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. 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.

```
TCACGGCGGACA--GTG
||||| || |||  |||
TCACGACGCACACAGTG
```

2. Give a definition of a locally optimal alignment for two sequences.

3. Assume some BLAST search is repeated after a year. The same similarity hits are produced, but the E-values changed. What are the possible reasons for that? Will these E-values increase or decrease?

4. Two DNA sequences A and B of the lengths n and m nucleotides respectively, have been aligned by some program for global alignments. In order to estimate statistical significance of this global alignment, its score is compared to the scores produced by some number of alignments of comparable random sequences. Suppose the following procedure is used for generating random sequences:

   - generate a random number between 1 and 4 by random generator;
   - ascribe to this number a character according to some rule, e.g. A to 1, G to 2, C to 3 and T to 4;
   - repeat the procedure n or m times so that to get a random sequence of n or m nucleotides.

   Is this suitable approach? Give an explanation to your conclusion.

5. Assume you have discovered a new gene coding for a protein in some organism. Among BLAST program versions, what is the best to use to find possible similar unannotated genes in other organisms?

   (1) blastp: protein query vs. protein database
   (2) blastn: nucleotide query vs. nucleotide database
   (3) blastx: translated nucleotide query vs. protein database
   (4) tblastn: protein query vs. translated nucleotide database

6. Which of the following sequences contains the pattern G-H-E-x(2)-G-x(5)-[GA] from the PROSITE database?

   seq. A: GHKNGVLVYLGA
   seq. B: GHEKRGKVYLVG
   seq. C: GHEGGRYVKRGA
   seq. D: GVLYVKGRKARV

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

```
TTCGAC
GTGGAC
GTCGAC
GTCAAC
```

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

_______PSSM-1______						
A	0	0	0	0	1	72
C	0	0	0	0	74	1
G	75	0	75	75	0	1
T	0	75	0	0	0	1

_______PSSM-2______						
A	0	0	0	3	75	0
C	0	0	73	0	0	74
G	73	0	2	72	0	0
T	2	75	0	0	0	1

_______PSSM-3______						
A	0	0	2	0	75	0
C	0	0	0	2	0	74
G	0	75	73	73	0	0
T	75	0	0	0	0	1

8. In a threading algorithm, a query sequence Q has been aligned to the sequence S with a known structure. Below the fragment of the alignment is shown, with deletion of two amino acid residues: (here dots denote identical residues)

```
Q: ...--V...
S: ...LAI...
```

Obviously, in principle here there are two alternative alignments, where valine (V) from sequence Q is considered to be homologous either to leucine (L) or to alanine (A) instead of isoleucine (I) from structure S. Nevertheless, the program did select homology V-I. What is the most likely reason? (choose one of the variants given below).

a) the V I is more conservative substitution than V A or V L, according to a substitution matrix used in the algorithm;
b) the V I is more conservative substitution than V L, and V A substitution is rejected because this configuration leads to two deletions instead of one;
c) there is no reason, it is just a random choice out of the three possibilities;
d) if the residue V in sequence Q is substituted by I from S, the most favourable combination of pairwise knowledge-based potentials is computed;
e) if the residue I in structure S is substituted by V from Q, the most favourable combination of pairwise knowledge-based potentials is computed;
f) if the residue V in sequence Q is substituted by I from S, some known structural motif is formed;
g) if the residue I in sequence S is substituted by V from Q, some known structural motif is formed.

9. In the so-called united residue approximation (UNRES) for ab initio protein structure prediction the following key approximations are used:

- (a) The length of a virtual bond between successive units is constant;
- (b) The angles formed by two virtual bonds have a fixed value of 90 ;
- (c) Side chains are represented by spheres of different sizes.

What happens with each of this assumptions in the framework of the simple cubic lattice model?

10. In the approximation of simple cubic lattice model, a chain of 8 hydrophobic residues (HHHHHHHH) can occupy 8 vertices of the cube. What is the energy of this structure? Is it possible to substitute 4 residues by polar ones (P) and to lose less than one half of the energy value?

11. Multiple alignments of related RNA molecules may be improved by using special objective functions. Apart from the sequence similarity scores, what are the possible RNA-specific parameters for such objective functions?

12. Below the alignment of 5 RNA sequences and the "bracket view" of consensus secondary structure is shown. Is it possible to suggest also non-canonical base-pairing in this structure?

```
GGGGACCCAGGGGAAACCCAGGGGACCC
GCGAGCCCAGGGGAAACCCAGGGAGCGC
GGGGACGCAGGGGAGACCCGGCGGACCC
CAAAGCGCAGGGGAGACCCCGCGAGUUG
GCGGACCCAGGGGAAACCCAGGGGACGC
(((..(((.(((....))).)))..)))
```

13. Give a short 'algorithmic' description of the Viterby algorithm for computing the most probable state sequence for a given observation for an Hidden Markov Model.

14. Five 'taxa' have the following distance table, under the assumption of a uniform molecular clock.

	1	2	3	4	5
1	–	18	18	8	18
2		–	12	18	12
3			–	18	4
4				–	18
5					–

Observe that the nodes 1 and 4 have maximal distance 18 to the nodes 2,3 and 5. What can you conclude about the tree for this matrix? How would you continue?

15. Build a Trie for the strings AACG, ACA, CACG, and CG. Extend the Trie with failure links to make a tree for the Aho-Corasick algorithm in order to search for this set of patterns.

16. In an experiment of physical mapping we have probes A,B,C,..., H. We find the clones $\{B, F, H\}$, $\{A, C, D, G\}$, and $\{A, B, C, E, F\}$. Give a representation of solutions to this problem as a PQ-tree, which has circular nodes to represent permutations, rectangular nodes to represent linear orders.