PSO Feature Selection and ELM Algorithm for Protein Classification based Secondary Structure and Hydropathy Profile

خوارزمية انتقاء الخصائص (PSO) ومصنف (ELM) لتصنيف البروتين مبنية على التركيب الثانوي والصورة المائية للبروتين

WATHIQ LAFTAH AL-YASEEN
Kerbala Technical Institute, Al-Furat Al-Awsat Technical University, 56001, Kerbala, Iraq
wathiqpro@gmail.com

Abstract

It is important to recognize protein classes in order to understand folding patterns. In this paper, we have proposed a method to extract the features based on secondary structure sequence and hydropathy profile. A feature selection algorithm that combines particle swarm optimization and extreme learning machine was employed to select a total of 25 features. The selected features were fed to the classifier in order to classify each protein to an appropriate class. The well-known data sets, i.e. ASTRAL$_{training}$, ASTRAL$_{test}$, 25PDB, 640 and 1189 were used to evaluate the proposed method. Upon comparing the current approach against other approaches based on the same data, it is evident that the proposed method shows higher efficiency in the prediction of structural class of protein, and its overall accuracy reaches up to 1.5%. Moreover, the extracted secondary and hydropathy features are important for us to differentiate the α/β and α+β classes.

الخلاصة

التعرف على اصناف البروتينات تلعب دور اساسي في فهم وتمييز الانماط. يقترح هذا البحث طريقة لتصنيف البروتينات من خلال استخراج مجموعة من الخصائص المبنية على سلسلة التركيب الثانوي للبروتين والصورة المائية للحامض الاميني للبروتين. كما ويقترح خوارزمية لانتقاء افضل الخصائص من خلال الدمج بين طريقتي particle swarm optimization وextreme learning machine لاستخلاص 25 خاصية من بين العدد الكبير من خصائص البروتين والتي تم الحصول عليه بالمرحلة الاولى. تغذى الخصائص التي يتم انتقاءها الى المصنف extreme learning machine لغرض تصنيف البروتين الى الصنف الصحيح من الفئات. تم استخدام مجموعات مختلفة من بيانات البروتينات المعروفة في الاعمال السابقة مثل ASTRAL$_{training}$، ASTRAL$_{test}$، 25PDB، 640 و 1189 لتقييم الطريقة المقترحة. بينت النتائج ان الطريقة المقترحة تمتلك كفاءة عالية من حيث دقة التصنيف تصل الى 1.5% بالمقارنة مع الطرق السابقة في تنبأ نوع البروتين. كما بينت النتائج ان الخصائص المعتمدة على سلسلة التركيب الثانوي والصور المائية للحامض الاميني للبروتين استطاعت التميز وبشكل حاسم بين انواع البروتينات من الصنف α/β و α+β.

Keywords
Protein structural classes, Extreme learning machine, Particle swarm optimization, Feature selection, Secondary structure sequence, Hydropath profile.

1. Introduction

The information of structural classes of protein plays a vital role in bioinformatics field for performing protein fold analysis and recognition, protein function prediction, and DNA prediction [1-4]. The first definition of protein structural class was presented by Levitt and Chothia in 1976 [5]. Accordingly, there are four major classes, i.e. (1) all-α class has small amount of strands; (2) all-β class has small amount of helices; (3) α+β class has both helices and strands, where the strands are commonly anti-parallel; and (4) α/β class has both helices and strands, where the strands are commonly parallel. Currently, the Structural Classification of Protein (SCOP) database is commonly used to classify protein structural classes [6]. The protein structures of SCOP are classified manually depending on the known tertiary structures of proteins. With the rapid development of sequencing technologies, the number of revealed protein sequences is increasing exponentially, thus enlarging the gap between the sequence-known and the structure-known proteins. Therefore, the manual methods are unable to cope with the increasing demand of classification. In other words, it is required to improve the current computational techniques in order to reduce the computational time and to enhance the determination accuracy of protein structural class. One of the shortcomings of the existing approaches is inaccuracy of datasets that have low-similarity sequences [2, 7-8], probably due to the usage of information that is extracted solely from the amino acid sequences [7-12]. Recently, several feature methods have been proposed to improve the prediction accuracy of the low-similarity sequences by using secondary structural information [13-18], such as SCPRED [19] and MODAS [20] that are designed based on the use of secondary structural sequences obtained from PSI-PRED [21]. In SCPRED, there are 9 selected features where 8 of them are based on secondary structure predictions and the remaining feature is based on the collection of Leucine and Glycine. In MODAS, the evolutionary and information profiles of the predicted secondary structure are employed for prediction. The extracted feature vector is used to train and evaluate different machine learning algorithms such as support vector machine (SVM) [22-25], neural network [26], fuzzy clustering [27], rough sets [28], etc. The feature representation and classification algorithms have been extensively reviewed [2, 8]. Although numerous methods have been developed based on secondary structures, it is challenging to develop high-quality prediction methods for low-similarity sequences.

In this study, we aim to enhance the prediction accuracy based on the secondary structure sequence and the hydropathy information obtained from amino acids sequence. The 25-dimensional integrated feature vectors were selected based on the feature selection algorithm that combines Particle Swarm Optimization (PSO) and Extreme Learning Machine (ELM). The multi-class ELM was then implemented to predict the protein structural class. In order to demonstrate the efficiency of the proposed prediction method, the 10-fold cross-validation test (10-CV) was conducted on 5 low-similarity data sets.

The rest of the paper is organized in the following manner. Section 2 summarizes the related works of protein classification methods. Section 3 explains the proposed method consisting of feature extraction and feature selection with classification by using PSO-ELM. Some performance measures are highlighted as well. Section 4 discusses the experimental results. Finally, Section 5 concludes the paper and recommends some possible future works.

2. Related works

The open literatures related to the feature extraction and selection methods for improving the performance of protein classification have been reviewed. The outcomes are presented in this section.

Wang et al. [4] proposed a model to extract features based on PSSM and secondary structure sequence for protein classification using SVM. The prediction precision has been improved by 3%–5%. SVM has been found efficient in improving the generalization performance, solving high-dimensional problem, and avoiding local minimum problem. However, SVM is particularly sensitive to missing data and the kernel function must be carefully chosen for processing. Furthermore, Wang et al. [29] extracted feature information based on PseAA structural properties and secondary structure patterns, using SVM in the protein structural class data sets. In general, its overall prediction accuracy is promising.

The method proposed by Yang et al. [13], i.e. RKS-PPSC is based on secondary structure sequence. A total of 24 features are obtained using recurrence quantification analysis, k-string based information entropy and segment based analysis. In this method, the prediction accuracy of α+β is lower than that of our current method for all data sets.

Ding et al. [14] predicted the protein structural classes based on the predicted secondary structure. Their method is reasonably accurate; however, it is computationally expensive due to the associated high dimensional space with SVM.

In additional, Zhang et al., [30] introduced a method to predict the protein based on secondary structure sequence. They used SVM-based Jackknife cross-validation test to classify the data sets of protein. However, the prediction accuracy of α+β is somewhat low.

Finally, Wang et al., [31] proposed a deep recurrent encoder-decoder neural network called Secondary Structure Recurrent Encoder-Decoder Network (SSREDN) to solve some problems related to predicted secondary structure. This model is not recommended as information leakage cannot be prevented.

3. Proposed approach

In this section, the details of the proposed approach are given. Figure 1 shows the proposed methodology in this study. The benchmark cases (ASTRAL, 25PDB, 640 and 1189) are employed to evaluate the performance of the proposed model. The scenario of the proposed method is described as follows:

In the feature extraction stage:

1. Convert the protein amino acids into the secondary structure sequence by using PSIPRED [21].
2. Extract 38 features from the secondary structure sequence based on the elements (H, E and C) shown in Table 2.
3. Remove the C element from the secondary structure sequence.
4. Extract 10 additional features from the obtained secondary structure sequence based on the elements (H and E) shown in Table 3.
5. Represent the protein amino acids in the form of hydropathy sequence with three elements (I, E and A) by using the formulation proposed by Wang et al. [29].
6. Extract 25 features from the hydropathy sequence as shown in Table 4.

Hence, a total of 73 features can be extracted from this stage.

In the feature selection stage:

1. Apply PSO technique to generate the swarm of feature subsets randomly.
2. Evaluate the subsets by using ELM algorithm.
3. Repeat the operations of PSO and step 2 to improve the subsets of features (*n* iteration).
4. Select the best global solution of feature subsets based on the maximum cost.

In the classification stage:

1. Determine the optimum parameters for ELM algorithm by using the grid search technique.
2. Build new training and testing data sets based on the best selection of feature subset.
3. Apply ELM to classify the data set into one of the classes (α, β, α/β, α+β).
4. Validate and compare the classification results with other related works.

Additional details about the proposed method are provided.

3.1 Protein data sets

In order to perform a comprehensive comparison experimentally, five commonly low-similarity benchmark data sets were employed. The selected ASTRAL data set (including 7 classes) has sequence similarity of < 20% and it comprises of 6424 sequences [20]. In our study, only four major classes (all-α, all-β, α/β, α+β) that contain 5626 sequences were used. The ASTRAL data set was arbitrarily separated into two equivalent subsets; one was used as the training data set ($\text{ASTRAL}_{\text{training}}$) and the other was used as the test data set ($\text{ASTRAL}_{\text{test}}$). The selected 25PDB comprises of 1673 protein domains with similarity of sequence of < 25% [2]. The 640 data set consisting of 640 sequence with 25% similarity of sequence was taken from [14]. The final data set named 1189 comprises of 1092 sequences with 40% sequence similarity [32]. The details of the five data sets are shown in Table 1.

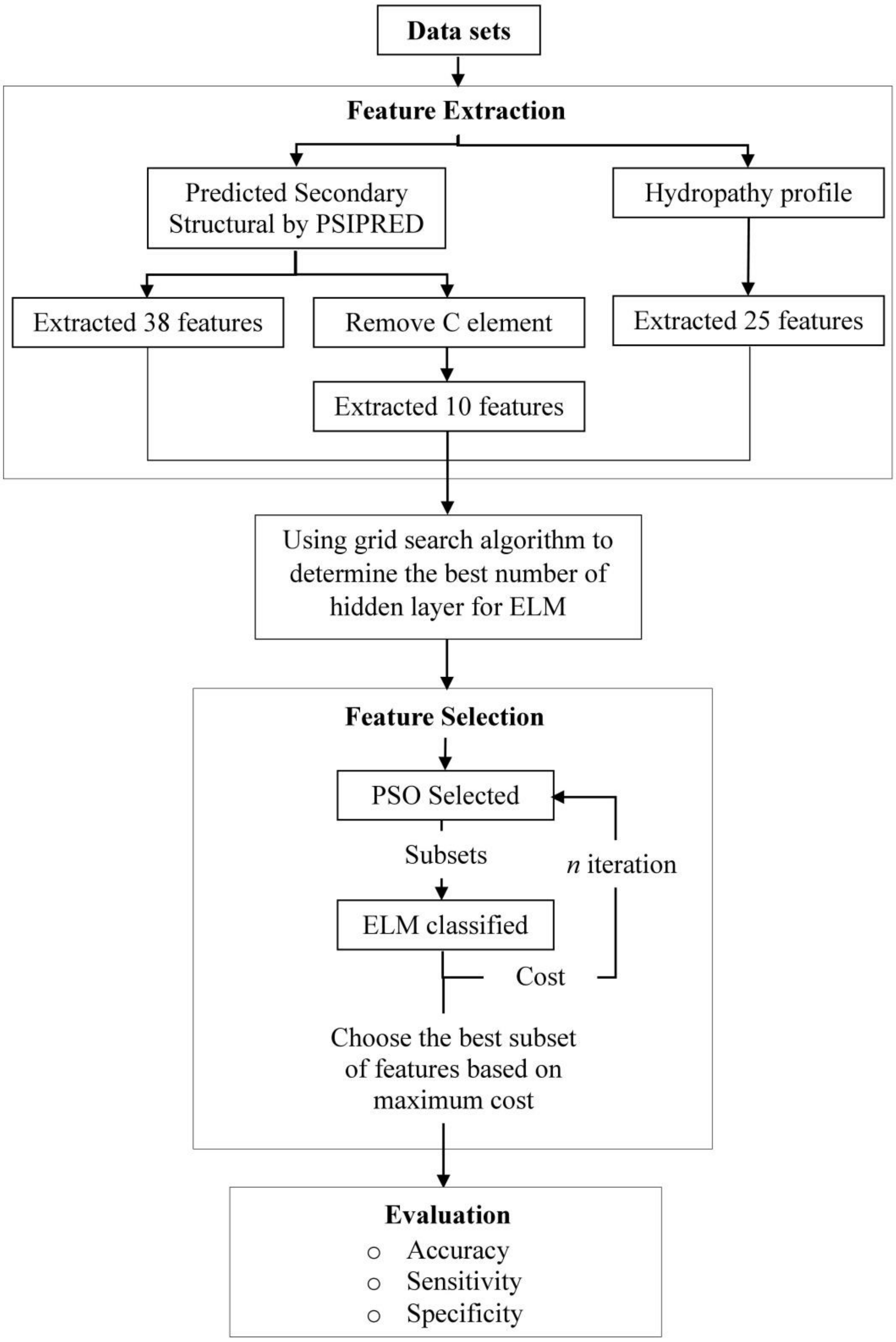

Figure 1. The methodology of proposed approach

Table 1. The characteristics of data sets employed to evaluate the proposed method

Data set	All-α	All-β	α/β	α+β	Total
$\text{ASTRAL}_{\text{training}}$	640	662	748	763	2813
$\text{ASTRAL}_{\text{test}}$	640	662	747	764	2813
25PDB	443	443	346	441	1673
640	138	154	177	171	640
1189	223	294	334	241	1092

3.2 Feature extraction

Based on the secondary structure and hydropathy sequences of amino acids, a total of 73 features are extracted. These features are detailed as follows:

3.2.1. Features extraction of secondary structure sequence

By using PSIPRED (version 2.6) [21], the prediction outcome of the protein amino acid is one of the following secondary structural elements: H (helix), E (strand), and C (coil). Hence, the protein can be categorized within one of the structural classes according to these elements. The features obtained from the above structures may be immediately applied to the protein structural class prediction. Accordingly, a total of 48 features were derived to reflect the common contents and spatial arrangement of the secondary structural elements of the specified protein sequence. Some of them have been applied previously [19, 33-34]. Table 2 summarizes the predicted secondary structural features, where the length of the secondary structural sequence is denoted as L.

α-helices and β-strands are normally separated in α/β proteins and they are usually interspersed in α+β proteins. Therefore, in order to reflect the distributions of α-helices and β-strands effectively, the C element was removed from the secondary structural sequence to obtain the new sequence H-E. A total of 10 features based on the new secondary structure were extracted, where the length of the H-E sequence is denoted by L_{new}. The features of the new H-E sequence are detailed in Table 3.

3.2.2. Features extraction of hydropathy sequence (HS)

The hydropathy profile of the protein sequence was selected based on the assumption that it would significantly influence the protein folding process. Hydropathy features define the hydrophilic and hydrophobic natures of the sequence [29]. Twenty types of amino acids of protein were classified into three groups based on their respective hydropathy profiles, namely *Internal (I), External (E)* and *Ambivalent (A)* (see Wang et al., [29]):

$$F\big(S(i)\big) = \begin{cases} I & if \quad S(i) = F, I, L, M, V \\ E & if \quad S(i) = D, E, H, K, N, Q, R \\ A & if \quad S(i) = S, T, Y, C, W, G, P, A \end{cases}$$

In the above equation, $S(i)$ denotes the i^{th} amino acid in the primary sequence of the protein, while $F(\mathrm{S(i)})$ denotes its constant substitute based on the nature of the hydropathy. For example, a sequence of the amino acid for a given protein can be expressed as:

$S = MDPFLVLLHSVSS$ is denoted by

$$F(S) = IEAIIIIIEAIAA$$

Therefore, based on the new hydropathy sequence (HS), $i \in \{I, E, A\}$, a total of 25 new hydropathical features were extracted. The hydropathical features were combined with the previous features extracted from the predicted secondary structural sequence. The hydropathical features are shown in Table 4, where L_{hydro} denotes the length of the new hydropathy sequence.

3.3 Feature selection

In this paper, many features have been extracted, signifying that a large computational cost will be required for machine learning. Furthermore, those irrelevant and redundant features could affect the prediction accuracy. Thus, feature selection was employed to select those more essential features in order to improve and accelerate the prediction process. Several feature selection approaches have been proposed in different bioinformatics studies [35]. These methods can be divided into two main groups: filter and wrapper. Filter approach adopts the statistical properties of features for selecting the good features. On the other hand, wrapper approach combines the feature selection method with a specific classifier to estimate the worth feature subsets by calculating its accuracy. Thus, better result can be obtained using wrapper approach. Moreover, the wrapper feature selection approach uses the cross-validation function to prevent overfitting during the course of calculation. Finally, PSO technique was used with 10 cross-validation ELM to construct 25-dimensional feature vector from the above 73 features. PSO was adopted to improve the quality of selected features using ELM. The process was terminated upon obtaining the optimal feature vector.

Table 2. The features derived from secondary structural sequence

No	Feature	Description
1	$f_1 = N(H)/L$	Normalized count of H in secondary structural sequence
2	$f_2 = N(E)/L$	Normalized count of E in secondary structural sequence
3	$f_3 = N(C)/L$	Normalized count of C in secondary structural sequence
4	$f_4 = maxSegH/L$	Normalized length of the longest segment of H
5	$f_5 = maxSegE/L$	Normalized length of the longest segment of E
6	$f_6 = maxSegC/L$	Normalized length of the longest segment of C
7	$f_7 = avgSegH/L$	Normalized average of length for segments H
8	$f_8 = avgSegE/L$	Normalized average of length for segments E
9	$f_9 = avgSegC/L$	Normalized average of length for segments C
10	$f_{10} = \sum_{j=1}^{L} PosH_j \,/L(L-1)$	Composition moment vector of H (PosH$_j$ is the *j*th position of H)
11	$f_{11} = \sum_{j=1}^{L} PosE_j \,/L(L-1)$	Composition moment vector of E (PosE$_j$ is the *j*th position of E)
12	$f_{12} = \sum_{j=1}^{L} PosC_j \,/L(L-1)$	Composition moment vector of C (PosC$_j$ is the *j*th position of C)
13	$f_{13} = nSegH/L$	Normalized number of segments H
14	$f_{14} = nSegE/L$	Normalized number of segments E
15	$f_{15} = nSegC/L$	Normalized number of segments C
16	$f_{16} = Pn_E/(Pn_E + APn_E)$	The ratio of parallel β-sheets to both parallel and anti-parallel β-sheets
17	$f_{17} = APn_E/(Pn_E + APn_E)$	The ratio of anti-parallel β-sheets to both parallel and anti-parallel β-sheets
18	$f_{18} = maxDistHE/L$	Normalized maximum distance between H and E
19	$f_{19} = maxDistEH/L$	Normalized maximum distance between E and H
20	$f_{20} = nExchangeHE/L$	Normalized number of exchanges between H and E
21	$f_{21} = varSegH/L$	Normalized variance of segment lengths H
22	$f_{22} = varSegE/L$	Normalized variance of segment lengths E
23	$f_{23} = varSegC/L$	Normalized variance of segment lengths C
24	$f_{24} = varPosH/L$	Normalized variance of positions H
25	$f_{25} = varPosE/L$	Normalized variance of positions E
26	$f_{26} = varPosC/L$	Normalized variance of positions C
27	$f_{27} = stdDevHE/L$	Normalized standard deviation of count between H and E
28	$f_{28} = stdDevHC/L$	Normalized standard deviation of count between H and C
29	$f_{29} = stdDevEC/L$	Normalized standard deviation of count between E and C
30	$f_{30} = stdDevHEC/L$	Normalized standard deviation of count for H, E and C
31	$f_{31} = N(HE)/L$	Normalized count of HE
32	$f_{32} = N(EH)/L$	Normalized count of EH
33	$f_{33} = \sum_{j=1}^{L} PosHE_j \,/L(L-1)$	Composition moment vector of HE
34	$f_{34} = \sum_{j=1}^{L} PosHC_j \,/L(L-1)$	Composition moment vector of HC
35	$f_{35} = \sum_{j=1}^{L} PosEH_j \,/L(L-1)$	Composition moment vector of EH
36	$f_{36} = \sum_{j=1}^{L} PosEC_j \,/L(L-1)$	Composition moment vector of EC
37	$f_{37} = \sum_{j=1}^{L} PosCH_j \,/L(L-1)$	Composition moment vector of CH
38	$f_{38} = \sum_{j=1}^{L} PosCE_j \,/L(L-1)$	Composition moment vector of CE

Table 3. The features derived from the secondary structural sequence upon removing the C element

No	Feature	Description
1	$f_{39} = N(H)/L_{new}$	Normalized count of H in new H-E sequence
2	$f_{40} = N(E)/L_{new}$	Normalized count of E in new H-E sequence
3	$f_{41} = maxSegH/L_{new}$	Normalized length of the longest segment of H in H-E
4	$f_{42} = maxSegE/L_{new}$	Normalized length of the longest segment of E in H-E
5	$f_{43} = avgSegH/L_{new}$	Normalized average length for segments H in H-E
6	$f_{44} = avgSegE/L_{new}$	Normalized average length for segments E in H-E
7	$f_{45} = \sum_{j=1}^{L_{new}} PosH_j/L_{new}(L_{new} - 1)$	Composition moment vector of H in H-E sequence
8	$f_{46} = \sum_{j=1}^{L_{new}} PosE_j/L_{new}(L_{new} - 1)$	Composition moment vector of E in H-E sequence
9	$f_{47} = nSegH/L_{new}$	Normalized number of segments H in H-E
10	$f_{48} = nSegE/L_{new}$	Normalized number of segments E in H-E

Table 4. The features which derived from hydropathy profile

No	Feature	Description
1	$f_{49} = N(I)/L_{hydro}$	Normalized count of I
2	$f_{50} = N(E)/L_{hydro}$	Normalized count of E
3	$f_{51} = N(A)/L_{hydro}$	Normalized count of A
4	$f_{52} = maxSegI/L_{hydro}$	Normalized length of the longest segment I
5	$f_{53} = maxSegE/L_{hydro}$	Normalized length of the longest segment E
6	$f_{54} = maxSegA/L_{hydro}$	Normalized length of the longest segment A
7	$f_{55} = avgSegI/L_{hydro}$	Normalized average length for segments I
8	$f_{56} = avgSegE/L_{hydro}$	Normalized average length for segments E
9	$f_{57} = avgSegA/L_{hydro}$	Normalized average length for segments A
10	$f_{58} = \sum_{j=1}^{L_{hydro}} PosI_j/L_{hydro}(L_{hydro} - 1)$	Composition moment vector of I
11	$f_{59} = \sum_{j=1}^{L_{hydro}} PosE_j/L_{hydro}(L_{hydro} - 1)$	Composition moment vector of E
12	$f_{60} = \sum_{j=1}^{L_{hydro}} PosA_j/L_{hydro}(L_{hydro} - 1)$	Composition moment vector of A
13	$f_{61} = nSegI/L_{hydro}$	Normalized number of segments I
14	$f_{62} = nSegE/L_{hydro}$	Normalized number of segments E
15	$f_{63} = nSegA/L_{hydro}$	Normalized number of segments A
16	$f_{64} = varSegI/L_{hydro}$	Normalized variance of segment lengths I
17	$f_{65} = varSegE/L_{hydro}$	Normalized variance of segment lengths E
18	$f_{66} = varSegA/L_{hydro}$	Normalized variance of segment lengths A
19	$f_{67} = varPosI/L_{hydro}$	Normalized variance of positions I
20	$f_{68} = varPosE/L_{hydro}$	Normalized variance of positions E
21	$f_{69} = varPosA/L_{hydro}$	Normalized variance of positions A
22	$f_{70} = stdDevIE/L_{hydro}$	Normalized standard deviation of count between I and E
23	$f_{71} = stdDevIA/L_{hydro}$	Normalized standard deviation of count between I and A
24	$f_{72} = stdDevEA/L_{hydro}$	Normalized standard deviation of count between E and A
25	$f_{73} = stdDevIEA/L_{hydro}$	Normalized standard deviation of count for I, E and A

The processing steps of PSO are outlined as follow:

1. Choose the optimum values of parameters (ω, $\varphi1$, $\varphi2$) for PSO using the grid search technique.
2. Generate the positions and velocities of particle swarms randomly.
3. Set the global best solution to zero.
4. Evaluate the particle swarms (i.e. computing costs) by using ELM model.
5. Search the global best solution based on the objective function, i.e. maximization of cost.
6. Recalculate the velocities of particle swarms using previous velocities and positions of the best particle swarms.

$$Velocity = \omega \times velocity + \varphi1 \times rand \times best\ particle\ swarm\ + \varphi2 \times rand \times global\ best\ solution$$

7. Recalculate the positions of particle swarms based on the previous positions and new velocities.

$$Position = position + Velocity$$

8. Repeated steps (4-7) until convergence is attained for the positions of particle swarms.
9. Select the global best solution as the best subset of features.

3.4 The ELM Classifier

The recognition of protein structural classes is indeed a multi-class classification problem. Here, an extreme learning machine tested successfully by Huang et al. [36] was employed as a single-hidden layer-feed-forward neural network (SLFNN). It is also named as ELM as it can exactly learn N distinct observations, i.e. almost any nonlinear activation function with at most N hidden nodes. Hence, the essential difference between ELM and traditional training of a feed-forward network is that the tuning of hidden layer of ELM is unnecessary (i.e. parameters of hidden layer are randomly chosen). However, the input weights and hidden neurons biases, as well as the output weights of the hidden layer, are assigned randomly in order to minimize the training error. ELM transforms the learning problem into a simple linear system where the output weights can be analytically determined. The results reported in [36] implies that ELM performs better and its implementation is easy. For kernel-based ELM, several nonlinear kernel functions can be used to calculate the hidden layer feature mapping of ELM. One of the popular kernel functions is Gaussian radial basis function (RBF) which was used in the current work. Moreover, the grid search algorithm was used to determine the optimum number of hidden layers for ELM. The optimum number of hidden layers was 30 and it was used in all experimental results.

3.5 Evaluation Measures

The 10 cross-validation (10-CV) test is a popular method used to validate the results of different classifiers [2]. Hence, it is exploited to assess the stability and reliability of our new approach. In order to perform extensive evaluation, parameters such as individual sensitivity (Sens), individual specificity (Spec), Matthew's correlation coefficient (MCC) of the four structural classes, and overall prediction accuracy (OA) of the whole data set were calculated. These parameters are defined as follows:

$$Sens_j = \frac{TP_j}{(TP_j + FN_j)}$$

$$Spec_j = \frac{TN_j}{TN_j + FP_j}$$

$$MCC_j = \frac{(TP_j \times TN_j - FP_j \times FN_j)}{\sqrt{(TP_j + FP_j)(TP_j + FN_j)(TN_j + FP_j)(TN_j + FN_j)}}$$

$$OA = \frac{TP_j}{(TP_j + TN_j + FP_j + FN_j)}$$

where TP, TN, $FP$and FN are the number of true positives, true negatives, false positives and false negatives in the structural class C_j, respectively.

4. Experimental Results and Discussion

The proposed method was tested with $\text{ASTRAL}_{\text{training}}$, $\text{ASTRAL}_{\text{test}}$, 25PDB, 640 and 1189 data sets by using 10-CV test. The predicted results for proteins from all-α, all-β, α/β and α+β classes were compared against those of other approaches using the same data sets. The predicted results of our current approach are reported in Table 5. The computer with the following specification was used: Windows 10 OS, Core i5 2.60 GHz CPU, and 12 GB RAM. The freeware package ELM [38] was coded using Java.

Table 5. The predicted results using the current method upon conducting the 10-CV test on five data sets

Data set	Class	Sens (%)	Spec (%)	MCC (%)
ASTRAL$_{training}$	All-α	94.22	97.93	91.75
	All-β	83.08	96.56	81.32
	α/β	84.22	92.98	76.35
	α+β	72.08	89.46	61.47
	OA	82.94		
ASTRAL$_{test}$	All-α	93.75	98.62	92.89
	All-β	81.72	96.65	80.53
	α/β	85.94	92.64	77.11
	α+β	72.38	89.12	61.2
	OA	83.04		
25PDB	All-α	95.49	95.45	88.76
	All-β	83.75	97.48	83.88
	α/β	81.79	95.78	78.16
	α+β	78.91	91.48	69.79
	OA	85.18		
640	All-α	90.58	98.41	90.2
	All-β	85.06	96.71	83.1
	α/β	90.96	92.87	81.59
	α+β	74.85	91.9	67.39
	OA	85.16		
1189	All-α	93.27	96.89	88.45
	All-β	86.39	97.99	86.77
	α/β	84.43	93.01	77.37
	α+β	70.54	90.36	59.95
	OA	83.7		

As shown in Table 5, the overall accuracies for all data sets are high (> 82.9 %), indicating that our predictor is reliable. Also, the proposed method is stable even though the size of data set and similarity are different. Moreover, the percentages of Sens, Spec and MCC for all-α class are the highest for all data sets, while the percentages of α+β class are the lowest. For example, the MCC is only 59.95 % for the 1189 data set due to the fact that it is difficult to distinguish the α+β class as there is an overlap with other classes [32]. Thus, it is challenging to identify anti-parallel sheets.

Moreover, the proposed method was compared against other popular methods such as SCPRED [19], MODAS [20], and the methods developed by Zhang et al. [38] and Ding et al. [14].

Table 6. Comparison of our proposed method with other methods

Data set	Reference	Accuracy (%)				OA
		All-α	All-β	α/β	α+β	
ASTRAL$_{training}$	[38]	94.06	81.72	79.55	73.79	81.80
	The proposed	94.22	83.08	84.22	72.08	82.94
ASTRAL$_{test}$	[38]	95.16	80.97	83.94	72.51	82.69
	[14]	94.53	77.49	87.28	71.47	82.33
	[19]	93.13	78.33	83.38	64.27	79.14
	The proposed	93.75	81.72	85.94	72.38	83.04
25PDB	[38]	94.81	82.39	81.21	77.32	84.1
	[14]	95.03	81.26	83.24	77.55	84.34
	[19]	92.6	80.1	74	71	79.7
	[20]	92.3	83.7	81.2	68.3	81.4
	The proposed	95.49	83.75	81.79	78.91	85.18
640	[38]	92.75	81.82	89.27	74.27	84.22
	[14]	94.93	76.62	89.27	74.27	83.44
	[19]	90.6	81.8	85.9	66.7	80.8
	[20]	89.1	85.1	88.1	71.4	83.1
	The proposed	90.58	85.06	90.96	74.85	85.16
1189	[14]	93.72	84.01	83.53	66.39	81.96
	[19]	89.1	86.7	89.6	53.8	80.6
	[20]	92.3	87.1	87.9	65.4	83.5
	The proposed	93.27	86.39	84.43	70.54	83.7

As shown in Table 6, our method exhibits the highest overall accuracy, i.e. accuracies are improved by 1.14 %, 0.35 %, 0.63 %, 0.94 % and 0.2 % for ASTRAL$_{training}$, ASTRAL$_{test}$, 25PDB, 640 and 1189 data sets, respectively. Moreover, the accuracies of ASTRAL$_{training}$ for all-α, all-β and α/β class are 0.16 %, 1.36 % and 4.67 % higher than those of Zhang et al. method [38]. For the ASTRAL$_{test}$ data set, the accuracy of all-β class is 0.75 % higher than the previously reported best result [38]. As compared with those best methods reported previously, the accuracies of all-β and α+β classes are improved by 0.05 % and 1.36 % respectively for 25PDB. In addition, the accuracies of α/β and α+β class are improved by 1.69 % and 0.58 % respectively as compared with the best results reported for 640 data set [14, 38]. Finally, for the 1189 data set, the accuracy of α+β class is 4.15 % higher than the previous best-performing result obtained by Ding et al. [14]. From Table 6, some results obtained using our method are slightly inferior to those predicted using other methods. Figure 2 compares the overall accuracy of our proposed method with the best method reported previously for each data set. In addition, Figure 3 compares the sensitivities of the proposed method and the best method reported in open literature for all data sets. As deduced from these figures, our method is more promising than others.

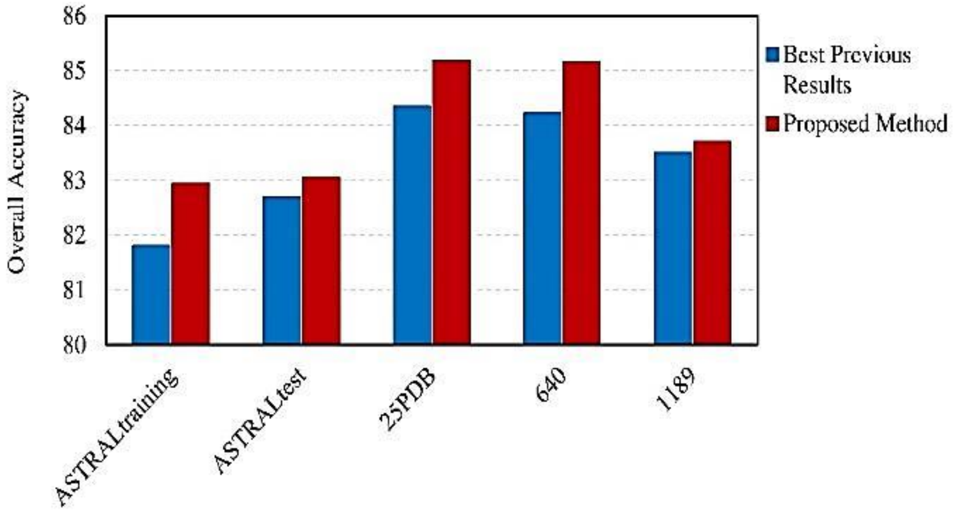

Figure 2. Comparison of proposed method with the other best methods in terms of overall accuracy

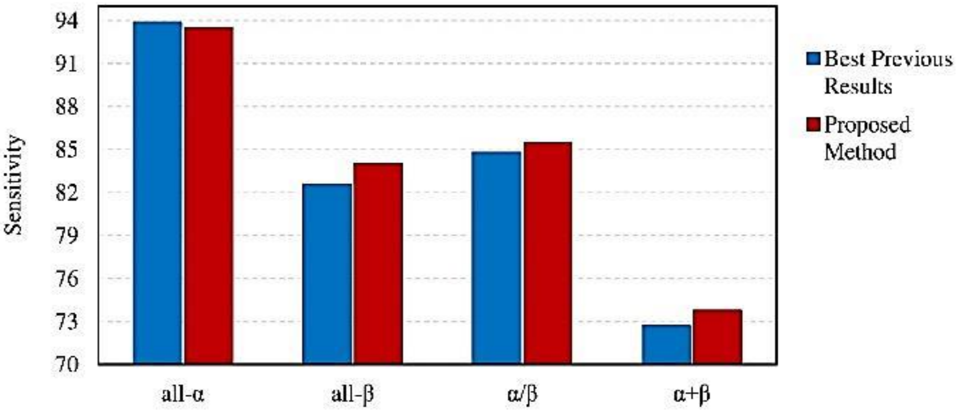

Figure 3. Comparison of proposed method with the other best methods in terms of average sensitivity

Based on the previous comparisons, our proposed method performs the best in predicting the α/β and α+β classes which are hard to be predicted precisely. For the α/β class, our approach shows promising prediction accuracy in spite of the fact that α/β class always produces non-promising results as each protein of α/β can be assigned to more than one class. Moreover, some less popular elements of secondary structural, e.g. β-turns and β-bugles are excluded. These elements could be included in future work for accuracy enhancement purpose.

The proposed approach selects features that are more related to the protein structural class. Moreover, the t-test shows that the proposed method improves the overall accuracy significantly as shown in Table 6, where the *p*-value is 0.008996.

On the other hand, the proposed method is computational cheaper than SVM, as the employed ELM is faster than SVM. Besides that, the computational time of PSO algorithm in the selection process is shorter than those of other optimization methods.

5. Conclusions

This study has presented an approach to predict the protein structural class based on secondary structural sequence and hydropathy profile. Firstly, the best feature vector was chosen using the combined method of PSO and ELM. Then, the selected feature vector was fed into the ELM technique in order to predict the protein structural classes. From the experiments, the accuracy of the proposed method can reach up to 85.18%. The main contribution of this study is the extraction of new features from secondary structure and hydropathy profile of protein. The current method performs better than other popular methods such as SCPRED, MODAS, etc. The proposed approach is about 1.5 % more accurate than the existing best-performing methods. Our method performs exceptional well in predicting α/β and α+β classes. From the current work, the influence of hydropathy feature on the prediction accuracy has been found to be significant.

References

[1] K. Chou, "Structural bioinformatics and its impact to biomedical science and drug discovery", *Frontiers in medicinal chemistry,* Vol.455, No.502, pp.455-502, 2006.

[2] L. A. Kurgan, and L. Homaeian, "Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy", *Pattern Recognition,* Vol.39, No.12, pp.2323-2343, 2006.

[3] P. Ferragina, G. Raffaele, G. Valentina, M. Giovanni and V. Gabriel, "Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment", *BMC bioinformatics,* Vol.8, No.1, pp.1-20, 2007.

[4] J. Wang, W. Cong, C. Jiajia, L. Xiaoqing, Y. Yuhua, and D. Qi, "Prediction of protein structural classes for low-similarity sequences using reduced PSSM and position-based secondary structural features", *Gene*, Vol.554, No.2, pp.241-248, 2015.

[5] M. Levitt, and C. Cyrus, "Structural patterns in globular proteins", *Nature,* Vol.261, No.5561, pp.552-558, 1976.

[6] A. G. Murzin, E. B. Steven, H. Tim and C. Cyrus, "SCOP: a structural classification of proteins database for the investigation of sequences and structures", *Journal of molecular biology,* Vol.247, No.4, pp.536-540, 1995.

[7] K. Chou and Z. Chun-Ting, "Prediction of protein structural classes", *Critical reviews in biochemistry and molecular biology,* Vol.30, No. 4, pp.275-349, 1995.

[8] K. Chou, "Progress in protein structural class prediction and its impact to bioinformatics and proteomics", *Current Protein and Peptide Science,* Vol.6, No.5, pp.423-436, 2005.

[9] K. Chou, "Structural bioinformatics and its impact to biomedical science", *Current medicinal chemistry,* Vol.11, No.16, pp.2105-2134, 2004.

[10] K. D. Kedarisetti, K. Lukasz and D. Scott, "Classifier ensembles for protein structural class prediction with varying homology", *Biochemical and Biophysical Research Communications*, Vol.348, No.3, pp.981-988, 2006.

[11] S. Costantini and M. F. Angelo, "Prediction of the protein structural class by specific peptide frequencies", *Biochimie*, Vol.91, No.2, pp.226-229, 2009.

[12] J. Y. Yang, P. Zhen-Ling, Y. Zu-Guo, Z. Rui-Jie, A. Vo and W. Desheng, "Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation", *Journal of Theoretical Biology,* Vol.257, No.4, pp.618-626, 2009.

[13] J. Y. Yang, Z. L. Peng and X. Chen, “Prediction of protein structural classes for low-homology sequences based on predicted secondary structure”, *BMC bioinformatics,* Vol.11, No.1, pp.1-10, 2010.

[14] S. Ding, Z. Shengli, L. Yang and W. Tianming, “A novel protein structural classes prediction method based on predicted secondary structure”, *Biochimie,* Vol.94, No.5, pp.1166-1171, 2012.

[15] S. Zhang, S. Ding and T. Wang, “High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure”, *Biochimie,* Vol.93, No.4, pp.710-714, 2011.

[16] L. Zhang, X. Zhao and L. Kong, “A protein structural class prediction method based on novel features”, *Biochimie,* Vol.95, No.9, pp.1741-1744, 2013.

[17] Q. Dai, L. Yan, L. Xiaoqing, Y. Yuhua, C. Yunjie and H. Pingan, “Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position”, *BMC bioinformatics,* No.14, No.1, pp.1-14, 2013.

[18] L. Kong, L. Zhang and J. Lv, “Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition”, *Journal of theoretical biology,* Vol.344, No.2014, pp.12-18, 2014.

[19] L. Kurgan, K. Cios and K. Chen, “SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences”, *BMC bioinformatics,* Vol.9, No.1, pp.1-15, 2008.

[20] M. J. Mizianty and L. Kurgan, “Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences”, *BMC bioinformatics,* Vol.10, No.1, pp.1-24, 2009.

[21] D. T. Jones, “Protein secondary structure prediction based on position-specific scoring matrices”, *Journal of molecular biology,* Vol.292, No.2, pp.195-202, 1999.

[22] A. Anand, G. Pugalenthi and P. Suganthan, “Predicting protein structural class by SVM with class-wise optimized features and decision probabilities”, *Journal of theoretical biology,* Vol.253, No.2, pp.375-380, 2008.

[23] Y. D. Cai, L. Xiao-Jun, X. Xue-biao and C. Kuo-Chen, “Prediction of protein structural classes by support vector machines”, *Computers & Chemistry,* Vol.26, No.3, pp. 293-296, 2002.

[24] C. Chen, T. Yuan-Xin, Z. Xiao-Yong, C. Pei-Xiang and M. Jin-Yuan, “Using pseudo-amino acid composition and support vector machine to predict protein structural class”, *Journal of theoretical biology,* Vol.243, No.3, pp.444-448, 2006.

[25] J.-D. Qiu, L. San-Hua, H. Jian-Hua and L. Ru-Ping, “Using support vector machines for prediction of protein structural classes based on discrete wavelet transform”, *Journal of computational chemistry,* Vol.30, No.8, pp.1344-1350, 2009.

[26] Y.-D. Cai and G.-P. Zhou, “Prediction of protein structural classes by neural network”, *Biochimie,* Vol.82, No.8, pp.783-785, 2000.

[27] H.-B. Shen, Y. Jie, L. Xiao-Jun and C. Kuo-Chen, “Using supervised fuzzy clustering to predict protein structural classes”, *Biochemical and biophysical research communications,* Vol.334, No.2, pp.577-581. 2005.

[28] Y. Cao, L. Shi, Z. Lida, Q. Jie, W. Jiang and T. Kexuan, “Prediction of protein structural class with Rough Sets”, *BMC Bioinformatics,* Vol.7, No.1, pp.1-6, 2006.

[29] J. Wang, L. Yan, L. Xiaoqing, D. Qi, Y. Yuhua and H. Pingan, "High-accuracy prediction of protein structural classes using PseAA structural properties and secondary structural patterns", *Biochimie*, Vol.101, No.2014, pp.104-112, 2014.

[30] S. Zhang, D. Shuyan and W. Tianming, "High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure", *Biochimie*, Vol.93, No.4, pp.710-714, 2011.

[31] Y. Wang, M. Hua and Y. Zhang, "Protein secondary structure prediction by using deep learning method", *Knowledge-Based Systems,* Vol.118, No.2017, pp.115-123, 2017.

[32] K. Chen, L. A. Kurgan and J. Ruan, "Prediction of protein structural class using novel evolutionary collocation-based sequence representation", *Journal of computational chemistry,* Vol.29, No.10, pp.1596-1604, 2008.

[33] L. A. Kurgan, Z. Tuo, Z. Hua, S. Shiyi and R. Jishou, "Secondary structure-based assignment of the protein structural classes", *Amino acids,* Vol.35, No.3, pp.551-564, 2008.

[34] T. Liu and C. Jia, "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information", *Journal of theoretical biology,* Vol.267, No.3, pp.272-275, 2010.

[35] Y. Saeys, I. Iñaki and P. Larrañaga, "A review of feature selection techniques in bioinformatics", *Bioinformatics,* Vol.23, No.19, pp.2507-2517, 2007.

[36] G.-B. Huang, Z. Hongming, D. Xiaojian and Z. Rui, "Extreme learning machine for regression and multiclass classification", *IEEE Transactions on Systems, Man, and Cybernetics,* Vol.42, No.2, pp.513-529, 2012.

[37] D. Li, "Extreme Learning Machine". http://www.ntu.edu.sg/home/egbhuang/, 2019.

[38] L. Zhang, Z. Xiqiang, K. Liang and L. Shuxia, "A novel predictor for protein structural class based on integrated information of the secondary structure sequence", *Biochimie*, Vol.103, pp.131-136, 2014.