

Tentamen Bioinformatics-I Questions

All answers require some (short) argumentation

1. Calculate the score of the DNA sequence alignment shown below using the following scoring rules: +1 for a match, -2 for a mismatch, -3 for opening a gap, and -1 for each position in the gap.

```

AACCTGTTGTGTACGGCTCG
|||||  ||||  ||||  |||
AACC---TGTGAACGGATCG

```

2. If a match from a sequence database search is reported to have an E-value of 0.0, should it be considered highly insignificant or highly significant?
3. The program BLAST for database sequence similarity search is based on an idea of searching for "words", that is, stretches of matching residues in sequences.

How do mismatches and small gaps appear in BLAST search, for instance like in a fragment of BLAST hit shown below?

```

gatgacgagctgg-tcgggctccgcac
||||||| | |||||||||
gatgacgagtcgactcgggctccgcac

```

4. Below a fragment from the PAM250 (point-accepted-mutation) substitution matrix is shown:
(...)

| | | | | |
|---|----|----|----|---|
| T | 3 | | | |
| W | -5 | 17 | | |
| Y | -3 | 0 | 10 | |
| V | 0 | -6 | -2 | 4 |
| | T | W | Y | V |

What is the meaning of diagonal elements (3, 17, 10, 4)? Why are they different?

5. What is the idea behind "weighting" sequences in the package CLUSTAL W for multiple alignment? How are these weights calculated?
6. Below the alignment of four DNA sites for a protein binding is shown.

```

CAACTG
CAGCTG
CAGGTG
CAGCTT

```

Which of the following three position-specific score matrices (PSSM) is more likely to be correct?

| | | | | | |
|---|---------------|----|----|----|----|
| | <u>PSSM-1</u> | | | | |
| A | 0 | 1 | 1 | 1 | 54 |
| C | 54 | 0 | 1 | 53 | 1 |
| G | 0 | 1 | 52 | 0 | 52 |
| T | 0 | 52 | 0 | 0 | 0 |

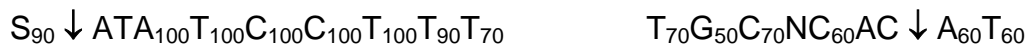
| | | | | | |
|---|---------------|----|----|----|----|
| | <u>PSSM-2</u> | | | | |
| A | 1 | 54 | 1 | 1 | 1 |
| C | 53 | 0 | 1 | 53 | 1 |
| G | 0 | 0 | 52 | 0 | 53 |
| T | 0 | 0 | 0 | 0 | 52 |

| | | | | | |
|---|---------------|----|----|----|----|
| | <u>PSSM-3</u> | | | | |
| A | 1 | 54 | 1 | 1 | 0 |
| C | 53 | 0 | 1 | 53 | 0 |
| G | 0 | 0 | 52 | 0 | 53 |
| T | 0 | 0 | 0 | 0 | 1 |

7. The program BLASTP (protein-protein database search) detects the presence of putative conserved motifs in a query sequence, prior to the output of sequence similarity hits. What kind of algorithm and database can be used for such an efficient search?

Z.O.Z. / P.T.O.

8. Below two consensus sequences for splice sites of RNA processing (5'- and 3'- ends of introns) are given for so-called AT-AC group introns:



(N: any nucleotide; S: C or G; arrows ↓ show the locations of splice cut points)

What is the most likely value of intron length in the following sequence?

(for simplicity, the part of the intron of 100 nucleotides is not shown)

5'...ACGCTGAACCATATCCTTTG-(100 nucleotides)-AACCGCTCACTGGCCCAGCT...3'

9. Which of the following sequences contains the pattern [AG]-x(4)-G-K-[ST] from the PROSITE database?

seq. A: VAGWGKST
 seq. B: GVLKRGKS
 seq. C: AGVLKGRT
 seq. D: AGVGKSTP

10. For translation initiation signals, a position-specific score matrix (PSSM) is given below:

| position: | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|-----------|-----|-----|----|-----|-----|-----|-----|
| A | 12 | 3 | 0 | 14 | -40 | -40 | -1 |
| C | -18 | 3 | 7 | -40 | -40 | -40 | -18 |
| G | -7 | -11 | -6 | -40 | -40 | 14 | 9 |
| T | -32 | 0 | -7 | -40 | 14 | -40 | -8 |

Determine the most likely initiation site in the following sequence: **TCTATGCACCATGGC**

11. What are the main signals used for gene finding in prokaryotic genomes? How are these signals introduced into the search algorithms?
12. What are the main approaches of predicting protein interactions using genomic context analysis?
13. What is the main idea of maximum parsimony in phylogenetic tree construction? What are the drawbacks?
14. *Bootstrap analysis evaluates evolutionary trees by sampling columns from the original alignment with replacement (multiplying or removing some of them) and computing a proportion of times that a particular branch appears in the resulting trees.*

What is the main idea behind this approach?

15. The expression of three genes A, B, and C has been measured in 2 experiments. Logarithms of expression ratios are given below in the table:

| experiments: | 1 | 2 |
|--------------|----|----|
| (Genes) | | |
| A | -2 | 2 |
| B | 1 | -2 |
| C | 1 | 3 |

Assuming Euclidean metrics in the "expression space" and a hierarchical clustering approach (e.g. UPGMA), determine what will be the first pair of the closest expression patterns:

A-B, A-C or B-C.