BioE/MCB/PMB C146/246, Spring 2005
Problem Set 4

1. Some acceptable answers:

- The length of K-tuples used for the initial search phase.
- Retention of only the 10 highest-scoring K-tuples.
- Score cutoff of greater than one standard deviation from the expected score of random alignments.

2. Some acceptable answers:

- Considers gapped regions, which extend its biological relevance.
- Ungapped extension via score-bounded dynamic programming.
- Lowering the similarity threshold from T = 13 to T = 11 leads to higher sensitivity.
- Requirement for two hits along the same diagonal speeds up the search time.
- Use of EVD statistics better re¤ects the score distribution for aligments.

3. Assessing the signi£cance of alignments gives us a measure of how likely a given match would appear by chance. With large databases, one may obtain spurious matches that do not imply any common evolutionary ancestor. By assessing the signi£cance of an alignment, we can determine how likely it is that a given much is spurious.

4. Heuristics are able to identify most biologically relevant alignments. First, although they may miss some optimal alignments, these optimal alignments are unlikely to appear by evolutionary processes. For example, a long sequence of perfectly alternating matches and mismatches would not be found by FASTA, yet this unlikely to arise evolutionarily. Second, once a region of similarity is identi£ed, the programs are able to go back and perform on optimal alignment over this region and assess the signi£cance of the alignment based on the optimal alignment, rather than the heuristic alignment.

5. The extreme value distribution (EVD) represents the distribution for the maximum of a large collection of random variables. In the case of sequence database searching, the random variable is best score (i.e., sum of substitution scores) for a local alignment.

The EVD is used to assess the signi£cance of a high-scoring sequence pair (HSP). The expected number of HSPs (E-value) with a score greater than or equal to S is given by the formula $Kmne^{-\lambda S}$, where K and λ are parameters associated with a substitution matrix, m and n are the lengths of the database and query sequence, respectively. Lower E-values re¤ect more signi£cant hits to the database.

The EVD parameters **u** and λ represent the location and scale parameters, respectively. They can be computed analytically for a given matrix using the formula:

$$\lambda = x \;\; \text{s.t.} \;\; \sum_i \sum_j p_i p_j e^{S_{ij} x} = 1$$

and $u = \ln Kmn/\lambda$ where K is a function of S_{ij}, p_i, and p_j.

The EVD parameters may also be determined empirically using an empirical null distribution, consisting of thousands of random alignments. Suppose that μ and σ represent the mean and standard deviation of an empirical null distribution. Then **u** and λ can be computed using the formulae:

$$\begin{aligned} \mathbf{u} &= \mu - \frac{\gamma}{\lambda} \\ \lambda &= \frac{\pi}{\sigma\sqrt{6}} \end{aligned}$$

6. We know that $S = 8.0$, $\mu = 1.3$, $\sigma = 2.296$, $u = \mu - \frac{\gamma}{\lambda}$, and $\lambda = \frac{\pi}{\sigma\sqrt{6}}$

$$\lambda = \frac{\pi}{\sigma\sqrt{6}} = \frac{1.2825}{\sigma} = \frac{1.2825}{2.296} = \mathbf{0.559}$$

$$u = \mu - \frac{\gamma}{\lambda} = \mu - \frac{\gamma\sigma\sqrt{6}}{\pi} \; = 1.3 - (0.45 * 2.296) = \mathbf{0.267}$$

Then compute the p-value:

$$P(S \geq x) = 1 - e^{-e^{-\lambda(x-u)}} = 1 - e^{-e^{-0.559(8.0-0.267)}} = \mathbf{0.0132}$$

7. See which λ satis£es

$$\sum_i \sum_j p_i p_j e^{S_{ij}\lambda} = 1$$

to £nd that $\lambda = 0.19$, then plug λ into $\frac{\ln Kmn}{\lambda}$ with $K = 0.15$ to get $\mathbf{u} = 38.49$.

8. Sequences in SwissProt:

Year	Sequences	Amino acids
1986	4,160	969,641
1992	25,044	8,375,696
1995	43,470	15,335,248
2001	101,602	37,315,215

BLAST: $E = KNne^{-\lambda S}$

$3.2 \times 10^{-7} = 969641Kne^{-\lambda S}$, so $Kne^{-\lambda S} = 3.3 \times 10^{-13}$

1992 $\mathbf{2.76 \times 10^{-6}}$

1995 $\mathbf{5.06 \times 10^{-6}}$

2001 $\mathbf{1.23 \times 10^{-5}}$

FASTA: $E = KDmne^{-\lambda S}$

$7.2 \times 10^{-8} = 4160Kmne^{-\lambda S}$, so $Kmne^{-\lambda S} = 1.7 \times 10^{-11}$

1992 $\mathbf{4.33 \times 10^{-7}}$
1995 $\mathbf{7.39 \times 10^{-7}}$
2001 $\mathbf{1.73 \times 10^{-6}}$

9. The formula for a BLAST E-value, given a bit score S^*, is $E = mn2^{-S^*}$ (remember that K and λ are already included in the bit score!). So, the E-value is $786,882,512 \times 15 \times 2^{-21.2} = 4899.7$. Not very signi£cant!

10. Some BLAST options:

-f	score threshold	Use a lower threshold
-W	word size	Use a shorter word size
-S	cutoff score	Use a lower cutoff score
-A	mutliple hits window size	Use a higher window size

11. To compare the same pair of distant homologs, the best BLOSUM matrices for BLAST are higher (typically BLOSUM62 BLOSUM80) than the BLOSUM45 used for Smith-Waterman. These matrices have greater relative entropy, which allows the triplets of residues seen by BLAST to be more signi£cant. The high BLOSUM matrices are particularly important for BLAST with ungapped alignments to ensure the possibility that an ungapped (and thus usually short) alignment is signi£cant.

12. An optimal alignment will be missed by FASTA or BLAST when it doesn't £t the model embedded in the heuristics.

For FASTA, this primarily means that you must have two identities on a diagonal without too much intervening sequence. FASTA can also fail if the optimal alignment involves a gap that is too large to be allowed by the heuristics.

For BLAST, the two sequences must have triplets of residues that are substantially similar (above a threshold) and these must be extendible into seeds above a particular threshold. For BLAST2, these must be on a diagonal within a speci£ed distance. For BLAST1, the optimal alignment will only be found if it is ungapped. For BLAST 2.2, so long as a signi£cant alignment between the two sequences is found by the heuristics, the optimal alignment will be found by the subsequently performed full Smith-Waterman alignment.

13. You should use TBLASTX, since it translates both nucleotide sequences in all six reading frames and compares them at the level of protein sequence. Protein sequence comparisons are more informative than nucleotide comparisons, when possible, because the protein sequence changes more slowly than the underlying DNA sequence.

14. Some options: regulatory regions, introns, RNA genes. They should be searched using BLASTN, which compares sequences at the nucleotide level. Translating non-coding regions into protein is nonsensical.