**Structure Function Bioinformatics 2012-2013 brief answers.**

1) Harry is trying to purify a transcription regulation protein, but he keeps loosing very much protein during the purification procedure, so he decides to introduce some mutations to make his protein more stable.

John is working for an Irish beer brewer. To make green beer for St Patricks day he needs to add some b-lytic protease very early in the process, exactly in the high temperature clearing step.

a) Neither Harry nor John has a 3D structure of the protein. And they can both at best make a very poor homology model that at best gives them ideas about which residues are in the core, in the active site, or at the surface. So their plans to stabilize their proteins must be based on rather general principles. What types of mutations do you think they should make (please add some examples and details to the answer)?

They can choose for entropic or enthalpic stabilisation (albeit that in practice this separation is less clear than in theory). Enthalpic stabilisation is based on the introduction of interactions in the folded form of the protein that are absent in the unfolded form. Examples are the introduction of salt-bridges of filling cavities with hydrophobic residues, etc. Entropic stabilisation is based the reduction of flexibility of the unfolded protein by mutations of the type Gly -> Xxx, and Xxx -> Pro.

b) Harry and John should approach their problem differently. Why?

John works in beer. That is a liquid with thousands of other ‘things’ swimming around. So John should care about destruction of his protein by the protease itself, or by other proteases, or by yet other processes. So John should focus on preventing degradation while Harry can simply do whatever works fine. This is rather much the same story as was told for erabutoxin in which we find many cys-cys bridges and prolines in every loop to protect it against the inate immune system.

c) And how does that reflect itself in the mutations they each will try first?

So, Harry can do whatever is good for stability and John should focus on things that reduce the (surface) mobility.

2) What are the major differences between solving a structure with X-ray diffraction and solving it with NMR spectroscopy in terms of:

a) The way the samples are prepared?

Crystals versus (highly concentrated) solution.

b) The type of data that is collected for those samples?

X-ray: reflections. NMR interatomic distances and some torsions angles.

c) The way the structures are calculated from the data?

X-ray: phases need to be obtained for the reflections but then a Fourier transform of those reflections gives you an electron density map. NMR: shake the molecule around (with MD) till atomic positions are both chemically sensible and in agreement with the data.

d) The way errors result from the whole process?

X-ray has Fourier transform between data and structure. So each reflection has influence on every atom, and vice versa. So error somewhere will make everything else a bit worse too. NMR errors tend to be, and stay local in nature.

3) Explain in less than ten words per topic:

a) HSSP file: PDB file based MSA

b) Ramachandran plot: Plot of psi versus phi to describe backbone torsion angles.

c) Rotamer: Either side chain conformation, or preferred side chain conformation.

d) Torsion angle: Angle between four atoms, rotation angle over middle bond.

e) Helix capping: Process of placing charged groups to compensate helix dipole.

f) R factor: Indicates how well X-ray coordinates agree with the reflections.

g) B factor: Indicates how mobile (or disordered) is an atom.

h) Z score: Number of standard deviations an observation is away from the mean.

i) Force field: Data plus algorithm to describe a system and predict its future.

j) Crystal packing artefact: Artefact in structure caused by packing of neighbours in crystal who are not neighbours in biology

4) We observe certain residues much more often than others in active sites. For example, I see Ser many more times in an active site than Thr, or Asp many more times than Glu. Below I list three such mutations. Explain those asymmetries.

a) Ser > Thr. Thr has extra C-gamma. That makes it less flexible to place th O-gamma at exactly the right place to do its quantum chemistry. C-gamma is also ‘in the way’.

b) Asp > Glu. Upon activation we normally see immobilisation. That is one degree of freedom more expensive for Glu.

c) His > Phe. Enzymatic activity is always associated with moving protons around. His can do that; Phe not.

5) The entropy of water is part of the answer of three of the four following sub-questions. But for the one question where entropy of water is not part of the answer, part of the answer looks very much like “the entropy of water”.

a) I want to stabilize a protein by mutating a buried Val into an Ile. This mutation can either fail or work out fine. Explain how it can fail, and if it works well, explain why. It can fail if it doesn’t fit well, but if it fits well, then the number of waters around the side chain in the unfolded form is bigger for Ile, and thus the EoW gain upon folding is bigger too.

b) Most metabolites bind to enzymes in a deep cleft or pocket. Why? In the pocket the metabolite touches more enzyme atoms which gives it better specificity. In the pocket the waters are badly restricted in their freedom so that their release leads to much EoW.

c) Transmembrane helices are rather hydrophobic. So many of them can happily swim around in a membrane just by themselves. Nevertheless, most transmembrane proteins consist of bundles of transmembrane helices that nicely stick together. Why? Transmembrane helices are kept together by interactions (just like everything), but also by the entropy of lipids. This gets worse if helices swim away from each other. At a high level this resembles the role of the EoW in ‘normal’ protein folding, but it is different as lipids are involved and not waters.

d) If a small cytosolic vesicle bumps into a big cytosolic vesicle and their membranes merge, then the content of the small vesicle rapidly flows into the big vesicle. Why? Water inside a big vesicle has more freedom and on average less contact with the vesicle wall. So waters are on average happier in EoW terms in a big than in a small vesicle.

6) I have a series of datasets. These data were collected to answer biomedical questions. In all cases listed here some machine learning method can be used to answer the question. Can you for each case explain to me which would be your machine learning method of choice?

In this question often multiple correct answers exist. Their correctness depends on the reasoning...

a) A medical doctor has been working on kidney cancer. He collected for hundreds of patients their age, sex, smoking behaviour, weight, medical history in terms of previous cancer treatment, three genetic markers, and the success of a series of treatments. He wants you to tell him for the next patient which treatment to choose. Feels most like work for a decision tree.

b) A colleague in cell biology has collected a few thousand short (35-65 base-pair long) RNA fragments She expects that about 2/3rd of these fragments contain a similar binding motif for the transcription factor protein bagatelle X. She wants to know the motif and which sequences hold that motif. Typical case for HMM (Hidden Markov Model).

c) A colleague in biophysics has been solving the structure of shikimate transferase in the apo form. This is a two-domain protein. He also solved many structures of this enzyme each with another inhibitor bound in the active site. His hypothesis is that inhibitor binding will cause the two domains to move toward each other over the active site cleft, thereby closing this cleft. He superposed all his structures, but the process seems more complicated than originally thought. He still wants to know what happens structurally upon inhibitor binding. Very much data from which a consistent subset should be selected. Shouts for Random Forest.

7) These sub-questions all relate to membrane proteins, not to water soluble proteins.

a) Give two significantly different classification schemes for membrane proteins.

Extracellularly anchored – transmembrane – intracellularly anchored

Functional – structural

Signal transmission – material transmission

Pumps - pores

b) Mention the two major classes of molecules that transfer information through a membrane.

GPCRs, Tyrosine kinases, but there are more.

c) Most transmembrane proteins consist of helix bundles, far fewer consist of beta barrels, and hardly any contain loops and turns. Why?

Inside the membrane we need to have all H-bonds satisfied. That is easiest in helices, doable in barrels of strands, and impossible for loops. If loops are observed (like we saw in the course in aquaporin), then the free backbones tend to have a functional role.

8) A threonine is buried deeply inside the protein protalionase. Its Oγ doesn't make a hydrogen bond. Which is the best mutation for improving the stability of protalionase? Briefly describe why. Make it a Val. Same size so it fits, and the gain comes from the extra gain in EoW upon folding as described above.

9) How do you make an antibody against the toxin of the Texan dessert snake? Describe which bioinformatics tools are needed in the process. BLAST against PDB to find homolog (or its own structure if Murphy is on vacation). Use homology modelling to get structure. Find nice stretchy at surface that is a bit extended. Use software that predicts antigenicity for aa types to find best stretch. Inject this stretch in a rabbit and hope that antibodies will also recognize the full protein.

10) Why do all good secondary structure prediction methods use multiple sequence alignments as input? So you can use more information. You sometimes find a Ser or Asp or so in a helix, but hardly ever will you find them conserved in the whole family, unless they have a functional role.

11) Why does a salt bridge care (much) less about the inter-atomic distance than a Van der Waals interaction? Use formulas to illustrate your answer.

Van der Waals interactions go with 1/r^6 and 1/r^12 electrostatic interactions with 1/r so the latter changes much less when you make a half Angstrom error.

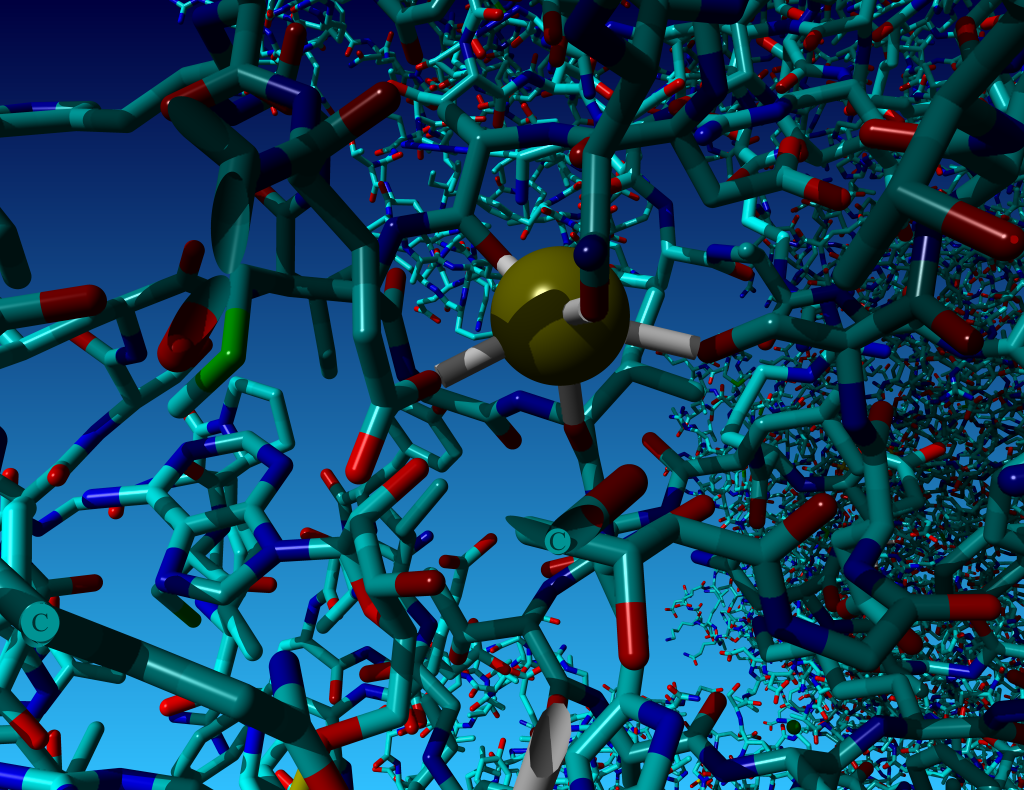
12) Many bacterial extracellular protein sequences start with a rather hydrophobic helix that is about 20 amino acids long. Some of these proteins use this helix to anchor themselves to the membrane. Some other, though, loose this helix (by proteolysis) upon secretion out of the cell.

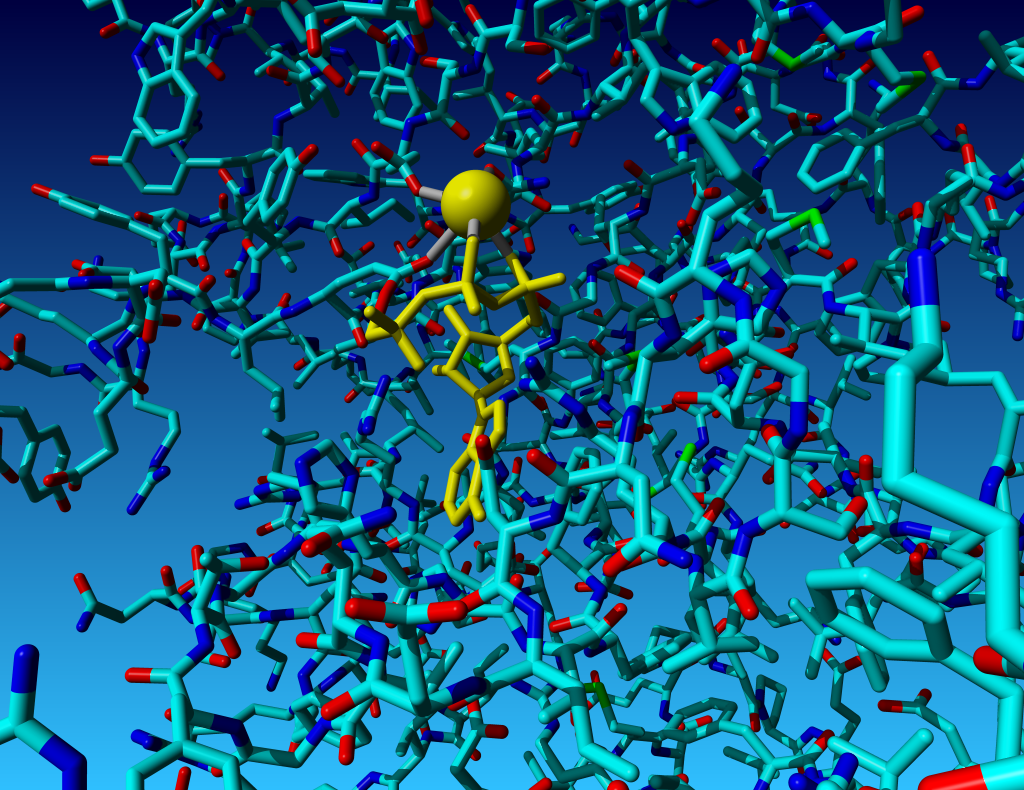
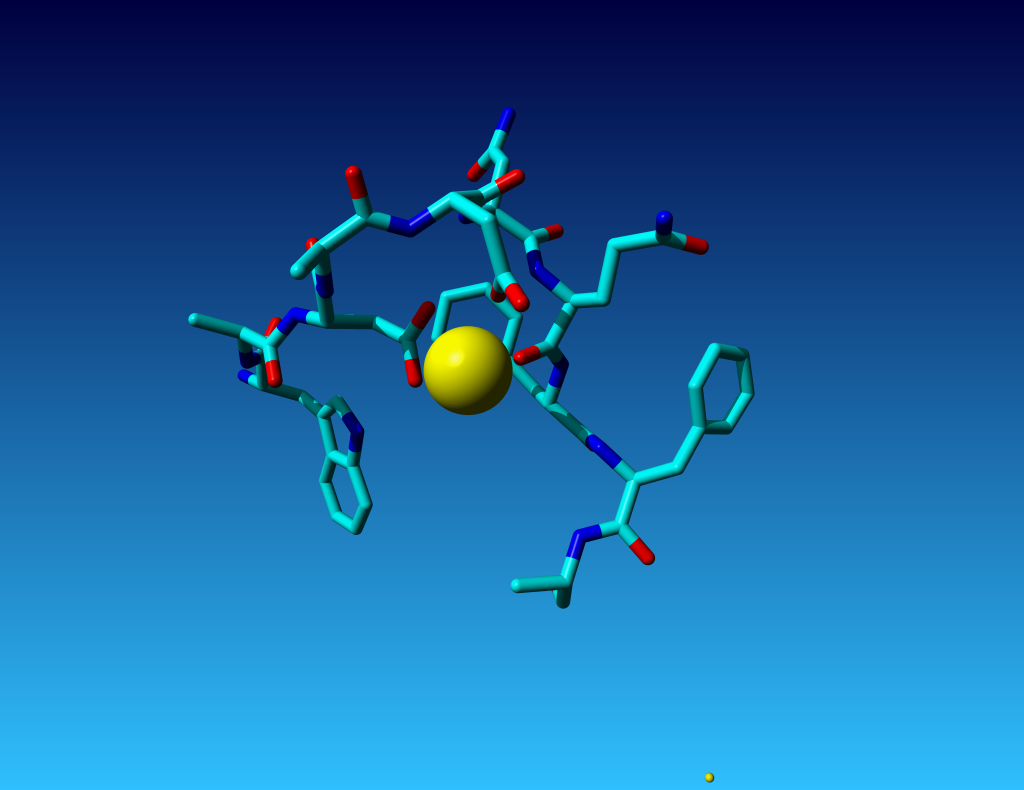
We have sequences a few thousand proteins of each of these two classes. Can you design a force field that helps us predict what the behaviour (anchoring or cleavage) will be of these helices in extracellular proteins of a recently discovered bacterium that we are sequencing right now? Try to first think of the main steps in the process and try to cleanly structure your answer along these main steps.

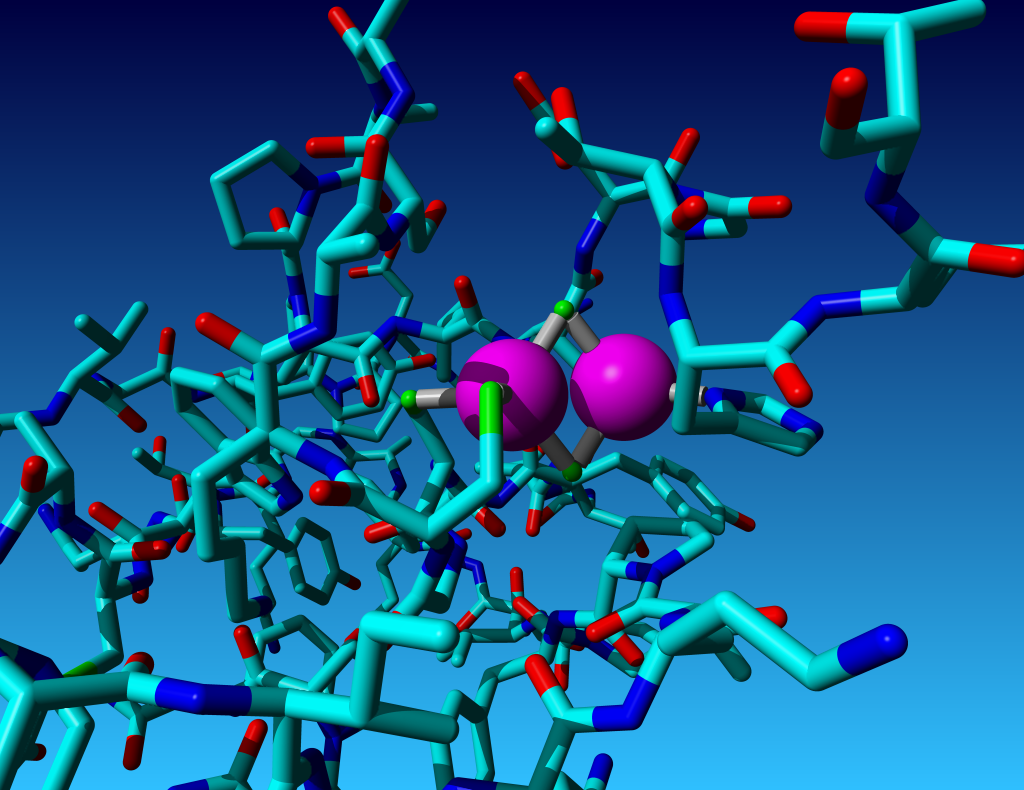
Get datasets one with the known helices and one ‘other’ dataset. The precise question determines how you select that other set.

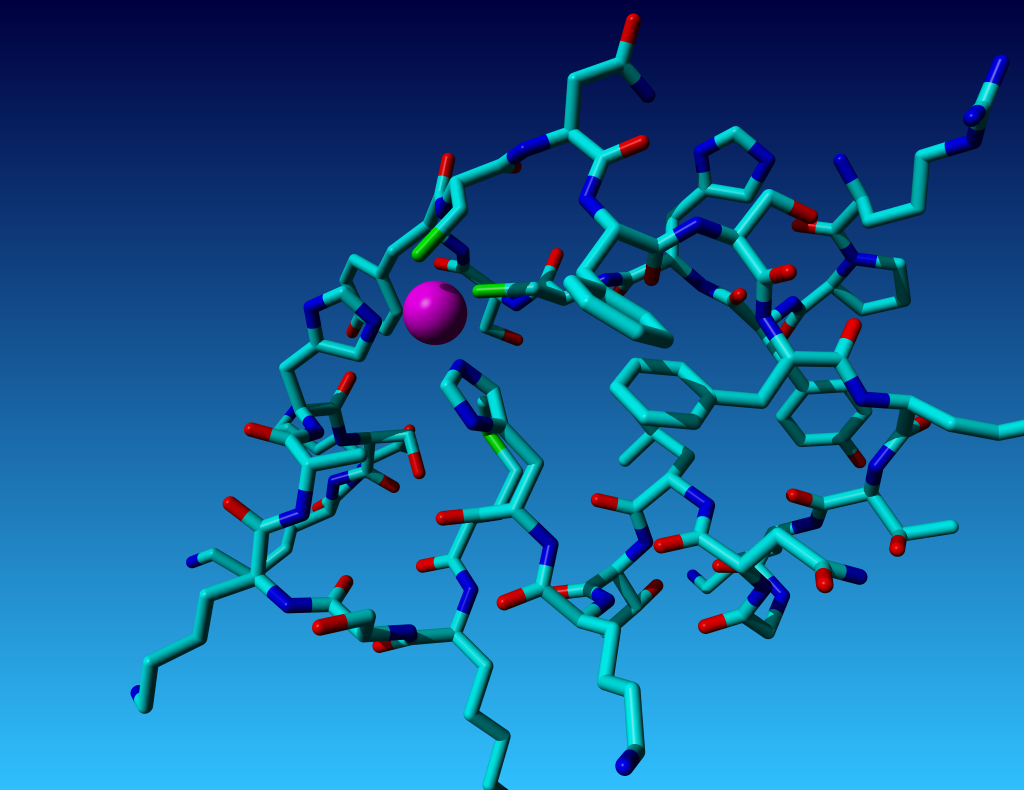
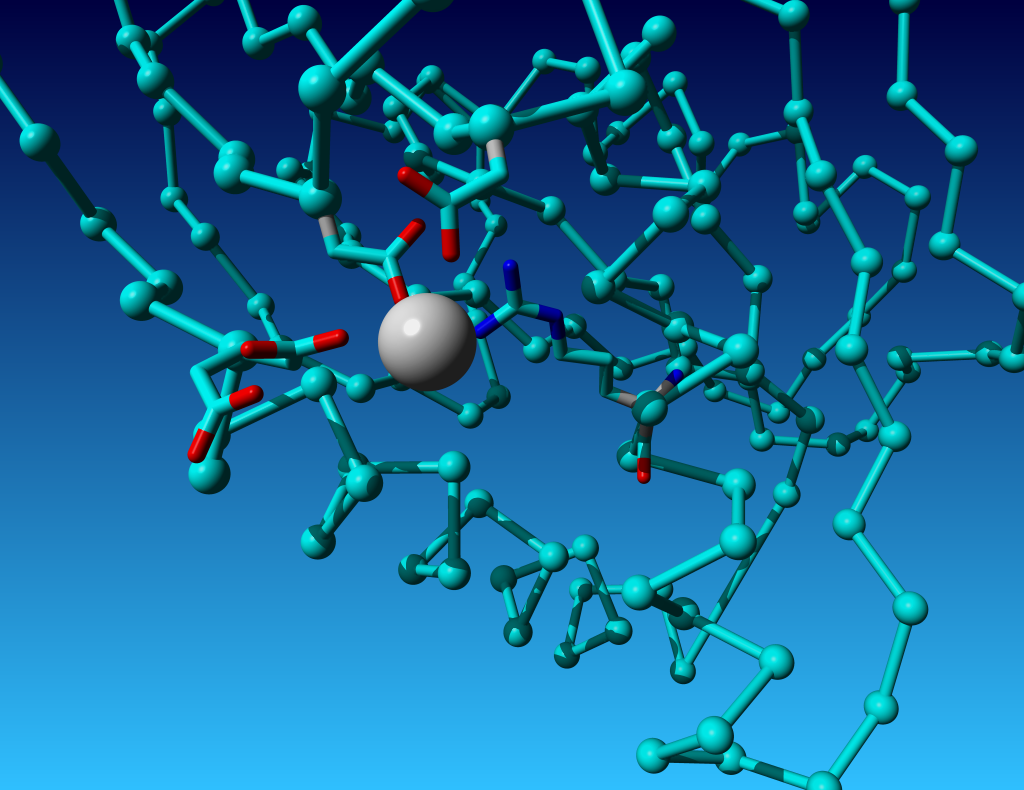
Determine the amino acid frequency at, for example the five spots before and the five spots after the cleavage spot (harder to define in the non-cleaved dataset, but you can simply take a whole series of 10-mers around the helix end). The frequency distribution in the latter set is your null-model. Now check how much more or less each residue type is seen at each position in the 10-mer in the positives than in the null-model data. Take fractions and take logarithm to convert changes into energies. Now think of some algorithm you can use. Simplest seems to score 10-mers. Optimise that method with the data used so far, and finally test the method against some positives and some controls that you put aside at the beginning of the study.

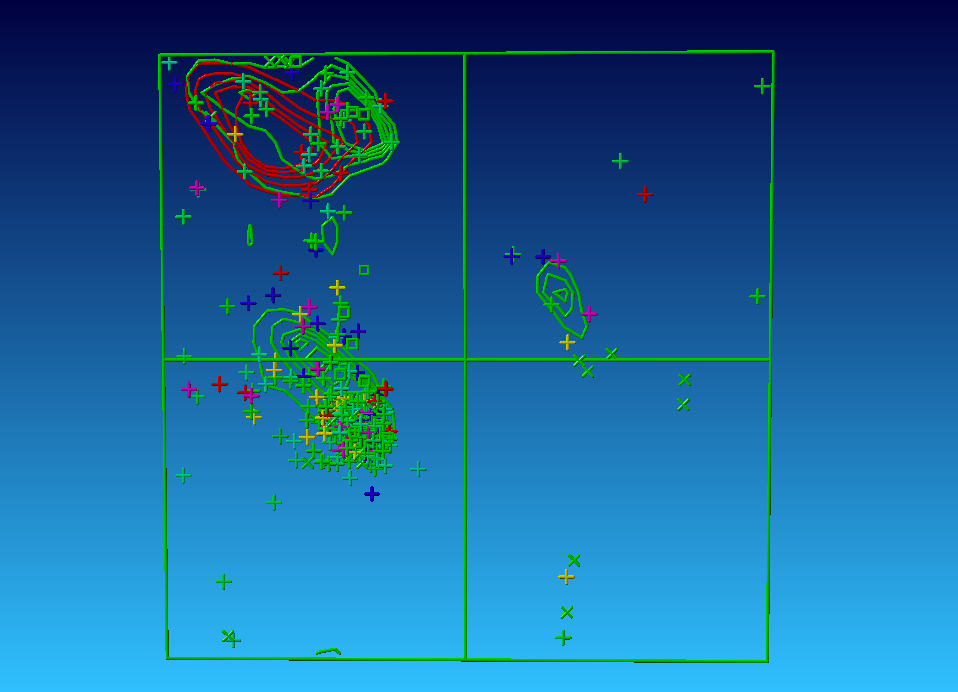
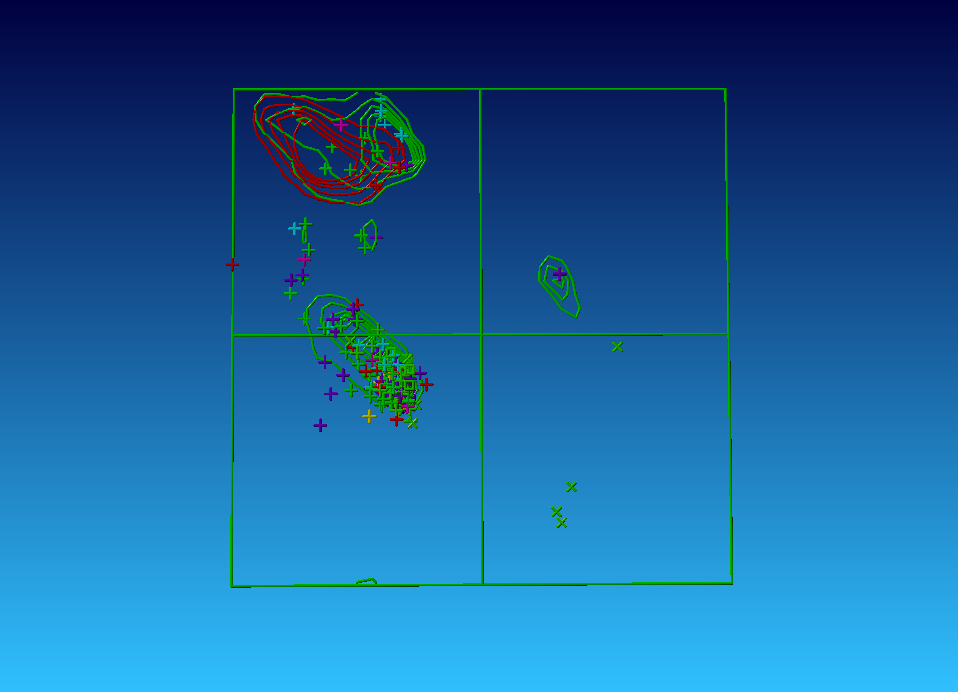
The buzzwords needed to score the maximum number of points are: datasets, amino acid types, null-model, logarithm, algorithm, testing against non-used data

13) I made a series of pictures of typical environments for metal ions in proteins. The ions are, in alphabetical order, Ca, Fe, K, Mg, Sm, and Zn. Unfortunately, I lost the figure captions ☺. Can you help me and find out which ion is shown in which picture (in the picture with two ions both are of the same type). From left to right from top to bottom: Mg sits against nucleotide phosphate. Ca is bound by oxygens including two Asp. K is bound by oxygens including one Asp. Zn is bound by two His and two Cys. The highly complicated one that we don’t know is even bound to an Arg, must be Sm. Fe (two of them) bound by many S in Fe-S cluster.



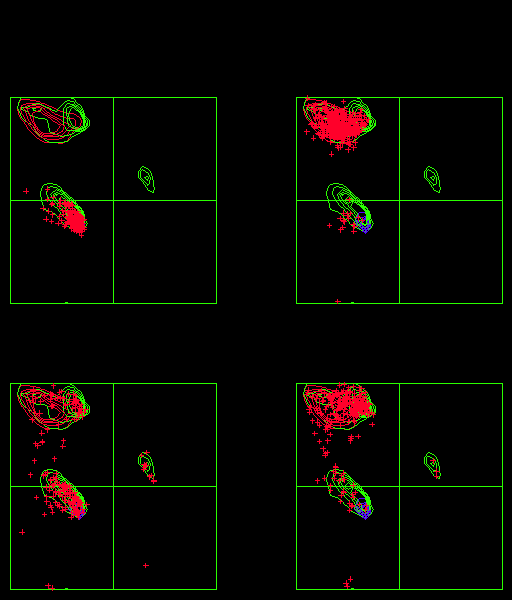
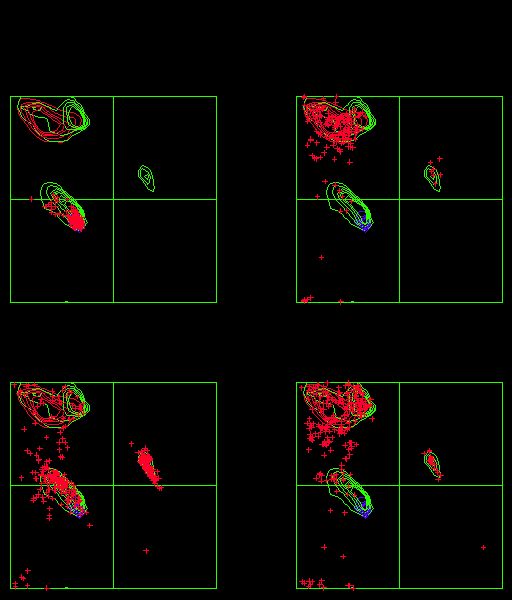




14) Below you see two times two Ramachandran plots.

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?

Left one is more chaotic and membrane proteins are hard to determine with great accuracy. Further right one has more blue and red stuff in the helices, and finally opsin is bigger and left one has more crosses.



b) The bottom two Ramachandran plots hold only residues found in b-strands in 100 high resolution X-ray structures. The one plot contains only asparagines and the other one only leucines. Unfortunately I forgot which one is which. Can you figure that out?

Asp makes H-bonds with local backbone. This is a force. Force means displacement from the ideal situation. So right one is Asp.