

On Predicting Secondary Structure Transition

Raja Loganantharaj and Vivek Philip
Bioinformatics Research Lab
University of Louisiana
Lafayette, LA 70504

Abstract

A function of a protein is dependent on its structure; therefore, predicting a protein structure from an amino acid sequence is an active area of research. Optimally predicting a structure from a sequence is NP-hard problem, hence several sub-optimal algorithms with heuristics have been used to solve the problem. When a structure is predicted by an approximate algorithm, it must be validated and such validation invariably involves validating the secondary structure using the predicted locations of all the residues. To improve the accuracy of validation of secondary accuracy, we are studying the predictability of secondary structure transitions using the following machine learning algorithms: naïve Bayes, C4.5 decision tree, and random forest.

The outcome of any machine-learning algorithm depends on the quality of the training set; hence it must be free from any errors or noise. Absolute error free training data set is not possible to construct, but we have created a data set by filtering out possible errors that are indicated by disagreement of secondary structure assignments or inconsistent with the annotations in PDB, DSSP and STRIDE. We have demonstrated that predicting structure transition with high degree of certainty is possible and we were able to get as high as 97.5% of prediction accuracy.

Keywords: Secondary Structure, Proteins, Model validation

1. Introduction

Proteins play an important role in bodily activities of any living organism taking the role of either a tissue building block or enzymes. Since the function of a protein is dependent on its structure, considerable amount of resources have been devoted to obtain the structure using either experiments or computations. The experimental methods are time consuming and very expensive, while the computational methods for finding protein structure from an amino acid sequence is considered to be a “holy grail” in bioinformatics. A

polypeptide chain of amino acid forms into a sequence of secondary structures namely α -helices, β -sheets, reverse turns and hairpin loops, which we will refer to as coils. The arrangement of these secondary structures folds and conforms into a native structure so as to minimize the energy of the newly formed structure. Prediction of a protein structure from an amino acid sequence is computationally intractable and therefore only approximate methods and heuristic techniques have been used. Once the structure is predicted, several algorithms are used to validate the conformation. The validation methods invariably validate the secondary structures too. In this paper we investigate the secondary structure transition problem and study the predictability of such transitions. The results obtained show high predictability of secondary structure transitions, which will improve the overall validation techniques of a putative structure of a protein.

The experimentally determined structures are deposited and maintained in the Protein Data Bank (PDB) [1]. While the information in PDB is mostly accurate, it is susceptible to human and experimental errors. DSSP [2, 3] and STRIDE [4] are among the popular programs that assign secondary structure for each of the files in the PDB. DSSP assigns secondary structure assignments using hydrogen bond energy and main chain dihedral angles. On the other hand, STRIDE recognizes secondary structural elements in proteins from their atomic coordinates.

We have been searching for a error free data set to train machine learning algorithms to predict secondary structure transitions. The secondary structure assignments of PDB files by DSSP and STRIDE do not always agree and even if they agree it may not be the same as the one annotated in PDB. In the absence of clear winner among the three, we have decided to collect the structures that have the same assignment by DSSP and STRIDE, and are consistent with the annotations of PDB. This set has a very high probability of being correct and we take it as a “gold standard” for this study.

This paper is organized as following. We provide the basis of secondary structure in the section on

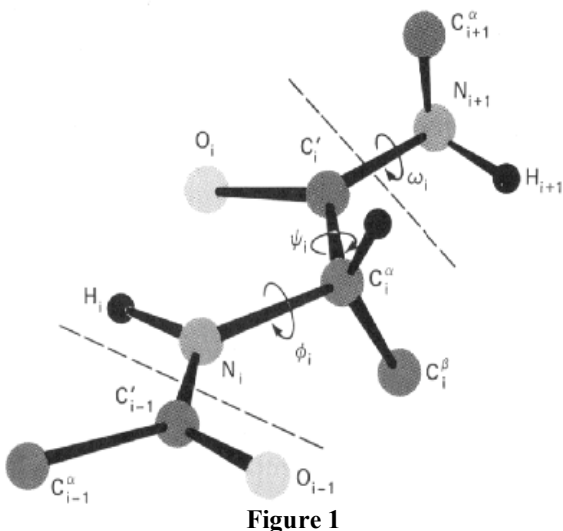
preliminaries. It is followed by a section on data preparation problem. In section 4, we briefly outline machine learning algorithms and provide the details of the experiment. It is followed by results. The paper is concluded by a section on summary and conclusion.

2. Preliminaries

In this section we will describe some basis of parameter measurements to characterize secondary structure and any other relevant materials.

2.1 Secondary structure and phi/psi angles

Figure 1 shows a schematic view of a short polypeptide chain consisting of sequences of residues or side chains each attached to an alpha carbon. A peptide chain is formed by a hydrogen bond between the amide and the carboxyl groups. There are two abundantly recurring conformations at this level. The α -Helix and the β -Sheet conformations are dependent on the angles between the α carbon and the nitrogen of that amino acid, and the α carbon and the carboxyl group. The angles are called phi (ϕ) and psi (ψ) as illustrated in Figure 1. Ramachandran [5] had investigated the relationship between the angles phi (ϕ) and psi (ψ) with the secondary structure assignment of the corresponding residue and found most stable structures of α -helix and β -sheet are confined into two disjoint regions as has been illustrated in Table 1.



Out of the three helix structures namely Right-handed α , 310 and π , α -helix occurs in abundance. In addition to the phi and psi angles, the side chains of the amino

acids also contribute in the formation of structural conformations.

	Bond Angle / Degrees			Res/ Turn
	ϕ	ψ	ω	
Antiparallel β -sheet	-139	+135	-178	2.0
Parallel β -sheet	-119	+113	-178	2.0
Right-handed α -helix	-57	-47	180	3.6
310 helix	-49	-26	180	3.0
π helix	-57	-70	180	4.4

Table 1: Relationships between secondary structures and the angles.

In order to calculate the phi and psi angle for each residue in a protein, we need the three-dimensional structure of that protein. We have downloaded the necessary PDB files from the Brookhaven Protein Data Bank. Once we have the PDB file, the Phi angle for a residue, say r , can be obtained by calculating the dihedral angle formed by four atoms: the carbonyl atom (designated C in a PDB file) of residue $r-1$ and atoms r , $C\alpha$, and C of residue r . Psi is the dihedral angle formed by atoms r , $C\alpha$, C of residue r , and atom r of residue $r+1$. The angle phi for the first residue and psi for the last residue is set to 360, since the first residue will not contain a carbonyl atom, and the last residue will not have the nitrogen atom.

3. Data preparation problem

Experimentally determined structures from the PDB as well as those structures assigned by STRIDE and DSSP are known to have errors or noise. When applying machine-learning algorithms for prediction purposes, the training data set must be free from any errors. Unfortunately, there is no such protein structural data set. Hence, we have created a data set that is relatively free from errors by selecting the structures from PDB, STRIDE and DSSP such that all three of them are in agreement with the structural assignments. We have used PDBSELECT database [5], which is a subset of the structures in the PDB that contains sequences of lower similarity. From this dataset we used structures with less than 25 % sequence similarity. This resulted in a representative dataset of 1990 protein structures. From these sequences with structural information, we have selected the structures that are in agreement with the assignment of both STRIDE and DSSP and as well as consistent with the annotations in PDB.

4. Edge Transitions and Prediction Accuracies

4.1 Edge Transitions

The commonly known secondary structures are α -helices, β -sheets, and coils. Therefore there are six transitions from one secondary structure into another one. Out of the six possible transitions, helix followed by sheets is structurally not feasible; hence we consider only the following five transitions.

1. Helix conformation followed by a coil (HC).
2. Coil followed by a helix (CH).
3. Sheet Conformation followed by a coil (EC).
4. Sheet Conformation followed by a Helix (EH).
5. Coil followed by sheet (CH).

To study the predictability of structure transitions, we have considered amino acid sequence of length 6 residues of which three are in the upstream and the other three are in the downstream of a transition. We have to capture the texture of the secondary structures of these six residues. We are enumerating the following features to capture the texture of the structures: the residues themselves, their chemical properties, their angles phi and psi.

In this study, we are using Naïve Bayes, C4.5 decision tree and random forest and we will describe each one of them briefly.

4.2 Naïve Bayes

Naïve Bayes have been successfully used to solve many problems in various domains including bioinformatics. Given a representation of a sequence, the probability of a class C is given by $P(C | \text{features})$. The Naïve Bayes uses conditional independence of the features and the conditional probability becomes proportional to $\prod P(f_k | C) \cdot P(C)$ for all $k = 1$ to all the features. F_k denotes k^{th} feature.

Once the naïve Bayes is trained with the features of the training set consists of sub structural sequence around the structure transition, it will make a decision on a new and probably unknown set of features as to what transition it belongs to using a technique called maximum likelihood.

4.3 Decision Tree C4.5

A decision tree is a simple, but a powerful machine learning algorithm that has been used successfully for classification problems. Each leaf node represents a class, while each internal node represents an attribute. When classifying an instance, a series of decisions was made during the traversal from root to a leaf node and the instance was classified to the one associated with the leaf node at the end of the traversal. Each internal node is a decision node and a value of a given instance is compared to the decision function to decide which branch to follow. A decision tree is build using a training data set so as to reduce the average depth of each path from root to leaf node and to be flexible enough to avoid data over fitting. The popular algorithms for decision tree include id3 [6, 7] and its successor C4.5 [7, 8]. Both algorithms are using changes in entropy to achieve overall shorter depth of all the paths from root to leaf nodes. The algorithm C4.5 is an improved version of ID3 by allowing continuous variable and partition optimization for them and as well providing tree pruning techniques to avoid data over fitting.

4.4 Random Forest

Breiman [9] has proposed random forest that can be used as a classifier and as well as a clustering algorithm. From each training example, the algorithm builds a tree starting with a feature that is selected randomly and grow to its maximum depth. When classifying, the decision is made by aggregating (majority vote for classification and average for regression) the prediction of all the trees built during training. Since each tree in the ensemble is built non-deterministically, the classifier is not sensitive to noise and it seems to be outperforming single tree classifier such as CART and C4.5. One major advantage of random forest classification is that it has a built in out of bag estimator that enables to perform cross validation during the forest building phase which in turn makes the classifier to improve its accuracy without over fitting.

4.5 Feature Selection and their influence in Predictability

To get a quick assessment of the features and their effect on the predictability, we use Naïve Bayes algorithm, which is considered to be a good base-line learning algorithm. WEKA[10,11] is a popular data mining workbench written in Java and is available for free downloading. Weka consists of many popular machine learning algorithms and it provides several

options for testing data sets. We have used the latest version of Weka to compare the performance of Naïve Bayes, random forest, and Decision tree C4.5. Each of these algorithms was tested with 10 fold cross validation, that is, the data set was randomly divided into 10 equal parts and the algorithm was trained with all but one partition, which was used for testing. The training and testing were repeated 10 times so that each partition was tested exactly once.

We have started with the residues as the only features and the results of prediction accuracy for each transition class with 10 fold cross validation are shown in Table 1. The performance is measured with true positive rate, false positive rate and precision. The overall performance is not satisfactory.

Table 1: Performance of Naïve Bayes with residues as input feature

Class	HC	CH	EC	CE	EH
TP Rate	0.623	0.24	0.74	0.605	0.011
FP Rate	0.228	0.046	0.279	0.055	0.002
Precision	0.559	0.434	0.609	0.606	0.257

Secondary Structure and Hydrophobic Property

The tendency for non-polar molecules to aggregate in water, is widely believed to be the main driving force behind the folding of globular proteins [10]. When proteins fold it is thermodynamically favorable to bury the hydrophobic residues [11], and as a consequence non-polar amino acids tend to be clustered in the interior of proteins [12]. An amino acid within a protein can be located either on the outside or the inside. The hydrophobic property of an amino acid determines its location in the final structure of the protein[13]. If the protein is a globular protein, then the hydrophobic R groups will be located on the inside of the protein, away from the water. On the other hand if the residue is hydrophilic, then the hydrophilic R group is located on the outside of the protein, interacting with the water[13]. We have obtained the hydrophobic property of each of the 20 amino acids from “A Review of Amino Acids” [14]. The results of naïve Bayes using the residues and their hydrophobicity are shown in Table 2. Adding the chemical property of the residues did not improve the performance significantly.

Table 2: naïve bayes using the residues and their hydrophobicity

Class	HC	CH	EC	CE	EH
TP Rate	0.597	0.261	0.719	0.595	0.029
FP Rate	0.226	0.06	0.278	0.056	0.007
Precision	0.551	0.393	0.603	0.6	0.221

We ran naïve Bayes classifier with 10% cross validation on phi and psi angle of the six residues and have obtained the overall prediction accuracy of 85.15%. When we add the residues along with the phi/psi angles the prediction accuracy had increased to 85.31%. When we add the hydrophobic property along with the previous two parameters the prediction accuracy have gone up to 85.95% though the increase is not much significant. Since a Naïve Bayes provides the best prediction accuracy with all three parameters namely amino acids, their corresponding phi and psi angles and their hydrophobic properties, we use the data set to compare the performance of other classifiers.

The dataset has approximately 14000 instances and each instance has 6 residues. The Figure 1 illustrates a snapshot of an instance of our dataset. The first six are the residues followed by six phi values, six psi values, information of whether the residue is hydrophobic or hydrophilic and lastly the transition class.

Table 3: Performance of naïve Bayes on structure transition

Class	TP Rate	FP Rate	Precision
HC	0.952	0.023	0.95
CH	0.89	0.027	0.831
EC	0.824	0.045	0.914
CE	0.891	0.035	0.783
EH	0.468	0.045	0.404

The performance of the three machine learning algorithms is compared side by side with respect to the prediction metric true positive and precision. The results are shown in Figures 2 and 3. The accurate prediction of sheets to helix seems to be difficult and at best is come close to 61% true positive rate.

Table 4: Performance of random forest on structure transition

Class	TP Rate	FP Rate	Precision
HC	0.975	0.027	0.944
CH	0.901	0.012	0.92
EC	0.975	0.068	0.893
CE	0.86	0.01	0.925
EH	0.431	0.002	0.923

Table 5: Performance of J48 on structure transition

Class	TP Rate	FP Rate	Precision
HC	0.965	0.027	0.944
CH	0.901	0.016	0.892
EC	0.928	0.046	0.922
CE	0.856	0.013	0.901
EH	0.61	0.02	0.663

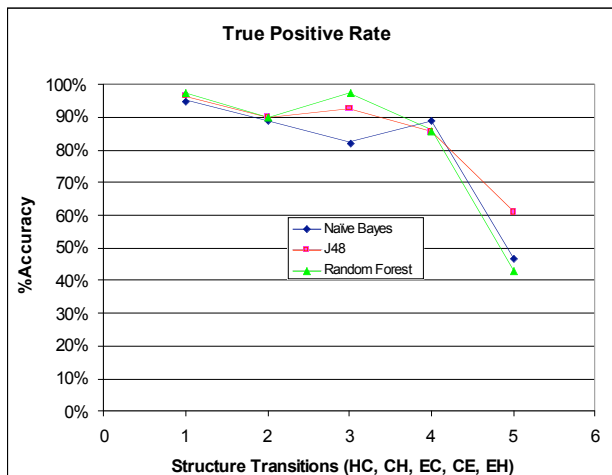


Figure 2: True positive rate of class transitions

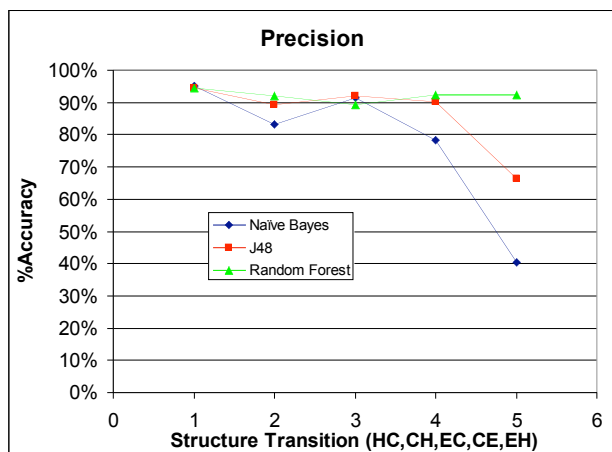


Figure 3: Precision of class transitions

5. Summary and Conclusion

When we started our efforts towards structure validation, we were faced with the problem of noisy data set of secondary structures. Even the very high-resolution PDB files have errors. Furthermore the software DSSP and STRIDE that assign secondary structure to a PDB file may have disagreements in

some structures. Since there is no “golden standard” dataset for secondary structures, we have created a data set that has high probability of being noise free. We have started with sequences from high resolution PDB that has less than 25% sequence similarity and have agreement with the assignment given by STRIDE and DSSP.

We have used three classifiers namely Naïve Bayes, C4.5 decision tree [15] and random forest [9] for testing the predictability of structure transition. The residue alone was a poor parameter for prediction. We then tried with angles phi and psi associated with six residues of which three of them are in the leading and the rest of the three residues are in the following structure. The phi and psi angle alone produced a prediction accuracy of 85.15% and it seems to be a good parameter for predicting the structure transition. When we include other parameters such as amino acids, and the hydrophobic property of each of the amino acids, the prediction accuracy increased to 85.95%. All the classifiers have produced excellent predictability for all but sheet to helix transition. For example, naïve Bayes predicted the following true positive rate with 10% cross validation: 95% for helix to coil transition, 89% for coil to helix transition, 82% for sheet to coil, 89% coil to sheet and 46.8% for sheet to helix transition.

We have demonstrated with this experiment that the secondary structure transition can be predicted with high accuracy for all but the sheet to helix transition. We have used the following features for training and testing: residues, their properties and phi and psi angles. Out of all the algorithms, the random forest seems to have better performance. By having a better prediction on the structure transition, we can improve the validation process of secondary structures.

Acknowledgement: This work was partially supported by the Governor’s IT initiatives.

References

1. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures*. Arch Biochem Biophys, 1978. **185**(2): p. 584-91.
2. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. Biopolymers, 1983. **22**(12): p. 2577-637.
3. Kabsch, W. and C. Sander, *How good are predictions of protein secondary structure?* FEBS Lett, 1983. **155**(2): p. 179-82.
4. Frishman, D. and P. Argos, *Knowledge-based protein secondary structure assignment*. Proteins, 1995. **23**(4): p. 566-79.
5. *PDB Select*, <http://swift.cmbi.kun.nl/swift/pdbsel/>.
6. Quinlan, J.R., *Simplifying decision trees*. International Journal of Man-Machine Studies, 1987. **27**: p. 221-234.
7. Winston, P.H., *Artificial intelligence*. 3rd ed. 1992, Reading, Mass.: Addison-Wesley Pub. Co. xxv, 737 p.
8. Quinlan, J.R., *C4.5: Programs for Machine Learning*. 1993: Morgan Kauffman.
9. Breiman, L., *Random Forest*. Machine Learning, 2001. **45**: p. 5-32.
10. Dill, K.A., *Dominant forces in protein folding*. Biochemistry, 1990. **29**(31): p. 7133-55.
11. Malakauskas, S.M. and S.L. Mayo, *Design, structure and stability of a hyperthermophilic protein variant*. Nat Struct Biol, 1998. **5**(6): p. 470-5.
12. Chothia, C., *Structural invariants in protein folding*. Nature, 1975. **254**(5498): p. 304-8.
13. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein*. J Mol Biol, 1982. **157**(1): p. 105-32.
14. Acevedo, O.E. and L.R. Lareo, *Amino acid propensities revisited*. Omics, 2005. **9**(4): p. 391-9.
15. Witten, I.H. and E. Frank, *Data mining: practical machine learning tools and techniques*. second ed. 2005, San Francisco, Calif.: Morgan Kaufmann. xxv, 524.

Figure 1: Representation of features for machine learning algorithms

