SFB exam April 2014. Name:

Dear students, take your time. It looks like a lot of questions, but you will see that it isn't. So think first and don't start to write answers immediately. If everybody writes legible, the exams can be graded (nagekeken) by Monday.

You can answer in Dutch, German, or English. If you are answering in one language but don’t know one word, use that word from another of these three languages.

**Every answer must be motivated! So implicitly each question ends with "Explain your answer".**

The rest of this sheet is scratch paper.

1) A colleague comes to you with the sequence:

TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN

and asks if you can find out which of the cysteines form bridges with each other. Obviously, you cannot see just from the sequence which of the six cysteines are bridged.

Can you list a series of things you might do (starting with the most trivial solutions and ending with "and if this all fails, we do an experiment with Eppendorf tubes and pipets etc") to find this out?

2) What can you tell about this peptide: MNNSAKALTRRGGALTLLAIVLLTLWAIVFMLLLIAFFGGSADAA A proteomics experiment indicates that this peptide is 79.9 Daltons too heavy (i.e. it is phosphorylated). Which residue holds this PO4  group (and why)?

3) We all have been vaccinated against a series of diseases. Let's take flu as an example. If a few particle enters my body, then my body starts probably the same day already with the production of antibodies, and after 48 hours my immune system is in high gear and prevents that I get sick. It is not possible to vaccinate against most toxins. Let's take scorpion toxin as an example. If scorpion toxin enters my body, I need an injection with ready-to-go antibodies soon.

a) Why can I not be vaccinated against scorpion toxin?

b) How do we produce the ready-to-go antibodies against scorpion toxin? (And don't forget to explain with some detail the bioinformatics needed in the process!)

4) Below you see two Ramachandran plots. They are for (very many) Glu and Asp residues of which the secondary structure determination software DSSP thinks that they are in one particular type of secondary structure.



a) What is a torsion angle?

b) What is a Ramachandran plot? So, which torsion angles are on which axes in the plots above?

c) From which secondary structure were the residues selected?

d) Which of the two plots is for Asp and which is for Glu?



4) Above you see four bases.

a) Can you please write their one-letter codes in the white circles.

b) Which parts of the bases are pointing to the major groove and which to the minor groove?

5) During the course we have seen the plot:

a) Where, why, and how do we use these equations?

b) Try to explain the terms in these equations (and if you don't know the names of the funny characters, just copy them as good as you can).

6) A colleague has found a peptide that circulates in the blood of certain fish. He knows that this protease needs to bind one calcium atom to become active. And now he asked me if I can predict where this calcium binds, and if I can predict just one point mutation in this fish protease that makes that the calcium no longer binds. My colleague needs that to continue experimenting with the protein intact but inactive, but in the presence of calcium ions.

 In the PDB I can find about 27000 short loops that wrap around a calcium atom. I want to use this wealth of information and I want to write a computer program that predicts calcium binding loops in protein sequences.

a) What is the PDB?

b) What is a the definition of a Force Field?

c) Describe what I must all do to arrive at a computer program to predict calcium binding loops in protein sequences.

7) I made four mutations in four different proteins. 1) Ile -> Asp at the surface. 2) Asp -> Ile at the surface. 3) Ile -> Asp in the core. 4) Asp -> Ile in the core. In all four cases I measured the melting (unfolding) temperature difference compared to the wild type (un-mutated) protein.

I observed that one protein was significantly more stable, one was very much less stable, one was nearly equally stable as the wild type, and one was a bit more stable than the wild type.

Can you tell me which of the four stability measurements (most likely) belong to which of the four mutations?

8) When we design mutations that make proteins more stable, we often talk about entropic versus enthalpic stabilisation.

a) What do we mean with enthalpic stabilisation when designing a mutation?

b) What do we mean with entropic stabilisation when designing a mutation?

c) Give at least two examples for enthalpic stabilisation.

d) Give at least two examples for entropic stabilisation.

9) In the course we have focused on protein sequence, structure, function, and dynamics. But there are many other types of molecules in cells.

a) List more than 5 very different non-protein molecule types commonly found in human cells. (Feel free to start with water...).

b) Please list five roles for water in a human cell.

c) How do membrane proteins differ from water soluble proteins?

d) What is the positive-in rule, where does it originate from, and how do we use in bioinformatics?

e) Mention five roles for sugar in biology.

f) So, sugar is important. Probably just as important as proteins. How come we know much more about proteins than about sugars?

10) In the PDB we find structures solved either with X-ray crystallography or by NMR.

a) Describe in about 25 words how X-ray crystallography works.

b) Describe in about 25 words how NMR structure solving works.

c) What is the difference between accuracy and precision?

d) It is a bit farfetched, but I might call one of the two techniques more accurate and the other one more precise. Which is the more accurate one?

e) X-ray and NMR are experimental techniques. The structures solved with these techniques therefore contain experimental errors. Can you describe how the differences in the experimental techniques lead to differences in the experimental errors?

f) What are side chain flips; for which residues do we observe them; and why do we see so many of them in structures solved by X-ray crystallography.

11) Describe in at most ten words per term what the term means:

Z-score

B-factor

R-factor

Force Field

MD

EM

BLAST

Homology Modelling

PDB

CSD

Salt Bridge

Bond angle

Torsion angle

NMR

ROC

MRS

Trajectory

SCOP, CATH, DALI

HSSP

DSSP

Occupancy

Resolution

NOE

Crystal packing artefact

12) I have a series of scientific problems that seem addressable with machine learning techniques. Can you for each problem indicate which technique seems most appropriate, and why?

Problem 1. There are many errors in PDB files. One type of error is that ions often are different from what is given in the PDB file. Our master student Bart has therefore been analysing metal ions in PDB files. He collected a large series of parameters, including the atom types of the surrounding ligand atoms and their distances to the ion. He calculated the net charge at the location of the atom, and at least 20 more parameters. Obviously not all parameters are equally important... Barts problem is that he wants to know which ions should be replaced by other ones, and he wants to know which other one to take in those cases.

Problem 2. When binding medicines to proteins, we realised a long time ago already that the hydrophobicity of the surface in the binding pocket is important. We collected for thousands of pockets the hydrophobicity, the size, and whether a medicine is known to bind there or not. If we now get a protein sequence, and we can build a good homology model, we want to know if that protein has a putative binding site for medicines.

Problem 3. Bacterial DNA contains a wide variety of signals just before ORFs (Open reading Frames). I have a collection of 350 sequences of 50 bases that sit just before the start codon of a bacterial protein. I am reasonably sure that these 350 sequences are all binding sites for the *mofope peda* activation protein. I want to use these data to predict if a given bit of sequence observed just before an ORF contains a binding sites for the *mofope peda* activation protein.

13) Let's go back to problem 1 in question 12. Can you think of five rather different parameters Bart included in that list of twenty that I didn't list all?

14) Homology modelling is a technique that consists of a series of steps that can be executed consecutively. Please list these steps in the same order as in the actual modelling process. With each step, indicate which (type of) software you need, and what that software should do.

15) Part of the structure of the mofope peda protein looks like:

Why does the yellow tyrosine have two side chains? What is going on here?