COMP0082 Coursework V2.01

Predicting the subcellular location of proteins

Over the last decade, the complete sequence has been determined for thousands of genomes. This has created the need for fully automated methods to analyse the vast amount of sequence data now available. The assignment of a function for a given protein has proved to be difficult where no clear homology to proteins of known function exists. Knowing the subcellular location of a protein (i.e. where in the cell it is found) in may give some clue as to its possible function, making an automated method that assigns proteins to a certain subcellular location a useful tool for analysis.

For eukaryotes, it is reasonable to define 4 major subcellular locations:

Cytosolic - i.e. within the cell itself, but not inside any organelles

Extracellular/Secreted - proteins which are transported out of a cell

Nuclear - proteins found/used within the cell's nucleus

Mitochondrial - proteins transported to the cell's mitochondria

Your task in this coursework is to develop a simple method for classifying eukaryotic protein sequences into the 4 categories above, plus "Other" i.e. none of the above 4 locations.

You are free to use any appropriate machine learning technique for this task - e.g. neural nets, SVMs, decision trees - whatever you like. *THIS IS NOT A SOFTWARE ENGINEERING CLASS, so you are free to use off-the-shelf machine learning methods or libraries e.g. Scikit-learn as long as you give enough information for someone to replicate your approach. Any method that you develop should be SELF-CONTAINED, however. It should not rely on external web services or even require a working internet connection to run it on a new sequence - you should in theory be able to just hand someone a USB stick with all the code and data needed to run your method on their own machine with nothing else required, other than e.g. a working Python + BioPython + Scikit-learn installation. You should also try to select open-source tools and libraries wherever possible (though use of Matlab is acceptable).*

Although you are free to select whichever method you like, extra credit will be given to being able to make the results from your method explainable. This can be achieved by simply doing analysis on the relative importance of input features of your method e.g. via ablation tests. If you opt for a method that does not use explicit features, then you could look at things like integrated gradients to show which parts of the raw sequence inputs are contributing most to the different output labels.

Predictions must be returned along with some measure of confidence. This can be a probability estimate, but should be convertible to an ad hoc measure of confidence e.g. HIGH, MEDIUM, LOW for the challenge sequences (see below). You must explain the basis of your confidence estimates.

Appropriate cross-validation studies should be carried out on the datasets. Details on the selection of test and training sets must be given in the write-up, and appropriate measures of success must be given e.g. ACC, F1, MCC etc. Results for the "blind" protein set must be included in the report.

The method and its evaluation must be written up as a report in the style of a *short journal paper*. As a guide that's around 2500 words max. Although you are quite free to use other formats, you might like to use the templates for the Bioinformatics journal, but you can use your own formatting if you wish. See **HERE**. Supplementary materials e.g. technical summary of the software you wrote (don't include the source code itself!) can also be submitted as an appendix to the paper. In your paper you should introduce the problem, discuss reasons for selecting the method you did, describe the method and the experiments you have carried out, present the results, discuss the results and give your conclusions.

Coursework assessment will be mostly focussed on the *quality of the report* and the thoroughness of the evaluation. You are expected to include a range of benchmarking results, using both graphical and tabular formats as appropriate. Bonus marks (up to 10% of the marks) will be awarded for especially successful (judged by the blind test set) or, in exceptional circumstances, particularly creative approaches if the report is also good. However, don't sweat blood to come up with the most sophisticated possible approach! Marks will be deducted for poor quality writing or poor quality, uninformative or irrelevant graphs.

NOTE: this coursework is worth 40% of the course marks, so it's worth taking a little care over it.

Hints

Although this list is by no means exhaustive, the following sequence "features" might well correlate with subcellular location:

- Sequence length
- Global amino acid composition (i.e. percentages of all 20 amino acids present in whole sequence)
- Local amino acid composition (i.e. over first 50 amino acids or last 50 amino acids)
- Isoelectric point & molecular weight (e.g. see HERE or HERE)
- Specific sequence patterns near the start or near the end of the sequence

It is likely that a good predictor will make use of some of these features in combination, or you are free to come up with your own ideas.

The Data

The sequences are provided as flat files in Fasta format

Note that every sequence in these sets can be assumed to be unique (non-homologous) i.e. you can break the sets into any division of testing and training sets you like when you are carrying out your cross-validation. *You should use ONLY the provided data to train/test your model. If it's found that you have made use of other data resources in building models or making your predictions, then marks may be deducted. Use of generic unsupervised pre-trained protein language models such as the Meta ESM models is fine, however.*

Cytosolic Proteins (2463 examples)
Extracellular/Secreted Proteins (1236 examples)
Mitochondrial Proteins (1023 examples)

Nuclear Proteins (2736 examples)
Other Proteins (2002 examples)

The "other" set are examples of prokaryotic proteins that sometimes contaminate samples during sequencing, and which should be labelled as "none of the above" in terms of subcellular location. You might decide to treat this data set differently e.g. as a binary classification problem against the other sets of sequences. That's entirely up to you.

The following set of proteins should be used as a final "blinded Challenge" test set for your method to see how it performs on true unknowns. When you include these results in your report, remember to include the sequence identifier in the table of results and make sure you format your results *exactly* like the following example:
SEQ01 Nucl Confidence High
SEQ02 Othr Confidence Medium
SEQ03 Extr Confidence Low
SEQ04 Mito Confidence Low
SEQ05 Cyto Confidence High
SEQ06 Nucl Confidence High
SEQ07 Othr Confidence Medium
SEQ08 Extr Confidence Low
SEQ09 Mito Confidence Low
SEQ10 Cyto Confidence High
SEQ11 Nucl Confidence High
SEQ12 Othr Confidence Medium
SEQ13 Extr Confidence Low
SEQ14 Mito Confidence Low
SEQ15 Cyto Confidence High
SEQ16 Nucl Confidence High
SEQ17 Othr Confidence Medium
SEQ18 Extr Confidence Low
SEQ19 Mito Confidence Low
SEQ20 Cyto Confidence High

This allows the challenge score to be calculated automatically, and any deviation will mean that you'll likely be awarded no bonus marks. *Do make sure the blind results are included as text and not pasted in as an image.* You will get no challenge bonus marks if the table cannot be read as text e.g. using pdf2text.

"Blind Challenge" Test Proteins

Please only include one set of predictions for the blind test cases - don't include multiple "guesses" for each one.

Note that although being highly ranked on the Challenge sequences will afford some extra credit (10%), it's obviously better to focus on getting a good report written than worry unduly about winning the challenge. It's much easier to get a decent result in the challenge than to produce a good quality report - hence the mark ratio.

NOTE: the confidence scores will be considered in the scoring as follows. High confidence predictions will get +1/-1 points (correct/incorrect). All medium confidence predictions will score +0.5/-0.5 points, and all low confidence predictions zero points. The point scores will be clipped so that you can't lose marks based on your blind test results i.e. you can only get bonus marks for being both accurate and confident.