

CS196-1 HW4

Due: Thursday, March 6th 2:30pm

Problem Reading

- *Unintelligent Design*, Charles Siebert.
<http://www.discover.com/issues/mar-06/cover/>
- *Non-Coding RNA Genes and the Modern RNA World*, S.R. Eddy.
<http://www.psi.utoronto.ca/%7Eefrey/tcb/papers/ncRNAReview.pdf>
- *How to Lie with Statistics*, Chapter 4.

Problem Alignment History (30)

- *Viral src Gene Products are Related to the Catalytic Chain of Mammalian cAMP-Dependent Protein Kinase*, W. C. Barker and M. O. Dayhoff.
<http://www.pnas.org/cgi/reprint/79/9/2836.pdf>

This is a landmark study in using Sequence Alignment. Read this paper, find the sequences it uses, and BLAST them against each other. Compare the results in the paper to the results you find.

Problem Mimivirus (40)

To provide some background for this problem, please read the article on “Unintelligent Design” located online: <http://discover.com/issues/mar-06/cover/>

We would like you to support some of the claims presented in the article using bioinformatics tools. More specifically, we would like you to show that the Mimivirus is related to:

1. the Avian Bird Flu virus (Influenzavirus A)
2. Ebola virus (ebolavirus)
3. one other virus of your choosing
4. a bacteria, or in general, a more complex cellular organism

The entire genome of the Mimivirus is conveniently located in the BLAST database. (You can find it by searching for “Mimivirus” under the category “Genomes” on NCBI’s main page.) The Influenzavirus A genome and the ebolavirus genome are also in the BLAST database. We would

like you to choose amino acid and/or protein sequences from the Mimivirus genome that you can use to show that the Mimivirus is related to both other viruses and bacteria. Feel free to use the Mathematica code, dot plots, BLAST, or any other tools we have introduced so far. You will be graded on the strength of your argument.

Problem Affine Gap Penalty (30)

Write a program that performs Global Alignment with an affine gap-penalty function. It should take as input sequences s_1 and s_2 , similarity matrix M , and gap-opening penalty w_g ; and it should output the alignment o with the highest score s .

The true DNA sequence of BRCA1, including introns, is more than eighty thousand bases long. If you tried to allocate a dynamic programming matrix of 4-byte floats to compute the optimal alignment for BRCA1 and its roughly six-thousand-base cDNA sequence, you would need more than 1.7 gigabytes of memory (and you would need to write every byte of it as well – taking quite some time).

Therefore we have posted on <http://www.cs.brown.edu/courses/cs196-1/hw4.html> a “simulated” BRCA1 sequence with introns that we generated, with a total length of about fourteen thousand bases. This is short enough that all three matrices you need to use an affine penalty function can be allocated in less than one gigabyte of RAM.

Run these two sequences against each other in your program. Tweak the scoring parameters to get the “best” results and display your results. Explain what makes your results the best.