A Novel Weighted Approach to Predict Protein Fold Type

Eeshwar Krishnan[1]

[1]Pennsylvania Leadership Charter School, 1585 Paoli Pike, Westchester PA 19380, USA

***Abstract*— Prediction of protein fold type is the first step in determining protein folding of amino acids. Predicting fold type is a difficult multi-class machine learning problem. Past work in predicting fold types has shown poor overall accuracy, although the methods worked well for determining some protein fold types. In this paper, we describe a novel framework to predict protein folding using a weighted approach, combining different machine learning approaches in a principled manner. This approach uses a weighted voting method to combine results from different machine learning methods to improve the accuracy of predicting fold type over individual methods. Results show an increase in the accuracy of protein fold measurements. Furthermore, the framework can be expanded to include new and emerging deep learning methods, and can serve to enable protein folding prediction.**

I. Introduction

Protein folding is a difficult and intricate problem in molecular biology. Proteins are created in unfolded, random shapes of amino acids. Through interactions with hydrogen bonding and other Intermolecular Attractive Forces (IMAFs), the protein folds into a different shape, known as the native shape. The shape of the protein helps define its function and how it interacts with the rest of the cell. Changing the folding pattern of the protein can make it become toxic or malfunction in the cell. This could lead to different conditions including Alzheimer's, Parkinson's, cystic fibrosis, and many others [1]. Understanding how these proteins fold can advance treatment of these diseases by helping the development of new medicines that have particular impacts on the cell.

However, determining folds of a protein is computationally intensive. Each protein can consist of 20 different amino acids that combine and adopt one of several trillion shapes [2]. To determine how a particular sequence of amino acids will fold, one can compare the sequence of proteins to a known set, and if the sequences are similar, the protein folds will likely be similar as well. There are many sequences that are not similar to others, and as a result the folds cannot be determined by a search. In these cases, one proposed method to determine protein folds is to first predict the fold type, such as Globin-like, Long alpha-hairpin, and Cytochrome c. After the fold type is determined, the specific fold shape can then be found through a local search [3].

The key to the success of this approach is predicting the fold type. The general problem for predicting fold type is a multi-class machine learning problem, where the goal is to label a sample into one of several defined types of fold. Several studies have attempted to use machine learning to predict the fold type of a protein molecule [4,5,6,7,8]. The results of these approaches are shown in Table 2. Each of these studies used the same set of data for training and testing, to allow for comparison of methods. While none of these approaches are accurate enough to be used in many general cases, the results show that each method worked for some subset of the cases. For example, the hierarchical classifier proposed by Lin *et al* [6] did fairly well with small protein fold types (average accuracy 70.7%) but much more poorly on membrane and cell surface protein types (41.9%). This suggested that a combination of these methods could be successful in accurately predicting many different protein folds. In this paper, we propose a weighted framework that combines different machine learning algorithms to predict fold type, and demonstrate how this approach can significantly improve accuracy over existing methods.

II. Methods

In this paper, a weighted approach is proposed to combine the output of different machine learning approaches to optimize classification performance. Based on the work of previous researchers, different machine learning approaches work well for different fold types. The idea is to weight the output of different algorithms based on their accuracy for different fold types, so that algorithms that perform well for identifying a particular fold type are weighted higher for those fold types. This framework is extendible to new classification approaches, thus allowing for continued improvement from new researchers while leveraging the strengths of existing approaches.

First, 20 different machine learning algorithms were implemented to predict fold type. These algorithms were selected to represent a wide cross-section of machine learning algorithm types. Each algorithm was developed using Weka [9], with a custom Java wrapper around the algorithm.

J48 Decision Tree	Random Committee	Weighted Instance	Bagging
Voted Perception	Replacement-Based Decision	Sequential Minimal Optimization	Additive Regression
Logistic Model Tree	Bayes Net	Partial Decision Tree	Random Subspace
Linear Regression	Multinomial Naïve Bayes	Apriori with Subgroups	Neural Network

Table 1: List of Algorithms Used

Training and test data to develop and test the model was taken from data on-line, and made freely available to researchers [12]. The 20 fold types correspond to the fold types in [3]. The training and test sets had 238 and 290 samples respectively, across the different folds. Each sample had 125 features from 6 different parameter sets: amino acid composition, predicted secondary structure,

hydrophobicity, normalized Van der Wahl volume, polarity, and polarizability. This was the same training and test data used by other researchers, and so the methods can be directly compared. Full details of the training and test set data, as well as the definitions of the folds, are available at [3]. Each algorithm was optimized using the training set. Feature selection was used to minimize the number of features per model and reduce the impact of overfitting. The final accuracy of each algorithm, defined as the number of times the algorithm selected the correct fold divided by the total number of samples, was computed on the training set.

The results from the training set confirmed the results of previous studies that different algorithms work very well for certain folds, and not others. For example, the J48 decision tree works very well to identify fold type 12 (95.9% accuracy) but very poorly for fold type 15 (7.4% accuracy). Conversely, local weighted learning works well for fold type 15 (95.8% accuracy) but poorly on fold type 12 (34.5% accuracy). In selecting the algorithms for this study, we were careful to ensure that each fold type worked well (greater than 80% accuracy) on the training set for at least one algorithm.

The results from the training set were used to develop a weighted voting system. Each fold in the test set was run against all of the algorithms. For each algorithm *i*, it computes the fold *j* that has the maximum likelihood. It then casts a vote for that particular fold. The value of the vote is weighted to the accuracy for that algorithm for that fold, $\mathbf{x}_{ij}$, as measured in the training set. After all the algorithms are run, the votes for each fold are added up, and the fold with the highest number of votes is declared the "winner" and that fold is selected.

III. Results and Discussion

The weighted voting method resulted in an overall accuracy of 83.2% on the unseen test set, where accuracy is defined as the percentage of the test set folds that were correctly classified in the appropriate fold type. The results of the weighted voting method, compared to other methods in literature using the same training and test set, are shown in Table 2 below.

Reference	Algorithm used	Accuracy
Ding *et al* [4]	Support Vector Machines and Neural Networks	20.5%
Chinnasamy *et al* [5]	Naïve Bayes	58.8%
Lin *et al* [6]	Hierarchical Classifier	60.1%
Jo *et al* [7]	Random Forest	40.8%
Gromiha *et al* [8]	Linear Regression	57.1%
Weighted Voting (this paper)	**Multiple**	**83.2%**

Table 2: Accuracy of Weighted Voting Versus Other Methods on Protein Fold Prediction Using Same Test Set

The results clearly demonstrate that a weighted voting method can substantially improve the accuracy of predicting a fold given parameters of that fold over current published methods. The accuracy of this method suggests that it can be used as a first step to determining the specific protein fold for a set of amino acids. The approach leverages the ability of certain approaches to work well for certain fold types, and combines them in a principled manner to optimize classification performance.

Furthermore, this approach can be expanded to leverage new methods to classify protein folds. Recently, Google has recently published results on using a deep learning approach, AlphaFold, to predict protein folds [11,12]. Applying these newer deep learning method methods, and combining them with current approaches, could further enhance the accuracy of the predictions. By measuring the accuracy of new approaches, such as Google's AlphaFold, on each fold type, these newer methods can be incorporated into this voting scheme. The approach described in this paper thus provides a framework for incorporating new predictive models to further improve performance.

Determining fold type is just a first step in predicting the overall protein fold. If a reliable method can be developed to address this problem, a new generation of treatments can be developed for a host of diseases, and bring hope to many people.

References

[1] Selkoe, Dennis J. "Cell Biology of Protein Misfolding: The Examples of Alzheimers and Parkinsons Diseases." *Nature Cell Biology*, vol. 6, no. 11, 2004, pp. 1054–1061.

[2] Everts, Sarah. "Protein Folding: Much More Intricate than We Thought." CEN RSS, cen.acs.org/articles/95/i31/Protein-folding-Much-intricate-thought.html.

[3] Dubchak, Inna, et al. "Recognition of a Protein Fold in the Context of the SCOP Classification." *Proteins: Structure, Function, and Genetics*, vol. 35, no. 4, Jan. 1999, pp. 401–407.

[4] Ding, C H, and I Dubchak. "Multi-Class Protein Fold Recognition Using Support Vector Machines and Neural Networks." Bioinformatics (Oxford, England), U.S. National Library of Medicine, Apr. 2001.

[5] Chinnasamy, A, et al. "Protein Structure and Fold Prediction Using Tree-Augmented Naive Bayesian Classifier." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, U.S. National Library of Medicine, 2004.

[6] Lin, Chen, et al. "Hierarchical Classification of Protein Folds Using a Novel Ensemble Classifier." PloS One, Public Library of Science, 2013.

[7] Jo, Taeho, and Jianlin Cheng. "Improving Protein Fold Recognition by Random Forest." BMC Bioinformatics, BioMed Central, 2014.

[8] Gromiha, M Michael, et al. "FOLD-RATE: Prediction of Protein Folding Rates from Amino Acid Sequence." Nucleic Acids Research, Oxford University Press, 1 July 2006.

[9] Holmes, Geoffrey; Donkin, Andrew; Witten, Ian H. (1994). "Weka: A machine learning workbench". Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia.

[10] http://www.nersc.gov/~cding/protein

[11] "AlphaFold: Using AI for Scientific Discovery." Deepmind, deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

[12] https://www.bloomberg.com/news/articls/2020-11-30/deepmind-s-alphafold-corssesthreshold-in-solving-protein-riddle