Protein Structure Prediction Using a Support Vector Machine

By: Luc DeGagne

The Problem and its Domain

Oftentimes, when people hear the term protein, they think of chicken or beef. However, more specifically, proteins are complex molecules that provide both structural and functional support within the cell and it's surroundings[1]. This means that proteins can do anything from maintaining cell structure, like a column in a building, to actually playing a role in the cell, such as transporter proteins that act like bouncers, ensuring only the right things get into the cell. Even the molecule that carries oxygen from the lungs to the body is made from a protein[2]. This quickly makes proteins a value valuable pharmaceutical commodity. Being able to turn on or off the functions in the body you desire to affect is the backbone of the pharmaceutical drug industry[3]. An industry that makes over $400 billion dollars annually[4].

All proteins are primarily composed of amino acids, which are organic molecules that are able to form long chains that fold into unique structures[5]. The sequence of amino acids in a protein is called its primary structure. They work by all having the same backbone structure and then having different sidechains depending on the task or function they need to complete. In order to create the chain they form, they have one negative end, containing a carboxylic acid (COO^-) and one positive end, containing an amine (NH_3^+) which are attracted to each other's opposite charge and form a covalent bond by sharing the extra hydrogen on the amine. This allows the amino acids to fit together like interchangeable puzzle pieces based on what sort of function the sidechain is being asked to perform.

Using the interactions that the sidechains have with each other, the overall chain of amino acids is pushed into different types of structure, which all perform different types of abilities. These structures are also known as secondary structures. Among those unique structures, there are three major forms: α-helices, β-sheets and random coils. By combining these three structures into different lengths and sizes, the amino acid chains are able to create the actual structure and shape of the protein they are working towards becoming.

Even as early as the 1960's, scientists had been able to prove that the amino acid sequence itself was what dictated the structure of the overall protein[6]. This meant that whoever could figure out how to predict the structure of amino acid sequences would be able to predict the function of the protein. The problem of structure

prediction is fundamentally different from protein design, something commonly able to be accomplished today. It would then be of great value to answer the question: based off of the amino acid sequence of a protein, is it possible to predict the secondary structure?

The Motivation

The field of secondary protein structure determination is a primary focus of computational structural biology[7]. In fact, so important that since 1995, a large group of protein biochemists have been getting together at a summit called the Critical Assessment of Protein Structure Prediction (CASP) so the scientific community can demonstrate how far along research is in predicting protein structure[8]. As previously mentioned, protein structure prediction is of great value in the pharmaceutical industry however would also have implications in the design and creation of synthetic proteins. This could be used in multiple fields, including but not limited to biotechnology.

The globular protein family, a group of proteins involved with enzymes, hormones, and transport molecules has been understood and had multiple of the family's structures been determined[9]. These proteins are extremely important as all three functions serve a vital process in the system of living organisms. In order to determine how globular proteins fold, the amino acid sequence, considered the primary structure of a protein will be examined to understand the secondary structure of a protein which is represented by the α-helices, β-sheets and random coils.

AI Techniques and Algorithms

In order to develop a program, that can predict protein structure, unique or uncommon methods of computational protein prediction will be used. As protein structure prediction is multivariate and complex a support vector machine (SVM) will be used to help predict the secondary structure. SVMs are an excellent supervised learning machine that can be used to solve classification problems like the one above. Since there are so many different characteristics that an amino acid can be measured by, the ability of an SVM to handle multiple variables will be quite helpful. However, the following three traits will be looked at in further detail: the static charge of the protein, the sequence being able to conform to the desired structure's characteristics and the probability those amino acids will be expected to be found in that particular structure.

An SVM with a linear kernel will be used on 5 variables with the help of Sci-Kit learn that will enable the creation of a weight vector[10]. A training set will be used to

predict the weight vector that will solve the linear discriminant function (wx+ b = 0) by determining the values of the support vectors given the equation is equal to +1 or -1. These support vectors where wx + b = 1 and wx + b = -1 are the points of the dataset such that they are closest to the hyperplane on each side of it. Without these support vectors, it would alter the location of the hyperplane. This will allow for the selection of a hyperplane and also the discriminant value with the greatest possible margin between itself and any support vectors on either side. The hyperplane is then used to separate and classify the various types of amino acids into their structures based on proteins from a test set.

Design Choices

The regularization parameter, C, was set to 100 to produce moderately soft margins, this helped compensate for the crossover between some of the data and the fact the SVM is linear. All other components were set to default. One notable component is the decision function shape, which is set to compare one variable to all of the variables, as opposed to comparing each variable one to one. This saves on computer power but is affected by imbalanced datasets.

In order to get a clear understanding of the data, 5 variables (and therefore 5 dimensions) will be used to help determine the hyperplane of the SVM. Due to the fact that the choice of protein variable components can quickly become extremely complex, many of the possibilities for variables were used to maintain simplicity. As a consequence, the amino acids were looked at in groups of 3 and each amino acid's structure would be considered based on the average created by itself and its two neighbours. The 5 variables considered were the hydrophobicity scale provided by Kyte and Doolittle[11] (a scale that measures an amino acid's ability to interact with water vs its ability to interact with a substance like oil), the polarity scale, a scale that determines the overall charge of an amino acid[12], and three frequency measurements of each amino acid and its likelihood to be found in each of the three structures. In order to come up with the frequency measurements, a database of known globular protein structures was used to measure how often each amino acid was in either the α-helix, β-sheet or random coil structure[13].

As most globular proteins have similar environments, it would be expected that hydrophobic (lipid-interacting) and hydrophilic (water-interacting) proteins would be grouped together, respectively, playing a role in whether they would be in a more concrete structure or a random coil. As far as polarity, a similar logic can be taken, for the basis of how polar a protein can be. This is based off of whether or not the protein itself has a positive or negative overall charge, as these charges are able to interact with each other, having the ability for the charges to be grouped together may pose an advantage to some types of structures and be more common or less

common in those. Lastly, creating a list of frequencies from a database of proteins will provide an understanding of which proteins belong in which structures after taking the average of the 3 proteins side by side[13]. These values will then be used to get averages of the three protein window from using a training database containing over 18,000 amino acids from over 300 proteins[14]. This training set, in which approximately 3,600 amino acids found in each structure, totaling just over 10,000 amino acids will then be compared to a unique testing set, provided from the same database.

Results

Before running the SVM, a linear discriminant analysis was completed as shown in **Figure 1**. The clustering of all three points shows that the accuracy of the SVM will be moderately low overall, however there are some areas when comparing the random coil values to the helices and sheets, it would be likely a line could be put diagonally, however separating the helices from the sheets would prove to be a much more complicated activity. So before running the SVM we can already notice that the points are not easily separable and there could be much confusion between the three types of structures, however there are also clear outlined areas for one or two of the structures where it could easily be expected that an amino acid found in that region is a random coil.

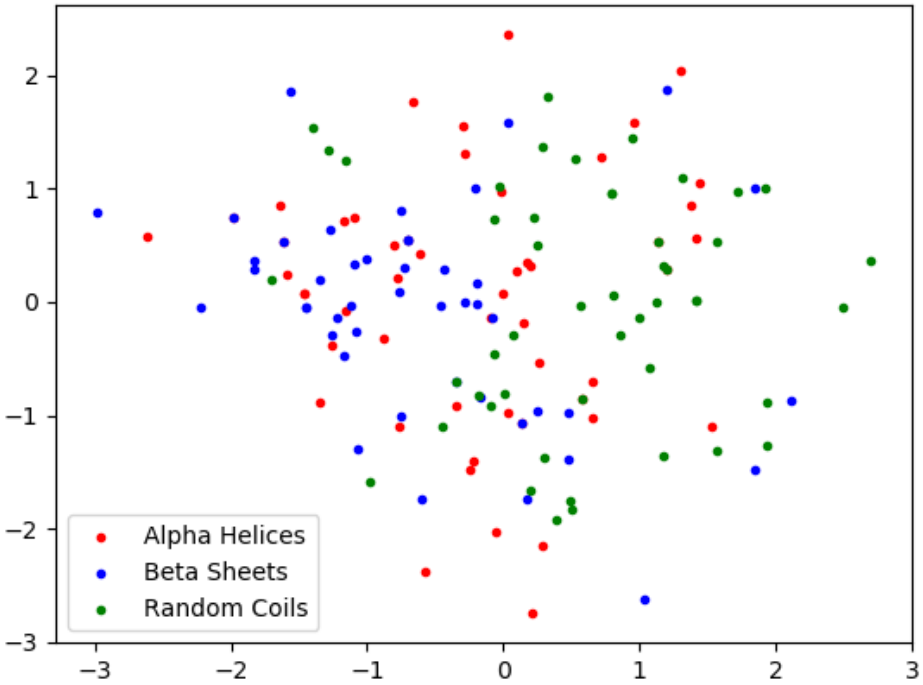

Figure 1. Using linear discriminant analysis (or LDA) the 5 variables originally used have been reduced to 2 dimensions in order to see where the points lie. Of 10,000 amino acids, only every 75th values was graphed in order to reduce clutter.

After running the SVM, the values have been compared to 2 neural networks that have previously been created to predict secondary structure[15][16]. Although the SVM performed approximately 10% worse than the neural networks, as shown in **Table** 1. It did show a much higher ability to correctly estimate β-sheets compared to both neural networks. It did, however, struggle to correctly predict random coils compared to the other two networks and had it been better able to do so would have been a very effective method in protein structure prediction. It is worth noting however, that the values of the SVM are not far off the values of the neural networks and therefore are heading in the right direction.

Table 1. A comparison of the results for the SVM in contrast to two previously researched neural networks, of which, the FSkbann is the most advanced, combining highly tactical biochemistry strategies with an advanced neural network.

Method	**Accuracy(%)**			
	Total	Helix	Sheet	Coil
SVM	49.0	34.3	52.5	54.2
ANN	61.8	43.6	18.6	86.3
FSkbann	63.4	45.9	35.1	81.9

Future Enhancements

Due to the nature of the SVM, having unclear margins of separation to the extent that the protein structures did (as seen in **Figure 1**), was probably the number one factor in the low accuracy of the results. In order to improve this, more complicated and advanced variables need to continue to be applied to the SVM to continue to increase those margins of separation in further dimensions. It may also benefit to increase the size of the window in which the amino acids are being looked averaged with, which could alter the overall values of the groups and provide better results.

Another solution would be to use a learning machine that can better handle unclear margins, such as logistic regression or a decision forest/jungle. It did however provide the necessary level of efficiency needed for a project of this complexity, but due to the fact that there is quite a lot of noise in protein structure prediction, its performance will continue to be subpar until such a time as there are either enough variables or the right variables to properly map out the structures. This is most likely due to the complex nature of protein folding, of which much of it spontaneously happens in nature, the next steps to add to the SVM would be the likelihood of hydrogen bonds forming, as well as the application of Van der Waals forces to better understand the intramolecular interactions occurring in those structures[17].

References

[1] J. PALAU, P. ARGOS, and P. PUIGDOMENECH, "Protein secondary structure," *Int. J. Pept. Protein Res.*, vol. 19, no. 4, pp. 394–401, Jan. 2009.

[2] M. Brunori, "Hemoglobin: Structure, function, evolution, and pathology," *Trends Biochem. Sci.*, vol. 9, no. 5, p. 247, May 1984.

[3] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there?," *Nat. Rev. Drug Discov.*, vol. 5, no. 12, pp. 993–996, Dec. 2006.

[4] D. W. Light and J. Lexchin, "Pharmaceutical Research and Development: What Do We Get for All that Money?," *SSRN Electron. J.*, Aug. 2012.

[5] K. O'Neil, W. DeGrado, and B. Matthews, "A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids," *Science (80-.).*, vol. 250, no. 4981, pp. 646–651, Nov. 1990.

[6] C. H. HIRS, S. MOORE, and W. H. STEIN, "The sequence of the amino acid residues in performic acid-oxidized ribonuclease.," *J. Biol. Chem.*, vol. 235, pp. 633–47, Mar. 1960.

[7] S. Lee, B. Lee, and D. Kim, "Prediction of protein secondary structure content using amino acid composition and evolutionary information," *Proteins Struct. Funct. Bioinforma.*, vol. 62, no. 4, pp. 1107–1114, Dec. 2005.

[8] J. Moult, "A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction," *Curr. Opin. Struct. Biol.*, vol. 15, no. 3, pp. 285–289, Jun. 2005.

[9] S. Provencher, J. G.- Biochemistry, and undefined 1981, "Estimation of globular protein secondary structure from circular dichroism," *ACS Publ.*

[10] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[11] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein.," *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–32, May 1982.

[12] P. Wang *et al.*, "Prediction of Antimicrobial Peptides Based on Sequence Alignment and Feature Selection Methods," *PLoS One*, vol. 6, no. 4, p. e18476, Apr. 2011.

[13] R. Sowdhamini, S. D. Rufino, and T. L. Blundell, "A database of globular protein structural domains: clustering of representative family members into similar folds," *Fold. Des.*, vol. 1, no. 3, pp. 209–220, Jun. 1996.

[14] R. Maclin and J. W. Shavlik, "Using Knowledge-Based Neural Networks to Improve Algorithms: Refining the Chou–Fasman Algorithm for Protein Folding," *Mach. Learn.*, vol. 11, no. 2/3, pp. 195–215, 1993.

[15] P. Y. Chou and G. D. Fasman, "Prediction of the secondary structure of proteins from their amino acid sequence.," *Adv. Enzymol. Relat. Areas Mol. Biol.*, vol. 47, pp. 45–148, 1978.

[16] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *J. Mol. Biol.*, vol. 202, no. 4, pp. 865–884, Aug. 1988.

[17] P. O. Freskgård, N. Bergenhem, B. H. Jonsson, M. Svensson, and U. Carlsson, "Isomerase and chaperone activity of prolyl isomerase in the folding of carbonic anhydrase," *Science*, vol. 258, no. 5081, pp. 466–8, Oct. 1992.

Appendix A

In order to run the code provided, Python 3 is required as well as the following libraries:

- Numpy
- Sci-Kit Learn
- Pandas
- MatPlotLib

Once all libraries have been added, the code is run by typing:

"python3 proteinSVM.py"

NOTE: both the train.txt and test.txt files are required to run the SVM.