Final test | Name:
Introduction to Bioinformatics, 2014 | Neptun code:

The point value of each question is listed in parentheses.

Total points: 100

1. Name the 4 **core data types** in bioinformatics! (4)

2. What is the **structure** and **function** in a system? Give a definition of both of them! (4)

3. Fill in the missing elements! (4)

Models	**Entities**	**Relationships**
	Atoms	
	Enzymes	
Genetic networks		Co-regulation
Protein structure		
	Amino acid	Sequential vicinity

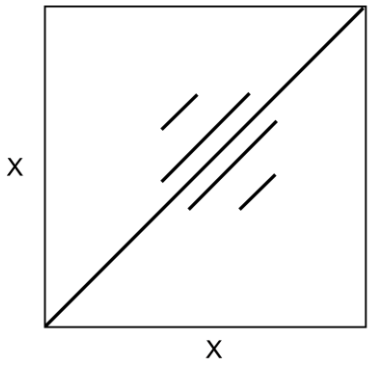
4. Define the **Levenshtein edit-distance**! (2)

   What is the Levensthein edit-distance of the following sequences: x=CATACCTA, y=ACATTACC? Write down each step of your computation! (2)

5. Define the **Jaccard coefficient** that expresses the similarity of property sets A and B! (2)

   Calculate the Jaccard coefficient, if:
   A=[‘cat’,’dog’,’horse’,’duck’,’cow’,’chicken’]
   B=[‘dog’,’dolphin’,’duck’,’dinosaurus’]! (2)

6. Draw the schematic sequence X based on the following dotplot: (3)



7. Calculate the log odds ratio of residues SERINE (S) and THREONINE (T) using the following multiple alignment! (4)

```
T V S R A G
S T S V E T
T G T S K A
T K T N R H
```

8. Fill in the missing words! (10)

Informally, an _____ is any well-defined computational procedure that takes some value or a set of values as _____ and produces some value or a set of values as _____. An algorithm is thus a sequence of computational steps that transform the input into the _____. It is important to predict the resources that the algorithm requires for a given input size. The resources are _____ and _____. It is based on a simple ideal model machine: instructions are executed one after another, with no concurrent operations (RAM model), thus this measure is machine independent. The _____ is defined as the number of steps to be carried out. It is assumed that the instructions are executed one after another, with no concurrent operations (RAM model). _____ is the number of items (numbers, strings etc.) that one needs to store during the calculations. It is measured in terms of megabytes of disk space or RAM space. Furthermore there is a trade off between _____ and running time.

The worst case analysis is when one assumes inputs that lead to the _____ limit of the resources. _____ is when one uses an average input (problem-dependent, not trivial). In biocomputing this is the "idealized average" database, average (random) sequence of given length, etc.

In biocomputing, we mostly deal with sequences and databases as special kinds of input, both represented as character strings. The input size is often defined as the number of amino acids or nucleotides, described as the number of characters in a string. Another important parameter is the alphabet and database size. The alphabet size for proteins is_____, and for DNA it is _____.

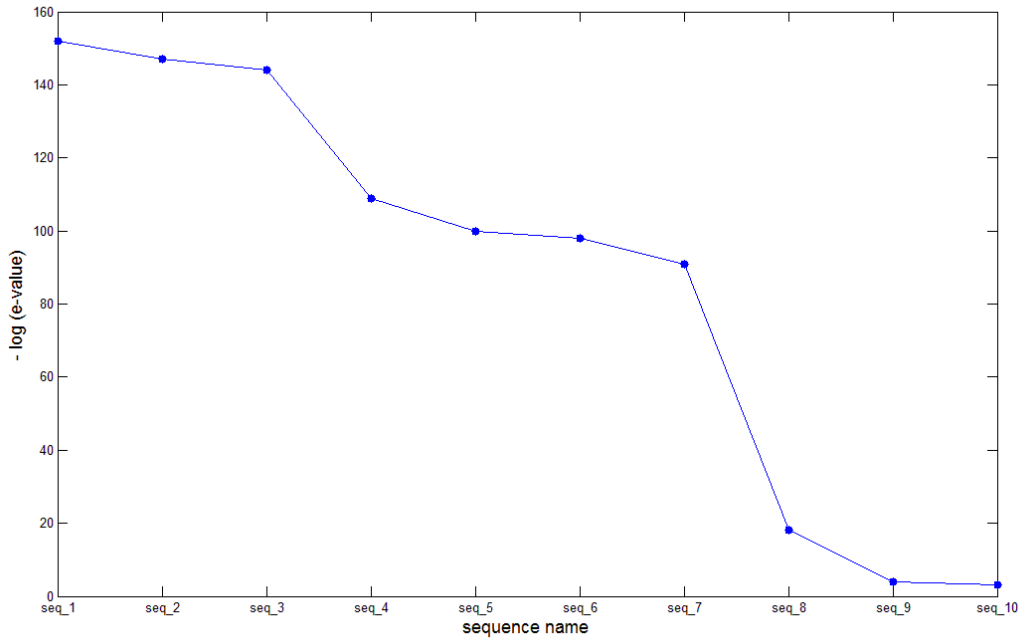
The _____ algorithms (O(l)) are the dream target. The $O(l^2)$, $O(l^3)$ may be used in practice, but not for all problems.

If the order is greater than $Q(l^n)$, i.e. it can not be expressed in a polynomial form, the problem is np-complete and _____.

The Needleman-Wunsch and the Smith-Waterman algorithms solve similarity search problems and they have the time complexity _____ The CLUSTALW2 algorithm has the time complexity _____ where the parameter n is the number of _____ and the parameter l is the _____ . This program is frequently used for computing _____.

9. What is the meaning of **p-value**? What does it mean if it is close to 0? (3)

10. Given the following **e-value plot** (generated after an hmmsearch) what would be your conclusion, which sequences would you consider belonging to the same protein family, how would you (if you would) further analyze the results? (5)



11. Calculate the **entropy** of the following multiple alignment and draw the **entropy plot**! (6)

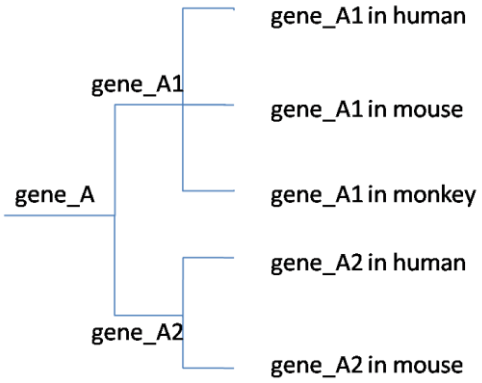
```
ATGC
ATCG
CTGA
CTCT
```

12. Prepare the **local alignment** of the words ’ACGC’ and ’GATTGA’ if we define the following point values:
    gap penalty = -2,
    mismatch point = 0,
    match point = 2 ! (6)

13. What is the difference between PAM and BLOSUM matrices? (4)

14. What are the advantages and disadvantages of parsimony? (5)

15. Give some examples of **orthologs** and **paralogs** in the following figure! Define the two terms! (4)



16. What are the biological and computation heuristics ("tricks") that Blast uses? Describe each of them briefly! (8)

17. What are the non-annotated sequence clusters? How can you use them to find a function of an unknown protein sequence? (4)

18. Define the following terms (3):

read:

coverage:

scaffold

19. What is **whole genome sequencing** and **targeted sequencing**? Give an example (clinical/scientific) when we can use them! (4)

20. What is the difference between the following terms: **SNP, mutation**? (2)

21. What are the three methods of **DNA fragmentation**? (3)

22. Given the reads {ACTT, TGCT, GCTT, TTTG, TTGC, CTTG, CTTT}. What is the original sequence? Use the de Bruijn graph based algorithm for the computation! (6)