Final test

Name:

Introduction to Bioinformatics, 2015

Neptun:

The point value of each question is listed in parentheses.

Total points: 100

1. **True or False? Fill the table. (16)**

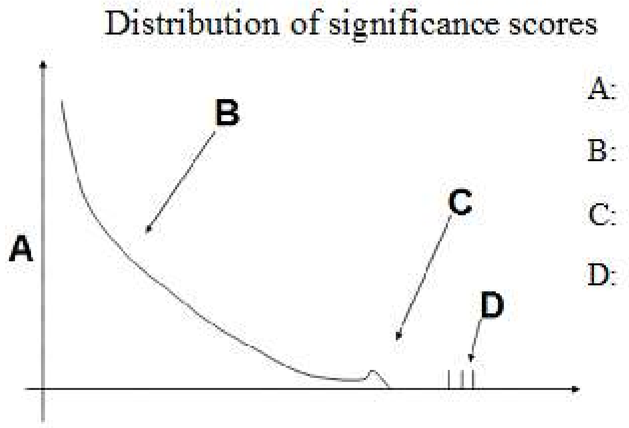
	T	F
The PAM matrix is derived from automatically created multiple alignment of related proteins.		
Higher BLOSUM number is better for detecting more remote similarities.		
The time complexity of the Needleman-Wunch algorithm when affine gap penalty is allowed is O(mn), where m and n are the length of the sequences.		
The input of a profile hmm builder program (such as *hmmbuild*) is a set of related sequences.		
Variable regions in a multiple alignment should map to the buried parts of the protein.		
Swiss-Prot is a manually curated non-redundant database.		
Ontology terms are used to help find the meaningful hits of a similarity search.		
When BLAST algorithm records words in a query, it includes similar words.		
TBLASTX program compares a nucleotide query to a translated nucleotide database.		
The Juces-Cantor DNA sequence evolution model is a good approximation of reality.		
ChIP-seq is a technology based on immune precipitation and NGS.		
MACS is a software tool for RNA-seq data analysis.		
Mutations are genetic changes that are present in more than 1% of a population.		
BRCA I and II are tumor suppressor genes and their mutation increases the risk of acute lymphoid leukemia.		
Structured representations can be described as graphs where entities are the nodes and relationships between them are the edges.		
Molecular biology usually studies well defined biological processes, while system biology concentrates on bigger networks using high throughput techniques.		

2. **Fill in the missing words/expressions! (8)**

Bioinformatics is an interdisciplinary field of science. It synthesizes knowledge and approaches mainly from three disciplines: _______________________, _________________________, _________________________. Narrow and broad definition of bioinformatics can be formulated. According to the narrow definition, bioinformatics is the ________________________, and its main tasks are ___________________________and_______________________ of its data. The board definition considers all computer applications in biology to bioinformatics including _________________ and _________________.

In most cases biological data is annotated. Annotation means ________________________. Annotations can be _________________, for example the name of the source organism, or ____________________, for example the position of a phosphorilation site. Possible annotations of a protein sequence record could be for example: _______________________, ________________________ ,_________________________, _______________________. The standardized collections of annotation terms are called ____________________. The use of them makes data comparable and searchable in the database.

3. **Which label express the meaning of A, B, C and D the most? (4)**



(p-value, random scores, e-value, significance, best hits, exact match with the database, non-random similarities, number of matches in the database with a given significance, number of exact matches in the database, S-score)

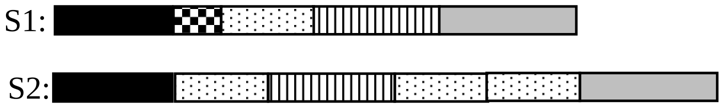
4. **Which item has the highest p-value (B, C or D)? (1)**

5. **Which item has the highest E-value (B, C or D)? (1)**

6. **Define EAV scheme (what does E,A,V stand for)? Give an example for that! (2)**

7. **Define Levenstein edit distance and calculate it between A=[AATGCTTAA] and B=[CTGCATCAAT] (4)**

8. **Draw a dot-plot for the following two protein sequences! Briefly explain how you created the graph. (5)**



9. **Calculate the log odds ratio of residues PROLINE (P) and THRYPTOPHAN (W) using the following multiple alignment! (4)**

```
W P G M R W
T W P M W P
W P P W S P
W L H P N W
```

10. **Prepare the local alignment of the words ‘MLLVAM’ and ‘GIIV’ using the following matrix and a penalty -1 for indels! Briefly describe your calculation! (6)**

	A	**G**	**I**	**L**	**M**	**V**
A	**4**					
G	0	**6**				
I	-1	-4	**4**			
L	-1	-4	2	**4**		
M	-1	-3	1	2	**5**	
V	0	-3	3	1	1	**4**

11. **Calculate the entropy of the following multiple alignment (4), draw the entropy plot and highlight the most conserved regions on it (2)!**
(0*$\log_2$(0):=0; $\log_2$(1/4)= -2; $\log_2$(1/2)= -1; $\log_2$(3/4)= -0.415; $\log_2$(1)=0)

```
AATTC
AATCC
CATCG
CAGAA
```

12. **Mention some cases when the constructed multiple alignment is definitely not valid/ biologically relevant! (2)**

13. **What are the main steps of a typical phylogenetic analysis? Describe them briefly. (6)**

14. **What are the two parts of a BLAST output? What values do you look at to analyse the results? (At least 2) What do they mean? (6)**

15. **Why do we need heuristics in similarity searching? What are the two kinds of it? Give at least 1-1 example! (5)**

16. **What is the correct order of steps in NGS workflow to assembly a whole genome? (6)**

a: isolate DNA, b: assemble contigs, c: PCR amplification, d: quality control of the DNA, e: break the cells, f: adapter ligation, g: draft genome, h: collecting the sample, i: quality control of the reads, j: DNA fragmentation, k: NGS sequencing, l: assemble scaffolds

17. **Given the reads {AATCGA, GATCGA, CGATCG, ATCGAG, ATCGAT, TCGATC}. What is the original sequence? Use the de Bruijn graph based algorithm for the computation! Define the Bruijn graph (what are the nodes and edges) and briefly describe each step of the calculation. Hint: use the largest possible k-mers. (5)**

18. **Fill in the missing words/expressions! (8)**

Sequencing is the identification of the order of nucleotides in ______________ or _________ molecules. The human samples can have various sources, for example: ___________, ________________ , _________________ or ________________. After extracting _______________ molecules from the cells quality control have to be performed typically with ___________________. In order to sequence them the long molecules have to be ___________________.

19. **Considering the following Burrows-Wheeler Transform, restore the original sequence! Briefly describe the algorithm you used. (5)**

```
$  .. G
A  .. $
A  .. C
C  .. T
C  .. T
G  .. C
T  .. A
T  .. A
```