Towards Better Understanding of Protein Secondary Structure: Extracting Prediction Rules

Minh N. Nguyen, Jacek M. Zurada*, Fellow and Jagath C. Rajapakse

Abstract— Although numerous computational techniques have been applied to predict protein secondary structure (PSS), only limited studies have dealt with discovery of logic rules underlying the prediction itself. Such rules offer interesting links between the prediction model and the underlying biology. In addition, they enhance interpretability of PSS prediction by providing a degree of transparency to the predicting model usually regarded as a black-box. In this paper, we explore the generation and use of C4.5 decision trees to extract relevant rules from PSS predictions modeled with two-stage support vector machines (TS-SVM). The proposed rules were derived on the RS126 dataset of 126 nonhomologous globular proteins and on the PSIPRED dataset of 1923 protein sequences. Our approach has produced sets of comprehensible, and often interpretable, rules underlying the PSS predictions. Moreover, many of the rules seem to be strongly supported by biological evidence. Further, our approach resulted in good prediction accuracy, few and usually compact rules, and rules that are generally of higher confidence levels than those generated by other rule extraction techniques.

Index Terms— protein structure, secondary structure prediction, support vector machines, multi-class SVM, C4.5 decision trees, rule extraction.

1 Introduction

Information on secondary structures of amino acid residues in proteins provides valuable clues for the prediction of their three dimensional (3-D) structure and function. Knowledge of a protein's structure, in turn, contributes to our understanding of the functions of the protein and is vital to many aspects of living organisms such as those of enzymes, hormones, and structural material, etc. It also helps in designing new drugs for combating disease. Hence, prediction of 3-D structure of a protein from its amino acid sequence has become one of the major goals of bioinformatics.

Unfortunately, the protein structure prediction problem is a combinatorial optimization problem, and hence it has so far eluded an effective solution because of the exponential number of potential solutions. One of the current approaches is to first predict protein secondary structure (PSS) assuming a linear representation of the full knowledge of the 3-D structure, and to use it to predict the 3-D structure. The goal of secondary structure prediction is to assign a pattern of residues in amino acid sequences to a class of protein secondary structure elements most often labeled as an a-helix (a), β-strand (β) or coil (ζ), the remaining type.

Many computational techniques have been proposed in the literature to deal with the PSS prediction problem. The statistical methods are mostly based on the likelihood of each amino acid sequence being one of three types of secondary structures [1]-[3]. Neural networks use residues in a local neighbourhood as inputs and compute an arbitrary non-linear mapping [4]-[7]. The Bayesian approach provides a framework to account for non-local interactions among amino acid residues [3], where the inferences are based on the generalized probability distributions incorporating prior probabilities of segments of secondary structure elements. The consensus approaches combine different classifiers in parallel to achieve a single superior predictor. Cuff and Barton employed a majority voting scheme to combine predictions from different techniques [8]. More complex approaches for combining different methods based on neural networks and linear discrimination have also been studied [9]. Sen *et al.* proposed a consensus algorithm for protein secondary structure prediction by combining two complementary methods: fragment database mining (FDM) was used to exploit the Protein Data Bank structures and the GOR V based on information theory and Bayesian statistics [10]. Recently, Meiler and Baker proposed using the information of 3-D structure and PSI-BLAST [11] profiles as inputs to a neural network [12]. Support Vector Machines (SVM) have been applied to PSS prediction, in combination with several binary classifiers [13].

- *Minh N. Nguyen is with BioInfomatics Institute, Singapore. Email: minhn@bii.a-star.edu.sg.*
- *Jacek M. Zurada is with BioInformatics Research Centre, Nanyang Technological University, Singapore; and Department of Electrical and Computer Engineering, University of Louisville, USA. Email: jacek.zurada@louisville.edu.*
- *Jagath C. Rajapakse is with BioInformatics Research Center, Nanyang Technological University, Singapore; and Singapore-MIT Alliance, Singapore; and Department of Biological Engineering, Massachusetts Institute of Technology, USA. Email: ASJagath@ntu.edu.sg.*
- * To whom correspondence should be addressed.

Digital Object Indentifier 10.1109/TCBB.2010.16 1545-5963/10/$26.00 © 2010 IEEE

The accuracy of the single-stage approaches to PSS prediction, however, has been found insufficient. Rost and Sander proposed the PHD approach using Multi-Layer Perceptrons (MLP) in cascade, with the second layer of MLP improving the accuracy of the prediction by capturing the contextual relations among the secondary structures at the output of the first layer [4]. We proposed a two-stage SVM (TS-SVM) for the prediction of PSS [14], of relative solvent accessibility [15], and of accessible surface area of amino acids [16], all with inputs from PSI-BLAST profiles. These techniques are able to incorporate useful information from multiple sequence alignments or PSI-BLAST profiles and contextual information among secondary structures in the prediction scheme.

Despite the success of many computational approaches, not much research has been done to discover what underlying general patterns of amino acid sequences are associated with specific secondary structure elements. Recently, He *et al.* proposed a rule-extraction method for PSS prediction by combining SVM and decision trees [17]. The method uses one-stage of binary SVM, which is unable to capture contextual relationships among the secondary structures and can not assign directly a pattern of amino acid sequences to a class of PSS outputs with sufficient accuracy.

To alleviate this shortcoming, we propose combining the PSS predictions from the TS-SVM with C4.5 decision trees to extract useful rules possibly governing PSS prediction. This not only increases the accuracy of prediction by decision trees or SVM along with decision trees, but it also renders a set of PSS prediction rules, which are more confident and more evident biologically as compared to rules reported so far. These rules describe amino acid patterns that are likely to produce specific secondary structures in a particular context. The rules can therefore imply protein structures likely to be produced by short segments of amino acids in a specific context, and may be useful in guiding biological experiments for determining protein structures, or, inversely, for inferring amino acid patterns from secondary protein structures.

The input to the TS-SVM is based on the position-specific scoring matrices generated by PSI-BLAST profiles of the input amino acid sequence. We use the output of TS-SVM to generate PSS prediction rules by C4.5 decision trees. We extracted three sets of rules for PSS prediction based on whether the prediction is purely on amino acid patterns, or it uses structural types of residues in the vicinity of predicted output. Furthermore, many rules extracted by our method were more confident and clearer supported by evidences from biological literature than any rules reported so far. Our method resulted in an improvement of 2.5% as compared to the best results on the RS126 dataset of 126 nonhomologous globular proteins [4], achieved previously by a rule extraction method.

The paper is organized as follows: Section 2, Methods, describes TS-SVM and C4.5 decision tree technique. Section 3, Experiments and Results, describes the datasets used, simulations made, and discusses extracted rules. Section 4, Discussion, provides concluding remarks for the analyses made.

2 METHODS

2.1 Two-stage SVM

Let $\mathbf{u} = (u_1, u_2, ..., u_n)$ be a given amino acid sequence where $u_i \in \Omega_U$ and Ω_U denotes the set of 20 amino acid residues, and $\mathbf{t} = (t_1, t_2, ..., t_n)$ be the corresponding secondary structure sequence where $t_i \in \Omega_T$ and the set of secondary structures, $\Omega_T = \{\alpha, \beta, \zeta\}$; n is the length of the sequence. The prediction of PSS sequence is the problem of finding the optimal mapping from the space of Ω_U to the space of Ω_T. Let $\mathbf{v}_i$ be the 21-dimensional feature vector representing residue u_i where 20 values are from raw matrices of PSI-BLAST profiles ranging in [0, 1] and the remaining unit is used for padding to indicate an overlapping end of the sequence [10]. Let $\mathbf{r}_i = (\mathbf{v}_{i-h_1}, ..., \mathbf{v}_i, ..., \mathbf{v}_{i+h_1})$ be the input to the multi-class SVM at site i of the sequence where h_1 denotes the width of a symmetric neighbourhood window of residues on one side. TS-SVM uses two multi-class SVMs in cascade for the prediction of protein features from amino acid sequences [14]-[16]. We use a multi-class SVM proposed by Crammer and Singer for both stages [18].

The first-stage constructs three discriminant functions for three secondary structures by solving the single optimization problem:

$$\arg\min_{\mathbf{w}_1} \frac{1}{2} \sum_{k \in \Omega_T} (\mathbf{w}_1^k)^T \mathbf{w}_1^k + \gamma_1 \sum_{j=1}^{N} \xi_j^1$$

subject to the constraints

$$\mathbf{w}_1^{t_j} \phi_1(\mathbf{r}_j) - \mathbf{w}_1^k \phi_1(\mathbf{r}_j) \geq c_j^k - \xi_j^1 \qquad (1)$$

where t_j is the secondary structural type at site j corresponding to input vector $\mathbf{r}_j$. N is the size of the training data, and $\mathbf{w}_1^{t_j}$ and $\mathbf{w}_1^k$ are weight vectors corresponding to secondary structures t_j and k, and

$$c_j^k = \begin{cases} 0 & \text{if } t_j = k \\ 1 & \text{if } t_j \neq k \end{cases}$$

The above optimization is simplified by solving the following quadratic programming problem [18]:

$$\max_{\alpha_j^k} -\frac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} K_1(\mathbf{r}_j, \mathbf{r}_i) \sum_{k \in \Omega_T} \alpha_j^k \alpha_i^k - \sum_{j=1}^{N} \sum_{k \in \Omega_T} \alpha_j^k c_j^k \qquad (2)$$

$$\text{such that } \sum_{k \in \Omega_T} \alpha_j^k = 0 \text{ and } \alpha_j^k \leq \begin{cases} 0 & \text{if } t_j \neq k \\ \gamma_1 & \text{if } t_j = k \end{cases}$$

where $\mathbf{w}_1^k = \sum_{j=1}^{N} \alpha_j^k \phi_1(\mathbf{r}_j)$ and $K_1(\mathbf{r}_i, \mathbf{r}_j) = \phi_1(\mathbf{r}_i) \phi_1(\mathbf{r}_j)$ denotes the kernel function. The input vectors, derived from a window of $2h_1+1$ amino acid residues, are transformed into a higher dimensional space via kernel function K_1. Once the optimal parameters α_j^k are obtained, the discriminant function of structure k, f_1^k for an input $\mathbf{r}_i$ is given by

$$f_1^k(\mathbf{r}_i) = \sum_{j=1}^{N} \alpha_j^k K_1(\mathbf{r}_i, \mathbf{r}_j) = \mathbf{w}_1^k \phi_1(\mathbf{r}_i) \qquad (3)$$

The second stage uses another SVM to predict PSS from the output of the first stage SVM to enhance prediction accuracy by capturing the contextual dependences of secondary structures, for example, β-strands span over at least three residues and α-helices composed of at least four residues [4], [14].

The input to the second SVM at site i is obtained from a neighbourhood, $\mathbf{d}_i^1 = (d_{i-h_2}^{1k}, ..., d_i^{1k}, ..., d_{i+h_2}^{1k} : k \in \Omega_T)$ where $d_i^{1k} = 1/(1 + e^{-f_1^k(\mathbf{r}_i)})$ and h_2 is the size of the neighborhood

on one side. The logistic sigmoid function is selected to normalize the inputs to the second stage to [0, 1]. The input patterns to the second stage are converted to a higher dimensional space by using a mapping ϕ_2 and a kernel function: $K_2(\mathbf{d}_i^1, \mathbf{d}_j^1) = \phi_2(\mathbf{d}_i^1)\phi_2(\mathbf{d}_j^1)$. The outputs in the higher dimensional space are linearly combined by a weight vector $\mathbf{w}_2^k$ to produce the final prediction. The vector $\mathbf{w}_2^k$ is obtained by solving the following convex quadratic programming problem, over all secondary structure sequences predicted by the first stage in the training stage. The secondary structural type $\hat{t_i}$ at site i of input sequence is estimated by

$$\hat{t_i} = \arg\max_{k \in \Omega_T} f_2^k(\mathbf{d}_i^1) \quad (4)$$

where $f_2^k(\mathbf{d}_i^1) = \mathbf{w}_2^k \phi_2(\mathbf{d}_i^1)$ is the discriminant function at the second stage given by as in Eq. (3).

2.2 Decision Trees

SVMs perform well compared to other statistical or machine learning techniques in predicting protein features [14]-[17] because of their generalization capabilities. However, SVMs alone yield black box models and provide no biologically meaningful prediction rules [17]. Decision trees, on the other hand, are capable of explicitly describing the nature of prediction since they capture rules as prevailing regularities governing the prediction process. Prediction rules offer useful guidance for wet-lab experiments and a basis for advanced inference of biological features correlated to specific structures.

Decision tree learning provides a means of approximating discrete-valued target functions, in which the learned function is represented by a decision tree [19]. In order to improve human comprehensibility, learned decision trees can be re-represented as sets of if-then rules. We use C4.5 tree induction algorithm at the output of TS-SVM to generate rules for PSS prediction. C4.5 was chosen because it has shown to extract more accurate rules in many applications including bioinformatics problems, for example generating automatic rules for protein annotation, mining protein sequences in SWISS-PROT, and PSS prediction [17]. It uses the gain ratio criterion based on the information theory to select the attribute at the root of the tree and produces suboptimal trees by learning heuristically from input [20]. The important rules are generated by first creating a decision tree on a training set, and then pruning the tree by replacing a whole of subtree with a leaf node if a decision rule establishes a greater expected error rate in the subtree than that in the single leaf. Rule sets are then derived from writing a rule for each path in the decision tree from the root to a leaf. The leaf-hand side is easily built from the label of the nodes and the labels of the arcs.

Let the training set of exemplars for C4.5 decision tree be $\Gamma^2_{\text{train}} = \{(\mathbf{a}_j, t_j) : j = 1, ..., N\}$ where the input at site j is $\mathbf{a}_j = (d^{2k}_{j-h_2} ... d^{2k}_j, ... d^{2k}_{j+h_2}, \mathbf{v}_{j-h_1}, ... \mathbf{v}_j, ... \mathbf{v}_{j+h_1})$ and t_j is the desired secondary structure where $d_j^{2k} = 1/(1+e^{-f_2^k(\mathbf{d}_j^1)})$. The training set is used to train the decision tree and to extract the corresponding rule sets. The rules are then tested with the same data set for evaluation of the performance of the algorithm.

3 Experiments And Results

The presented approach was implemented using position-specific scoring matrices generated by PSI-BLAST as inputs and tested on benchmark with RS126 dataset. The results were compared with other prediction methods and with other rule extraction results for PSS on this dataset. Separate rule extraction using the TS-SVM methods and C4.5 was also performed on PSIPRED dataset of 1923 proteins. Subsequently, a set of common rules for both datasets was identified as discussed below.

3.1 Datasets

The set of 126 nonhomologous globular protein chains, used by Rost and Sander and referred to as the RS126 set [4], was used to evaluate the accuracy of the predictors and relevance of extracted rules. Many current PSS prediction methods have been developed and tested on this dataset. To refer to most relevant reports in the literature, experiments on this dataset were performed with 7-fold validation. The dataset contained 23349 residues with 32% *a*-helix, 23% *β*-strand, and 45% coil. The RS126 set is available at

http://www.compbio.dundee.ac.uk/~www-jpred/data/pred_res/126_set.html.

A much larger dataset PSIPRED [7] has also been used in this project to evaluate the TS-SVM predictors with and without C4.5. After eliminating 322 sequences with some unknown amino acids, experiments were performed on remaining 1923 sequences using 10-fold validation. The dataset is available at

http://bioinf.cs.ucl.ac.uk/downloads/psipred/old/data.

The type of the secondary structure of each residue in training and testing sets was assigned from DSSP [4]. This reduction approach was adopted since it is commonly used for comparison of PSS prediction accuracies by other researchers [4]-[9].

3.2 Implementation

The mostly used SVM software includes LibSVM (*http://www.csie.ntu.edu.tw/~cjlin/libsvm/*), Pls-SVM (*http://www.esat.kuleuven.ac.be/sista/lssvmlab/*), and SVM-Light (*http://svmlight.joachims.org/*). In this paper, the multi-class SVM method was implemented using BSVM library which is known to show fast convergence for large optimization problems [21]. The Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\sigma\|\mathbf{x}-\mathbf{y}\|^2}$ showed superior performance over linear and polynomial kernels for predicting PSS [14], relative solvent accessibility [15], accessible surface areas of amino acids [16], and gene classification [22]. The sensitivity parameter γ and the Gaussian kernel parameter σ were determined by using the grid-search method [21]. Grid-search provides good parameter estimates for multi-class SVM in a relatively short time. The parameters of the Gaussian kernel and TS-SVM, as σ_1=0.0625, σ_2=0.0156 and γ_1= γ_2=0.5, and the neighborhood window h_1 = 7, and h_2 = 3 were empirically determined for optimal performance. We implemented the decision tree C4.5 by using Weka software [23]. For C4.5, the confidence factor of 60% was chosen, and an appropriate value for the minimum number of instances per leaf within [1, 60] was selected

based on cross-validation results.

3.3 Prediction Accuracies

We used Q_3 accuracy to measure the percentages of correctly predicted residues of three types of secondary structures [8]:

$$Q_3 = \frac{\sum_{t \in \Omega_T} \eta_t}{\sum_{t \in \Omega_T} \upsilon_t} \times 100 \quad (5)$$

where η_t is the number of correctly predicted residues and υ_t is the total number of residues observed of secondary structure type t. We also used a rule's *confidence* to indicate its accuracy verified on the whole dataset. The confidences C_α, C_β, and C_ζ represent the percentages of correctly predicted residues of each type of secondary structure.

The performance of secondary structure prediction on the RS126 dataset of 126 proteins by using TS-SVM and C4.5 is shown in Table 1. It can be seen that the combination of TS-SVM and C4.5 predicted PSS with the highest average accuracy (75.0%) in comparison to C4.5 alone (58.6%), and to the combination of SVM with C4.5 (73.7%). It should be noted that the combination of TS-SVM and C4.5 decision trees generates fewer rules than SVM + C4.5 or C4.5 alone, while also yielding higher accuracy of prediction even with fewer rules. In addition, a decision tree with smaller number of leaves makes rules more comprehensible and at the same time more reliable. Less leaves means that on the average there are more examples per leaf that give the rule better statistical conformation.

Table 2 shows comparative performance of different PSS on the RS126 set without C4.5 rules (rows 2, 4) and with C4.5 rules (rows 1, 3, 5). The results indicate that after combining SVM or TS-SVM with C4.5 decision trees, the accuracy of the combinations drops by 3% and 1.6%, respectively, because of pruning of less significant rules. This occurs when some amino acid patterns contributing to the prediction, but occurring rarely in the dataset, are filtered out during pruning when building decision trees. Table 2 also shows an improvement of 2.5% in prediction accuracy (see row 1 and 3) of our approach compared to the method of He *et al.* produced on RS126 by combining single-stage binary SVM with C4.5 [17]. Though the method only reported the accuracies of three binary SVM classifiers, we computed the overall prediction accuracy from the counts of structure types as 72.5%.

To validate the rules extracted by our approach on different training and testing datasets, we have used the rules produced with a confidence level above 60% on PSPIRED and computed the accuracy on RS126 dataset. The accuracy of 76.1% shows that our approach has the ability to perform well on different training and testing datasets.

3.4 Extracted Rules

Subsequent description concerns rules extracted using the C4.5 method from TS-SVM predicting model. Overall number of rules produced with confidence level above 60% was 56 and 523, on RS126 and PSPIRED datasets, respectively. This count includes all rules of three different types as defined below. In the discussion to follow, we have, however, reduced the set of all rules to the subset being intersection of both sets. It jointly identifies 32 common rules that cover both sets for which we are seeking biological relevance and potential interpretability across both datasets.

Validation Run	C4.5		SVM + C4.5		TS-SVM + C4.5	
	Acc	Rules	Acc	Rules	Acc	Rules
1	56.6	148	72.4	91	74.4	45
2	59.0	159	75.2	79	76.9	41
3	58.4	169	74.2	79	74.6	61
4	57.5	166	72.2	75	73.3	49
5	58.9	163	73.1	78	73.7	45
6	61.6	159	76.0	100	78.2	52
7	58.5	167	72.9	79	73.6	53
Average	58.6	161	73.7	83	75.0	49

Table 1: Generated rules and accuracies (Acc) of different type for rule-based classifiers, obtained using seven-fold cross validation on a RS126 dataset of 126 proteins, 60% confidence level.

Dataset	**Method**	C_α	C_β	C_ζ	**Acc**
RS126	Binary SVM + C4.5[17]	72.8	79.6	69.3	~72.5
	TS-SVM[14]	73.1	65.7	83.8	78.0
	TS-SVM + C4.5	77.9	69.3	75.3	75.0
PSIPRED	TS-SVM[14]	78.4	74.6	83.5	80.9
	TS-SVM + C4.5	84.7	80.7	77.0	79.3

Table 2: Performance comparison between SVM and TS-SVM alone vs. with C4.5 for PSS prediction on the RS126 and PSIPRED datasets.

Below we partitioned the common rules into three categories I, II, and III, based on whether TS-SVM already predicted the specific secondary structure. The rules are shown in Tables 3, 4, and 5. The bold amino acid indicates the position of the secondary structure. The symbol 'x' indicates that a 'do not care' condition for the amino acid in that site.

The confidence of the rules on each dataset is given in third and fourth column of the tables, respectively. The co-occurrences of such patterns with a specific secondary structure were the basis of prediction of PSS in GOR methods [1]. As can be seen from all the tables, the presented method resulted in more accurate predictions than those based on linear associations in the GOR method. This is because of the complex non-linear mapping by TS-SVM and extraction of relevant rules transforming patterns of amino acids to secondary structures. To show the usefulness and biological relevance of the rules, we interpret some of the derived rules by discussing evidence found in the literature.

Type I Rules

Type I rules extracted by the presented method are shown in Table 3. Listed are rules common for two datasets with variable confidence above 60%, indicating which amino acid patterns lead to the prediction of specific protein secondary structures. The first two rules indicate that the method predicts an *α*-helix when patterns **L**xxM and **V**xAL are present, with 66.7% and 60.0% confidence, and 60% and 64% confidence, respectively. As seen, Leucine (L, Leu) and Methionine (M, Met) are present at three sites downstream of the site. Amino acids L and M are non-polar R group (hydrophobic) and tend to form *α*-helix, and their presence at three sites downstream proves to be helix-stabilizing.

It has been previously reported that L-L, L-V, L-I, F-M, and L-M pairs at the local site and occurs commonly three and four sites downstream in *α*-helices and contribute to protein's structural stability [24]. Experimental and theoretical studies on natural and synthetic peptides and proteins indicate that individual side chains differ in their potential of helix-forming. Four aliphatic side chains occur in the standard complement of amino acids: L and A are helix stabilizing whereas V and I are weakly destabilizing helices [25]. From position-specific amino acid preferences in *α*-helices [26], there is a peak preference for hydrophobic amino acids L and V in positions N4 (N-cap + 4) and C3 (C-cap - 3) and M in position C4 (C-cap - 4). Helix boundary residues (the first and last helical residues) are called N-cap and C-cap at the N- and C-terminus, respectively. Positions N4 and C4 are underneath the polypeptide chain leading the helix, and also usually on its interior face as the chain at each end must connect to the rest of the protein [26].

As also seen from Table 3, patterns **D**VxLG, **S**VxVG, **W**VxIG, and TV**T**V, predict *β*-strands with 100% confidence for RS126 and somewhat lower confidence levels for PSPIRED as shown in the right-most column. Rule 3 shows that if Aspartic acid (D, Asp) is present at a site and Valine (V, Val), Leucine (L, Leu), and Glycine (G, Gly) at one, three, four sites downstream, respectively, then the secondary structure at the site will be a *β*-strand. This rule suggests that negatively charged (hydrophilic) amino acid D at the local site and non-polar R group (hydrophobic) amino acids V, L, and G downstream, prove to be sheet stabilizing. Colloc'h and Cohen focused their attention on the conformational and structural properties of residues that initiate or terminate a *β*-strand [27] and are referred to as *β*-breakers because of their role in breaking the regular geometric structure of the strand. They found a preference for D, T, and R as the N-terminal *β*-breaker and G and S as the C-terminal *β*-breaker. Interestingly, our previous work found that hydrophobic amino acids V and I strongly tend to be *β*-strand [14]. Moreover, in rules 6 and 7 in Table 3, the weakly hydrophilic amino acid T is two sites upstream, the non-polar R group (hydrophobic) amino acid V is one site upstream, then another non-polar R group (hydrophobic) amino acids I or weakly hydrophilic amino acid T is the local site, and finally another hydrophobic amino acid V. If this forms a sheet, then the two hydrophobic amino acids C and V moves in the same direction (possibly into the core of the protein), and the hydrophilic amino acid T could then face the solvent [17].

Prediction		Rule	Confidence on RS126	Confidence on PSIPRED
α	1	**L**xxM	66.7	60.0
	2	**V**xAL	60.0	64.0
β	3	**D**VxLG	100	60.0
	4	**S**VxVG	100	93.8
	5	**W**VxIG	100	80.0
	6	TV**T**V	100	88.0
	7	TC**I**V	66.7	100
ζ	8	A**V**P	100	72.7
	9	M**x**P	72.2	70.4
	10	**D**xY	65.2	60.0

Table 3: Type I rules extracted in predicting PSS: confidences of amino acid patterns with the secondary structures (bold symbol indicates the site where the prediction of secondary structure is made; x denotes any one of the 20 amino acids)

Further, in rule 8 in Table 3, pattern A**V**P predicts a coil with 100% and 72.7% confidence, respectively. Amino acid Proline (P, Pro) invariably shows a high frequency of occurrence at neighbouring positions of all coil sites. Given the unique structural feature of amino acid P where its side-chain is bonded to the main-chain N atom, the conformation of the polypeptide backbone is often perturbed by the presence of amino acid P and, therefore, is induced to form coils in proteins [28]. The rule 10 in Table 3 shows that if Aspartic acid (D, Asp) is present at a site with Tyrosine (Y, Tyr) two sites downstream, then a coil is predicted with 65.2% and 60% confidence, respectively. The amino acid D in negatively charged R group (hydrophilic) and Y in aromatic R group (hydrophobic) tend to create coil, spanning over at least three adjacent residues [14], and making the likelihood of a presence of

the secondary structure stronger. Crasto and Feng found that amino acid D has a moderate preference for coil conformation and the coil propensities of amino acids Y and P have significant variations in coils of different sizes [28]. Also, charged amino acids D and K have lower frequencies of occurrence in the interior than in the surface coils.

Type II Rules

Table 4 lists type II rules or the amino acid patterns that enhance the prediction of a secondary structure by C4.5 if the presence of the secondary structure is already known for TS-SVM prediction. The prediction accuracy of *α*-helices by TS-SVM alone stands at 73.1% (see Table 2). The decision tree predicts an *α*-helix for patterns G**x**xY, Mx**x**S, **D**xxxxxxY, and PxN**x** if TS-SVM predicts the site to be an *α*-helix (rules 11-14). The confidence level of the decision tree prediction is 100% for RS126 and exceeds 88% for PSIPRED. The above rules can be given a different interpretation: when one of the four above amino acid patterns appear, then the surrounding patterns of amino acid make the confidence of prediction to be at least 88%.

For illustration, consider rule 16 in Table 4, which indicates that if hydrophilic amino acid Serine (S, Ser) is at one site upstream, Proline (P, Pro) is present at the local site, Aspartic acid (D, Asp) is at two sites downstream, and TS-SVM predicts the local site to be an *α*-helix, then the pattern S**P**xD is present with 89.3% and 83.3% confidence, respectively. For this pattern, hydrophilic amino acid S followed the hydrophobic amino acid P and another hydrophilic amino acid D at two sites downstream prove to be helix stabilizing if the amino acid P forms an *α*-helix. From position-specific amino acid preferences in *α*-helices [26], the N-cap position is dominated by amino acid S. This is because when amino acid S does occur in *α*-helix, its OH often forms a second H bond to a backbone CO on the previous helical turn. The preference distribution for amino acid P indicated that amino acid P in the first turn are almost exclusively in the N1 position (the first residue after the N-cap) [26]. This rule concurs with the findings of Richardson *et al.* that amino acid P prefers to be a helix-initiator than a helix-breaker [26]. Also, there is a peak of preference for hydrophilic amino acid D in positions N2 and N3 (the second and third residue after the N-cap).

Moreover, results in Table 4 indicate that the presence of the amino acids with the known secondary structure type at the local site improves the confidence of the secondary structure prediction.

Type III Rules

The patterns expressed by rules of type III listed in Table 5 make use of the secondary structures predicted by TS-SVM not only at the site but also at adjacent sites. Interestingly, the presence of a secondary structure at a particular location is associated here with the presence of the same structure in the vicinity.

For instance, rule 29 shows that if Cystine (C, Cys) is at four sites upstream, Arginine (Arg, R) is at four sites downstream, and an *α*-helix is predicted by TS-SVM at two sites upstream, at two sites downstream, and at the local site, then the secondary structure is predicted as an *α*-helix with 100% and 93.3% confidence. This result suggests that hydrophilic amino acids C at four sites upstream and R at four sites downstream prove to be helix-stabilizing if the amino acids at two sites upstream, two sites downstream, and the local site form an *α*-helix.

Prediction		Rule	Confidence on RS126	Confidence on PSIPRED
α	11	G**x**xY	100	90.4
	12	Mx**x**S	100	88.5
	13	**D**xxxxxxY	100	88.2
	14	PxN**x**	100	88.0
	15	D**x**N	91.7	85.6
	16	S**P**xD	89.3	83.3
β	17	Gxxxxx**K**	100	90.8
	18	**T**xxxxxR	100	91.0
	19	**I**xE	91.7	90.7
	20	**E**xY	89.3	88.0
	21	Hxx**x**N	86.1	84.8
	22	**x**xxxMxR	85.7	88.0
	23	**L**xxxxA	85.3	91.3
ζ	24	**G**P	93.7	89.1
	25	I**x**xM	81.1	84.0
	26	M**x**xY	80.0	78.7
	27	**x**xxxG	78.6	81.2
	28	**L**xxxxxC	75.0	78.1

Table 4: Type II rules generated by TS-SVM and C4.5 approach when the secondary structure is already predicted by TS-SVM.

Prediction		Rule	Confidence on RS126	Confidence PSIPRED
α	29	Cxx/*α*x**x**xx/*α*xR	100	93.3
β	30	Kxxx**V**x/*β*xx/*β*	94.7	96.3
	31	Q/*β***x**	84.9	89.1
ζ	32	Gx**L**x/*ζ*	71.1	78.1

Table 5: Type III rules generated by TS-SVM and C4.5. The secondary structure at the site as well as the structures /*α*, /*β*, /*ζ* have been predicted by TS-SVM.

As shown in Table 5, the pattern of rule 29 given the secondary structural type of *a*-helix at the local site indicates they stabilize helix. Also in rule 29 in Table 5, amino acid R in positively R charged group (hydrophilic) strongly tends to be *a*-helix and helices consist of at least four consecutive residues [14], and, therefore, the secondary structure of the target is a strong *a*-helix.

The rule 32 in Table 5 demonstrates that if hydrophobic amino acid Glycine (G, Gly) is two sites upstream, Leucine (L, Leu) is present at the local site, and TS-SVM predicts the local site and one site downstream to be coils, the pattern Gx**L**x/ζ is present with 71.1% and 78.1% confidence, respectively. For this pattern, hydrophobic amino acid L is known to have low coil propensities, however, these amino acids have high propensities at neighboring positions of G in coils [28]. Amino acid G is commonly found at neighboring positions of hydrophobic Isoleucine (I, ILE), Leucine (L, Leu) and hydrophilic coil residues.

4 Discussion

Utilizing the PSS predictions made by TS-SVM approach, we employed C4.5 decision trees to generate prediction rules. As manifested by the experiments, we were able to extract three types of joint prediction rules on RS126 and PSIPRED datasets. To generate a set of prevailing rules that can also be interpreted, we used empirically preset confidence threshold of 60%. The number of rules derived was relatively small and they showed higher confidence levels compared to those derived by other approaches. The rules were divided into three types based on whether the secondary structures predicted by TS-SVM were already included in the prediction rule. The number of structures in each rule in the datasets range from 11 to 813. The final overall prediction accuracies of TS-SVM+C4.5 were lower than of TS-SVM alone because of pruning of less confident rules in the decision tree. However, our overall accuracies were better than earlier methods combining binary/multi-class SVM with C4.5 [17]. The reason is that the contextual effects on structural formations on a secondary structure at a particular site by the neighbouring structures are taken into account for prediction by TS-SVM. As can be seen from Table 2, lower overall accuracies after incorporating C4.5 are mainly due to the low prediction accuracies of coil structures. Interestingly, by narrowing down the patterns responsible for PSS prediction, improvements of accuracies of *a*-helices and *β*-strands were seen which carry important structural formation. Coils account for remaining structural patterns of proteins.

Recently, the rules based on decision tree algorithms have been effectively extracted from a thermodynamic database of proteins and mutants to explore potential knowledge of protein stability prediction [29]. The performance of decision trees is better than the other methods for predicting changes of protein stability on a thermodynamic dataset consisting of 1615 mutants [30]. The results of the experiments for protein structure prediction show that the rules extracted by decision trees have meaningful biological interpretation and their comprehensibility is better than that of other methods [31].

Most rules extracted by the presented approach have significant and meaningful biological interpretation. As seen, the presence of specific amino acids improves the confidence of the secondary structure prediction. This could be interpreted as the confidences of the existence of a secondary structure pattern due to the presence of a particular amino acid pattern in the neighbourhood. After assignment of secondary structures, structural classes can be assigned to domains that provide an excellent source for further analysis. Structural classes divide proteins according to secondary structure content and organization. Amino acids and structural patterns based on type III rules can be extended for domains with predominantly *a*-helices or *β*-sheets.

The inspection of the prediction rules has offered interesting new insights into stabilizing *a*-helix, *β*-strand, and coil structures. Our results concur with the findings of Lyu *et al.* that amino acid L (Leu) tend to be helix stabilizing.[34] The preferences for amino acids T (Thr), R (Arg), and G (Gly) in *β*-strand prediction rules indicate their role in breaking the regular structure of the strands [28]. The rules of prediction of coils confirm that the most influential amino acids (the affecters) in coils are P (Pro) and G (Gly) [28]. The analysis of the prediction rules also shows that the neighbouring residues could have a profound effect on the preference of certain amino acids adopting *a*-helix, *β*-strand, and coil structures. Recently, the rule-extraction method of He *et al.* uses one-stage of binary SVM, which is unable to capture contextual relationships among the secondary structures. Therefore, the rules for β-strand prediction were less than 90% confidence on RS126 dataset [17].

Furthermore, these rules could be useful for guiding biological experiments aimed at satisfying the sequence conditions to produce a certain protein structure. There are several proteins that bind to membranes via a small amphipathic helix with one face made of hydrophobic residues [32]. In the study of Mad1 and mSin3A interaction, Eilers *et al.* determined that the hydrophobic face of amphipathic alpha-helical structure makes key contacts with mSin3A [33]. With our protein secondary structure prediction, the amphiphilicity of the predicted helix can be inferred by calculating the helix hydrophobic moment [34].

Acknowledgments

The authors wish to thank Dr. Jaume Torres for his help in finding biological relevance of the rules.

In addition, support of Nanyang Technological University and its School of Computer Engineering is acknowledged by the second author who served there as a Nanyang Professor from 2007-08.

References

[1] Garnier J, Gibrat JF, and Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 1996;266:541-553.

[2] Salamov AA, Solovyev VV. Prediction of protein secondary

structure by combining nearest-neighbor algorithms and multiple sequence alignments. Journal of Molecular Biology 1995;247:11-15.

[3] Schmidler SC, Liu JS, Brutlag DL. Bayesian segmentation of protein secondary structure. Journal of Computational Biology 2000;7:233-248.

[4] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology 1993;232:584-599.

[5] Riis SK, Krogh A. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignment. Journal of Computational Biology 1996;3:163-183.

[6] Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G, Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999;5:937-946.

[7] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 1999;292:195-202.

[8] Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;4: 508-519.

[9] Ouali M, King RD. Cascaded multiple classifiers for secondary structure prediction. Protein Science 1999;9: 1162-1176.

[10] Sen TZ, Cheng H, Kloczkowski A, Jernigan RL. A consensus data mining secondary structure prediction by combining GOR V and Fragment Database Mining. Protein Science 2006; 15 (11): 2499-2506.

[11] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research 1997; 25 (17): 3389-3402.

[12] Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Protein Science 2003;100(21):12105-12110.

[13] Kim H, Park H. Protein secondary structure prediction based on an improved support vector machines approach. Protein Engineering 2003;16:553-560.

[14] Nguyen MN, Rajapakse JC. Prediction of protein secondary structure with two-stage multi-class SVM approach. International Journal of Data Mining and Bioinformatics 2007; 1(3):248-269.

[15] Nguyen MN, Rajapakse JC. Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins: Structure, Function, and Bioinformatics 2005;59:30-37.

[16] Nguyen MN, Rajapakse JC. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. Proteins: Structure, Function, and Bioinformatics 2006; 63:542-550.

[17] He J, Hu H, Harrison R, Tai PC, Pan Y. Rule generation for protein secondary structure prediction with support vector machines and decision tree. IEEE Transactions on Nanobioscience 2006;5(1):46-53.

[18] Crammer K, Singer Y. On the Learnability and Design of Output Codes for Multiclass Problems. Machine Learning 2002;47: 201-233.

[19] Mitchell MT. Machine Learning, McGraw-Hill, New York; 1997.

[20] Quinlan JR. C4.5: Programs for Machine Learning, Morgan Kaufmann, SanFrancisco, CA; 1993.

[21] Hsu CW, Lin CJ. A comparison on methods for multi-class support vector machines. IEEE Transactions on Neural Networks 2002;13:415-425.

[22] Ma JM, Nguyen MN, Rajapakse JC. Gene Classification using codon usage and support vector machines. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2009;6(1):134-143.

[23] Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. 2nd Edition. Morgan Kaufmann, San Francisco; 2005.

[24] Padmanabhan S, Baldwin RL. Tests for helix-stabilizing interactions between various nonpolar side chains in alanine-based peptides. Protein Sci. 1994; 3:1992-1997.

[25] Lyu PC, Sherman JC, Chen A, Kallenbach NR. *a*-Helix stabilization by natural and unnatural amino acids with alkyl side chains. Proc. Nati. Acad. Sci. USA 1991;88:5317-5320.

[26] Richardson JS, Richardson DC, Amino acid preferences for specific locations at the ends of *a* helices. Science 1988;240(4859): 648-1652.

[27] Colloc'h N, Cohen FE. β-Breakers: An aperiodic secondary structure. Journal of Molecular Biology 1991; 221(2): 603-613.

[28] Crasto CJ, Feng JA. Sequence codes for extended conformation: A neighbor-dependent sequence analysis of loops in proteins. Proteins: Structure, Function, and Genetics 2001; 42(3):399-413.

[29] Huang LT, Gromiha MM, Ho SY. Sequence analysis and rule development of predicting protein stability change upon mutation using decision tree model. Journal of Molecular Modeling 2007;13(8):879–890.

[30] Huang LT, Gromiha MM, Hwang SF, Ho SY. Knowledge acquisition and development of accurate rules for predicting protein stability changes. Comput Biol Chem. 2006;30(6):408–415.

[31] He J, Hu HJ, Chen B, Tai PC, Harrison R, Pan Y.Rule extraction from SVM for protein structure prediction. Studies in Computational Intelligence 2008; 80: 227-252.

[32] Cornell RB, Taneva SG. Amphipathic helices as mediators of the membrane interaction of amphitropic proteins, and as modulators of bilayer physical properties. Curr. Protein Pept. Sci. 2006; 7:539–552.

[33] Eilers AL, Billin AN, Liu J, Ayer DE. A 13-amino acid amphipathic alpha-helix is required for the functional interaction between the transcriptional repressor Mad1 and mSin3A. J. Biol. Chem. 1999; 274, 32750–32756.

[34] Eisenberg D, Weiss RM, Terwilliger TC, The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature 1982; 299(5881): 371-374.