

Problem Set # 2: Due Wednesday, Feb. 18.

These are open book problem sets: any library or web resource is useable, but please reference those you use. I also don't mind if you talk with other students in a general way, but I want you to finish (write up) the problems out themselves.

1. Distinct Regions of Similarity. Write an algorithm to detect the maximal scoring local alignment and then the maximal scoring alignment distinct from the first, and then the maximal distinct from the first two described, etc. [This is due to Waterman-Eggert; you should follow the pattern of the Smith-Waterman algorithm for local alignment.]

The next question, and question 4, have already been begun by those in the 548 lab, but not written down. This is to ask you to make a report of some analysis you perform on these cases. If you are not in the lab, you still have to do these, but you can speak with the lab students about the problem. They were basically on their own, with a little help about using the web servers. We will keep using these sequences for basic computational examples.

2. Possible Evolutionary Distances, I. Use the sequences given for various forms of cytochrome-c (file *ccdata* in the 548 Resources directory for the course), and run some of these pairwise through sequence alignment programs with p -value functions. Adjust the values of the affine gap penalty functions, and use *various* of the BLOSUM scoring matrices. [This can be done from the unix command line in the .math network via the application *pvlocalS* from the Waterman group which is supported by the local system. That function is not supported on the USC web-server.] The point of the exercise is to see whether we can detect any "movement" in the p -value as we go further apart evolutionarily in the represented species. This may not be a "clean" exercise, i.e., don't expect huge differences to jump out at you aggressively from the various runs. A nice extension of this, if someone in your group has any competence or interest in this direction, is to see whether you can compare the regions of high local alignment with known function areas in cytochrome-c. It might help to locate these in the structure databases which can give you displays of this sort. E.g., you could try <http://www.ncbi.nlm.nih.gov/Structure/> or <http://www.rcsb.org/pdb/>, or check out the sections on cytochromes in Branden and Tooze or Garrett and Grisham (or any other of the standard biochemistry references).

This is a very *ad hoc* or by hand method to estimate crudely evolutionary distances. Such estimates are regularly carried out by some practitioners of phylogeny, although the assumptions necessary to make such calculations are sometimes controversial. Part II of this exercise will come much later.

3. Penalties and Global vs. Local Behavior. Consider the pairwise alignment problem for coin flips, that is, we have an alphabet of two letters, H and T , and a scoring scheme

$$\begin{aligned}s(H, H) &= 1, s(T, T) = 1, \\s(H, T) &= -\mu = s(T, H), \\s(-, H) &= s(-, T) = s(H, -) = s(T, -) = -\delta.\end{aligned}$$

We are going to examine the role of various values of μ, δ . We will try to align a sequence x of H 's and T 's, with a reference sequence y which will be string of nH 's, where n is greater than the length of x .

a.) What is the local alignment score $S(x, y)$, if $\mu = 0, \delta = 0$? Does this pick out any local features of the sequence x ? How does it depend probabilistically with the length of x , if we assume x is a random sequence as earlier in the course?

b.) What is the local alignment score $S(x, y)$, if $\mu = +\infty, \delta = +\infty$? Does this show any local features of the sequence x ? How does the score depend probabilistically with the length of the random sequence x ? [You might want to consult Waterman's text, chapter 11, section 5, besides the class notes.]

The difference in behavior between examples a.) and b.) is part of what is called a *phase-transition* in the parameter space of possible penalty values μ, δ . For more detail, see Waterman, section 11.6.1, or the three papers of Arratia-Waterman (1985, 1989, 1994) referred to in the Waterman textbook. A very brief sketch of this is on the web page in the notes for "Significance". I asked this question loosely and in passing in the notes. This is to give more shape to the question.

4. tRNA and rRNA. This problem is based on the *E. coli* tRNA and rRNA data I showed you (and which is in the 548 resources directory: ECORRD – this is a GenBank locus name, you can use this to locate what we need in GenBank with this ID – and the file *EctRNAdata*) in connection with our discussion of p -values and significance of alignment scores. The data you received in class about significance or standard deviations were computed with a scoring scheme, as noted in class, giving 1 for an exact match, $-\mu = -0.9$ for each mismatch, and linear indel penalty $-\delta = -2.1$. This problem is a bit imprecise, but I would like you to vary the parameters μ and δ and see how the p -values of best local alignment score change. The naive idea is simply that the effect we are looking for (traces of a "pre-historic" RNA-based biosphere which conjecturally preceded the current DNA/protein biosphere) is very distant, evolutionarily, and involves nucleic acids directly as opposed to protein products. Nucleic acid evolution is much more indistinct, protein evolution much more conservative. Thus, one would hope to relax the penalties to allow for more distant signals (permitting more mismatches and indels), but then your given data, if it contains such a putative signal, is competing to be heard in a "noisier" arena. Do the subsequences of optimal score vary with change of parameters? Don't do this for all the tRNA sequences you have been given! Just pick out three tRNA sequences to compare to the rRNA, of which one tRNA should be that for cysteine. Work on varying the parameters for these three. Negative results (that is, "nothing interesting happened") are acceptable here, of course, but you do have to document what you tried and your interpretation of that.

For more information on the "RNA pre-history" hypothesis and sequence analysis, have a look at Sean Eddy's web page, in particular at some of his recent publications:

<http://www.genetics.wustl.edu/eddy/publications/>

The standard print reference (which has recently been updated) is *The RNA World*, 2nd edition, eds. R. Gesteland, T. Cech and J. Atkins, Cold Spring Harbor Press, 1999. A later problem will return to the general problem of establishing a PAM-like scale for nucleic acid data.

5. Alternate Architecture. Consider the CpG islands model in chapter 3 of Durbin.

Describe the HMM architecture of a model of CpG islands which has only two states in the hidden Markov process. Discuss this model: can you compute its transitions from data given in the textbook example? Do you think it would be as effective as the one in the textbook at discriminating between islands and non-islands? Why or why not, of course. (You will not really be able to settle this without the full raw data they used.) Since you don't have this text necessarily, I have put some of it on the web so you can make sense of this question.