` and contains a "\*" for positions conserved across several yeast species. We will count all motif matches in the genome and then only count motif matches that are completely conserved.

Fill in the template code at the three indicated areas and include a hard copy listing of only the code you added and the top 50 motifs of length six in decreasing order of appearance for each method. Also, give the top 50 most conserved motifs, where conservation is defined to be the number of conserved occurrences divided by the total number of occurrences.

How do the motifs ranked high on the basis of frequency alone compare to those with a high conservation ratio? On the basis of their sequence, which set would you say is more likely to represent biologically significant motifs and why? See the provided files `Yeast_known_motifs.txt` and `Yeast_motif_functions.html`. Do you recognize any known yeast motifs in your lists? What processes are they involved in?

4. **Random Projections.** In the fifth lecture we introduce the idea of random projections. This probabilistic tool allows us to perform inexact motif matching. We consider two sequences to match if a randomly selected subset of positions match. Here we will analyze some properties of random projections.

Given two sequences  $v$  and  $w$ , define the *Hamming Distance*,  $d_H(v, w)$ , as the number of positions in which  $v$  and  $w$  differ.

- (a) If the Hamming distance between two sequences  $v$  and  $w$  of length  $d$  is  $d_H(v, w)$ , what is the probability that  $k$  positions selected at random (with replacement) will agree?
  - (b) Plot the probability of two sequences having the same random projection when  $d = 10$  and  $k = 3$ . How does this plot change as  $k$  increases?
  - (c) We can also compute several random projections to compare sequences. Let  $l$  be the number of random projections we perform. Calculate the probability of  $v$  and  $w$  having at least one equal projection.
5. **6.985 Problem.** Find a journal or conference article published in the last five years that describes a new alignment algorithm or a modification to an old one. What heuristics does it use and what are the biological justifications for these? You can find appropriate articles by searching PubMed or computational biology journals such as bioinformatics. Major conferences in the area such as ISMB, RECOMB and PSB are also a very good source of articles.