SciLife Project

Sarah McComas

March 18, 2016

Abstract

Here will be my abstract it will be cool

1 Introduction

Understanding protein structure is crucial to deepening our knowledge of their function and how diseases affect them and us. One of the most abundant of these are known as transmembrane proteins, which exist across membranes and participate in a variety of function such as cell transport and signaling. Of these transmembrane proteins, most of them are in the form of an alpha helix because their H-bonding properties allow for stability inside the hydrophobic lipid bilayer. About 27% of all proteins produced by humans are classified as an alpha helical transmembrane protein (1), and while understanding their exact structure is difficult, predicting the topology (i.e. where and in what orientation the protein is located in the membrane) gives enough information about how the protein function and potential targets for diseases or drug delivery. Gaining this information with traditional research is difficult and costly, which lead to methods to predict protein structure given the properties of the protein's amino acids. The first method used to do this was known as the Chou Fasman method (2), which analyzed the frequencies of the amino acids' presence in alpha helices and other structures to predict a protein structure with about a 50% accuracy (3). Since the development of machine learning algorithms and incorporation of multiple sequence alignments, the accuracy of these predictions now can reach 90% or higher (4).

Here I have drawn inspiration from such state-of-the-art predictors in creating my own predictor which determines if an alpha helix is inside the membrane or not. As these predictors before mine have, I used a multiple sequence alignment (MSA) profile generated from PSI-BLAST (5) (as well as information from single sequences in order to compare predictor quality) with information about amino acid presence at a certain position in a structure. The MSA profile is crucial in understanding evolutionary information, and can analyze the certain amino acid through evolution and see how frequent it occurs in the structure. I could then use this information as a learning algorithm for a support vector machine, or SVM (6). SVM's use supervised or unsupervised learning to do binary classification, to determine what a positive and negative example for each datapoint look like. In this case, I used an SVM only with supervised learning techniques. This trained machine can then be implemented on new data where the structure is unknown in efforts to predict it. To use the MSA data on a machine learning algorithm is a successful and common practice in todays' top predictors for topology or secondary structure such as TOPCONS, TMHMM, PSIPRED, among many others (4, 7, 8) and is the main reason why these predictors are so successful in their accuracy and efficiency.

Another successful feature of modern computational predictors is the presence of open source information, web servers, and specific databases. Here, all groups can access information quickly and easily and leads to collaboration and

therefore an increase in database size and agreements on protein structure prediction. In this way, these groups can build off of each other to improve their own methods and the quality of the data presented. The two main limitations in these predictor qualities are time and accuracy. Now, accuracy is very high but there is still some disagreement between prediction models, which programs such as CASP (9) aim to solve. Many databases continue to separate that which has been computationally generated and what is experimentally proven by traditional methods, but with these improvement programs, one day we will no longer need traditional methods to validate structure.

2 Methods

I extracted features from a database of solely alpha membrane proteins with known topologies to train an SVM in supervised learning. This consisted of about 300 sequences and each entry consisted of the accession number of the protein, the amino acid sequence, and the corresponding position with regards to the membrane (I for inside, O for outside, and M for membrane). In this case, all positions with an 'M' would become a positive example, while 'I' or 'O' would be a negative one. I also assigned each amino acid letter a corresponding number so that it was readable by the SVM. This lead to a standard SVM format, as seen in Figure 1.

Figure 1- each line began with a target value (a positive or negative number depending on structural position). Show maybe the PSIBLAST output.

3 References

1-http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-7-50 (alpha proteins info) 2- http://pubs.acs.org/doi/abs/10.1021/bi00699a002 (chou fasman) 3- Kabsch W, Sander C (1983). "How good are predictions of protein secondary structure?". FEBS Lett 155 (2): 179?82. doi:10.1016/0014-5793(82)80597-8. PMID 6852232. 4- TOPCONS paper 5- (PSIBLAST) 6- SVM7- PSIPRED 7 - TMHMM 8- PSIPRED

9- CASP