

# A Novel Approach of Protein Secondary Structure Prediction by SVM Using PSSM Combined by Sequence Features

Yehong Chen<sup>1</sup>(✉), Jinyong Cheng<sup>2</sup>, Yihui Liu<sup>2</sup>, and Pil Seong Park<sup>3</sup>

<sup>1</sup> School of Graphic Communication and Packaging,  
Qilu University of Technology, Jinan, China  
chenyh@qlu.edu.cn

<sup>2</sup> School of Information, Qilu University of Technology, Jinan, China

<sup>3</sup> Department of Computer Science, University of Suwon, Suwon, Korea

**Abstract.** Knowledge of protein secondary structure is a useful step toward prediction of the 3D structure of a particular protein. In this paper, a support vector machine (SVM) based method used for the prediction of secondary structure is introduced in details. Protein sequence data is in a hybrid representation combining the Position-specific Scoring Matrix (PSSM), the Hydrophobicity Sequence Feature (HSF), and the Structural Sequence Feature (SSF). Protein sequences are obtained from CB513 dataset, corresponding PSSM profiles are obtained from PSI-BLAST Program and sequence features are computed based on amino acid scales offered by ExPASy website (<http://web.expasy.org/protscale/>). Basically, PSSM profiles are used as input data to the SVM-PSSM classifier of the secondary structure prediction. Furthermore, to construct more accurate classifiers, more than 40 SFs (sequence features) are examined as accessional input vector to SVM-PSSM classifier for feature selection. The most accurate classifier in this study is constructed using a combination of PSSM and few relevant sequence features. The experimental results show that relevant sequence features extracted from Hydrophobicity index and Structural conformational parameters can improve the SVM-PSSM classifier for the prediction of protein secondary structure elements. Our proposed final SVM-PSSM-SF method achieved an overall accuracy of 78%.

**Keywords:** Protein secondary structure prediction · SVM · Position specific scoring matrices · Sequence feature · Amino acid scale · ProtScale

## 1 Introduction

The study of protein structure and its function is one of the most important questions of molecular biology. Functionalities of proteins have been commonly believed to be determined by their unique 3-dimensional structures [1]. In 1973, Christian B. Anfinsen and his colleagues performed the definitive experiment showing that amino acid sequence determines protein shape [2]. It means that the primary structure of proteins determine their secondary structure. Furthermore, the secondary structure is a useful first step toward 3D structure prediction of a particular protein. Due to the gap between

the number of known protein sequences and the number of known protein structures is widening rapidly, therefore various significant efforts have been devoted to find computational methods to predict secondary structure of proteins automatically based on the protein sequences [3]. Until now, numerous efficient methods have been proposed, such as method based on probabilistic models (HMM) [3], dynamic Bayesian networks (DBN) [4], or machine learning-based methods such as, neural networks (NN) or support vector machines (SVM) [5, 6]. The Support Vector Machine algorithm has achieved high accuracy, however challenges remains.

Using position-specific scoring matrices (PSSM) [7] which encodes evolutionary information as the profiles of the protein sequences was proven to be the most helpful for building prediction model by SVM. Furthermore, the secondary structure prediction quality should be improved if the PseAA (pseudo amino acid) is combined with the evolutionary information as the hybrid representation of the primary structure profiles [8, 9]. Some artificial neural network and support vector machine are trained and tested using both the physicochemical properties and PSSM matrices generated from PSI-BLAST, and tests show that the PSSM+SF(sequence feature) model has made a significant improvement in the accuracy compared to other pure PSSM representation for SVM methods [5].

Our previous work proposed a protein secondary structure prediction method based on the support vector machine (SVM) with position-specific scoring matrix (PSSM) profiles. In this paper, PSSM which reflects evolutionary information is combined with sequence features, based on amino acid scale reference ProtScale as representation of protein sequence is used to predict three type of secondary structural elements for low-similarity sequences.

Firstly, we utilize PSSM matrix represent protein sequences. Secondly, the special amino acid scales are got from <http://web.expasy.org/protscale/>, including the hydrophobicity and 3 kind of secondary structure conformational parameters. Some amino acid scale based PseAA (pseudo amino acid) are formulated as extracted SFs (sequence features) for further selection. After feature selection, the most relevant SFs are selected to combine with PSSM for composing the hybrid representation of primary structure. Finally, the SVM-PSSM-FS classifier (in addition to the selected few SFs) based on the concept of a sliding window along the protein sequence is used to predict the secondary structure states. The experimental results show that our proposed SVM-PSSM-FS classifier improves the prediction accuracy of SVM-PSSM classifier build in our previous work.

## 2 Data and Method

### 2.1 Dataset

The dataset used in this study was derived from the CB513 database [10]. CB513 includes 513 protein sequences with similarity less than 25%. In order to remove the uncertainty in this paper, Proteins with length less than 30 or with uncertainty components B Or X from DSSP are not included. Or cut off a couple of amino acid involved components B or X. Then there are 493 proteins from CB513 are used in this

paper. All these 493 protein data from CB513 is encoded in PSSM matrix by PSI-BLAST as the protein sequence representation. The concept of slid window is used for pilling up these PSSM as protein sequence data [9], and the window size used here is 13. For example, 30 residues cut off 18 protein subsequence instance. We also add 6 zeros in the head and the tail of a protein so that the first and last 6 residues can be presented in the same way. Then the total number of protein samples (protein subsequence of 13) is 82309. Take into consider the time consuming and convenience, we use a sub-sampling test method to evaluate classifier performance. In this study, the partition of the training data and test data is 70000 to 12309, and all samples were randomly indexed before division for every running. Hence, the mean results from a couple of running can present the general classifier performance intuitively.

## 2.2 DSSP

Protein secondary structure can be assigned from experimentally determined tertiary structures by algorithms such as DSSP [10]. DSSP file of the proteins have eight classes: H ( $\alpha$ -helix), G (310-helix), I ( $\pi$ -helix), E ( $\beta$ -strand), B ( $\beta$ -bridge), T (turn), S (bend) and C (rest random coil). However, many computational approaches have been developed in the past decades to predict the 3-state secondary structure from protein sequences. There are different 8–3-state reduction schemes were adopted. Our work use DSSP files, and adopt H, E and C denotes  $\alpha$ -helix,  $\beta$ -strand and all other elements include coil. This strategy usually results in lower prediction accuracy than other definitions [11].

## 2.3 PSSM

Homology modeling method by homologous sequence analysis and pattern matching is proved the most reliable method to predict protein spatial structure unit or structure domain. The theoretical basis is that the most reliable way to predict protein secondary structure is by homology to a protein of known structure. It is due to the observation that the sequence alignment of homologous proteins according to their structural alignment and aligned residues usually have similar secondary structure [12].

BLAST is the abbreviation of basic local alignment searching tool. Its function is to compare the amino acid sequences of different biological protein in the corresponding database, looking for the same or similar sequences sequence as similarity search. PSSM (position-specifics scoring matrices) is built by the result of BLAST as a matrix. PSI-BLAST is through multiple iterations to find the best results. Using the first search results to build PSSM, and then this PSSM is used in the second search, the second search results also used for the third search again, and so on, until find out the best search results. All current high-performance methods, make use of the iterative databank-searching tool position specific iterative BLAST (PSI-BLAST) (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) select homologous sequences in the form of PSSM for predicting the secondary structure.

PSSM introduce evolution information of protein for prediction model learning. We use each protein sequence as a seed to search and align homogenous sequences from NCBI's NR database (<ftp://ncbi.nih.gov/blast/db/nr>) using the PSI-BLAST program with three iterations and a cutoff E-value 0.001. In this study, BLOSUM62 Substitution Matrix is adopted measures as score matrix to reflect the similarity between the amino acid. Finally got PSSM profile of a protein sequence is a  $L * 20$  matrix,  $L$  is the length of the protein instance.

## 2.4 Sequence Features

More than forty sequence features are used to code each amino acid residue in a data instance. The amino acid scales are obtained from Protscale [13] (<http://expasy.org/tools/protscale.html>). ProtScale also use slid window for SF (sequence feature) formulation. The linear combination of residue scales within the window was assigned to the center residue as its sequence feature values, so long as they can reflect some sorts of sequence-order effects [9]. These amino acid scales used in this paper fall into the following two classes:

1. Hydrophobicity index (H) which is important for amino acid side chain packing and protein folding [14]. Sequence feature  $H_i^j$  is for the  $i$ th amino acid residue in the sequence formulated as the mean of H scale among  $j$  consecutive residues.

$$H_i^j = \frac{1}{j} \sum_{w=i-j+1}^{i+j-1} H(a_w) \quad L \geq w \geq 1 \quad (1)$$

$L$  is the length of protein sequence

$i$  is positive integer,  $1 \leq i \leq L$

$j$  is positive integer, and  $1 \leq j \leq \min(i, 11)$

This feature extraction formulation (1) can be explanted that a region of Hydrophobic interactions make non-polar side chains to pack together inside proteins [8] which usually related to folding or sheet pattern. So the mean of H among consecutive residues will be more robust for feature extraction.

2. *Structural features*: including the conformational parameters for alpha-helix (A), beta- sheet (B) and coil (C); this 3 features assign an amino acid by a different tendency to form one of the three types of secondary structures. In this study, the conformational parameters reported by Deléage and Roux [15] were used for features A, B and C. These three conformational parameters bonded together as one sequence feature S.

$$S_i^j = \prod_{w=i-j+1}^{i+j-1} S(a_w) \quad L \geq w \geq 1$$

S bond 3 amino acid scales: A, B, T and C

L is the length of protein sequence

i is positive integer,  $1 \leq i \leq L$

j is positive integer, and  $1 \leq j \leq \min(i, 11)$

(2)

This feature extraction formulation (2) for S is as the same as (1) for H, except formulate multiplier instead of mean of the consecutive residues scale. Our consideration of multiplier is due to the probability meaning of these three amino acid scales, hence multiplier helps to extract more frequent relevant pattern for sequence features, and the robustness of this formulation is proved by experiment results.

All above amino acid scale based SFs reflect some sorts of sequence-order effects in the chain. We got  $H^{1-11}$ ,  $H^{19}$  and  $S^{1-10}$ , totally 41 sequence features.

## 2.5 SVM (Support Vector Machine)

The machine learning problem can be specified as follows: given the amino acid sequence of a protein and the definition of amino acid scale, the task is to predict the secondary structure states corresponding to residues in the protein. The Support vector machine (SVM) was first proposed by Vapnik [16], its main idea is to create a hyperplane as decision surface so that the gap isolation between the two class examples is maximized. Given  $x$  is in the input vector space the dimension of which is  $m_0$ .  $\{\varphi_j(x)\}_{j=1}^{m_1}$  represents a nonlinear transformation from input space to a feature space of  $m_1$  dimension. For every  $j$ ,  $\varphi_j(x)$  was defined according to the prior knowledge worked as feature extraction function. For non-linear classifiers that are generally applicable to biological problems, a kernel function can be used to measure the distance between data points in a higher dimensional space.

$$K(x, x_i) = \varphi^T(x) \varphi(x_i) = \sum_{j=1}^{m_1} \varphi_j(x) \varphi_j(x_i), \quad i = 1, 2, 3, \dots, N$$
(3)

Here,  $x_i$  is support vectors and  $x$  is data instance. This allows the SVM algorithm to fit the maximum-margin hyperplane in the transformed space, in another words, based on kernel function, nonlinear inseparable instance vectors being transformed from  $m_1$  into  $N$  dimensional separable vector space. Thus, how to choose the kernel function is an related issue. The publicly available LIBSVM software was used to process the SVM regression [17] in this paper. There are four kinds of kernels are commonly used in SVM: “linear”, “polynomial”, “sigmoidal tanh” and “radial basis”. This paper used the radial basis function (RBF) as the kernel, it is the same as Gaussian kernel function.

$$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2); \quad (4)$$

where  $\gamma$  is a training parameter. A smaller  $\gamma$  value makes decision boundary smoother. The regularization factor  $C$  is another parameter for SVM training, controls the tradeoff between low training error and large margin. For SVM-PSSM classifier, we get the most advantage parameters by genetic algorithm ( $C = 1, \gamma = 0.065$ ), which remain unchanged in all our experiments in this paper for good.

The Slid window strategy was used to cut every protein instance into overlapped subsequences. SVM classifier use the concept of sliding window strategy for the combination of PSSM and SF as input representation, and output the class labels which present the secondary structure states of the middle residue. Maintaining the Integrity of the Specifications.

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

### 3 The Experimental Results

The generated PSSM matrix of CB513 within the search scope of nr dataset is used as input, and the protein secondary structure three states are H [1 0 0], E [0, 0, 1] and C [0, 0, 1]. Sliding window length is set to 13, so before and after 6 amino acids will be taken into account, so totally 260 dimension are for PSSM. All SFs (sequence features) are assigned to the residue positioned in the middle of the subsequence of  $j$  size, and the other  $(j - 1)$  neighbor residues provided context information for the sequence feature by (1) and (2).  $j$  can be more than window size, so that sequence feature can encode long distant influence beyond the limitation of sliding window. In our experiments, we got  $H^{1-11}$ , and  $S^{1-10}$ . All Protein samples was a subsequence of 13 consecutive residues, so 260 dimension are for PSSM,  $13 * 11$  for HSF, and  $13 * 30$  for SSF for every sample representation.

#### 3.1 Evaluation Measure

Various measures can be used to asses secondary structure prediction method, the most common one being  $Q_3$  [18], which defines accuracy as the percentage of correctly identified states, as in (5).

$$Q_3 = \frac{\sum_{i \in \{H,E,C\}} \# \text{ of residues correctly predicted}_i}{\sum_{i \in \{H,E,C\}} \# \text{ of residues in class } i} \times 100 \quad (5)$$

In this paper, we also use  $Q_3$  to evaluate our proposed SVM classifier and feature selection. In statistical prediction, three methods that are independent dataset test, sub-sampling test, and jackknife test often can be used to examine a predictor for its effectiveness in practical application [8]. Take into consider the time consuming and convenience, we use a simulated eight fold cross-validation method to evaluate classifier performance. Every evaluation result is the mean of  $Q_3$  value over eight times running under fivefold partition, actually 7 in 8 randomly picked up as training data, and all remaining as test samples.

### 3.2 Feature Selection

From experiments, combining all the 41 sequence features for input encoding surprisingly drop the accuracy of SVM-PSSM classifier. The possible explanation is that all the 41 features contain redundant or correlated information, which may cause classifier performance degradation. Thus, further feature selection is needed, and additional sequence feature selection by our proposed method was chosen using the  $Q_3$  achieved in prediction as criteria.

Then the input vector include 260 PSSM profile with the addition of one or more biological sequence features. Every HSF is 13 dimension, and SSF is a bond of three structural features with 39 dimension. Feature selection among 11 HSF and 10 SSF were examined to optimize SVM-PSSM classifier performance. Let  $H^n$ ,  $S^n$  note the index of sequence features that model respectively Hydrophobicity and the bond of structural feature ( $\alpha$ -helices,  $\beta$ -strands and coils).  $n$  is related to the context size for features, for example,  $H^{10}$  means the HSF is got from 21 consecutive residues, so it response 21 distant hydrophobicity characteristics. At the beginning, we choice HSF one by one as the addition to PSSM for SVM-PSSM classifier, all experimental results are in following Table 1.

**Table 1.** Experimental results from SVM-PSSM classifier in addition with HSF

SVM_PSSM classifier in addition with	$Q_3$ (%)	Feature description
PSSM	77.19	260 PSSM profiles
$H^1$	77.35	Hydrophobicity of single residue
$H^2$	77.60	Mean of 3 consecutive residues
$H^3$	77.69	Mean of 5 consecutive residues
$H^4$	77.73	Mean of 7 consecutive residues
<b><math>H^5</math></b>	<b>77.800</b>	Mean of 9 consecutive residues
<b><math>H^6</math></b>	<b>77.80</b>	Mean of 11 consecutive residues
<b><math>H^7</math></b>	<b>77.82</b>	Mean of 13 consecutive residues
<b><math>H^8</math></b>	<b>77.81</b>	Mean of 15 consecutive residues
$H^9$	77.74	Mean of 17 consecutive residues
$H^{10}$	77.76	Mean of 19 consecutive residues
$H^{11}$	77.72	Mean of 21 consecutive residues

Well-known Hydrophobicity is closely related to the secondary structure. A region of Hydrophobic interactions make non-polar side chains to pack together inside proteins which usually is relevant to folding or sheet pattern. However our experiments find that only use  $H^1$  as an addition to PSSM, the resulting classifier was not as accurate as the SVMs trained only with PSSM features. Table 1 data also show that the mean Hydrophobicity sequence features are more related to the secondary structure. It can be explained that the mean of H among consecutive residues will result in more robust for feature extraction, and from  $H^1$  to  $H^8$ , it show increase trend of SVM\_PSSM classifier accuracy. From our experiments, one bond HSF( $H^{5:8}$ ) also recorded in Table 1. Among all the selected features,  $H^7$  seems to be most effective in improving the accuracy of SVM-PSSM classifier.

Then we do experiments on all SSF features. Every S responding the conformational parameters for alpha-helix (A), beta-sheet (B) and coil (C), for example,  $S^4$  is for  $A^4$ ,  $B^4$  and  $C^4$  combining together. We find S as the bond of three FSF with the same consecutive residues can improve the accuracy from SVM\_PSSM classifier.  $S^n$  from 3 to 5 are found the most related features to secondary structure, it means special sub-sequence of 5 to 9 in some pattern appears more frequently related to the secondary structure. The following Table 2 is the records from our experiments. All the above experiments (Table 2) show that Structural features got more improvement in accuracy of SVM\_PSSM classifier.

**Table 2.** The experiment results from SVM-PSSM classifier in addition with conformational parameters of secondary structures one at a time

SVM-PSSM classifier in addition with	$Q_3$	Feature description
PSSM	77.19	260 PSSM profiles
$S^1$	77.79	A+B+C for single residue
$S^2$	77.89	Mean of 3 consecutive residues
$S^3$	<b>77.95</b>	Mean of 5 consecutive residues
$S^4$	<b>77.96</b>	Mean of 7 consecutive residues
$S^5$	<b>78.04</b>	Mean of 9 consecutive residues
$S^6$	<b>77.90</b>	Mean of 11 consecutive residues
$S^7$	77.84	Mean of 13 consecutive residues
$S^8$	77.721	Mean of 15 consecutive residues
$S^9$	77.66	Mean of 17 consecutive residues
$S^{10}$	77.53	Mean of 19 consecutive residues

In the next experiments, we combine HSF and SSF together for further feature selection. From experiments, long-term HSF combine with middle-term SSF show effective to improve the accuracy of SVM-PSSM classifier. Rectangular frame with red border in the Table 3 show these region involved effective features. From Table 3, the selected best combination of HSF and SSF( $H^{11} + S^5$ ), as the addition with PSSM for SVM classifier, gets less accuracy than the best selected SSF( $S^5$ ) from Table 2 dose. On the other hand,  $H^{11} + S^5$  is more effective than the best selected HSF ( $H^7$ ) in Table 1 as addition with PSSM from SVM classifiers. All in all below feature subset should be

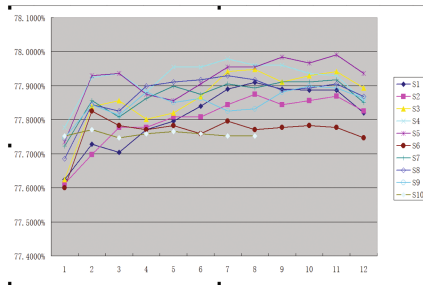


**Table 3.** The experiment results from SVM-PSSM classifier in addition with combination of HSF and SSF

SVM-PSSM classifier combine H with S	Q3 (%)							
	S1	S2	S3	S4	S5	S6	S7	S8
H <sup>1</sup>	77.62	77.61	77.62	77.78	77.73	77.60	77.72	77.68
H <sup>2</sup>	77.73	77.70	77.84	<b>77.92</b>	<b>77.93</b>	77.82	77.86	77.84
H <sup>3</sup>	77.70	77.78	77.86	<b>77.94</b>	<b>77.94</b>	77.78	77.81	77.82
H <sup>4</sup>	77.77	77.77	77.80	77.88	77.87	77.77	77.86	<b>77.90</b>
H <sup>5</sup>	77.79	77.80	77.81	<b>77.95</b>	77.85	77.78	<b>77.90</b>	<b>77.91</b>
H <sup>6</sup>	77.84	77.81	77.87	<b>77.95</b>	<b>77.90</b>	77.76	77.87	<b>77.92</b>
H <sup>7</sup>	77.89	77.84	<b>77.94</b>	<b>77.98</b>	<b>77.95</b>	77.79	<b>77.91</b>	<b>77.93</b>
H <sup>8</sup>	<b>77.91</b>	77.87	<b>77.95</b>	<b>77.96</b>	<b>77.95</b>	77.77	77.89	<b>77.92</b>
H <sup>9</sup>	77.89	77.84	<b>77.91</b>	<b>77.96</b>	<b>77.98</b>	77.77	<b>77.91</b>	77.88
H <sup>10</sup>	77.88	77.85	<b>77.93</b>	<b>77.93</b>	<b>77.97</b>	77.78	<b>77.91</b>	77.89
H <sup>11</sup>	77.89	77.87	<b>77.94</b>	<b>77.94</b>	<b>77.99</b>	77.77	<b>77.92</b>	<b>77.90</b>

selected as the addition features to PSSM for SVM classifier respectively: S<sup>5</sup>, H<sup>7</sup> + S<sup>4</sup> H<sup>11</sup> + S<sup>5</sup>, and H<sup>9</sup> + S<sup>5</sup> so that to get accuracy of SVM-PSSM classifier around 78%.

Figure 1 is a plot for feature selection illustration. It says that S<sup>3</sup>, S<sup>4</sup> and S<sup>5</sup> are top three lines, at the same time all series show the increase tendency along the horizontal axis except the last point responding to 11, however finally decrease at 12, which is a long term feature from the mean of 39 consecutive residues.



**Fig. 1.** Feature selection illustration. The horizontal axis show HSF, line series show SSF, and the vertical axis show the related accuracy of PSSM-SVM with addition of selected SFs

### 4 Conclusions

In this paper, support vector machine (SVM-PSSM-SF) based on sliding window (13) was used to study protein secondary structure prediction from amino acid sequences. Firstly, the novelty of our method lies in the combination of multiple sequence features with PSSM profile in order to improve the accuracy of SVM-PSSM to around 78%. Secondly,

sequence features formulated by consecutive amino acid units help to take into consider the long-distant influence to slid windows limitation of 13. Thirdly, the novel formulations (1) and (2) of sequence features are proposed, and the most related features by these formulations are picked up for more effective classifier construction. Finally, our experiments show that structure features which SFs are more related to secondary structure than Hydrophobicity features, and long term HSF combine with middle term FSF also show effective to the accuracy of SVM-PSSM classifier. In short, we finally identify the optimal subset of selected features to get the performance improvement of SVM-PSSM classifier. Since the previous studies did not utilize the biological knowledge for classifier construction, our method can be used to complement the existing methods.

Our study also provides some information about the secondary structure characteristics of the important structural information. The combination method can be further attempt for protein structure prediction and feature analysis.

In the future, we may take more sequence features into the consideration for SVM-PSSM-SF classifier, For example, Feature Aa which estimates a residue's average area buried in the interior core of a globular protein [19]; Bulkiness (Bu), the ratio of the side chain volume to the length of an amino acid, may affect the local structure of a protein [20].

**Acknowledgements.** The research work is supported by the National Natural Science Foundation of China (Grant No. 61375013); and the Natural Science Foundation of Shandong province (Grant No. ZR2013FM020) China.

## References

1. David, W.: *Proteins: Structure and Function*. Wiley, Hoboken (2013)
2. Raven, P.H., Johnson, G.B.: *How Scientists Think*. WCB/McGraw-Hill, New York (1997)
3. Martin J., Gibrat J.F., Rodolphe F: Analysis of an optimal hidden Markov model for secondary structure prediction. *BMC Struct. Biol.* **6**(25) (2006)
4. Yao, X.-Q., Zhu, H., She, Z.-S.: A dynamic Bayesian network approach to protein secondary structure prediction. *BMC Bioinform.* **9**(49), 25 (2008)
5. Kunal, J.: Prediction of ubiquitin proteins using artificial neural networks, hidden Markov model and support vector machines. *Silico Biol.* **7**(6), 559–568 (2007)
6. Chen, C., Tian, Y., Zou, X., Cai, P., Mo, J.: Prediction of protein secondary structure content using support vector machine. *Talanta* **71**(5), 2069–2073 (2007)
7. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**(2), 195–202 (1999)
8. Ding, S., Li, Y., Shi, Z., Yan, S.: A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile. *Biochimie* **97**, 60–65 (2014)
9. Teng, S., Srivastava, A.K., Wang, L.: Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genom.* **11**(Suppl. 2), S5 (2010)
10. Cuff, J.A., Barton, G.J.: Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins Struct. Funct. Genet.* **40**(3), 502–511 (2000)
11. Qu, W., Sui, H., Yang, B., Qian, W.: Improving protein secondary structure prediction using a multi-modal BP method. *Comput. Biol. Med.* **41**(10), 946–959 (2011)

12. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**(12), 2577–2637 (1983)
13. Gasteiger, E.H.C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A.: *The Proteomics Protocols Handbook*. Humana Press, New York (2005)
14. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982)
15. Deleage, G., Roux, B.: An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* **1**(4), 289–294 (1987)
16. Vapnik, V.N.: *The Nature of Statistical Learning Theory*, 2nd edn. Springer, Heidelberg (2000)
17. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**(27), 1–27 (2011)
18. Gibrat J.F., Rodolphe F: Analysis of an optimal hidden Markov model for secondary structure prediction. *BMC Struct. Biol.* **6**(25) (2006)
19. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H.: Hydrophobicity of amino acid residues in globular proteins. *Science* **229**(4716), 834–838 (1985)
20. Cho, M.K., Kim, H.Y., Bernado, P., Fernandez, C.O., Blackledge, M., Zweckstetter, M.: Amino acid bulkiness defines the local conformations and dynamics of natively unfolded alpha-synuclein and tau. *J. Am. Chem. Soc.* **129**(11), 3032–3033 (2007)