Structural Bioinformatics
Assessed Coursework (2022)

The aim of this assessed task is to allow you to demonstrate some of the practical skills and understanding you have gained during the module. You will have the opportunity to apply your skills to study **a real example of a protein-ligand complex**. The first part concentrates on the protein, the middle parts concentrate on the small molecule bound in the complex, and the final part on a hypothetical modelling scenario involving the complex. **Coursework contributes 50% of the total marks for this module**.

You are expected to use your initiative in completing this coursework, we deliberately do not tell you how to go about answering these questions as there are many ways to answer to each one. Resources have been used in previous tutorials or mentioned in lectures so you have to use this information (please listen to the lectures and look at the tutorials again), but you are free to use different ones. You can use any resource on the web you choose to get an answer to these questions, you should use Chimera to illustrate or calculate at least some of your answers. **In each case**, you should explain, in as much as possible the details, of how you have obtained the answer and **identify the resources or methods you used to find the answer**. Show details of any calculation and analysis.

For all web-services used provide a web-reference and, where one is available, a reference to the original publication of the methods. Follow the instructions given by the authors of the method on how to cite their work. Where you have used information from a publication, also include that in a reference list at the end of your coursework. References in the text and the reference list should follow the format of the journal *Bioinformatics*.

To get full marks you should expect to write (with correct answers) **~ 1500 words**, a total of **12 figures & tables** (that you have to prepare yourself using software packages), a list of **references** and an **appendix** containing code or commands used in the analysis.

Queries regarding the content of the task can be addressed by e-mail to Mark Williams - ma.williams@bbk.ac.uk. There is no guarantee of response to queries after the 16th of December 2022.

Plagiarism: In preparing their answers, students should take note of the School's policy on plagiarism. If you have any doubt about whether part of your coursework may be considered plagiarism, please discuss it with the module organiser prior to submission. **Your coursework must include the cover sheet declaration with respect to plagiarism found on the Moodle site for the module.**

Deadline: Midnight on Sunday, January 8th, 2023

The electronic copy (PDF format) of the coursework and the requested PDB format file of the model should be submitted on Moodle.

In the interests of fairness, the deadline will be strictly enforced. Extensions will only be granted under exceptional circumstances (e.g., possibly if you are ill and have a certificate from your doctor) via the College's "mitigating circumstances" procedures. In the absence of mitigating circumstances, if you hand your coursework in after the agreed deadline, your mark will be reduced by 10% for the first week late and capped at 50% for the second week late.

Do not add links to temporary results web pages in the report that you have generated while carrying out work for the report – these will not work for someone else trying to open them later!

Analysis of a Protein-Ligand Complex

You have been allocated the structure of a 'target' complex of a protein with one or more ligands. The protein belongs to the super-family of short-chain dehydrogenases/reductases. This is a super-family of enzymes sharing the same fold but varying widely in their function. Most (but not all) are dehydrogenase or reductase enzymes, but their substrates can be very different.

Part A: Analysis and description of the structure [22 marks]

a) What is the enzymatic reaction that this protein catalyses in vivo? Write the relevant equation **[2 Marks]**

b) What is the CATH structural classification of your protein? **[1 Mark]**

c) Describe the structure of the biologically active form of the protein, accompanied by relevant illustrations including your own Chimera figures and a plot indicating the secondary structure. **[5 Marks]**

d) Is the structure of good quality? Support your opinion with argument and evidence, including supporting plots. **[4 Marks]**

e) What small molecule ligands are found in your structure? One is the cofactor. One of the others is a small molecule that may have some resemblance with the reaction substrate for this enzyme (NOTE: this resemblance might not be great!) or acts as an inhibitor to substrate binding. What is this substrate/inhibitor-like ligand? (you will be using this ligand in Parts B & C). **[1 Mark]**

f) Identify the functional groups of your protein that interact with this substrate-like ligand and the type of interaction (hydrogen bonds, hydrophobic interactions etc.). Present this information in a table and as a figure (e.g., using ligplot/Chimera). **[5 Marks]**

g) Which of the interactions identified in (f) are the most important to the binding affinity and which to the catalysed reaction? Support your opinion with argument and evidence. **[4 Marks]**

Part B: Identifying potential inhibitors [12 marks]

Now you will need to compare your ligand to a representative dataset of drug or drug-like molecules and try to find molecules that could act to inhibit the enzyme. It is up to you what sources you use and how you build this dataset, but some obvious sources are DrugBank, ChEMBL and ZINC. Full collections of drug molecules are very large, so ideally we are looking for a reduced set of representative structures selected according to some sensible criterion/criteria.

a) Create a connection table or SMILES string for the selected ligand in your complex (do not select the cofactor!). You may find this molecule in .sdf or or .sd or .mol format. (see: http://www.ebi.ac.uk/chebi). Alternatively, you can find or create your own SMILES string for it. You can use any resource on the web, but explain how you got your connection table/SMILES string and copy the file as text in your answer. **[2 Marks]**

b) Does your selected ligand obey the Lipinski rule of 5? Explain your answer. **[2 Marks]**

c) Create a set of drug or drug-like molecules for subsequent searching. Document how and why you chose your 'search set' and how many molecules you have ended up with. You should have either SMILES strings or an .sdf file for these molecules. Do NOT include the whole list in the coursework you hand in - the first page of the list/file will be fine in the appendix. **[2 Marks]**

d) Fingerprint the drugs in the set and calculate the Tanimoto similarity scores for the comparison of your selected ligand against each one of the drugs. You can do this using the OpenBabel ready-made programs (i.e. the binaries you can access as advised in the chemoinformatics tutorials) or by writing your own program. No extra marks will be given for writing code as opposed to using ready-made binaries. Produce a histogram of the scores and submit this together with a list containing the actual commands you used or the source code of your program (in the appendix if the code is longer than a few lines). What are the mean and standard deviation of your scores? **[4 Marks]**

e) Did you find any drug-like molecules that significantly resembles your ligand – if so, give examples. If not why do you think nothing similar was found? [**2 marks**]

Part C: Docking to a model structure [16 marks]

Imagine for a moment that the structure of your target protein is not known and that there are no very similar homologues with a known structure. Use the sequence of your target protein to make a model structure using homology modelling (in making the model be sure to exclude any template with more than 70% sequence identity).

a) Using your method of choice (or multiple methods) identify a reliable 'template' protein that is homologous to your target. Explain how you chose the template and what score/s and other criteria you used to reach this decision. **[2 marks]**

b) Generate a comparative (homology) model of your target protein using the template structure you've selected in Part C (a). You can use any method you wish, but you need to explain your choice and how you have obtained your model (including model assessment score/s). Be careful to document any settings of the modelling process (especially non-default ones). Give a brief evaluation of the quality of the model structure. Include figures illustrating your results. **[4 Marks]**

c) Compare the best scoring model of your model to the actual PDB structure of the target. Describe any differences in the protein structure between model and the actual structure. Use assessment scores (e.g RMSD and TM-score). Illustrate global and/or local structural differences in a figure (or figures) **[4 marks]**.

d) Dock the ligand molecule identified in Part A (e) into the *comparative model* using *SwissDock* (or any other program/server of your choice, eg. *haddock*/*autodock vina*). Present the results in a table/figure **[2 marks]**

e) Describe and illustrate the similarities and differences of the ligand pose and interactions between protein and ligand in the model of the complex and the actual structure from the PDB. Why do you think any differences have arisen? If there was no PDB structure of the target complex, how useful would your model structure be in understanding ligand binding? **[4 Marks]**

Submit a PDB format file containing the docked ligand complex with the best scoring model with your report (via Moodle/Turnitin).