

CS195L HW4

Due: Thursday, March 3rd 2:30pm

Problem Reading

- *Unintelligent Design*, Charles Siebert.
<http://discovermagazine.com/2006/mar/unintelligent-design>
- *Non-Coding RNA Genes and the Modern RNA World*, S.R. Eddy.
<http://www.psi.utoronto.ca/%7Eefrey/tcb/papers/ncRNARReview.pdf>

Problem Alignment History (20)

- *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/COBALT them against each other. Attempt to reconstruct the same alignment as in the paper, and document your methods. Find any significant differences between your alignment and Dayhoff's and propose a biological or algorithmic reason for the difference.

Lastly, suggest a formalization for the description of the RELATE program described in the paper. Write pseudocode for how you would implement it. Then propose a modification of the algorithm that would make it produce arguably *more* biologically relevant results.

Problem Mimivirus (25)

To provide some background for this problem, please read the article on "Unintelligent Design" located online.

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 each of:

1. the Avian Bird Flu virus (Influenzavirus A)
2. Ebola virus (ebolavirus)
3. a bacteria of your choosing
4. a more complex cellular organism, possibly animal

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 alignment code, dot plots, BLAST, and NCBI Genome/Nucleotide/Proteins/Genes/PubMed databases, etc. You will be graded on the strength of your argument – we will approach it with skepticism! Your argument should include at least statistics, similarity, and biology as bases for plausibility.

Problem Affine Gap Penalty (25)

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 the website 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 ($14000 \times 6000 \times 4 \text{ bytes} \times 3 = 961\text{MB}$).

Run the TA DNA against the BRCA1 cDNA with your program. Tweak the scoring parameters to get the “best” results and display your results. Argue what makes your results the best (why the parameter choices are good). Then, explain the biological significance of this highest scoring alignment o in two sentences or less.

Your grade is based on the algorithm design and correctness primarily. 5 points for the argument above. Since there are a variety of languages/environments being used, performance is not assessed; however, if your program requires more than 2.5GB of RAM (about the amount free on a Sunlab computer) to execute, we will subtract points. It’s important to note that not all programming environments may be able to handle the high memory demand for this problem.

Problem Hirschberg Algorithm (30)

Apolipoprotein L, 3 (APOL3) and apolipoprotein L, 4 (APOL4) are believed to be paralogs. Two genes are paralogous if they are related by a gene duplication event. One hypothesis concerning gene duplications is that the new copy of the gene can evolve freely while the old copy can preform its

normal functionality; however the genes should still be related at the sequence level. To examine this hypothesis, we want to align these two gene sequences and analyze the results. APOL3 and APOL4 are very long genes and are difficult to align using the standard Needleman-Wunsch algorithm (using 1 Byte to represent each cell would yield a DP table of $> 3.2\text{GB}$).

In order to align these two sequences, you are tasked to write a program that performs Global Alignment using Hirschberg's algorithm. Your program should take as input sequences s_1 and s_2 and similarity matrix M and should use no more than $O(n)$ space. Please implement Hirschberg's algorithm **without** affine gap penalties. We've uploaded the APOL3 and APOL4 gene sequences to the homework page. Analyze the alignment. What would you conclude about the relationship between the two sequences? Is global alignment the appropriate alignment algorithm to use in this case?

The sequence alignment applet linked from the resources page (<http://drp.id.au/align/2d/AlignDemo.shtml>) is a great resource for this problem. Because space is less of an issue with this algorithm, Mathematica implementations earn extra-credit (and probable Pastiche pie award!!!!).