Feature Identification and Reduction for Improved Generalization Accuracy in Secondary-Structure Prediction

Matt Seeley, Mark Clement, and Quinn Snell

Abstract—**Secondary structure prediction is an important step in understanding gene function. Several algorithms have been proposed for applying machine learning techniques to this problem. This research examines these algorithms and constructs a framework that is effective in providing accurate predictions.**

I. Introduction

The most effective modern algorithms for secondary-structure prediction almost universally use information from multiple-sequence alignments of homologous proteins with known structures. This is undoubtedly a sound approach for predicting structures of sequences that have many known homologues; good accuracy could probably be achieved by simply predicting that the test sequence's structural label at any given position in the sequence matches the most consensus label at the corresponding position in the multiple-sequence alignment. However, for proteins with no known homologues, it would be prudent to use a different approach—one that still leverages domain-specific knowledge in the context of a machine-learning model.

Ideally, predicting the secondary structure of a protein at a given position would be as simple as identifying unique, short subsequences whose central amino acids always have one specific label. This approach's effectiveness is limited, though, for two reasons: first, the number of possible permutations of 22 possible amino acids (with replacement) for a subsequence is exponentially large. There would, for example, be 22^{13} possible amino-acid subsequences of length 13aa. Even with all the data in the RCSB protein data bank, the number of subsequences of length 13aa with known labels is a very small fraction of the number of subsequences that are possible. More important, though, is the fact that the RCSB data demonstrates that many identical subsequences of length 13aa have different labels when they appear in different proteins or in different contexts. Thus, even if the search space of every possible subsequence were tractable, some subsequences could only be assigned tentative majority labels; this would limit the maximum theoretical accuracy of the model.

In order for a machine-learning model to generalize well to test instances that have little sequence identity with training instances, it must use some intelligent metric that can tell when dissimilar subsequences have amino acids with similar properties at identical positions. It must also be able to identify similar periodic patterns in those properties so that instances that are nearly identical, but whose attribute values are all shifted by a single position (like two successive sliding windows) can still be recognized as similar to one another.

This research identifies intelligent metrics that can recognize similarities between individual amino acids and patterns found in sequences of amino acids that are meaningful in the context of secondary-structure prediction. It also tests machine-learning algorithms that can suitably harness the information from those metrics to create a model that can accurately the predict secondary structure of a test sequence that has a low degree of sequence identity with the model's training sequence.

The next section discusses previous research and related work in secondary-structure prediction. The methods section discusses the metrics and models that were used and the rationale behind why they were chosen. This is followed by an experimental results section and a conclusion section that discusses what was learned from the research.

II. Related Work

Researchers have attempted to predict protein secondary structure using sequence data for at least five decades. In 1964, Straub published a thorough article describing the "widely accepted hypothesis" that that secondary and tertiary structure could be determined entirely based on primary structure [1]. The theory seemed attractive enough, especially given that some previous and subsequent studies demonstrated that many unfolded proteins can refold into their original conformations when placed in the proper environments [2, 3]. However, Straub wisely noted some observations that were "not in harmony with the theory of absolute determinism," thereby showing awareness of the problem's greater complexity [1]. By 1969, some had theorized that a protein would simply fold into the conformation corresponding to its globally lowest free energy; Cyrus Levinthal, however, presented the now-famous "Levinthal's paradox" in which he argued that a protein could not be randomly moving through all of the exponentially large number of its possible conformations quickly enough to find its global minimum in time to explain the experimental refolding speeds of some proteins [4]. Several years later, Afinsen presented a postulate that is now known as "Afinsen's Dogma": given a specific set of environmental conditions, a small globular protein's native conformation will be a unique, stable, and kinetically

Manuscript received July 1, 2013. This work was supported in part by NIH under Grant R01HG005692.

Mark Clement is a professor in the Department of Computer Science at Brigham Young University, Provo, UT 84602 phone: 801-422-7608; fax: 801-422-0169; e-mail: clement@cs.byu.edu).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

accessible structure (though it may only represent a local minimum relative to free energy) [5]. Simon also published useful research detailing some structural features that contribute to refolding ability [3].Though it is now known that there are exceptions to Afinsen's dogma, such as intrinsically disordered proteins [6], the principles it embodies are still useful in guiding research.

With the explosion of publicly available biological-molecular data in the past two decades, researchers have continued to incrementally discover new knowledge that is pertinent to the structure-prediction problem. Berezovsky and Trifonov, for example, presented evidence that that proteins fold into subunits of 25-30 amino acids in a local way [7]. Rost and Sander used evolutionary information from a multiple-sequence alignment to improve a neural network's ability to predict secondary structure [8]. Many other researchers also demonstrated the power of using multiple-sequence alignment information and position-specific scoring matrices to improve machine-learning models' prediction accuracy [9]. Some machine learning models that have been used to predict secondary structure have included feed-forward neural networks, support-vector machines [10], and recurrent neural networks [11].

While the efforts of many capable researchers continue to yield improving prediction accuracy, the holy grail of prediction based on primary sequence alone remains an elusive goal. The structure of a newly sequenced protein that lacks known homologues, for example, would likely be more difficult to predict because the multiple-sequence alignment approach would provide less information. Nevertheless, there are some key amino acid properties that have been shown to aid in secondary-structure prediction in in the absence of information from homologous proteins. Moments of hydrophobicity and residue conformational propensities are among those important properties [12].

III. Methods

When aiming to identify the secondary structure of any single amino acid in a sequence, it is important to remember that there is a high degree of dependence between its label and the labels of the amino acids immediately next to it. An amino acid that is part of an alpha helix, for example, is *always* next to at least one other amino acid that also has the same label because at least four consecutive amino acids are needed to form an alpha helix structure [13]. This principle applies to all secondary structures, since they are all formed as a result of bonding between the backbones of at least two amino acids. The data found in SS.txt, a file containing the known secondary structure labels for all RCSB Protein Data Bank files is consistent with principle. As an example, the distribution of lengths of all contiguous alpha helix structures found in ss.txt is shown in Figure 1.

Given this high degree of dependence between the labels of successive amino acids, the best machine-learning models for secondary structure prediction should some means of capturing the dependence information that is found in a given training set. Complex dependencies that could not be captured by simply looking back one step clearly existed in this project's data set; four amino acids with structure labels NHHH must be followed by an amino acid with label H, for example, but one amino acid with label H may or may not be followed by an amino acid with label H. Hidden Markov models and Elman nets can look back one step, but recurrent neural networks can look back an arbitrary number of steps and can therefore capture more complex dependencies between successive instances. As a result, a recurrent neural network was chosen as the principal model to be used for this research.

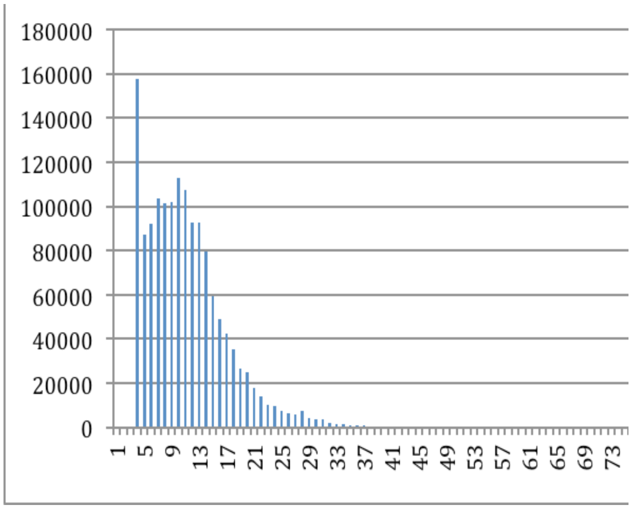

Figure 1: Distribution of Lengths of All Coniguous Alpha-Helix Sequences Found in the NCBI Protein Database.

Measuring how similar two amino acids are to one another is deceptively difficult because there are at least 237 known properties [14] that can be compared; some may be similar to each other with regard to one property, but dissimilar with regard to one another. While it is likely that many of these properties would not yield useful information for secondary structure prediction, it is difficult to define each property's relevance *a priori*. Fortunately, Venkatarajan used multidimensional scaling to condense the information from these 237 properties into five quantitative descriptors [14]. It seemed prudent to use these descriptors, since they contain a great deal of information that should help a machine-learning model quantify amino-acid similarity. It also seemed self-evident that helical propensity is a property that should be included [15].

It is generally known that the interaction of amino-acid residues with water strongly influences the native structure of proteins [12]. Amphiphilic helices are often situated in proteins such that one side of the helix interacts with the hydrophobic interior of the protein and the other side interacts with the hydrophilic surrounding solution. As a result, hydrophobic and hydrophilic residues are generally distributed in a non-random pattern that isolates them on opposite sides of the helix. The angle at which one residue is radially pointed outward from the center of an alpha helix is approximately 100 degrees greater than the angle of the previous residue in the helix. This principle is best illustrated with a Wenxiang diagram [12], a “conical projection of an α-helix onto a plane perpendicular to its axis” as shown in Figure 2 [17].

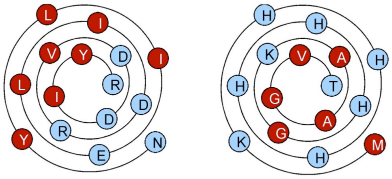

Figure 3: Diagrams of a true alpha helix (left) and a region without secondary structure mapped as though it was an alpha helix (right). Hydrophobic residues are colored red.

Some researchers have quantified this property by using the hydrophobic moment [12]. To calculate the hydrophobic moment, a descriptive vector is created for each amino acid. The direction of the vector points outward from the center of the helical axis toward the residue, while the magnitude of the vector equals the hydrophobic magnitude of the residue (which is, of course, negative for hydrophilic residues). The hydrophobic moment is calculated by adding the all the individual residue vectors. It has shown to be a helpful metric for secondary structure prediction [16].

In order to glean more information from the hydrophobicity patterns, though, a slightly modified approach was used. The cumulative moments of the hydrophilic and hydrophobic residues were calculated separately and the angle between them was determined. The inter-moment angle is a metric that the author invented independently and has not seen used in any of the literature, but it looked promising after data gleaned from ss.txt demonstrated that the distributions of inter-moment angles for alpha-helical regions and unstructured regions appear to be very distinguishable; that data is shown in the histograms found in Figures 3 and 4.

Given the promising results shown by the inter-moment angle metric, it was included along with the magnitudes of both the hydrophobic and hydrophilic moments as an attribute for each training instance.

Since the many specialized metrics described above expand the attribute space considerably, it seemed prudent to seek out a method to reduce the size of the feature space without losing most of the pertinent data contained therein. While many options exist, a stacked auto-encoder seemed like an ideal approach because it condenses information from all attributes into a new set of smaller set of dynamically defined attributes; other approaches that simply discard attributes that are only weakly to the label might lose the benefit (albeit small) of that discarded information.

First, the pdb file for ferritin from the pseudo-nitzschia series was chosen as the data set because ferritin is a large protein with intricate secondary-structure patterns. It was converted to an arff file using a Perl script; the resulting data set had thirteen attribute columns and one classification column. For every given instance, each of the attribute columns could have any single-letter value found in the set {A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X}, where each letter represented its corresponding amino acid (or, in the case of X, an unknown amino acid; X values are occasionally found in PDB files). The classification column of each instance could have any single-letter value found in the set {H,B,E,G,I,T,S,N}, where H = helix, B = residue in isolated beta bridge, E = extended beta strand, G = 3_{10} helix, I = pi helix, T = hydrogen-bonded turn, S = bend, and N = nothing. The classification column represented the secondary-structure classification of the middle amino acid (i.e., the seventh) in the instance. The instances represented all successive subsequences of length 13aa (i.e., a sliding window of size 13 was used). This arff file was meant to serve as a control, since it used none of the specialized metrics described in the preceding sections.

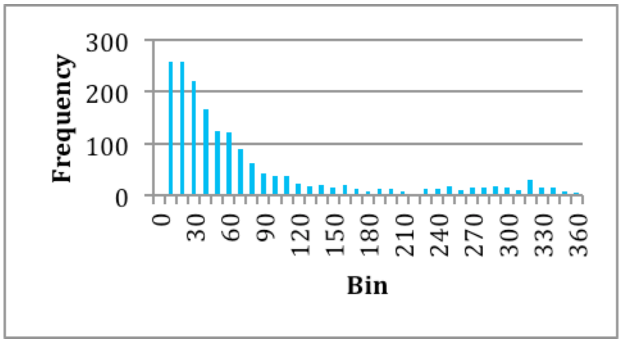

Figure 2: Distribution of Hydrophobic Inter-Moment Angles Found in Alpha-Helices of length 13 amino acids in NCBI Protein Database

Next, the first arff file was converted to a new arff file that replaced the original 13 attributes with an expanded set of 85 attributes. For each instance, this was done by exchanging each amino-acid letter for its five Venkatarajan quantifiers and its helical propensity (delimited by commas appropriately). The three hydrophobicity attributes were then added, followed by the output classes of the previous four instances in order to achieve the recurrent neural network

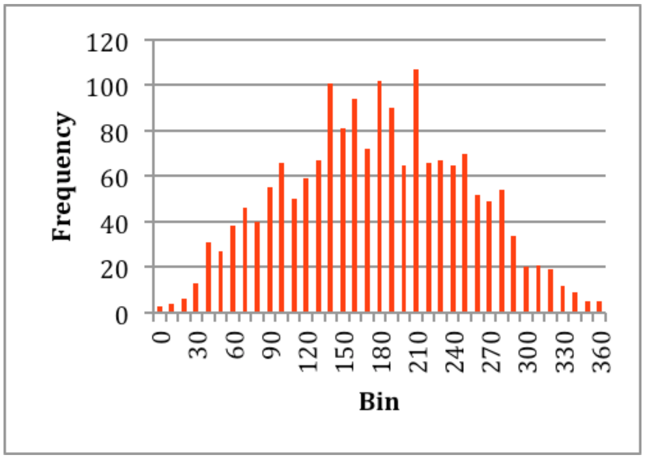

Figure 4: Distribution of Hydrophobic Inter-Moment Angles Found in Non-Structured Sequences of Length 13aa in NCBI Protein.

structure. This was also done with a Perl Script.

Both arff files were then tested using several different machine-learning algorithms in Weka. In addition, a stacked auto-encoder was implemented in Perl and used on the experimental arff file in order to reduce the feature space from 85 attributes to 43. The output from the auto-encoder was then used as input for a single-layer neural network.

IV. Results

The stacked auto-encoder reduced the feature space from 85 attributes to 43. When the auto-encoder's output was used as input for a single-layer neural network, the network achieved a prediction accuracy of 86.73%. A summary of the results is shown in Figure 5.

Algorithm	Prediction Accuracy on control	Prediction accuracy experimental data
Neural Network	69.05%	89.05%
Random Forest	65.23%	86.13%
Naïve Bayes	59.50%	71.96%
IBK (nearest neighbor)	66.73%	77.49%

Figure 5: Prediction Accuracies for several algorithms on 3E6R data 9Ten-Fold Cross Validation

The sizable increase achieved in prediction accuracy when using the experimental attribute set suggests that the three-pronged approach of using recurrence attributes, amino-acid similarity attributes, and whole-subsequence similarity attributes is more effective than the control approach. These results are encouraging, but it should be noted that some of the metrics used in this project—particularly those focused around hydrophobicity patterns—may be less effective if used on trans-membrane proteins that have a large number of non-amphiphilic α-helices.

The recurrent neural network approach probably achieved the highest accuracy out of all the models because the back propagation algorithm is a more natural fit for real-valued attributes than the random-forest and naïve-Bayes algorithms (which are generally a better fit for categorical attributes). Interestingly, the IBK algorithm had the smallest margin of improvement out of all the algorithms. Since nearest-neighbor algorithms like IBK are known to struggle with irrelevant attributes, the weak margin of improvement may indicate the presence of irrelevant attributes in the feature space.

The results from the auto-encoder also seem to suggest that the attribute set can be reduced without sacrificing too much accuracy. A three-percent decrease in accuracy seems very reasonable after going from 85 attributes to 43. Furthermore, the 86% accuracy achieved with a Perl-coded single-layer neural network on the auto-encoder output may not be the highest accuracy that could have been achieved, since the Weka neural network used on the larger data set had more hidden layers and was trained over more epochs.

V. Conclusions

This research demonstrates the effectiveness of machine learning algorithms in secondary structure prediction. Multidimensional scaling combined with a stacked auto-encoder are effective in reducing the feature space to a level compatible with the training data. Neural Networks appear to be most effective in extracting features that are important to accurate prediction.

Future research should focus on discovering additional relevant metrics that measure amino-acid similarity and whole-sequence similarity, identifying and removing irrelevant attributes, and defining the set of proteins for which these metrics are effective with greater specificity.

References

[1] F. B. Straub, *Formation of the Secondary and Tertiary Structure of Enzymes*, 236 Advances in Enzymology and Related Areas of Molecular Biology 89 (1964).

[2] Christian B. Anfinsen et al., *The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain*, 47 Proceedings of the National Academy of Sciences 1309 (1961).

[3] I. Simon, *Investigation of Protein Refolding: A Special Feature of Refolding Responsible for Refolding Ability*, 113 Journal of Theoretical Biology 703 (1985).

[4] Cyrus Levinthal, *How to Fold Graciously*, *in* Mössbaun Spectroscopy in Biological Systems Proceedings, 67 University of Illinois Bulletin 22 (1969)

[5] Christian B. Afinsen, *Principles that Govern the Folding of Protein Chains*, 181 Science 223 (1973).

[6] A.K. Dunker et al., *Intrinsically Disordered Protein*, 19 Journal of Molecular Graphics and Modeling 26 (2001).

[7] Igor N. Berezovsky & Edward N Trifonov, *Loop Fold Structure of Proteins: Resolution of Levinthal's Paradox*, 20 Journal of Biomolecular Structures 5 (2002).

[8] Burkhard Rost & Chris Sander, *Prediction of Protein Structure at Better Than 70% Accuracy*, 232 Journal of Molecular Biology 584 (1993).

[9] J.A. Cuff & G.J. Barton, *Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction*, 34 Proteins 508 (1999).

[10] Sujun Hua & Zhirong Sun, *A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach*, 308 Journal of Molecular Biology 397 (2001).

[11] Gianluca Pollastri et al., *Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles*, 47 Proteins 228 (2002).

[12] Kuo-Chen Chou, Chun-Ting Zhang, & Gerald M. Maggiora, *Disposition of Amphiphilic Helices in Heteropolar Environments*, 28 Proteins 99 (1997).

[13] David Eisenberg, *The Discovery of the α-helix and β-sheet, the Principal Structural Features of Proteins*, 100 Proceedings of the National Academy of Sciences 11207 (2003).

[14] Mathura S. Venkatarajan & Werner Braun, *New Quantitative Descriptors of Amino Acids Based on Multidimensional Scaling of a Large Number of Physical-Chemical Properties*, 7 Journal of Molecular Modeling 445 (2001).

[15] C. Nick Pace & J. Martin Scholtz, *A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins*, 75 Biophysical Journal 422 (1998).

[16] Ross D. King & Michael J.E. Sternberg, *Identification and Application of the Concepts Important for Accurate and Reliable Protein Secondary Structure Prediction*, 5 Protein Science 2298 (1996).

[17] Kuo-Chen Chou, Wei-Zong Lin, and Xuan Xiao, *Wenxiang: A Web Server for Drawing Wenxiang Diagrams*, 3 Natural Science 862 (2011).