

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.

A – wrong – we do not know this from the information in the question. In any case, global and local alignment algorithms usually scale in the same way.

B – wrong – there was nothing about memory storage in the question.

The expected time for sequences of length 2000 will be $2.5 \times (2000/500)^3 = 160 \text{ sec} = 2.67 \text{ mins}$, *i.e.* D is correct.

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.

Let's think of a few possible alignments and calculate their scores.

CAGCAT
CGACAT score 4.0

C-AGCAT
CGA-CAT score 4.6

CAG-CAT

C-GACAT score 4.6

Common sense says there are no other ones we need to consider (although you are welcome to calculate the full scoring matrix and check). Thus there are two optimal alignments with score 4.6, so C is correct.

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.

A – correct

B – wrong – g_{open} has an important effect on gap lengths

C – wrong – the point of the affine gap penalty function is that the scaling is the same as for a linear gap penalty, even though the scoring system is more realistic.

D – wrong – its hard to think of sensible wrong answers sometimes...

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.

A – wrong – it is possible, but the algorithms are usually too slow to be useful

B – correct

C – wrong – yes it does – see the chapter text.

D - wrong

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.

A – wrong – although the original version of BLAST uses ungapped alignments, this is not why you get an EVD.

B – wrong – again, this is not why you get an EVD

C – correct

D - wrong

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 $u = 23$.

You should remember that λ does not depend on the size of the database and that u increases with database size, so it must be C or D. The formula is

$$u = 22 + \frac{\ln(2)}{0.69} = 23$$

Therefore D is correct

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.

D is correct – all are true