Example questions for Bioinformatics, first semester half

Sommersemester 2010

Note

- The schriftliche Klausur wurde auf deutsch geschrieben
- The questions will be based on material from the Übungen and the Lectures.
- We missed most of the first lecture. If we do not finish the planned lectures, the last few questions are not relevant and the last topics will not appear in the final Prüfung.

Example questions

1. You are given two DNA sequences to align

   ACGTCCTTCATT and GTCTCATG

   You have a scoring scheme where a

   - match gives you +1
   - a mismatch gives you 0
   - gap opening costs −10

   Write down the best alignment of the two sequences
2. You have a scoring scheme where

   - A match gives you +1
   - a mismatch gives you −1
   - opening a gap costs you −1

   Write down the best alignment for the same two DNA sequences.

3. You are aligning protein sequences using a substitution matrix :

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Gap opening costs -8. Gap widening (extension) costs -1.

You are given an alignment

```
AACDQRST
A-CD-RST
```

What is the score of this alignment ?

4. Given

```
AACDQRST
A-CD--ST
```

What is the score of this alignment ?

5.
```
AACDQRST
A-CD-SST
```

What is the score of this alignment ?

6. You have calculated a score matrix for a pair of DNA sequences. You have performed the traceback calculation and found a result like:

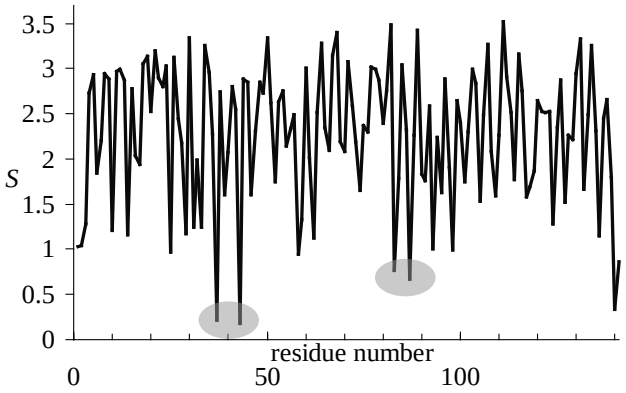
	A	C	A	C	C	T	T	A
C								
C								
A								
T								
C								
C								
A								
A								

Write down the corresponding sequence alignment with gaps in the correct positions.

7. Outline the steps used to find values for a BLOSUM amino acid similarity matrix.

8. What is the advantage of a Needleman-Wunsch alignment compared to a seeded alignment ?
9. What is the advantage of a seeded method like BLAST compared to a Needleman-Wunsch alignment ?
10. Name an application where you would use a method like BLAST and not a Needleman-Wunsch alignment.
11. Name an application where you would use a slow method like Needleman and Wunsch, rather than a BLAST-like method.
12. You have two proteins with weak, remote similarity. You know their sequences. You also know the sequences of the original DNA. Why would you expect a better alignment using the protein sequences ?
13. In protein alignments, we do not just look at match/mismatches. We look at the similarities between amino acids. How are these represented ?
14. There are several different substitution matrices used in protein alignments. Why would you prefer one over another ?
15. What is the difference between an iterated blast (psi-blast) search and a simple blast search ?
16. What is the advantage of an iterated blast search compared to a simple blast search ?
17. I have written a program that generates random DNA sequences. I expect to see 25 % sequence identity between pairs of random sequences in gapped alignments. When I try the calculation, I usually see more than 25 % sequence identity. Why ?
18. If I take random biological sequences from a data bank, I see even more sequence similarity. Why ?
19. I have two proteins with 20 % sequence identity. I ask you if this is likely to be significant. What other simple piece of information do you need to answer this question properly ?
20. If you use the program "chimera", what representation would you pick in order to see the secondary structure ?
21. I want to calculate a multiple sequence alignment for N_{seq} sequences. How many pair-wise alignments will I have to calculate ?
22. In a multiple sequence alignment, you want to maximise a score, $score = \sum_{b \neq a}^{N_{seq}} \sum_{a=1}^{N_{seq}} S_{a,b}$ What is this score in words ?
23. In a multiple sequence alignment, I want to build a "guide tree". What determines the order in which I join the nodes together ?
24. I have 3 sequences, A, B, C. The sequences B and C are both related to A, but I cannot get a good alignment score when I align B and C. What could be a reason ? Draw a diagram if it is easier to explain.
25. I have calculated a multiple sequence alignment. I want to find which sites in the alignment are conserved and which vary. I would like to make a plot of variability/conservation as a function of sequence position. You remember a formula $S = \sum_{i=1}^{N_{states}} p_i \ln p_i$ What is the meaning of p_i ?

26. From a multiple sequence alignment, I have calculated variability/conservation as a function of sequence position:



Residues 37 and 43 seem to be very conserved. Why might they be important residues ?

27. In the picture above, some sites are not very conserved. I say these residues cannot be important to the function of the protein. Why may I be wrong ?
28. I have calculated a sequence alignment of 400 tyrosine kinases and I find that very few sites seem to be conserved in evolution. How could I change my results, so that more sites seem to be conserved ?
29. We have a family of sequences and all pair-wise alignments. I can count the number of differences (mutations) between any two sequences and calculate the fraction of residues that have changed : $p_{mut} = N_{diff} / N_{length}$. I would like to estimate evolutionary time by saying $t = k\ p_{mut}$ for some constant k. This is not a good measure. Why not ?
30. I want to use aligned DNA sequences to build a phylogenetic tree. Name two reasons that the branches in the tree may not be reliable.
31. I have built a phylogenetic tree using a neighbour joining method. Describe a general approach I could use to see how reliable the tree is.
32. It is believed that protein sequence evolves and changes faster than protein structure. What could be an evolutionary explanation for this.
33. A structure determined by X-ray crystallography has a resolution of 1.5 Å. When I look at the coordinates, I find every backbone C-N distance is 1.32 Å. How can the coordinates seem so accurate if the measurement is so imprecise ?
34. I want to compare two protein structures using the root mean square difference of coordinates, given by $\left(\frac{1}{N_{res}} \sum_{i=1}^{N_{res}} \left| \vec{r}_i^{\,a} - \vec{r}_i^{\,b} \right|^2 \right)^{1/2}$ where N_{res} is the number of residues in a protein and $\vec{r}_i^{\,a}$ is the coordinate vector of the C^{α} of residue i in protein a. What steps do I have to perform on the coordinates of protein a or b before I can apply the formula ?

35. I have two closely related proteins from different organisms with similar structures and sequences. The proteins are of slightly different sizes. Describe a set of steps one would need to apply the root mean square difference formula given above.
36. Which graphical representation would you use in order to emphasize the secondary structure content of protein? all-atom, chaintrace, ribbon, ...?
37. You have a protein of unknown function from a bacterium. You have made a knock-out mutant, but the bacteria die immediately without the corresponding gene. You have sequenced the protein. What steps would you take to guess the function of the protein ? What kind of information would you look for ?
38. An welchem Merkmal erkennt man phylogenetische Bäume, die mit der UPGMA Methode erstellt wurden ?
39. Sie haben Multiple Sequenzalignment aus
    a) 10 Sequenzen aus sehr ähnlichen Quellen
    b) 967 ähnlichen Sequenzen
    c) 13 sehr diversen Sequenzen
    Welche Methode würden Sie in jeden oben genannten Fällen wählen um aus den Alignments phylogenetische Bäume zu berechnen ? Warum ?
40. Wie würden Sie vorgehen um in einem Multiplen Sequenzalignment potentiell katalytisch wichtige Seitenketten im aktiven Zentrum zu identifizieren ?