Extracting Decision Rules in Prediction of Protein Secondary Structure

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

***Abstract*— Information on secondary structures of amino acid residues in proteins provides valuable clues for the prediction of their 3-D structure and function. 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). Our approach has produced sizable 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. The proposed rules were derived and tested on the RS126 dataset of 126 nonhomologous globular proteins.**

I. INTRODUCTION

One of the major goals of bioinformatics is to predict three-dimensional (3-D) structure of a protein from its amino acid sequence. Information of a protein's structure provides valuable clues to the functions of a protein, vital for many aspects of living organism such as those of enzymes, hormones, and structural material, etc. It also helps in designing of new drugs for combating disease. Unfortunately, protein structure prediction problem is a combinatorial optimization problem, which so far has 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 the use of it to predict the 3-D structure [1]. 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; often as an α-helix (α), β-strand (β) or coil (ζ, the remaining type).

Many computational techniques have been proposed in the literature to solve the PSS prediction problem. The statistical methods are mostly based on the likelihood of each amino acid being one of three types of secondary structures [2], [3]. Neural networks use residues in a local neighborhood as inputs and find an arbitrary non-linear mapping [5-8]. The Bayesian approach provides a framework to account for non-local interactions among amino acid residues [4], 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 [9], [10]. Cuff and Barton employed a majority voting scheme to combine predictions from different techniques [9]. More complex approaches for combining different methods based on neural networks and linear discrimination [10] have also been studied. Support Vector Machines (SVM) have been applied to PSS prediction, in combination with several binary classifiers [11], [12].

The accuracy of the single stage approaches to PSS prediction is insufficient. Rost and Sander proposed the PHD approach using Multi-Layer Perceptrons (MLP) in cascade, with the second stage MLP improved the accuracy of the prediction by capturing the contextual relations among the secondary structures from the output of the first stage [5]. We proposed a two-stage SVM (TS-SVM) for the prediction of PSS [13], of relative solvent accessibility [14], and of accessible surface area of amino acids [15], which receives 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 find what patterns of amino acid lead to the prediction of PSS. Recently, He *et al.* proposed a rule-extraction method for PSS prediction by combining SVM and decision trees [16]. The method uses one-stage of binary SVM, which is unable to capture contextual relationships among the secondary structures and cannot assign directly a pattern of amino acid sequences to a class of protein secondary structure outputs with sufficient accuracy. To alleviate this shortcoming, we propose combining the PSS predictions from the two-stage SVM (TS-SVM) with C4.5 decision trees to extract useful rules for PSS prediction. This not only increases the accuracy of prediction by decision trees or SVM followed by decision trees, but it also renders

M. N. Nguyen is with the BioInformatics Institue and the BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore (e-mail: minhn@bii.a-star.edu.sg).

J. M. Zurada is with the BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore, and the Department of Electrical and Computer Engineering, the University of Louisville, Louisville, KY 40292 USA (e-mail:jacek.zurada@louisville.edu).

J. C. Rajapakse is with the BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore, Singapore-MIT Alliance, Singapore, and Biological Engineering Division, Massachusetts Institute of Technology, USA (e-mail: asjagath@ntu.edu.sg). (corresponding author; phone: +65 67905802; fax: +65 67926559; e-mail: asjagath@ntu.edu.sg).

a set of rules of PSS prediction, 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 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 rules for PSS prediction by C4.5 decision trees. We extracted two sets of rules for PSS prediction, based on whether the prediction is purely on amino acid patterns, or uses structural types of residues in the vicinity of predicted output. Furthermore, the rules extracted by our method were more confident and 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 [5], achieved previously by a rule extraction method.

II. Methods

A. Two-stage SVM

Let $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ denote the given amino acid sequence where $r_i \in \Omega_R$ and Ω_R is the set of 20 amino acid residues, and $\mathbf{t} = (t_1, t_2, \ldots t_n)$ denote 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_R to the space of Ω_T . Let $\mathbf{v}_i$ be the 21-dimensional feature vector representing the residue r_i where 20 units are the values from raw matrices of PSI-BLAST profiles ranging from [0, 1] and the remaining unit is used for padding to indicate an overlapping end of the sequence [8]. Let $\mathbf{r}_i = (\mathbf{v}_{i-h_1}, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_{i+h_1})$ be the input pattern 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 [13-15]. We use a multi-class SVM proposed by Crammer and Singer for both stages [17].

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 classes 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 [17]:

$$\max_{\alpha_j^k} - \frac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} \mathcal{K}_1(\mathbf{r}_j, \mathbf{r}_i) \sum_{k \in \Omega_T} \alpha_j^k \alpha_i^k - \sum_{j=1}^{N} \sum_{k \in \Omega_T} \alpha_j^k c_j^k$$
$$\text{such that } \sum_{k \in \Omega_T} \alpha_j^k = 0 \text{ and } \alpha_j^k \leq \begin{cases} 0 & \text{if } t_j \neq k \\ \gamma_1 & \text{if } t_j = k \end{cases} \qquad (2)$$

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

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

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 [5], [13].

The input to the second SVM at site i is obtained from a neighbourhood, $\mathbf{d}_i^1 = (d_{i-h_2}^{1k}, \ldots, d_i^{1k}, \ldots, d_{i+h_2}^{1k} : k \in \Omega_T)$ where $d_i^{1k} = 1/(1 + e^{-f_1^k(\mathbf{r}_i)})$ and h_2 is the size of the neighbourhood 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: $\mathcal{K}_2(\mathbf{d}_i^1, \mathbf{d}_j^1) = \phi_2(\mathbf{d}_i^1)\phi_2(\mathbf{d}_j^1)$. The outputs in the higher dimensional space are linearly combined by a weight vector $\mathbf{w}_2^k$ to obtain 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 [17]. The secondary structural type $\hat{t}_i$ at site i of input sequence is estimated by

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

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).

B. Decision Trees

SVMs perform well compared to other statistical or machine learning techniques in predicting protein features [15], [16] because of their generalization capabilities. Nevertheless, SVMs yield a black box model and provide no biologically meaningful prediction rules [16]. 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. In order to improve human comprehensibility, learned decision trees are re-represented as sets of if-then rules. We use C4.5 decision trees at the output of TS-SVM to generate rules for PSS prediction. C4.5 was chosen because it has shown to give 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 [16]. 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 [18]. 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 $\Gamma^2_{\mathrm{train}} = \{(\mathbf{a}_j, t_j) : \ j = 1, \ldots, N\}$ where the input at site j is $\mathbf{a}_j = (d^{2k}_{j-h_2}, \ldots, d^{2k}_{j+h_2}, \mathbf{v}_{j-h_1}, \ldots, \mathbf{v}_{j+h_1})$ and t_j is the desired secondary structure where $d^{2k}_j = 1/(1 + e^{-f^k_2(\mathbf{d}^1_j)})$. The rules are then tested with the same data set for evaluation of the performance of the algorithm.

III. Experiments and Results

The present approach was implemented using position-specific scoring matrices generated by PSI-BLAST as inputs and tested on benchmark datasets with seven-fold cross-validation. The results were compared with other prediction methods and with other results extracting amino acid patterns leading to the prediction.

A. Dataset

The set of 126 nonhomologous globular protein chains, used in the experiment of Rost and Sander [5] and referred to as the RS126 set, was used to evaluate the accuracy of the predictors. The dataset contained 23349 residues with 32% α-helix, 23% β-strand, and 45% coil. Many current generation secondary structure prediction methods have been developed and tested on this dataset. The RS126 set is available at *http://www.compbio.dundee.ac.uk/∼www-jpred/data/pred_res/126_set.html.*

B. Implementation

The multi-class SVM method was implemented using BSVM library which is known to show fast convergence for large optimization problems [20]. The Gaussian kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = e^{-\sigma\|\mathbf{x}-\mathbf{y}\|^2}$ showed superior performance over the linear and polynomial kernels for predicting protein secondary structure [13], relative solvent accessibility [14], accessible surface areas of amino acids [15], and gene classification [19]. The sensitivity parameter γ and the Gaussian kernel parameter σ were determined by using the grid-search method [20]. Grid-search provides useful parameter estimates for multi-class SVM in a relatively short time. The parameters of the Gaussian kernel and TS-MSVM, as $\sigma_1 = 0.0625, \sigma_2 = 0.0156$ and $\gamma_1 = \gamma_2 = 0.5$, and the neighborhood window $h_1 = 7$, and $h_2 = 3$ were experimentally determined for optimal performance. We implemented the decision tree C4.5 by using Weka software [21]. 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.

TABLE I

Generated rules and accuracies of different type for rule-based classifiers on RS126 dataset.

Validation	C4.5		SVM + C4.5		TS-SVM + C4.5	
Run	Accuracy	Rules	Accuracy	Rules	Accuracy	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

C. Prediction Accuracies

We used Q_3 accuracy to measure the percentages of correctly predicted residues of three types of secondary structures [9]:

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

where η_t is the number of correctly predicted residues and ν_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 *occurrence* of an amino acid pattern is the frequency of presence the amino acid pattern in the training dataset.

The performance of secondary structure prediction on the RS126 dataset of 126 proteins using TS-MSVM and C4.5 is shown in Table I. 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%). As shown in Table I, the combination of TS-SVM and C4.5 decision trees tends to generate fewer rules but also yields higher accuracy of prediction even with a smaller number of rules.

Table II shows an improvement of 2.5% in prediction accuracy of our approach compared to the method of He

TABLE II

Comparison of performances of combining TS-SVM with C4.5 with other methods for PSS prediction on RS126 dataset.

Method	C_α	C_β	C_ζ	**Accuracy**(%)
Binary SVM + C4.5 [16]	72.8	79.6	69.3	∼72.5
SVM + C4.5	76.3	67.7	74.2	73.7
TS-SVM + C4.5	77.9	69.3	75.3	75.0

TABLE III

Type I rules extracted in PSS prediction.

Prediction	Rule		Occurrence	Confidence
α	1	**L**xxM	43.4	66.7
	2	**V**xAL	39.1	60.0
β	3	**D**VxLG	34.2	100
	4	**S**VxVG	39.4	100
	5	**W**VxIG	43.1	100
	6	RxV**x**I	32.7	100
	7	TV**T**V	44.6	100
	8	TC**I**V	45.1	66.7
ζ	9	A**V**P	49.2	100
	10	Kxxxx**C**xxxxxxL	44.1	78.4
	11	M**x**P	55.8	72.2
	12	**D**xY	50.1	65.2

et al. produced on RS126 by combining single-stage binary SVM with C4.5. Futhermore, on the RS126 set, accuracy Q_3 after combining TS-MSVM with C4.5 approach on the PSI-BLAST profiles was significantly higher than the results produced by NNSSP (72.7%) [3], PREDATOR (70.3%) [22], DSC (71.1%) [23], the refined neural network (71.3%) [6], Jpred (74.8%) [9], PHD (70.8%) [5], and binary SVM (71.2%) [11].

D. Extracted Rules

Logical rules from amino acid sequences were decoded using the SVM-predicted output values [16]. We classified the rules into two categories, types I and II, based on whether TS-SVM already predicted the specific secondary structure. The rules are shown in Tables III and IV. 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 occurrences of such regularities and the confidence of the rules are given in second and third 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 [2]. 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 provided 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 rules derived by bringing evidences from the literature.

1) Type I Rules: Type I rules extracted by the presented method are shown in Table III. Listed are the rules with confidence above 60%, indicating amino acid patterns leading 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, 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 seen from Table III, patterns **D**VxLG, **S**VxVG, **W**VxIG, RxV**x**I, and TV**T**V, predict β-strands with 100% confidence. 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 [13]. Moreover, in rules 7 and 8 in Table III, 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 [16].

As seen in rule 9 in Table III, pattern A**V**P predicts a coil with 100% confidence. Amino acid Proline (P, Pro) invariably shows a high frequency of occurrence at neighboring 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 12 in Table III 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 with 65.2% confidence. 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 [13], 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.

2) Type II Rules: Table IV 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 by TS-SVM prediction. The decision tree predicts an α-helix with 100% confidence for pattens G**x**xY, Mx**x**S, Gx**x**P, **D**xxxxxxY, and PxN**x** if TS-SVM predicts the site to be an α-helix. The accuracy of prediction of α-helices by TS-SVM stands at 73.1%. Therefore, the above rule can be given a different interpretation: when the above amino acid patterns appear, then the surrounding patterns of amino acid makes the confidence of prediction to be 100%.

For illustration, consider rule 21 in Table IV, 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% confidence. In 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 IV 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.

TABLE IV
TYPE II RULES EXTRACTED IN PSS PREDICTION.

Prediction	Rule		Occurrence	Confidence
α	13	G**x**xY	45.9	100
	14	Mx**x**S	51.7	100
	15	Gx**x**P	39.2	100
	16	**D**xxxxxxY	49.7	100
	17	PxN**x**	48.2	100
	18	Kx**G**xxI	45.7	95.9
	19	D**P**	47.3	93.3
	20	D**x**N	51.1	91.7
	21	S**P**xD	44.3	89.3
	22	Sx**x**xK	51.0	83.3
	23	N**x**xxxP	45.7	77.8
β	24	Gxxxxx**K**	43.9	100
	25	**T**xxxxxR	44.7	100
	26	**x**xPxxxxR	42.8	100
	27	**x**GN	42.1	100
	28	**G**xxxF	41.0	100
	29	AxxMxx**x**	42.2	100
	30	AxxMxx**x**G	47.1	93.3
	31	**I**xE	46.8	91.7
	32	**E**xY	42.0	89.3
	33	Hxx**x**N	54.0	86.1
	34	**x**xxxMxR	41.6	85.7
	35	**L**xxxxA	44.0	85.3
	36	**x**xxxxxC	54.5	83.5
	37	AxxxxY**x**	45.2	83.3
	38	**x**xxxM	52.6	82.9
ζ	39	**G**P	76.6	93.7
	40	**x**SV	58.1	84.7
	41	**x**xxT	67.1	84.3
	42	**S**xI	58.9	83.3
	43	xRxxxx**x**xI	54.6	82.4
	44	**x**xD	68.1	82.3
	45	Gxxxxx**x**xxxG	54.1	81.7
	46	I**x**xM	56.9	81.1
	47	M**x**xY	59.8	80.0
	48	**x**xxxG	60.9	78.6
	49	**L**xxxxxC	55.8	75.0

IV. DISCUSSION

Following the predictions made by TS-SVM approach, we used C4.5 decision trees to generate prediction rules for PSS prediction. As manifested by experiments, we were able to extract two types of rules, which previous literature and physiochemical properties of amino acids seem to support. The number of rules derived was relatively small and they showed higher confidence levels compared to those derived by other approaches. To generate a set of prevailing rules that can also be interpreted, we used empirically preset confidence threshold of 60%. The rules were divided into two types based on whether the secondary structures predicted by TS-SVM were already included in the prediction rule.

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. The inspection of the prediction rules has offered interesting new insights into stabilizing α-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 [25]. 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 [27]. 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 α-helix, β-strand, and coil structures. These rules could be useful for guiding biological experiments aimed at satisfying the sequence conditions to produce a certain protein structure. Furthermore, we will apply our method to a much larger dataset for investigations that link the prediction model with the underlying biology.

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 in 2007-08.

References

[1] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2001.

[2] J. Garnier, J. F. Gibrat, and B. Robson, “GOR method for predicting protein secondary structure from amino acid sequence,” *Methods Enzymol*, vol 266, pp 541–553, 1996.

[3] A. A. Salamov and V. V. Solovyev, “Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments,” *Journal of Molecular Biology*, vol 247, pp 11–15, 1995.

[4] S. C. Schmidler, J. S. Liu, and D. L. Brutlag, “Bayesian segmentation of protein secondary structure,” *Journal of Computational Biology*, vol 7, pp 233–248, 2000.

[5] B. Rost and C. Sander, “Prediction of protein secondary structure at better than 70% accuracy,” *Journal of Molecular Biology*, vol 232, pp 584–599, 1993.

[6] S. K. Riis and A. Krogh, “Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignment,” *Journal of Computational Biology*, vol 3, pp 163–183, 1996.

[7] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, “Exploiting the past and the future in protein secondary structure prediction,” *Bioinformatics*, vol 15, pp 937–946, 1999.

[8] D. T. Jones, “Protein secondary structure prediction based on position-specific scoring matrices,” *Journal of Molecular Biology*, vol 292, pp 195–202, 1999.

[9] J. A. Cuff and G. J. Barton, “Evaluation and improvement of multiple sequence methods for protein secondary structure prediction,” *Proteins*, vol 4, pp 508–519, 1999.

[10] M. Ouali and R. D. King, “Cascaded multiple classifiers for secondary structure prediction,” *Protein Science*, vol 9, pp 1162–1176, 1999.

[11] S. Hua and Z. Sun, “A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach,” *Journal of Molecular Biology*, vol 308, pp 397–407, 2001.

[12] H. Hu, Y. Pan, R. Harrison, and P. C. Tai, “Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier,” *IEEE Transactions on Nanobioscience*, vol. 3, no. 4, pp. 265–271, 2004.

[13] M. N. Nguyen and J. C. Rajapakse, “Prediction of protein secondary structure with two-stage multi-class SVM approach,” *International Journal of Data Mining and Bioinformatics*, vol. 1, no 3, pp. 248–269, 2007.

[14] M. N. Nguyen and J. C. Rajapakse, “Prediction of protein relative solvent accessibility with a two-stage SVM approach,” *Proteins: Structure, Function, and Bioinformatics*, vol 59, pp. 30–37, 2005.

[15] M. N. Nguyen and J. C. Rajapakse, “Two-stage support vector regression approach for predicting accessible surface areas of amino acids,” *Proteins: Structure, Function, and Bioinformatics*, vol 63, pp. 542–550, 2006.

[16] J. He, H. Hu, R. Harrison, P. C. Tai, and Y. Pan, “Rule generation for protein secondary structure prediction with support vector machines and decision tree,” *IEEE Transactions on Nanobioscience*, vol. 5, no 1, pp. 46–53, 2006.

[17] K. Crammer and Y. Singer, “On the Learnability and Design of Output Codes for Multiclass Problems,” *Machine Learning*, vol 47, pp 201–233, 2002.

[18] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, SanFrancisco, CA, 1993.

[19] J. M. Ma, M. N. Nguyen, and J. C. Rajapakse, “Gene Classification using codon usage and support vector machines,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2007 (on-line and in press).

[20] C. W. Hsu and C. J. Lin, “A comparison on methods for multi-class support vector machines,” *IEEE Transactions on Neural Networks*, vol 13, pp 415–425, 2002.

[21] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques. 2nd Edition*, Morgan Kaufmann, San Francisco, 2005.

[22] D. Frishman and P. Argos, “Knowledge-based secondary structure assignment,” *Proteins: Structure, Function, and Genetics*, vol 23, pp 566–579, 1995.

[23] R. D. King and M. J.E. Sternberg, “Identification and application of the concepts important for accurate and reliable protein secondary structure prediction,” *Protein Science*, vol 5, pp 2298–2310, 1996.

[24] S. Padmanabhan and R. L. Baldwin, “Tests for helix-stabilizing interactions between various nonpolar side chains in alanine-based peptides,” *Protein Sci.*, vol 3, pp 1992–1997, 1994.

[25] P. C. Lyu, J. C. Sherman, A. Chen, and N. R. Kallenbach, “α-Helix stabilization by natural and unnatural amino acids with alkyl side chains,” *Proc. Nati. Acad. Sci. USA*, vol 88, pp 5317–5320, 1991.

[26] J. S. Richardson and D. C. Richardson, “Amino acid preferences for specific locations at the ends of α helices,” *Science*, vol 240 (4859), pp 1648–1652, 1988.

[27] N. Colloc’h and F. E. Cohen, “β-Breakers: An aperiodic secondary structure,” *Journal of Molecular Biology*, vol 221 (2), pp 603–613, 1991.

[28] C. J. Crasto and J. A. Feng, “Sequence codes for extended conformation: A neighbor-dependent sequence analysis of loops in proteins,” *Proteins: Structure, Function, and Genetics*, vol 42 (3), pp 399–413, 2001.