MOL204 – Exam Fall 2015

Exercise 1 – 15 pts

1. 1A. Define primary and secondary bioinformatical databases and mention two examples of primary bioinformatical databases and one example of a secondary bioinformatical database. For each of the databases you mention, explain briefly why they are primary or secondary. [3pt]

2. 1B. (i) What is the core data stored in UniProtKB?
(ii) UniProtKB is composed of two databases; explain what they are and what the relationship between the two are.
(iii) Each entry in UniProtKB has different types of annotation. Give 3 examples of different types of annotation
(iv) What are "sequence features" in UniProtKB? Give 2 examples of "sequence features" found in UniProt [9pt]

3. 1C. Explain briefly what Gene Ontology (GO) is and mention the three top-level categories that form the main (orthogonal) branches in GO. How can you use GO to find all cyclins found in the cell nucleus? [3pt]

Exercise 2 – 15 pts

You are given two protein sequences q: **MIGV** and d: **FFIGL** and the scoring matrix (excerpt from PAM250) shown at the top of the figure. The lower part of the figure shows the *matrix H,* partially filled, which is used to find the best global alignment(s) by dynamic programming. Use the information given here to answer the questions in this part of the exam (2A-2D).

4. 2A. Which gap penalty has been used in this matrix? Calculate and fill in the remaining values in the matrix *H* for the cells labelled *A, B, C, and D (give the answer as follows: A=<value>, B=<value>, etc)*. What is the score(s) for the best global alignment(s)? [5pt]

5. 2B. For the sequences and the scoring matrix given in 2A, find the best global alignment(s). *(Write the alignment with one sequence above the other and with dash(es) to indicate where gaps are)*. [4pt]

6. 2C. Explain briefly the procedure used in 2B to find the best alignment(s). [3pt]

7. 2D. Explain briefly the difference between the algorithms used for global and local alignment by dynamic programming. When would you prefer to use local alignment? [3pt]

Excerpt from the scoring matrix PAM250

	G	I	L	M	F	V
G	5	-3	-4	-3	-5	-1
I	-3	5	2	2	1	4
L	-4	2	6	4	2	2
M	-3	2	4	6	0	2
F	-5	1	2	0	9	-1
V	-1	4	2	2	-1	4

Partially completed matrix *H* which is used to find the optimal global alignment using dynamic programming. In question 2B you shall use dynamic programming with the scoring matrix PAM250 and the gap penalty you found in 2A to complete the matrix *H*. *Give the numerical values for the four cells labelled A, B, C and D.*

	sequence *q*	M	I	G	V
sequence *d*	0	-2	-4	-6	-8
F	-2	0	-1	-3	-5
F	-4	-2	1	-1	-3
I	-6	-2	3	1	3
G	-8	-4	1	*A*	*B*
L	-10	-4	-1	*C*	*D*

Exercise 3 – 19 pts

8. 3A. Describe briefly the key steps in the BLAST algorithm. Define and use the terms: "high-scoring pairs" (HSP) and "words". [4pt]

9. 3B. For each match in a BLASTP search, score and E-value is calculated. Explain briefly what these values are and how they relate to each other. [3pt]

10. 3C. Explain briefly why searches with the scoring matrix BLOSUM45 gives more sensitive searches than with BLOSUM62. [2pt]

11. 3D. Explain briefly how database searches with PSI-BLAST differ from ordinary BLASTP searches. Why are searches with PSI-BLAST more sensitive than ordinary BLASTP searches? [4pt]

12. 3E. Using a PSI-BLAST search you have found a match between your query sequence and the sequence of a cyclin. But the E-value is high and the match is only partial, and you doubt if the match is a true positive and reflects homology. Suggest three methods you could use to find support for the idea that your query sequence is a cyclin or not. For each of the methods, explain briefly what information you can obtain. [6pt]

Exercise 4 – 13 pts

13. 4A. Explain in brief the main steps in progressive multiple sequence alignment as in Clustal W. [4pt]

14. 4B. Explain briefly how one can evaluate how accurate different methods for multiple sequence alignment are (benchmarking) by using information from protein structures. [3pt]

15. 4C. You have 20 different amino acid sequences of human cyclins. Why is multiple sequence alignments better suited to identify globular domains in these cyclins than pairwise alignments? [3pt]

16. 4D. Explain briefly how you can use Jalview and information from UniProtKB to assess the quality of multiple sequence alignments generated with different parameters, just like you did in the course PC-lab. [3pt]

Exercise 5 – 12 pts

17. 5A – Explain what defines the fold of a protein. Explain then what the tertiary structure of protein is. [2pt]

18. 5B – Name one database of protein structures and one database of folds. Explain briefly what they contain and how the data is organized in each. [8 pt]

19. 5C – How is the information about tertiary structure written in a structural database? If you want to compare the tertiary structure of two or more proteins, which measure can you calculate? [2 pt]

Exercise 6 [12 pts]

20. 6A - Which secondary structure elements are easiest to predict from sequence and why? [2 pts]

21. 6B - In the table below we have reported the performance of six different methods for prediction of protein secondary structure from sequence. [3 pts]

Performance of various methods for secondary structure predictions

Method	**all-α**	**all-β**	**α+β**	**α/β**
Chou-Fasman	50.9%	42.6%	38.0%	52.4%

GOR I	59.0%	55.1%	39.2%	67.2%
GOR V	84.3%	60.3%	69.9%	73.8%
ZPred	76.9%	44.9%	60.8%	66.0%
PSIPRED	94.4%	65.4%	67.8%	82.8%
JNET	84.2%	74.3%	70.5%	84.4%

The performance of each method has been measured by applying it on sequences for which the tertiary structure is known. Each prediction could then be compared to the secondary structural elements present in the corresponding protein structure. The values reported are averages over several sequences. A score of 100% for one protein sequence means that the predicted secondary structure element predicted to each amino acid is correct. A score of 50% means that the predicted secondary structure element is correct for only 50% of the amino acids in the sequence.

For each method four scores are given; one for each of the four classes of protein structures:
- all-α: protein folds with predominantly alpha-helices
- all-β: protein folds with predominantly beta-strands
- α+β: folds with segregated alpha-helices and beta strands
- α/β: folds with alternating alpha-helices and beta-strands

Questions: For each of the six methods, list the two protein classes for which it performs the worse. Is the result consistent within the six methods and why?

22. 6C - Chou-Fasman and GOR I are both statistical methods but GOR performs somewhat better than Chou-Fasman. PsiPred and JNET perform even better and belong to the best performing methods for prediction of secondary structure.

Explain what statistical methods are and explain what is the main improvement introduced in GOR compared to Chou-Fasman. Further name which general technique is used in the best methods for secondary structure prediction and explain its basic principles. [4 pts]

23. 6D - Helices of transmembrane proteins can also be predicted from sequence but dedicated methods have been developed for these. Explain why there are dedicated methods for membrane proteins and list which information these methods are exploiting. [3 pts]

Exercise 7 [14 pts]

24. 7A – [5 pts] *Threading*, *Homology Modeling* and *Ab Initio* modeling refer to three categories of methods for the prediction of protein tertiary structure from their amino acid sequence. Explain briefly the basic principles of each of the three methods.

25. 7B – [5 pts] A researcher needs to predict the tertiary structure of a protein for which she knows the full amino acid sequence. According to the question above, she has three methods to choose in-between. Explain how she

should proceed to choose the method that will give her the most reliable model of her protein tertiary structure. Explain then how she can evaluate this model and possibly improve it.

26. 7C – [4 pts] In homology modeling, what is the most critical step and why? How can one improve the outcome of the critical step?