Sarah McComas

Started March 15, 2016

**Introduction**:

Understanding protein structure and function is crucial to deepening our knowledge of how our bodies work and how diseases affect them. A large portion of these proteins involved in these processes exist inside the cell membrane and act among many things as cell gates, signals, etc. Of these membrane proteins, a majority of them are comprised of alpha helices. In fact, about one fourth of all genes transcribed are alpha transmembrane (TM) proteins \*\*CITE TOPCONS1 \*\*\* From understanding the protein topology, we can further work on things like drug delivery. To do this with traditional techniques would not be possible given time and financial constraints but an in silico approach would allow for fast and accurate prediction of the structure of these proteins.

Since the beginning of computational biology, there has been a lot of predictors formed in attempts to accurately and quickly determine protein topology. These include TMHMM, TOPCONS, DAS and many others \*\*\*CITE\*\*\* . They use a variety of algorithms and factors such as HMM’s to determine protein structure and now are very accurate and fast. My predictor here draws inspiration from these predictors, using BLAST, and SVM to attempt to accurately and with relative speed predict topology for protein sequences. Another useful feature of these protein predictors is that they are accessible on the web, which greatly increases the number of proteins we have predicted today \*\*\*MNETION NUMBER/?? FROM PDB??\* when everybody can collaborate and use open resourses. In addition, predictors can draw inspiration from each other and see how their models work together. Predictors can also be used in conjunction with each other for other reasons. Topology knowledge can for example help with other understandings like in fold recognition

Current limitations with predictors is accuracy and time. Many databases such as PDB will only confirm the protein in their database with experimental backup ie xray crystallography bl. a. \*\*\*\*is this true???\*\*\*\* . But consensus based predictions greatly improve the reliability of these models and allow for more accurate prediction than ever. There is still a lot of disagreement between models which is something we need to work on \*\*TOPCONSPAPER1\*\*\* but models such as TOPCONS have attempted to answer that problem because it uses consensus prediction.

This model presented is a basic model, which uses programs such as BLAST to increase the accuracy and sensitivity of the model, multiple sequence alignment is better than single sequence alignment.

My model takes inspiration from these models and although it is not nearly as good as TOPCONS or TMHMM it does use a basic SVM trained on known alpha membrane structures to predict with about a 90% accuracy where in the sequence one can expect that the protein in inside the membrane. A basic local webserver is also available in which one can post a membrane sequence and receive a prediction based on that sequence and a BLAST output within a few minutes. An SVM is used here, which takes a set of data who already have a sequence annotated for structure and uses information from this to make a model and further classify new examples based on single sequence or evolutionary information using PSIBLAST.

A support vector machine (SVM) was trained in this predictor. These are particularly useful in problems where one must classify data based on properties it does or does not have. In this predictor, training data was performed on sequences of known structure and positive and negative examples for each amino acid at a certain position were calculated from this. From here, the SVM was trained with both single sequence and later on multi sequence (ie evolutionary) information which provided a slower yet more accurate model. These models from the SVM were then later on the webserver. The models could be optimized with several parameters to achieve the most successful accuracy as well as the highest MCC, indicative of a low amount of false positives and negatives and can also be reflected in the ROC curve which shows the TP and FP rates against each other.

Although it is already better than random accuracy when one trains an SVM in a single sequence, adding evolutionary information from BLAST yields much more accurate results because a position specific scoring matrix (PSSM) \*\*\*CITE \*\*\* which calculates the likelihood of the amino acid being at that specific position. This is much more specific than single sequence reads, which simply indicate a binary-type data of the amino acid being at that position or not. This means the SVM training for the BLAST outputs is much slower but also much more accurate.

There are many good predictors out there, now the only thing is just improvement with time. To produce more accurate reads to the point that we don’t have to physically confirm them with research is the general goal, and of course to increase computational speed is always desirable.

Hydrophobicity is usually a good indicator of topology and is usually the basis for why a certain amino acid would be in the membrane. The biochemical properties of cell membranes require that these proteins are mainly hydrophobic in the membrane. This means that although the SVM is not trained to look at the hydrophobicity of each amino acid, automatically more hydrophobic ones will be much more present in the membrane and receive a higher ‘score’ for that. (valine for example)

**METHODS:**

As mentioned, the predictor was formed based on SVM learning algorithms. To do this, the SVM must be trained first. This was accomplished in slightly different methods, for the SVM was trained on both single sequence information as well as the evolutionary information from PSIBLAST. Data was extracted from a dataset whos structural information was already known, in order to dictate target values (showing what is a positive example, inside the membrane, and what is a negative example, outside the membrane). The amino acid letters were also transformed into a number, 0-20, assigned to that amino acid. This was held consistent throughout the entire program. For both sequences there was a value appended to each amino acid number, for single sequence this was a binary value indicating the presence of that amino acid at the indicated position, and for evolutionary information this was a PSSM number, indicating the likelihood of the amino acid being at that position as indicated by PSIBLAST. Cross validation sets were also created from the database, 6 sets of about 50 sequences in each set in order to give more accuracy and precision to the SVM. One dataset was excluded in order to complete the optimization, which tested different parameters of the SVM models in order to increase the accuracy and MCC values and therefore decrease the amount of false positives and negatives found in the data.

what is a kernel and a parameter??? What do they mean?

Homologs were also considered in this model preparation. It is important to not have homologs in separate cross validation sets because otherwise the accuracy will be falsely high since they are nearly the same sequence. For this, CD-HIT \*\*\*\*CITE\*\*\*\* was used to determine homologs in the dataset before proceeding.

This prediction algorithm was developed and is important because it is often amino acids properties and predictions that are the same or similar in forming certain structures. This is especially true for transmembrane proteins because inside the protein requires hydrophobicity whereas outside the cell surface often calls for hydrophilicy. This also plays a role in alpha helix formation, as amino acid sequences inside the helix can only be a certain size and the sequence can only be a certain length to form the helix. Although the SVM does not account for the information such as this, this is how the SVM machines can accurately say that these amino acids at these positions are the most likely to be a transmembrane helix.

**RESULTS:**

Once optimized, this predictor was capable of predicting protein topology with approximately 85% accuracy. These results are reported in FIGURE \*\* in comparison with other kernels tested with the SVM. The running time however was quite slow. It took about 3 hours to train the SVM on a linear kernel, and about 7 on a radial kernel. For single sequence, this was significantly shorter, around 30 minutes to train the SVM. This is due to much less data to process since there is no PSSM in single sequence data.

The MCC and the AUC are perhaps the most important features to analyze in my predictor. A perfect MCC is +1 and my predictor is 0.5 \*\*WHAT are some others?? The AUC is also important. This can be observed in FIGURE \*\*. This, similarly to the MCC, shows the accuracy of the predictor. Again, a value near the +1 is perfection, anything under 0.5 means the predictor is as good as random and cannot successfully separate true and false positives. A ROC curve and a bar graph were generated using scikitlearn and matplotlib toolkits \*\*\*CITE\*\*\*, respectively. Both of these help to generate an MCC, AUC, which are distinct yet similar means of calculating predictor quality of results separation.

The accuracy is also reported by the SVM at the time of classification(have you mentioned what this is?) , which is different from the MCC and AUC BECAUSE>>>>

**DISCUSSION**:

My predictor is accurate, but only when the right data is applied. It is not nearly as ubiquitous as TOPCONS, TMHMM, or DAS, for it is fairly one dimensional, it only tests whether a sequence is inside the membrane or not, and assumes automatically that if it is inside the membrane, it is an alpha helix. Although this is certainly the most common scenario, alpha helices are much more abundant in the membrane than beta barrels (\*\*\*CITE PDB), it would be prudent for the predictor to also factor in what the structure might be, not just the topology. WHAT IS THE ACCURACY OF YOUR MODEL WHEN YOU TEST IT AGAINST OLIVERS DATA?

Another limitation of this model is time. Currently, TOPCONS can complete an entire proteome search in just several hours. This predictor can finish with a multi sequence analysis in \*\*\*\*\*\*\*\*\*.

There are most likely several limitations to this predictor. One could be the sheer computing power of the program. Running BLAST occupies a large majority of the long time it takes to run this predictor, mostly because searching a database takes quite a long time, especially as today’s databases become larger and larger. Predictors such as TOPCONS have achieved such a fast speed by optimizing between searching a large database in order to distinguish membrane proteins, and searching a smaller database which contain only membrane proteins.

When the predictor is just considering single sequence information and not running BLAST, the process is significantly faster, both for SVM classify (theres got to be a better way of saying this…) and by the actual running time because BLAST is not necessary. However, the model accuracy is decreased greatly and it is not nearly as good of a predictor. This is because a profile is not generated for the topologies, it is only a binary value for amino acid presence at a certain position or not, not a whole calculation for the likelihood of each amino acid being present.

Optimizing the SVM is also a good way of increasing accuracy but it has little to do with the speed of the program. It is true that the radial kernel is slightly slower in processing than the linear kernel but the accuracy increases with this.