

## Chapter 7 SELF-TEST Alignments and database searching

This test covers material in chapters 6 and 7

1. The time taken for a sequence alignment algorithm is thought to scale as  $O(N^3)$ , where  $N$  is the sequence length. For sequences of length 500 it takes approximately 2.5 seconds.

Which of the following statements is true?

- A. The algorithm must be a local alignment algorithm rather than a global alignment algorithm.
- B. The algorithm will be impractical to run because the memory storage required will be too large.
- C. For sequences of length 2000 it is likely to run in approximately 10 seconds.
- D. For sequences of length 2000 it is likely to take more than 2.5 minutes.

2. Consider the two DNA sequences CAGCAT and CGACAT. These sequences are aligned using a global alignment algorithm where the score is 1 for a match and 0 for a mismatch and there is a penalty of 0.2 for each gap. Which statement is true?

- A. There are no gaps in the optimal alignment.
- B. The score of the optimal alignment is 4.0
- C. There are two optimal alignments with equal score.
- D. The score of the optimal alignment is 4.8.

3. The affine gap penalty function used in sequence alignments has a cost  $g_{open}$  for gap opening and a cost  $g_{ext}$  for gap extension. Which of these statements is true?

- A. When  $g_{open} \gg g_{ext}$ , the gaps should be fewer but longer than when  $g_{open} = g_{ext}$ .
- B. The distribution of gap lengths produced by the alignment algorithm is independent of  $g_{open}$ .
- C. When the affine gap penalty is used instead of the linear gap penalty (where the total cost of a gap is proportional to the length of the gap) the scaling of the global alignment algorithm becomes  $O(N^3)$  instead of  $O(N^2)$ .
- D. An affine gap penalty function would be appropriate for aligning DNA sequences but not protein sequences.

4. Which of these statements about multiple alignments is correct?
- A. It is not possible to define a dynamic programming algorithm to align more than two sequences.
  - B. The guide tree in CLUSTALW is produced using a distance matrix method.
  - C. The guide tree in CLUSTALW does not influence the final alignment.
  - D. All three of the above statements.
5. The expected distribution of scores from the BLAST algorithm is an Extreme Value distribution because ...
- A. ... it uses ungapped alignments.
  - B. ... it is derived as an approximation to a dynamic programming algorithm.
  - C. ... it returns the highest scoring match from a database.
  - D. ... it uses a probabilistic alignment model.

6. The distribution  $F(S)$  of top-hit scores for a database search algorithm based on local sequence alignments follows an extreme value distribution

$F(S) = \lambda e^{-\lambda(S-u)} \exp(-e^{-\lambda(S-u)})$  with  $\lambda = 0.69$  and  $u = 22$ . A couple of years later, the database has doubled in size. The distribution of top-hit scores should now be given by an extreme value distribution with

- A.  $\lambda = 0.81$  and  $u = 22$ .
- B.  $\lambda = 1.38$  and  $u = 24$ .
- C.  $\lambda = 0.69$  and  $u = 44$ .
- D.  $\lambda = 0.69$  and  $u = 23$ .

7. Which of the following is correct?

- A. If a database search algorithm is set to return a large number of results with high  $E$  values, it is likely to have high sensitivity and low selectivity.
- B. The  $E$  value of a database search algorithm is sensitive to the statistical model used to calculate the probability of chance matches.
- C. The z-score is not useful as a measure of significance of search algorithms based on local sequence alignment because the z-scores apply to normal distributions not extreme value distributions.
- D. All of the above.