BIOL 5172M Assessment 2

Important Information

This assessment constitutes 70% of the total mark.

Introduction

You are required to analyse a protein sequence by bioinformatics methods. What is its likely function? If it does not have a structure can you determine a model for the structure? What does the model tell you about its function? It might also be a "hypothetical" protein, where the structure has already been determined by a structural genomics consortium. In which case can you use sequence and known structure to determine protein function? You are not restricted to methods covered in the lectures, but you should focus on methods with the general aim of prediction of protein function and/or structure from sequence.

You can either request a sequence to the module manager or chose your own.

Some useful guidelines when choosing the sequence

It should be a protein sequence for which there is only a small amount of functional information (e.g. in Swiss-Prot) and no structural information (preferably with some biological relevance such as one relating to a particular disease, or of interest as a drug target). For example, the annotation score at UNIPROT of 1-2 stars for this sequence and those in the Similar proteins section is a good indication that the sequence is relatively unknown.

It may be a sequence that it is of interest to you for some reasons and you may be unsure if your chosen protein fits any of the above routes, or you may even have an idea for a completely different taks - **If so can you email me a very brief outline (~200 words) of what you intend to do, e.g., the list of 5-6 tools you will use to investigate the protein (by Thursday December 14th) , so I can give advice before you proceed further.**

We have collected some sequences from colleagues in this Faculty which fit to

the description above: you may ask by email to the module manager a sequence if you prefer not to chose your own **(from Thursday December 14th)**

Requirements

If you chose your own sequence, this will be your first task. This may take you some time and some preliminary analysis using the bioinformatics tools you have learnt about in this module.

The second task is to find whatever you can on that sequence (see "hints" below). Somewhat more is expected from those who have not chosen their own sequence!

The third is a presentation of your results in the style of a scientific paper.

There should be an abstract of up to 250 words. An introduction, detailing what you have done, and why it is interesting, perhaps with a brief literature review if relevant. **You are not expected to give details for the methods you have used, but do cite primary references if you use them**. The remainder of the paper should be:

- Results and Discussion section giving relevant results and discussing their significance;
- Conclusions section where you review the significance of your results and comment on the usefulness of the methods used;
- References.

The deadline for this assignment is **12 p.m. on 25th January 2024**. Electronic submission is compulsory. This should be in the form of a Word, or pdf document.

Marks

Marks will be awarded as follows:

- Abstract (5%) - awarded for a clear and concise abstract of the paper.
- Introduction (10%) - awarded for a clear introduction to the study and its motivation.
- Methods (5%) - awarded for the choice of a suitable number of relevant investigations. **Do not include a literature review of the methods used**.
- Results and Discussion (30%) - awarded for the clarity of the presentation

of results, and the choice of an appropriate level of detail.
- Conclusions (20%) - awarded for a discussion showing theoretical insight into the methods chosen, the likely accuracy of any predictions, and the biological relevance of the results.
- References (10%) - awarded for appropriate and adequate use of references.
- Presentation (20%) - awarded for clear presentation in all sections. Over long papers will be penalised at 5%, just as they are when submitted to real scientific journals. Good marks will be obtained if the relevant information is given concisely, but with sufficient detail that the expert reader could repeat the investigations if necessary.
- Appendix - **should not be used**. Long lists of raw data are not acceptable and will not be considered.

Word Limit

The word limit is 2500-3000 words. At the top of the paper you should include a word count for the entire manuscript including figure/table legends, i.e. all sections except the References). **Failure to do this will result in return of the manuscript (with associated time penalties). A 5% penalty will be applied to the total mark per 100 words over the word limit. So stay within the word limit!!**

There is no lower bound on word limit, however the experience shows that good works covering the analysis in necessary details are usually in the range of 2500-3000 words,

Plagiarism

Every year we detect a few cases of plagiarism. After that a student has to resubmit the work. The resubmitted work is capped at 40%. It is better to try and fail and then resit, then to be caught plagiarizing. Do not copy and past while changing a few words! Read and rewrite with your own words!

Hints

The assessment is open ended, and is therefore more like a mini research project. Here are some ideas about the sorts of things you might investigate:

- Searching protein sequence databases for related sequences using BLAST, FASTA or Smith-Waterman algorithms.

- Prediction of likely function of the sequence by similarity methods.
- Deduction of the domain structure of the sequence from the results of sequence searches.
- Analysis of the appearance of the sequence, or domains from it, in other organisms, or other kingdoms of life.
- Analysis of the sequence using PROSITE, PRINTS, BLOCKS or Pfam.
- Doing database searches with PSI-BLAST.
- Making multiple alignments of the sequence (or domains from it) with related sequences.
- Making phylogenetic trees based on multiple alignments.
- In the case of a sequence of known structure, searching for related structures.
- Prediction of secondary structure for the sequence, or a domain from it.
- Prediction of tertiary structure - Comparative Modelling.
- Prediction of tertiary structure - Fold Recognition.
- Prediction of trans-membrane segments.
- Prediction of protein-protein interactions.

Bear in mind that this assessment is supposed to take about 40 hours to complete, so it will not be possible to do all of the above. To gain good marks you should expect to include 4-6 of the above ideas, or other equivalent, relevant investigations. Your analysis is not therefore expected to be exhaustive. However, you should expect to have to try more analyses than this in order to find an appropriate number of interesting analyses, to include in your final report.

It is not generally a good tactic to adopt a "kitchen sink" approach, try all possible methods, and simply list the results. More marks will be given for investigations with a clear purpose and direction.

Here are some questions you might want to think about before choosing a protein

- Use this assessment to do an in depth analysis of a protein (or protein family) that interests you and/or for which there is an interesting biological question.
- Use it to pull out new protein family members (e.g. to those of a well characterized family of proteins) that are not found by BLAST/FASTA.
- Use it to try out some new tools that we were unable to cover in the course.
- Use it to increase your depth of knowledge of tools that we have covered in

the course e.g. structure prediction methods (comparative modelling or fold recognition) and try to relate the structure you get to its known function - e.g. active site, binding site etc.
- Use it to study a protein that has recently appeared in the scientific press (Nature, Science, New Scientist) or even the daily press! If it is not already highly characterized (and in some cases it may well be) then there may be an interesting bioinformatics story to be told or new and different analyses that can be done.
- If you are interested in the protein family then there are family tools we have looked at (ClustalW, JalView Pfam)
- If you are interested in finding new proteins related to your protein of interest in other organisms then PSI-BLAST and Hidden Markov techniques such as HMMER or SAM may be useful
- If you have a transmembrane spanning protein- there are programs you can use to predict TM helices etc. (e.g. TMpred, Memsat, DAS etc.)

Do you need to get structural information in order to answer the biological questions you are asking? If so you should check either you can make a comparative model OR get meaningful information from fold recognition, before you proceed with your chosen protein.

You do not need to use ALL the tools at your disposal ONLY those that it is sensible to use for the analysis - this should be guided very much by the biological questions are you hoping to answer. And what is already evident in the sequence/structural databases or scientific literature.

You should use (this) practical session to find out what is possible for the sequence(s) you have chosen - it is easy to say you will make a multiple sequence alignment, make a comparative model or run fold recognition methods. BUT what if they do not prove fruitful? You may end up with a weak report.

Try DOING the things you intend to do very quickly! - if it is unlikely to yield interesting results then spend your time looking for a new protein of interest or ask the module manager to provide a sequence!.

Do not choose a highly characterised protein sequence (or one at random) e.g. an enzyme from glycolysis, a protease e.g thrombin, where there is a large amount of literature about it and a structure (or very near relative) in the PDB. It is a test of your bioinformatics skills not a literature survey!