This exam consists of 16 questions with which you can earn 8 points. This score is complemented with at most 2 points from the assignments. All answers require some (short) argumentation.

1. What is the best strategy to find on the Internet the reference to the original article describing sequencing of a gene ? (NB. A "straightforward" search in PubMed with the name of the gene as a query may yield hundreds of references).

2. 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.

```
CACGTGTGTGCGTCGTGA
|||   ||| || |||||
CAC---TGTCCGCCGTGA
```

3. Sometimes the BLAST program substitutes a stretch of consecutive identical residues in the query sequence (e.g. gggggggg) by n's (nnnnnnnn), considering these residues equivalent to unknown "unalignable" residues.
What is the reason for this?

4. In the program PSI-BLAST, a position-specific score matrix (PSSM) is constructed on the basis of the hits obtained from the standard BLAST search and is used as a query for the second search. Do you expect to find closer evolutionary related sequences or more distant ones? Give an explanation.

5. Below the fragments of three different multiple alignments for the same three protein-coding sequences are shown. Assuming that the adjacent parts of the sequences are the same (denoted by dots), try to estimate the most likely alignment. Which of the parameters, used in the alignment program, would mainly determine this result?

```
   (1)          (2)          (3)
...CGAA...   ...CGAA...   ...CGAA...
...---A...   ...---A...   ...--A-...
...---G...   ...-G--...   ...-G--...
```

6. Below a consensus pattern from the PROSITE database is given.

W - x - [DNH] - x(5) - [LIVF] - x - [IV] - P - W - x - H - x(9,10) - [DE] - x(2) - [LIVF] - F - [KRQ] - x - [WR] - A

Which of the following amino acid positions cannot be occupied by an arbitrary residue?

(a) position 10
(b) position 15
(c) position 20
(d) position 25

7. Below the alignment of four DNA sites for a protein binding is shown.

```
GTTGAC
GTCGAC
GTCCAC
GTCGAA
```

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

```
_______PSSM-1______
A  0 50  0  0  0  0
C  0  0 50  0 50  0
G 50  0  0 50  0  0
T  0  0  0  0  0 50

_______PSSM-2______
A  0  0  0  0 50 10
C  0  0 48  1  0 40
G 50  0  0 49  0  0
T  0 50  2  0  0  0

_______PSSM-3______
A 10  0  0  0  0 50
C  0  0 45  0 50  0
G 40  0  0 50  0  0
T  0 50  5  0  0  0
```

8. An analysis of DNA binding sites of a transcription factor yielded the following position-specific score matrix:

```
-----------------------------
A   5   2   0   1  27  21  13
C  19  24   0   0   0   5   0
G   1   1  24  27   1   0  14
T   3   1   4   0   0   2   1
```

What substitution is likely to increase the binding to the sequence given below?

```
GTTCGGCCCGGACGTACCCG
```

9. What is the purpose of mean force potentials used in protein structure predictions? Where are they derived from?

10. In the alignments used for fold recognition algorithms, a so-called "frozen approximation" is frequently used to define the scores of amino acid substitutions. In this approximation, an alignment matrix is calculated by replacing the amino acids in the template structure with amino acids from the target sequence one at a time. Why is such computation used rather than simply using a substitution matrix as in standard alignment algorithms?

11. Assume that in the framework of the HP lattice model (for simplicity, planar) you need to design a simple hexapeptide that would fold in the following structural element:

```
    X-X-X
    |   |
(1) X X-X    (X = any amino acid)
     (6)
```

Which of the sequences shown below is the most suitable (H = hydrophobic monomer, P = polar)? Explain.

(a) `PPPPPP`
(b) `HHHHHH`
(c) `PHPHHP`
(d) `HPHHPH`
(e) `HPHPPH`

12. Two RNA sequences (A and B) have been aligned by some program. A part of this alignment is given below:

```
sequence A     AUGGGCAAGCUCCGCUUGUCG
sequence B     AUGGGAAAGC--GACUU-UCC
```

Assuming that these sequences form conserved secondary structures and the structure of sequence A is as shown below, try to suggest a more likely alignment and the structure B.

```
    C
  U   C
   C-G
   G-C
   A-U   (structure A)       structure B?
   A-U
   C-G
   G-U
   G-C
AUG   G
```

13. Given a Hidden Markov Model and an observation for that model, what is the purpose of computing the 'posterior decoding'? How does this differ from the result of the Viterbi algorithm?

14. Describe a single step of the neighbour joining algorithm: given an additive distance matrix $d(x, y)$ for N nodes (taxa), a new distance matrix for $N - 1$ nodes is computed.

15. This exercise contains a typing error. Explain why a solution does not exist.

In the context of sequencing by hybridization the set of substrings of length $\ell = 3$ of a string S has been determined: $\{ACG, ACT, ATC, CAC, CCA, CTC, TCA, TTC\}$.

Construct a string S that has exactly these substrings of length 3.

16. Build a suffix tree for the string `GACGACG` (where as usual edges are labelled by substrings, and the leaves by positions in the string). Explain how we use the tree to find the positions of the substring `AC`.