SFB 2016-2017 Name

All questions are worth equally many points. The grade will be 10/13-th of your points. My answers are very short…. Often multiple good answers are possible. When in doubt contact me with your answer, and we discuss it.

7) Fill out the amino acid form. After that keep it 'nearby' because it will help you answer at least one of the questions.
Answer: see web pages.

1) Lipinski invented, a long time ago, his famous 'rule of five'. The original rule (it got modified a bit over the years) implies that good ligands to try as a drug in a drug design project will have a series of properties:

A. a molecular weight less than 500
B. fewer than 5 hydrogen-bond donors and less than 5 hydrogen-bond acceptors
C. fewer than 5 internal degrees of freedom
D. logP less than 5 (which means that the compound must be rather hydrophobic).

Now that you went through the whole SFB course, do you think this comes as a surprise? Let me help you, I think it does not come as a surprise. But why not? Can you explain for these four properties (A-D) why none of them is a surprise to you? (Some of the answers might overlap a bit...).

A) The simple answer is that most medicines need to replace (sit in the pocket of) an endogenous ligand. And those aren't very big either. If the medicine gets too big, it will either stick out in the solvent or bump into something.

B) When the ligand binds, it loses a lot of entropy (its own 6-dimensional motion, and probably some internal degrees of freedom). This must be compensated, and that goes best by the entropy of water, both the waters in the pocket, and the waters that are unhappy around the hydrophobic parts of the ligand.

C) Every degree of freedom is lost upon binding. Too many of that might be so much energy loss that the entropy of water (see B) can no longer compensate for it.

D) See B.

And the extra questions are: There must be fewer than 5 hydrogen-bond donors and fewer than 5 acceptors, but why would it be bad to have no hydrogen-bond donors and acceptors at all?
There is also a need for specificity and that can be achieved very well with H-bonds, but also, something totally hydrophobic will not dissolve and will get nowhere…

And altogether, we should also think about passing membranes to get to the target. Things should not be too hydrophobic to stay in the membrane and not be to hydrophilic (especially not to positively charged; things like arginine sidechains don't pass membranes very well) to refuse to pass the membrane.

It surprised me that half of the students failed to shout: "Entropy of water".

2) a) What is the positive-in rule? Where does it come from? Which parts of the cell are involved?
When (membrane) proteins are produced by ribosome they first move to ER membrane. Translocon that helps put it there is bad at crossing positive charge through ER membrane, so positive charge stays outside, in cytosol. Golgi buds of from ER and moves to outer membrane. Golgi merging with

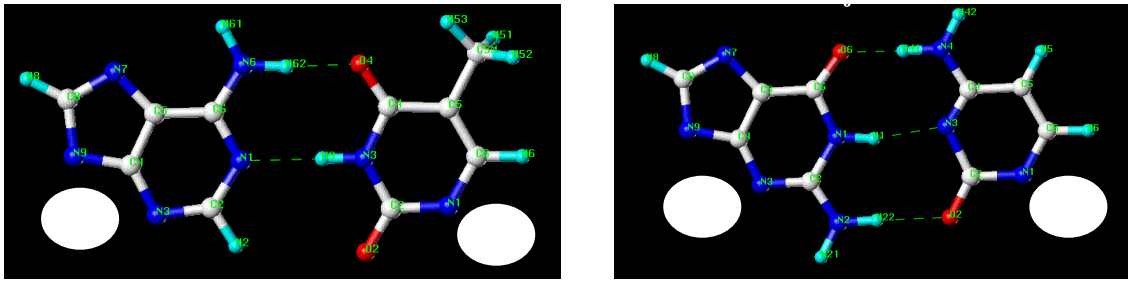
outer membrane leaves cytosolic side cytosolic side, so outside of ER becomes inside of outer membrane.

b) What can we do with this rule?
Predicting TM helices in a sequence is easy, but predicting in/out side is difficult. This rule helps with that.

c) The common household bacterium *bacterius dirtyus* can kill a person by causing an immune overreaction. The small bacterial peptide:

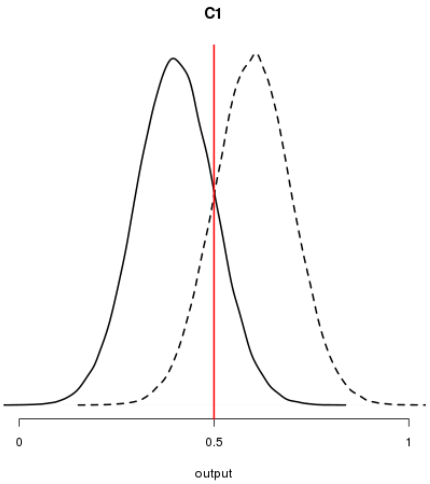
AGNYLHSPGPAGYAALAAYMFLLILVLFPVSFLTLYVKLQ**HKKPRT**PLNYILLLLAVAILFMVLAGFLALMYTSM

is known to elicit this immune response, and it is known that a histidine in this peptide is crucially important for this response. The company Vaccinus Inc wants to make a vaccine against *bacterius dirtyus* by injecting a goat with a dodecameric peptide. Which peptide is the most likely candidate?
I coloured (roughly) the two TM helices red. The little loop in-between (blue) is rather positive, and thus likely in the cytosol of the bacterium and will not be seen by any antibody. So it should be either the N-terminal, or the C-terminal loop, and the latter is way too short. So I suggest the underlined peptide (or take 1 more aa or 1 fewer).
I am always surprised when people do not think about predicting TM helices in question c after answering a and b (more or less) correctly.



3) Above you see four bases.
a) Can you please write their one-letter codes in the white circles.
Left to right: A T G C
b) Which parts of the bases are pointing to the major groove and which to the minor groove?
Top -> major groove.
c) All for bases were taken from the same molecule. Was that DNA or RNA?
Not nice of me to not show the sugars, I admit. But second from left is T and not U, so it is DNA.
Only two people lost points over this question. So it is not a good question. I'll nevertheless keep throwing it in occasionally so everybody will need to know the bases every year.

4) Suppose you have used some machine learning technique and created a two-class classifier for a bioinformatics classification problem, in which we have sequences that elicit a response (positives) and sequences that don't (negatives). The output of this classifier for the negative (solid line) and positive (dashed lines) classes is shown in the figure below. The red line indicates a decision threshold.



a) Given the decision threshold, indicate in the figure the true negative, true positive, false negative, and false positive fractions by writing in the figure TP, FP, TN, and FN at logical places. From left to right you see TN, FN, FP, and TP.

b) Macromolecular structure validation software criticizes structures solved by crystallographers and NMR spectroscopists. Suppose the output of a structure checking algorithm is shown in the figure above. Do you think the decision threshold needs to be set differently? Why (not) and how (not)? Ps, I can imagine that you can equally well defend that the red line must move to the right as that it should move to the left. Do you have equally much imagination as I have?
Obviously, the line should stay where it is to maximise the chance of a correct answer. But, … When validating somebody else's software, you might want to minimize the number of false negatives (= incorrect error messages). But when we are re-refining structures, or looking at our own work, then we want the software to point out anything that might be wrong.
Obviously, there are many other lines of reasoning that might give you points. Actually, this was not a very good question as it was hard to call many answers wrong (partly because I am too nice…).

5) 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?
Obviously, things will depend on the environment (e.g. salt bridge or not), but on average and everything else being neutral:
1) Ile -> Asp at the surface. Will do little, but might help a bit against protein aggregation.
2) Asp -> Ile at the surface. Will do little.
3) Ile -> Asp in the core. Most likely devastating.
4) Asp -> Ile in the core. Most likely very good for stability. But please ask why the Asp was there in the first place.
Many people answered this question correctly, but without any explanation. Grrr.

6) Describe in at most ten words per term what the term is/means:

Z-score Number of standard deviations a 'score' is away from mean

B-factor Mobility / positional error of atom in Xray structure

R-factor How well does Xray coordinate model describe Xray experimental data

Force Field data+rules to describe system and optionally predict its future

MD Molecular dynamics simulation

BLAST Tool to find homologous sequences in sequence database

Homology Modelling Predict protein structure from its similarity to a known structure

PDB Database for experimentally determined macromolecular structures

CSD Database for experimentally determined small molecule structures

Salt Bridge Interaction between oppositely charged atoms

Bond angle Angle between two covalent bond vectors originating from same middle atom

Torsion angle Rotation between to bonds connected to opposite ends of a middle bond.

NMR Nuclear Magnetic Resonance (technique to e.g. solve protein structures)

SCOP, CATH, DALI Databases of protein structure relations, family relations.

HSSP Database of multiple sequence alignments against PDB file sequences

DSSP Database of secondary structure assignments of proteins in PDB

Occupancy How much (time/location) atom is here.

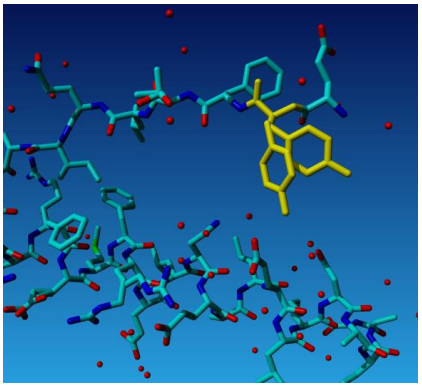
Resolution How well does crystal diffract, smallest distance observed.

NOE Nuclear Overhauser Effect. Main thing you obtain in NMR experiment.

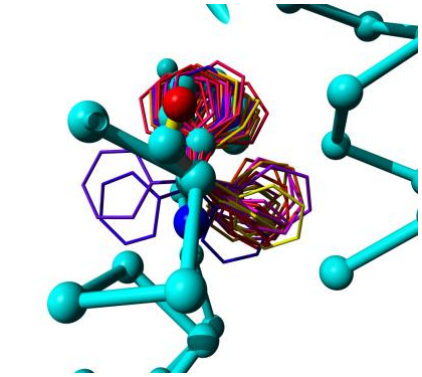
Crystal packing artefact. Something wrong caused by the fact that proteins touch in crystal

MD Time step Time in MD that atoms fly in a straight line (order of femto seconds)

8) Some questions about side chain conformations



a) In this picture the tyrosine in the top right seems to have two side chains. Can you explain why that is?
Both conformations are observed in a fraction of the molecules in the crystal. The final structure is the 'average' of all conformations in the crystal.



b) The rotamer distribution is shown for Phe at position 38 in PDB-file 6TBP (Phe 38 itself is also shown as a ball-and-stick model). What can we learn from this picture?
We see that the three primary rotamers (i.e. the three that have different chi-1 angles) are observed to different extends. The Phe-38 itself sits in the most populated rotamer cloud. So, probably, we can conclude that there is nothing to worry about.

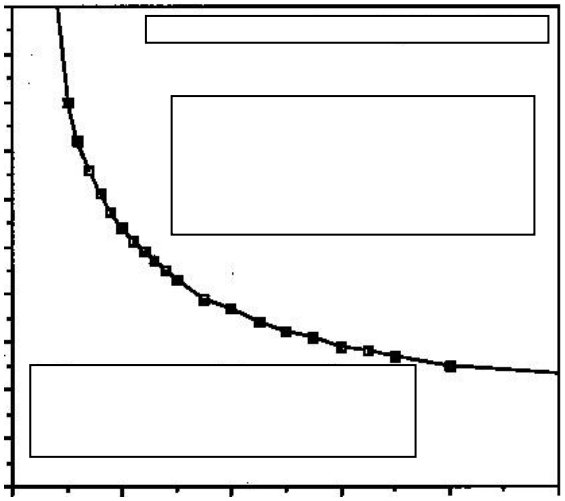
c) Whose tombstone is this? Boltzmann
d) Which formula is printed above his head?
The Boltzmann equation: Entropy = k* partition function
e) What is the rule of 10?
1 kCal/Mol gives about a factor 10 shift in equilibrium
e) How do we get from this formula to that rule of 10?
fill everything in (with K=10) in $\Delta G=-RT\ln(K)$
f) How is the rule of ten related to the questions 8a and 8b?
8a) Both positions are (perhaps only roughly, occupancies are not given) energetically equally likely in crystal, so both show up. YASARA then shows both possibilities at the same time. 8b) A rotamer density (=number of observed rotamers in same cloud) difference of a factor 10 indicates an energy advantage of 1 kCal/Mol for the more densely populated rotamer.

Funny how many people answered d) with S=-k.log(W), which is the correct answer, but didn't bring them many points…

9) ATP tends to be bound to proteins in a deep cleft from which only one or two phosphate groups protrude (stick out). Almost always do we observe a glycine rich loop making contacts with the protruding 2nd and 3rd phosphate; often these loops have a threonine directly adjacent to the two glycines, and there are a few more striking sequence features. These loops tend to be around 12 amino acids long. I want to build a force field that predicts if a glycine rich loop actually is an ATP binding loop. Can you describe how I should do that?

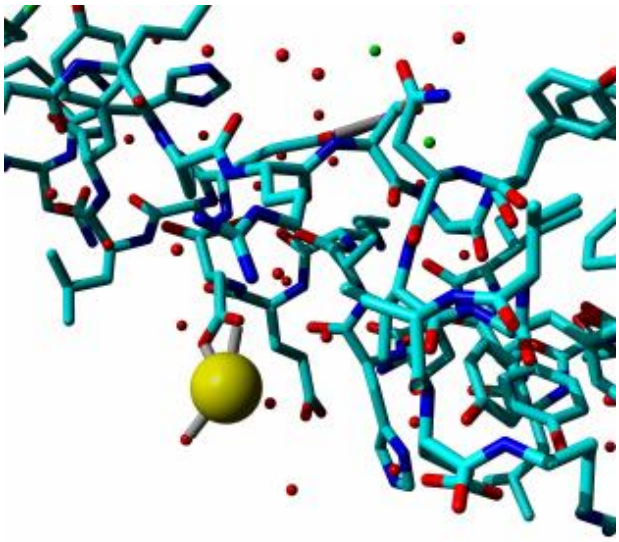
It's a long story, but summary:
1) Collect loops that are and loops that are not binding. (Keep a portion of the data aside for testing of the method at the end).
2) One way or another count residue types (small dataset: count just all; large dataset: count residue types per position).
3) Determine asymmetry in counts (counts in non-binders is null model, and those frequencies are called predicted, or pred; counts in binders are called observed, or obs).
4) Determine pseudo energies with $E_i=\log(obs_i/pred_i)$ for each amino acid type counted.
5) Use these energies E_i to calculate energies for binders and non-binders. This will, with some luck, give a plot as in question 4. And what you wrote for question 4 applies here.
6) Design a method/recipe (again as discussed in Q4).
7) Test the method on the data you put aside in step 1.

10) In an article about homology modelling I found this plot, but upon copying it, the annotation of the plot got lost. Can you give me the following texts:



Title of the plot (top text box)? Sander Schneider plot, or Hom Mod threshold plot, or…
Text for horizontal axis? Length of alignment
Text for vertical axis? % seq identity
Text in middle text box? Homology modelling is possible and will go OK.
Text in bottom text box? Alignment is too uninformative to start modelling without a lot of extra knowledge

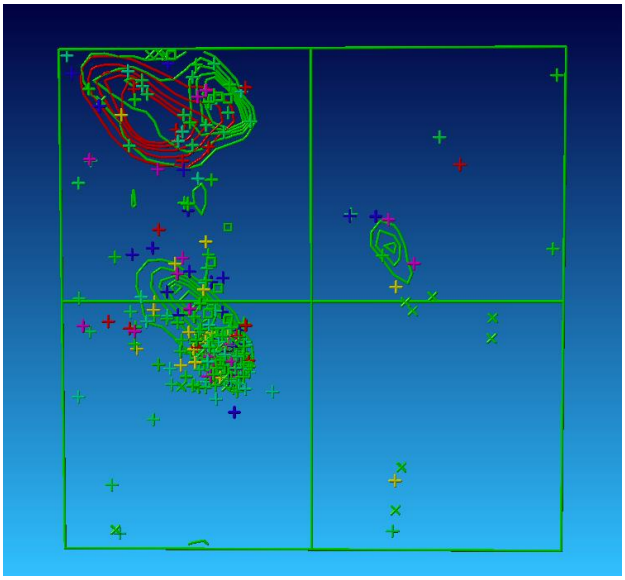
This article described homology modelling as a three step process. Can you tell me (briefly) what each of these three steps described? And in which of those three steps did I find this figure? (Figure in step 1)
1) Template detection (BLAST); alignment; alignment optimisation (also structure based).
2) Insertions, deletions, sidechain modelling, optimisation by specialized MD.
3) Validation, optional iteration of previous steps, interpretation.

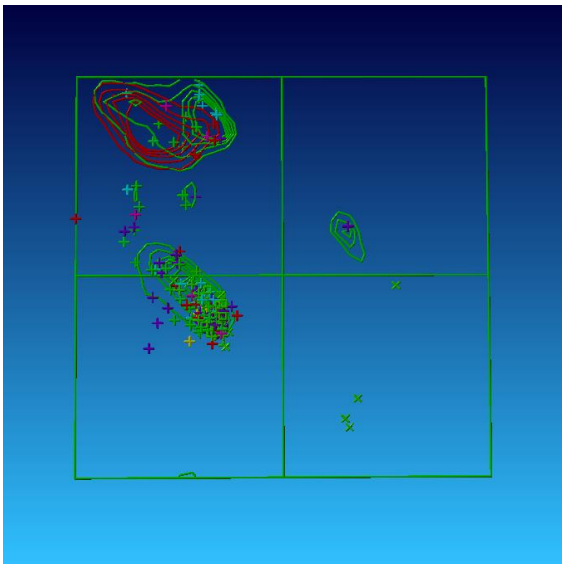


11) This picture shows a small part of the extra-cellular protein Extractase. I see one bound ion (the yellow ball…). Which ion is that most likely, and why?

Two correct answers:
Na because extra-cellular and bound to 1 negative residue, or
Ca because extra-cellular and bound to 1 negative residue but another negative residue (binding via water?) and the C-terminus are nearby.

12) Below you see two times two Ramachandran plots.

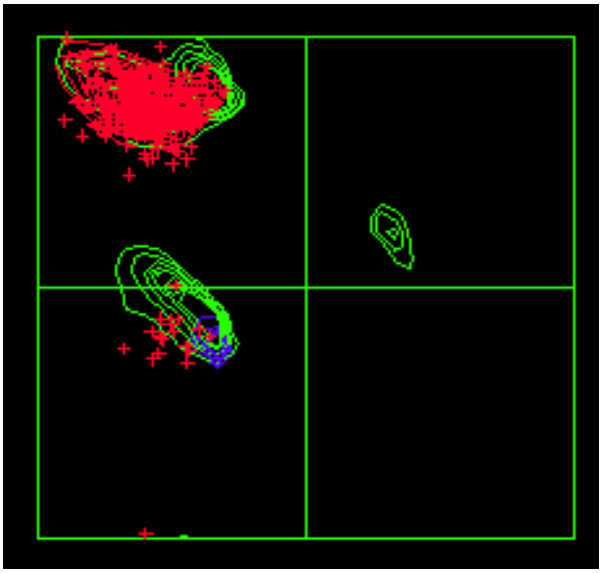


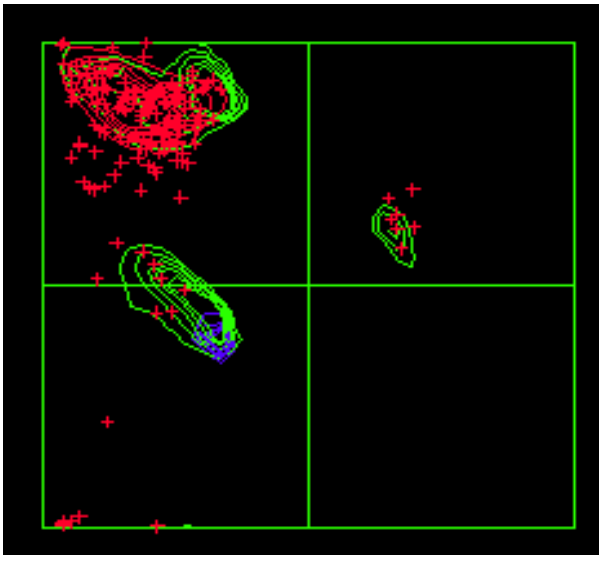


Rhodopsin Haemoglobin

a) The top two are for bovine rhodopsin, and haemoglobin. The residues are represented by crosses or squares and coloured by characteristic (positive=blue; polar=purple; negative=red; hydrophobic=yellow/green/light-blue). Unfortunately I forgot which one is which. Can you figure that out?

Rhodopsin is membrane protein. Those cannot be solved as well as water-soluble proteins. So, the lower quality Xray structure gives a less-clean Ramachandran plot.





Leucine Asparagine

b) The bottom two Ramachandran plots hold only residues found in β-strands in 100 high resolution X-ray structures, shown as red crosses. The one plot contains only asparagines and the other one only leucines. Unfortunately I forgot which one is which. Can you figure that out? Asn can make H-bonds with the local backbone. That creates a force which pulls at backbone and thus pulls atoms away from perfect phi-psi position. Asn thus has less-clean Ramachandran plot.

13) Given the sequence of the active site domain of the enzyme *somethingcutase*:

```
.GTHTVTIKVDGNSADLLKAVAQAAGNGSYIEITGDGNNPAELLMKAAIQLVMRAGNGDVKITVGGG.
..ssssssss....hhhhhhhhhhh....sssss.....hhhhhhhhhhhhhhhh....sssss....
```

What does *somethingcutase* cut: a sphingolipid, the small domain of the Spats protein, a polysaccharide, RNA, or perhaps even something else.

Rule 1. When you see a sequence, you predict its structure. Having done that, half of the points are yours already.
It is a strand-helix-strand-helix-strand domain. The active site residues are then expected at, or directly after, the C-terminal ends of the strands. I coloured the suspicious residues green. Two Ds and one S (by this time you have ¾ of the points, already). Active sites tend to come in three tastes: SHD for many, many activities; several histidines (and often a Glu or Asp or something else) around a Zn for many activities; or DD(E) for sugar cleavers/cutters/movers/pasters/etc. So, polysaccharide.

Kladblaadje 1.

Kladblaadje 2.