

HW 1: due Tuesday, September 28th in class

For this assignment, all answers are to be submitted in hardcopy in class.

1. Readings:

- Clote and Backofen, Chapter 3.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

2. There are 64 codons (4^3) out of which 61 encode the amino acids that make up the proteins in our body. However, only 20 such amino acids exist. Many amino acids can therefore be encoded in more than one way. Write the most divergent sequence you can construct that translates into the same amino acid sequence, using the genetic code in reverse. What percent identity do you achieve at the nucleotide level?

3. Score the alignment

GA-CGGATTAG
GATCGGAATAG

where a match gets a score of +1, a mismatch 0, and a gap -1.

4. Find the optimal global alignment (and the resulting score) between GAGC and CCG, with the scoring system used in class (i.e., +1 for a match, -1 for a mismatch, and -2 for a gap).
5. Find the best local alignment between ATACTCTCCTAAG and GACTCGTAACGTAT, with the scoring system used in class. Also turn in the entire scoring matrix which the dynamic programming algorithm would compute (writing this out should take less than 15 minutes). If you need to break ties between local alignments with the same score, choose the longer subsequences.

6. * Suppose we wish to perform a local alignment between two strings s and t , but we wish to penalize gaps in s by -1 and gaps in t by -3 . What should the update rule in our $O(mn)$ algorithm be changed to?
7. You are studying pollen allergies, in particular you are considering *Cryptomeria japonica* the Japanese Cedar, the most common pollen allergy in Japan. Luckily the two major pollen allergens have been sequenced and their sequences have been deposited in SWISSPROT. (One of the main protein sequence databases). However, their crystal structures have not been solved.
 - (a) Find the accession number of the two pollen allergen sequences from *Cryptomeria japonica* in the SWISSPROT databases, and get the sequences in FASTA format.
 - (b) Run BLAST on these sequences against the PDB to find possible homologs of solved proteins. Change some of the default settings (gap penalties, which PAM or BLOSUM matrix is used)– does the set of “hits” change?
 - (c) Look up the folds of the solved structures you found in either SCOP or CATH. What is the shape of their folds?
 - (d) Based on this information, can you make a prediction of the fold of these pollen allergens? How confident are you, and why?
 - (e) Find other (unsolved) sequences that display sequence homology to the cedar pollen allergens. Can you make predictions of any of their folds? How confident are you, and why?
 - (f) Do a full text search in SWISSPROT for “pollen allergen” and find a pollen allergen sequence from a different organism of your choice (such as ragweed pollen). Repeat all the previous parts of this question on the new pollen allergen you chose.