BioE/MCB/PMB C146/246, Spring 2005

Problem Set 4: Heuristic Alignment and Statistics

Due 14 February 2005 by midnight to lfl@compbio.berkeley.edu.

1. 6 points

Name 3 arbitrary thresholds used in the FASTA method.

2. 6 points

Name 4 ways in which BLAST2 is a better algorithm than BLAST.

3. 6 points

Why do we need to assess the significance of alignments?

4. 6 points

Why are heuristic search algorithms acceptable for database searching?

5. 10 points

What is the extreme value distribution? How is it used in database searching? How and why are its parameters derived analytically and empirically (be specific)?

6. 10 points

The best-scoring match for a BLAST search has score 8.0 bits. Given the mean (1.3) and standard deviation (2.296) for the distribution of all BLAST results, compute the *p*-value for the best-scoring match.

7. 10 points

Using the S_{ij}'s from BLOSUM62 and the amino acid frequency table found at http://www.tiem.utk.edu/~gross/bioed/webmodules/aminoacid.htm, choose the correct value of λ and u from the three choices below. Assume that K=0.15 and you are comparing two sequences of length 100. (Use the analytical, not empirical, method.)

1. $\lambda = 0.19$, u = 8.97
2. $\lambda = 0.32$, u = 8.45
3. $\lambda = 0.51$, u = 7.99

8. 10 points

You perform a database search in November 1986 against SWISS-PROT release 3 (just SWISS-PROT, not TrEMBL) which had 969,641 amino acids in 4,160 sequence entries. You find a matching sequence using BLAST with a reported E-value of 4.1×10^{-7}. Using FASTA, the same sequence matches had a reported E-value of 7.9×10^{-8}. What would be the E-values of alignments of these two sequences using the same versions of BLAST and FASTA if the search were performed in June 1992, March 1995, and November 2001?

9. 10 points

Imagine you performed a database search with the sequence NICETESTPEPTIDE using BLAST and BLOSUM62 and gap penalties of 11 for existence and 1 for extension and parameters gapped-K = 0.041, gapped-lambda = 0.267, effective database length = 786,882,512. You received the following alignment with a score of 21.2 bits:

```
ICETESTPEPTID
ICENDNLPKPVLD
```

Ignoring end penalties and other second order effects, what would the E-value be?

10. 6 points

Describe how you could make a BLAST search more sensitive (at the expense of being slower) by modifying BLAST’s parameters. (HINT: Look at the command-line options for BLAST). Give at least 3 options.

11. 5 points

The best substitution matrix for Smith-Waterman comparisons of distant homologs is often BLOSUM45. Which BLOSUM matrices would you use for BLAST comparisons of distant homologs? Why?

12. 5 points

When will an optimal alignment not be found by FASTA? By BLAST?

13. 5 points

If looking for similar protein-coding regions in two unannotated genome (nucleotide) sequences, what BLAST program would you use? Why?

14. 5 points

Name two features of genomes that are not protein coding regions. What BLAST programs would you use to find these similar features in two unannotated genome sequences? Why?

Extra Credit: 5 points

What is wrong with question 8?