Enhanced Protein Structural Class Prediction Using Effective Feature Modeling and Ensemble of Classifiers

Sanjay Bankapur and Nagamma Patil

Abstract—Protein Secondary Structural Class (PSSC) information is important in investigating further challenges of protein sequences like protein fold recognition, protein tertiary structure prediction, and analysis of protein functions for drug discovery. Identification of PSSC using biological methods is time-consuming and cost-intensive. Several computational models have been developed to predict the structural class; however, they lack in generalization of the model. Hence, predicting PSSC based on protein sequences is still proving to be an uphill task. In this article, we proposed an effective, novel and generalized prediction model consisting of a feature modeling and an ensemble of classifiers. The proposed feature modeling extracts discriminating information (features) by leveraging three techniques: (i) Embedding – features are extracted on the basis of spatial residue arrangements of the sequences using word embedding approaches; (ii) SkipXGram Bi-gram – various sets of skipped bi-gram features are extracted from the sequences; and (iii) General Statistical (GS) based features are extracted which covers the global information of structural sequences. The combined effective sets of features are trained and classified using an ensemble of three classifiers: Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machines (GBM). The proposed model when assessed on five benchmark datasets (high and low sequence similarity), viz. z277, z498, 25PDB, 1189, and FC699, reported an overall accuracy of 93.55, 97.58, 81.82, 81.11, and 93.93 percent respectively. The proposed model is further validated on a large-scale updated low similarity ($\leq$25%) dataset, where it achieved an overall accuracy of 81.11 percent. The proposed generalized model is robust and consistently outperformed several state-of-the-art models on all the five benchmark datasets.

Index Terms—Amino acid sequence, bi-gram, embedding, ensemble classifier, machine learning, protein structural class, skip-gram

1 Introduction

In 1976, biophysicist Michael Levitt *et al.* [1] proposed the concept of structural classes by visual inspection of the topologies of polypeptide chains from the 31 globular protein dataset. The identified protein structural classes are primarily of four classifications: All–α, All–β, α/β, and α+β. While the first two classes comprise secondary structures dominated by α–helices and β–strands, respectively, the other two classes consist of both α–helix and β–strand secondary structures with interspersed in α/β class structures and segregated in α+β class structures.

New protein structures discovered by diverse scientific communities have been submitted to protein databases, such as the Structural Classification of Proteins (SCOP). According to the latest extended version of the SCOPe 2.07 database http://scop.berkeley.edu/statistics/ver=2.07, the proteins are mainly categorized into seven classes, namely, (1) All–α, (2) All–β, (3) α/β, (4) α+β, (5) Multi-domain proteins, (6) Membrane and cell surface proteins and (7) Small proteins. Over the years, it was observed that 90 percent of these protein entries consistently belong only to the first four structural classes [2], [3], [4]. Therefore, this study mainly concentrates on predicting the first four structural classes.

Identification of protein structural class is one of the important activities of protein sequence analysis for mainly two reasons: (i) Prior knowledge of the structural class information of protein sequences enhances the prediction accuracy of several activities of sequence analysis such as DNA-binding sites [5], protein secondary structure [6], protein folds [7], [8], [9], [10], protein folding rates [11], tertiary structure prediction [12]; all these activities have potential applications in further analysis of protein functions and drug discovery [13]. (ii) Newly discovered protein sequences from the various scientific communities are consistently increasing due to rapid advancement of sequencing technology. Hence, to annotate the structural class information for newly discovered protein sequences, there is an imminent need of automated, accurate, and generalized structural class prediction models which works for all categories of sequence similarity proteins.

Earlier investigations on the identification of PSSC were carried out by experimental methods [14]. However, these methods are time-consuming and cost-intensive. To overcome the limitations of experimental methods, several computational methods have been proposed [15], [16], [17], [18], [19]. They are categorized under multiclass classification problems, which involve two major activities: (i) Feature modeling and (ii) Classification.

In feature modeling activity, the given sequences are transformed into fixed length feature vectors and relevant

- *The authors are with the Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangalore, Karnataka 575025, India. E-mail: sanjaybankapur.mit@gmail.com, nagammapatil@nitk.edu.in.*

Manuscript received 1 May 2019; revised 18 Nov. 2019; accepted 2 Mar. 2020. Date of publication 9 Mar. 2020; date of current version 8 Dec. 2021.
(Corresponding author: Sanjay Bankapur.)
Digital Object Identifier no. 10.1109/TCBB.2020.2979430

features are identified to predict the PSSC accurately. In literature, state-of-the-art models extract features mainly from sequences, structures, and evolutionary information. Sequence-based features are primarily extracted from information such as physicochemical properties of protein residues [6], [20], [21], amino acid composition (AAC) and their distribution [15], pseudo amino acid composition (PseAAC) [20], [22], [23], [24], and averaged chemical shifts [25]. The advantage of sequence-based features is that they exhibit significant discriminating information for high similarity datasets. In contrast, sequence-based features fail to discriminate classes correctly for twilight zone (low similarity) datasets [26], [27], [28].

Structure-based features are extracted from secondary structural sequences. The secondary structural sequence can be generated by mapping every amino acid residue from protein sequence to one of the secondary structure elements such as, α-Helix (H), β-Sheet (E), or Coil (C) [29]. Liu *et al.* [30] focused on designing features from structural sequences. Kong *et al.* [31] extracted features to characterize general contents and spatial arrangements of the secondary structural sequences. PSSC prediction methods [30], [31] using secondary structural sequences reported better prediction accuracy. However, these methods were not able to explore and extract highly discriminating features.

Evolution-based features are extracted from sequence profiles such as position-specific scoring matrix (PSSM) which are generated using PSI-BLAST [32]. To address PSSC prediction, various techniques are applied on evolutionary information; Zang *et al.* [33] extracted a large vector space and reduced it using the principal component analysis approach. Xia *et al.* [34] work focuses on transforming evolutionary features using the linear regression technique. Liu *et al.* [35] adopts auto-covariance transformation technique on PSSM. Dehzangi *et al.* [18] extracted features from both physicochemical properties and evolutionary information using overlapped segmented distribution and autocorrelation techniques. Zang *et al.* [36] extracted features based on evolutionary differences. Ding *et al.* [37] extracted long-range and linear correlation features from evolutionary information. Qin *et al.* [38] generated a fixed length feature vector by the linear predictive coding approach.

Other approaches to solve the PSSC problem include: Liu *et al.* [16] addressed the PSSC problem using a distance measure instead of extracting discriminating features. Yu *et al.* [17] parallelly extracted features from multiple views and fused them to form a complex feature space. Kavianpour *et al.* [19] transformed amino acid residues to binary codes based on the hydrophobicity index and then generates cellular automata images to extract features using image textural properties.

The latter activity of PSSC prediction i.e., for the classification, various state-of-the-art classifiers such as Bayesian classifier [39], Logistic Regression [27], [40], Artificial Neural Network [24], [41], [42], ensemble classifiers [18], [43], and Support Vector Machine [27], [33], [35], [36], [37], [38] have been developed for PSSC prediction. In the literature, the supervised learning techniques such as SVM and other ensemble classifiers have been widely adopted to solve the PSSC problem.

These previous studies have revealed that the protein sequences, structural sequences, and evolutionary profile information provide promising ways to improve the effectiveness of the PSSC prediction. However, these studies lack in extracting generic features since they have mainly focused on either high or low similarity datasets. Hence, there is an ample scope to extract a generic set of features from both high and low sequence similarity datasets. In this study, we proposed a generic PSSC prediction model which works effectively for all similarity datasets. The key research contributions are:

- An effective and novel feature modeling approach is proposed, consisting of three feature extraction approaches, namely, Embedding (E) technique, SkipXGram (SXG) technique, and General Statistical (GS) based feature extraction technique.
  - Embedding (E) technique extract local information of a query sequence as features based on the spatial arrangements of a given sequence in a close vicinity. A total of five embedding architectures are explored and their performances are analyzed in detail.
  - SkipXGram (SXG) technique also extracts various levels of local information of a query sequence, and the performances of various feature sets are analyzed in detail.
  - General Statistical (GS) based features cover the global information of a query sequence.
- An ensemble of three classifiers, namely, Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machines (GBM) is proposed to predict PSSC using probability-based voting.
- The performance of the proposed model is assessed against state-of-the-art models on five benchmark datasets, out of which two are high similarity, and the rest are low similarity datasets. Further, the performance consistency and robustness of the proposed model is validated on a large-scale updated dataset consisting of newly discovered sequences which are of low sequence similarity ($\leq$25%) and the prediction results are benchmarked.

The remainder of this paper is organized as follows: In Section 2, initially we discuss the datasets that are used in this study, and later we discuss on the proposed feature modeling and ensemble of classifiers in detail. In Section 3, we discuss the performance analysis of the proposed model in detail followed by statistical significance test and lastly, Section 4 concludes the paper with the prospects of future work.

2 Materials and Methodology

2.1 Datasets

To assess the performance of the proposed model against state-of-the-art models, we have considered five benchmark datasets. Further, the performance consistency and robustness of the proposed model is validated on a large-scale updated dataset.

A. Benchmark Datasets. The first dataset consists of 277 protein sequences and the second dataset consists of 498 sequences which are constructed by Zhou [44] and denoted as z277 and z498 respectively. Both the datasets, despite

TABLE 1
Data Characteristics of Five Benchmark Datasets

Dataset	Sequence Similarity	Number of Protein Sequences				
		All–α	All–β	α/β	α+β	Total
z277 [44]	High	70	61	81	65	277
z498 [44]	High	107	126	136	129	498
25PDB [43]	Low ($\leq$25%)	443	443	346	441	1673
1189 [39]	Low ($\leq$40%)	223	294	334	241	1092
FC699 [45]	Low ($\leq$40%)	130	269	377	82	858

possessing high similarity (about 80 percent), are widely used to validate the prediction models. To analyze the performance impact of the proposed model on low similarity datasets, we have considered three other benchmark datasets. 25PDB [43], 1189 [39], and FC699 [45] datasets contain 1673, 1092, and 858 protein sequences respectively and exhibit less than 40 percent sequence similarity. All the protein sequences from these five benchmark datasets are span across four classes of protein secondary structure such as All–α, All–β, α/β and α+β. Data characteristics and frequencies for each class are shown in Table 1.

B. Large-Scale Updated Dataset. To validate the performance consistency and robustness of the proposed model, we have derived a dataset consisting of high volume of newly discovered protein sequences based on two aspects, such as: (1) protein sequences that are extracted from the latest extended version of SCOP i.e., SCOPe 2.07 database [46], and (2) all the protein sequences that exhibit $\leq$25% sequence similarity. This dataset consists a total of 7906 protein sequences in which 1760, 1791, 2174 and 2181 number of sequences are from All–α, All–β, α/β and α+β classes respectively and henceforth referred as SCOPe_2.07 dataset.

2.2 Data Preparation

From a protein sequence, every amino acid residue can be predicted to one of the possible secondary structural elements such as Helix (H), Sheets (E), or Coil (C). By this, a secondary structural sequence can be generated from a query protein sequence, and several methods are available to perform the same. In this study, we adopt PSI-BLAST based secondary structure PREDiction (PSIPRED) method [29] to generate secondary structural sequences because PSIPRED (v3.2) is simple and accurate (reported 81.6 percent accuracy) in predicting secondary structural sequence [47]. The PSIPRED method incorporates two sequential feed-forward neural networks to predict the structural elements with the help of PSI-BLAST profiles [48]. An example of PSIPRED: consider a sample protein sequence, say S_1 = GEYFTLQIRGRERFEMFRELNEALELKDAQA and its corresponding structural sequence, say $StrS_1$ is CCEEEEEECHHHHHH HHHHHHHHHCCCHHCC. For all the five datasets, we prepared the structural sequence for every input protein sequence.

2.3 Feature Modeling

Feature Modeling is an important step to extract features by transforming raw protein sequences into feature vectors of fixed size which exhibit discriminating information in predicting the PSSC accurately. In this study, we propose an enhanced feature modeling approach to extract the global and local discriminating features from the amino acid sequences and generated structural sequences. The local-based features are extracted using two proposed techniques such as Embedding (E) and SkipXGram (SXG). The global-based features are extracted using General Statisctical (GS) technique.

A. Embedding Technique (E):

Embedding technique follows an unsupervised learning approach to train and generate the vector space. Word embedding is a word vectorization technique which transforms a word into a contiguous vector such that similar words are mapped in the vicinity in the vector space and the generated vectors are dense, real-values with limited lower dimensions.

In the recent past, the word embeddings have been an effective approach for extracting semantic information for various challenges of bioinformatics and health-informatics [49], [50], [51]. Moreover, in our earlier investigation to predict the protein structural classes using the Word2Vec skipgram architecture reported a satisfactory result [52].

In this study, we considered three popular word embedding models such as Word2Vec [53], GloVe [54], and fastText [55] and modified these models such that they return character embeddings, where each character is a residue of a protein sequence.

Word2Vec [53] and fastText [55] are predictive models where each model exhibits two architectures, namely, Contiguous Bag-Of-Words (CBOW) and Skip-Gram (SG). The training phase of CBOW architecture predicts the current word from a window of context words. In contrast, SG architecture predicts the window of context words from a current word. Global Vectors (GloVe) [54] is a statistical count-based model in which the model learns their vectors by training on non-zero entries in a word-word co-occurrence matrix. Both Word2Vec and GloVe models train the network by treating each word from the corpus as an atomic entity, whereas, fastText model trains by treating each word as a set of character n-grams.

To extract embedding based feature sets, we have explored a total of five embedding architectures, such as Word2Vec Contiguous Bag-Of-Words (W2V-CBOW), Word2Vec Skip-Gram (W2V-SG), GloVe, fastText Contiguous Bag-Of-Words (fastText-CBOW), and fastText Skip-Gram (fastText-SG).

Two sets of features are extracted using character embedding approach from a query protein sequence and its respective secondary structural sequence — each feature set consisting of 400 features.

The secondary structural sequence is further processed to generate a structural sequence code by removing the Coil elements (i.e., C) and by replacing the contiguous repetition of the same structural element with the combination of a total number of occurrences and its structural element. For example, the structural sequence, say $StrS_1$ = CCEEEEEECHHHHHH HHHHHHHHHHCCCHHCC after removing Coil elements will be EEEEEEHHHHHHHHHHHHHHHHH and processed to generate structural sequence code, $Str - Code_1$ = 6:E–17:H, where the contiguous repetition of segment E (i.e., EEEEEE) is replaced by its number of occurrences and the structural element (i.e., 6: E), and similar activity is performed for rest of the sequence. From the resulting structural sequence code (Str-Code) information, one more set of 400 features is extracted using the word embedding approach. The contiguous frequency and its structural element separated by a colon constitute a word, i.e., 6:E and 17:H are the two words of $Str - Code_1$. Using the embedding technique, a total of three sets, each consisting of 400 features is extracted, making it a 1200 feature vector. The 1200 embedding-based feature vector represent an effective spatial arrangements of amino acid sequences, secondary structural sequences, and structural sequence codes.

B. SkipXGram Technique (SXG):

The most common types of protein secondary structures are the α-helices (H) and the β-sheets (E). Both these structures are formed due to the hydrogen bond between two residues. Moreover, there are 3.6 residues per turn in an α-helix structure. To mimic these biological characteristics, we have extracted various skipped bi-gram feature sets by adopting SkipXGram Technique (SXG) [52]. Using SXG technique, we have extracted six sets of skipped bi-gram features from each protein sequence and secondary structural sequence.

C. General Statistical based Feature (GS):

Along with the features extracted using E and SXG techniques, a set of 9 general statistical (GS) based features are generated to cover the global information of a structural sequence and those are:

- $f_H(StrS_{H,E,C})$: The frequency of an element H from a structural sequence.
- $f_E(StrS_{H,E,C})$: The frequency of an element E from a structural sequence.
- $f_C(StrS_{H,E,C})$: The frequency of an element C from a structural sequence.
- $f_H(StrS_{H,E})$: The frequency of an element H from a structural sequence without the C elements.
- $f_E(StrS_{H,E})$: The frequency of an element E from a structural sequence without the C elements.
- $Max_H(StrS_{H,E,C})$: Ratio contributing to the maximum number of consecutive H elements in a structural sequence.
- $Max_E(StrS_{H,E,C})$: Ratio contributing to the maximum number of consecutive E elements in a structural sequence.
- $Max_C(StrS_{H,E,C})$: Ratio contributing to the maximum number of consecutive C elements in a structural sequence.
- $Length(StrSC_{H,E})$: The total length of a structural sequence code.

2.4 Ensemble of Classifiers

The prediction of PSSC is a multiclass classification problem. The relevant sets of features are extracted using the proposed feature modeling approach, and the extracted feature vectors are fed to a classification model as an input, to predict the PSSC. Majority of the existing works on the PSSC problem were carried out using a SVM classifier and have reported satisfactory prediction accuracy [37], [38].

In the literature, an ensemble of different classifiers are explored to address various challenges of protein sequence analysis [56], [57] and the recent work [18] has shown that the ensemble of classifiers is effective in addressing the PSSC problem. An ensemble of different classifiers facilitates prediction by combining the opinions of all classifiers via majority or probability-based voting. By this, the limitations of a classifier can be overcome with the strength of the other classifiers. Therefore, we explored various state-of-the-art classifier methods on all the benchmark datasets and proposed a generalized prediction model, which works for all the categories of sequence similarity datasets (i.e., high to low), by an ensemble of three classifiers in parallel such as SVM, RF (bagging) and GBM (boosting) to work as single classification model.

Based on the performance of various state-of-the-art classifiers, an ensemble of classifiers is proposed and the results are reported in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2020.2979430, under the topic "Performance Analysis of the Proposed Ensemble Classifier".

The predicted output of the proposed ensemble classifier is based on the highest probability-based voting, i.e., each classifier outputs four probability values (for four classes) for a given protein sequence. Output probabilities of each class across the three classifiers are averaged and the query protein sequence is classified to the highest average probability class.

In this study, we implemented the SVM classifier with the penalty parameter C=4.0 and the Radial Basis Function (RBF) as the kernel, since RBF relatively outperforms linear, polynomial, and sigmoid kernels due to its tolerance to input noise with generalizability. In RF classifier, it is well known that higher the number of decision trees, lower the risk of the model being subjected to over-fitting, and better the prediction accuracy. Hence, RF classifier is implemented with the number of decision trees equal to 350, a value which is identified empirically. In our experiments, it was observed that any further increase in the number of trees, did not improve the prediction accuracy. We implemented GBM classifier by choosing regression trees as weak predictors and negative gradient multinomial deviance as the loss function. Higher the number of weak predictors (boosting stages), gradient boosting is fairly robust to over-fitting. Therefore, the number of boosting stages is set to 350. It is worth noting that in all our experiments, the hyperparameters of the proposed ensemble classifier are constant, and then the prediction accuracies are recorded across all the datasets.

The proposed ensemble classifier is trained on the sets of features which are extracted and shortlisted using the proposed feature modeling for PSSC prediction. In this study, a combination of the proposed feature modeling and the proposed ensemble classifier constitute the proposed model. The overall framework of the proposed model is as shown in Fig. 1. The performance of the proposed model is discussed in the next Section.

3 Results and Analysis

As explained in Section 2, various sets of local and global features are extracted using the proposed feature modeling and are fed to the proposed ensemble of three classifiers for structural class prediction. The SXG and GS feature extraction techniques are implemented in Java, Eclipse Platform 3.8.1. The embedding approaches, such as, the Word2Vec, GloVe, and fastText are implemented in Python. The proposed

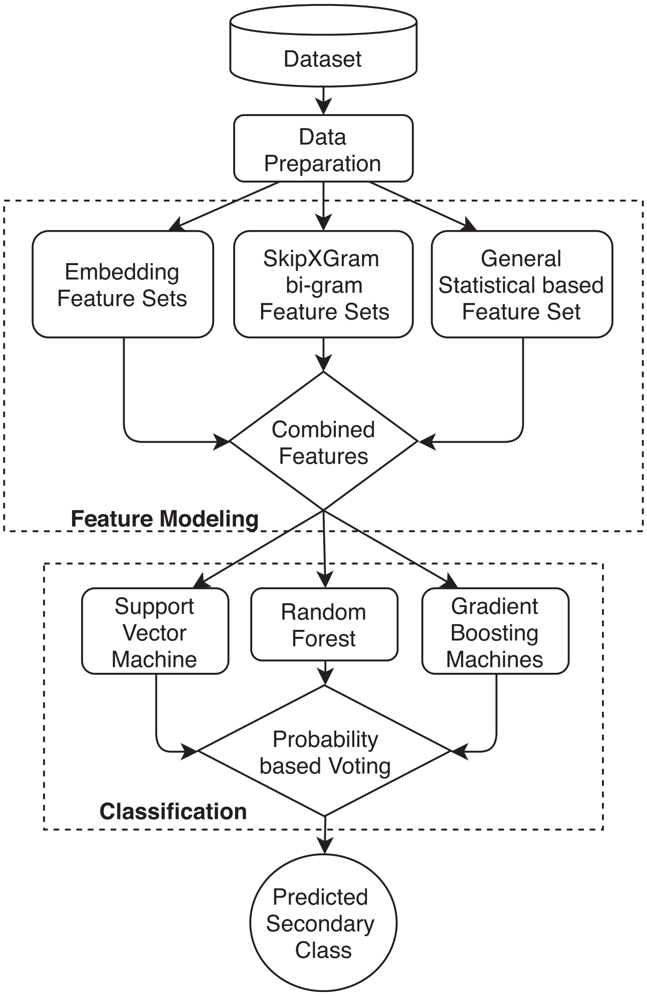

Fig. 1. Framework of the proposed model constituting feature modeling and ensemble of classifiers.

ensemble classifier is implemented in Python using Scikit-learn [58].

The performances of embedding and SkipXGram based features are analyzed separately. The results of the best performing sets of features with the proposed ensemble of classifiers on five benchmark datasets, i.e., z277, z498, 25PDB, 1189, and FC699 are benchmarked. The overall prediction performances of the proposed model on benchmark datasets are compared with the state-of-the-art methods. Further, the prediction performance of the proposed model is validated on the large-scale updated dataset, i.e., SCOPe_2.07 which consists of a high volume of newly discovered protein sequences exhibiting $\leq$25% sequence similarity.

Most of the published works on the PSSC problem are validated using the Jackknife approach [35], [36], [37]. It was observed that cross-validation evaluation approach produces results similar to Jackknife [59] and recent work on PSSC [36] showed that the overall prediction accuracy using 10-fold cross-validation on most of the datasets is slightly lesser when compared to Jackknife. Therefore, in this study, we have recorded all the performance measures using 10-fold cross-validation approach. Detailed discussion on the performance analysis of various feature sets are as follows:

TABLE 2
The Performance Comparison (in Percentage) of Various Embedding Architectures on the Benchmark Datasets

Dataset	Embedding Architectures				
	W2V- CBOW	W2V- SG	GloVe	fastText- CBOW	fastText- SG
z277	86.00	87.80	86.00	**91.10**	88.54
z498	96.70	**96.76**	95.95	96.55	94.54
25PDB	75.50	74.64	72.84	**75.58**	72.73
1189	**77.01**	76.62	76.17	76.61	72.87
FC699	88.90	89.03	86.52	**89.05**	89.04
SCOPe _2.07	74.30	74.52	73.90	**74.58**	73.57

3.1 Performance Analysis of Embedding Based Feature Sets

To effectively extract embedding based features, we explored a total of five embedding architectures, such as W2V-CBOW, W2V-SG, GloVe, fastText-CBOW, and fastText-SG. From each architecture, three sets of features are extracted in which each set consists of 400 features. These three sets of features extracted from protein sequences, structural sequences, and structural sequence codes respectively, constitute to 1200 features. The detailed analysis of all the five embedding architectures over six datasets is reported in the supplementary material, available online, under the topic "Detailed Analysis of Embedding-based Features."

The overall performance comparisons of all the five embedding architectures across six benchmark datasets are tabulated in the Table 2. The results are obtained using 10-fold cross validation approach.

We can observe that both the Word2Vec and fastText embedding approaches reported better prediction accuracy when compared to GloVe embedding approach. This is mainly due to the fact that being both a neural-network based prediction model train and extract comparatively highly discriminating sets of features than GloVe architecture.

From the predictive based embedding models, we observe that the CBOW architecture can extract better sets of features for low similarity datasets when compared to its respective SG architectures. The CBOW architecture, trains the embedding model by predicting the current word from a window of context words, thus minimizing the training error. For high sequence similarity datasets, both CBOW and SG architectures can extract highly discriminating features equally.

Further, it is worth noting that the performance of the fastText-CBOW architecture outperformed other embedding architectures for most of the benchmark datasets. This is mainly because the fastText treats every input word (in case of structural sequence code) as a set of characters and it is able to train the model effectively. Therefore, we have considered and shortlisted only fastText-CBOW sets of features (1200) for further analysis. The performance variations of all the five embedding architectures across six datasets are plotted, and is shown in Fig. 2.

3.2 Performance Analysis of SkipXGram Based Feature Sets

Six sets of features were extracted (for X: 0 to 5) on amino acid sequences using the SkipXGram technique. All the feature sets are evaluated individually with the proposed ensemble of classifiers. The overall prediction accuracy

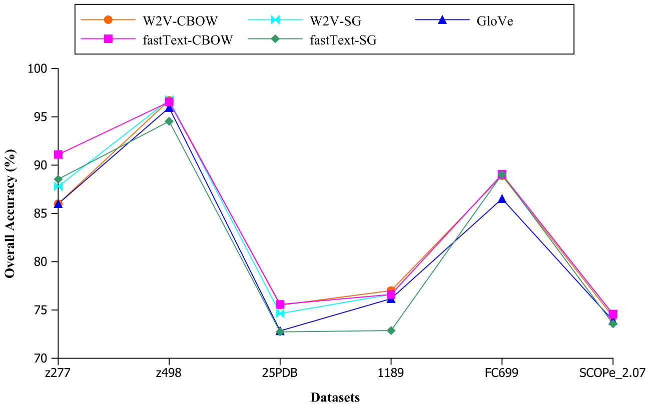

Fig. 2. The performance variations of embedding architectures on benchmark datasets.

obtained using 10-fold cross-validation on all the five datasets are shown in Fig. 3 where the x-axis represents different values of X (0 to 5), and the y-axis indicates the average prediction accuracy. From Fig. 3, it can be observed that the three skipped bi-gram feature set (i.e., X=3 or S3G) reported the highest overall prediction accuracy for all the five datasets.

A similar analysis was carried out for the other six sets of features that are extracted from secondary structural sequences using the SkipXGram technique. From the Fig. 4, it is observed that S3G feature set (i.e., X=3) reports the highest overall prediction accuracy when compared to other feature sets. Therefore, we can say that the three skipped bi-gram feature set (i.e., S3G feature set) exhibits highly discriminating features when compared to other skip-gram bi-gram features for the selected five benchmark datasets.

From Figs. 3 and 4, below observations are made: (i) the structural class prediction accuracy is high for the high similarity datasets when compared to low similarity datasets. This is mainly due to the fact that high similarity datasets exhibit more discriminating information than low similarity datasets. Further, it is worth noting that the proposed skip-gram bi-gram technique is able to extract discriminating features for FC699 dataset (which is of low similarity) and thus able to achieve higher prediction accuracy when compared to other two low similarity datasets. (ii) the feature sets extracted from secondary structural sequences reported higher prediction accuracy across all the benchmark datasets when compared to feature sets from amino acid sequences. Hence, we can say that the secondary structural sequences possess highly discriminating information of structural class when compared to amino acid sequences across all the benchmark datasets.

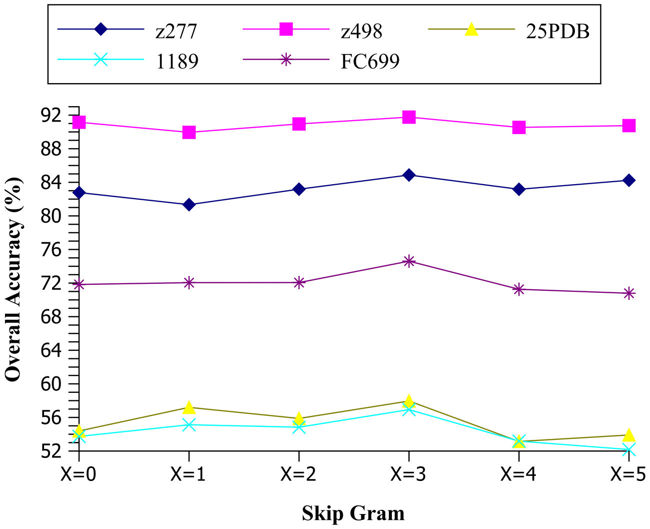

Fig. 3. Overall accuracy variations on feature sets extracted from the amino acid sequences for different skip-gram (i.e., X) values.

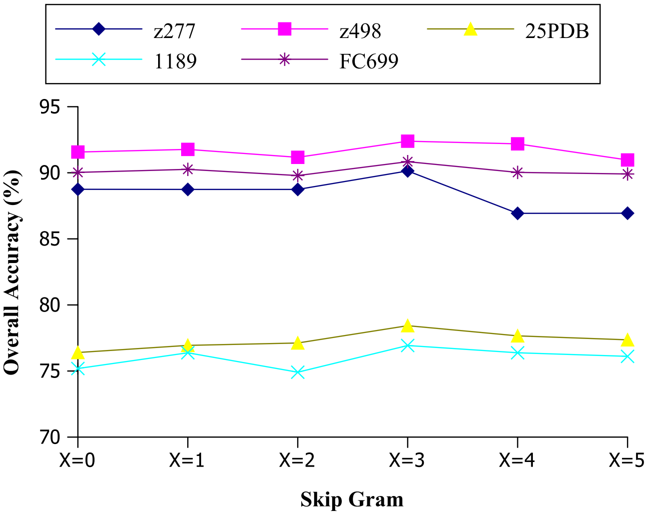

Fig. 4. Overall accuracy variations on feature sets extracted from the secondary structural sequences for different skip-gram (i.e., X) values.

3.3 Performance Analysis of the Proposed Feature Extraction Techniques

Along with the nine GS based feature set, five more feature sets were extracted and shortlisted using the proposed feature extraction techniques. Out of these five sets, three are from fastText-CBOW embedding (E) architecture consisting of 1200 features, and the other two sets of features are from SXG technique (for X=3) consisting of 409 features. We categorized these six feature sets into two groups as E (containing three feature sets) and SXG-GS (containing three feature sets). For both E and SXG-GS, the overall prediction accuracy using 10-fold cross-validation on all the five datasets are recorded and shown in the Table 3. From the Table 3 it is observed that the SXG-GS features consistently reported higher prediction accuracy when compared to E features for low sequence similarity datasets. However, E features consistently reported better prediction accuracy than SXG-GS features for high sequence similarity datasets. Since the E

TABLE 3
The Impact Analysis of the Proposed Feature Extraction Techniques on Benchmark Datasets Using Ensemble of Classifiers

Feature Extraction Techniques	Overall Accuracy (%)				
	z277	z498	25PDB	1189	FC699
SXG-GS	90.20	92.67	78.85	77.19	91.84
E	91.10	96.55	75.58	76.61	89.05
E-SXG-GS	**93.55**	**97.58**	**81.82**	**81.12**	**93.93**

TABLE 4
The Various Performance Metrics Result of the Proposed Model on Benchmark Datasets Using 10-Fold Cross Validation

Dataset	Class	Sens (%)	Spec (%)	MCC	Overall Accuracy
z277	All–α	94.29	94.29	0.9225	93.55
	All–β	95.09	95.08	0.9361	
	α/β	96.30	95.12	0.9381	
	α+β	87.70	89.06	0.8479	
z498	All–α	96.27	98.10	0.9640	97.58
	All–β	98.42	99.20	0.9840	
	α/β	99.26	96.43	0.9700	
	α+β	96.13	96.88	0.9530	
25PDB	All–α	91.43	88.62	0.8556	81.82
	All–β	82.40	89.24	0.8009	
	α/β	78.62	79.53	0.7291	
	α+β	74.15	70.32	0.6145	
1189	All–α	88.78	86.08	0.8345	81.12
	All–β	87.41	92.77	0.8576	
	α/β	82.33	81.12	0.7258	
	α+β	64.70	63.41	0.5336	
FC699	All–α	98.46	97.00	0.9727	93.93
	All–β	94.80	96.59	0.9364	
	α/β	95.76	94.01	0.9069	
	α+β	75.61	79.49	0.7517	

technique has the ability to determine similar residues that are in close vicinity in the spatial arrangements of protein sequences, it tends to perform better for high sequence similarity datasets.

Further, the combined set of features (i.e., E-SXG-GS), consisting of 1618 features, reported the highest overall prediction accuracy for all the five datasets when compared to individual feature sets. Thus, in this study, the proposed feature modeling combines all the six sets (E-SXG-GS) of highly discriminating features.

Along with the combined set of features (i.e., E-SXG-GS) from the proposed feature modeling, evolutionary features are also explored. Evolutionary features are extracted from the Position-Specific Scoring Matrix (PSSM) which are generated using the PSI-BLAST profiles [32]. The evolutionary features did not improve the prediction accuracy when compared to the proposed combined set of features (i.e., E-SXG-GS). The detailed analysis of this study is available in supplementary material, available online, under the topic “Evolutionary-Based Features: Position-Specific Scoring Matrix”.

The proposed combined set of features (i.e., E-SXG-GS) is further explored using the Feed-forward Deep Neural Network. It is worth noting that the overall accuracy of the proposed ensemble of classifiers outperformed the Feed-forward Deep Neural Network on all the benchmark datasets. The detailed analysis of this study is available in supplementary material, available online, under the topic “Deep Learning: Feed-forward Deep Neural Network”.

3.4 Performance Analysis of the Proposed Model With State-of-the-Art Models

The proposed model effectively combines the proposed feature modeling and an ensemble of three classifiers. The performance of the proposed model was evaluated and benchmarked using four standard metrics such as Sensitivity (Sens), Specificity (Spec), Matthews correlation coefficient (MCC), and Overall Accuracy. Sensitivity and specificity measures the proportion of actual positives and actual negatives that are correctly identified; Overall Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. These three metrics are measured in percentage. MCC takes into account of true positives, false positives, true negatives, and false negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC value ranges from -1 to 1, where 0 indicates random correlation, the higher negative value (i.e., -1) indicates poor prediction quality, and a higher positive value (i.e., +1) indicates better prediction quality. These evaluation metrics are mathematically defined as follows:

$$Sens = \frac{TP}{TP + FN} \tag{1}$$

$$Spec = \frac{TN}{FP + TN} \tag{2}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3}$$

$$OverallAccuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4}$$

Where TP, FP, TN, and TN are the total number of true positives, false positives, true negatives, and false negatives respectively. The detailed results that are obtained by 10-fold cross-validation on all the datasets are shown in Table 4.

The proposed model reported above 93 percent prediction accuracy for both high similarity datasets, as well as for FC699 dataset which is a low similarity dataset. For 25PDB and 1189 datasets, the overall prediction accuracy of the proposed model is consistently reported above 81 percent.

The performance of the proposed model is compared with more than 20 different state-of-the-art models across the five benchmark datasets. It is to be noted that none of the state-of-the-art models have benchmarked their performances on all the five datasets and neither have they made their implementations available to the community. Hence, the performance of the proposed model is compared with these models’ results from their respective papers.

The proposed model reported an overall accuracy of 93.55 and 97.58 percent for high similarity datasets, i.e., z277 and z498 respectively. From the Tables 5 and 6, it can be observed that the proposed model outperformed all the state-of-the-art models by a maximum margin of around 10 percent on z277 and around 4 percent on z498 datasets. The second-best performance for these datasets was reported by Kavianpour *et al.* [19] in 2017, where they convert amino acid sequences into binary codes to build cellular automata images. Further, the texture-based features were extracted using Harlick approach to predict PSSC. It is worth noting that the activities involved in Kavianpour *et al.* [19] to extract features from sequences via automata images is not only computationally expensive but also it is more suitable for high similarity datasets only. The proposed model outperforms the Kavianpour *et al.* [19] method for both the

TABLE 5
The Performance Comparison (in Percentage) of the Proposed Model Against State-of-the-Art Methods for z277 Dataset

Models	All–α	All–β	α/β	$\alpha+\beta$	Overall Accuracy
IGA-SVM (2008) [60]	84.30	85.50	92.60	70.70	84.50
IB1 (2008) [61]	89.70	88.10	92.20	80.00	87.70
CWT-PCA-SVM (2009) [23]	85.70	90.20	87.70	80.10	85.90
Information Theoretical (2010) [62]	87.10	80.30	93.80	67.70	83.00
NN-CDM (2010) [16]	80.00	86.40	91.60	81.80	85.20
AAC-PSSM-AC (2012) [35]	88.60	95.10	97.50	81.50	91.00
LZ-BMKL (2013) [63]	92.90	85.30	92.60	69.20	85.60
COMSPA (2013) [17]	86.10	87.30	91.30	82.30	87.00
Dehzangi *et al.* (2013) [18]	90.00	93.40	80.00	96.30	90.30
PSSM-LPC (2015) [38]	91.40	90.10	92.50	78.40	88.40
Kavianpour *et al.* (2017) [19]	92.07	93.35	93.47	90.46	92.34
PMCI-RFE (2018) [64]	-	-	-	-	84.43
Proposed Model (this study)	94.29	95.09	96.30	87.77	**93.50**

TABLE 6
The Performance Comparison (in Percentage) of the Proposed Model Against State-of-the-Art Methods for z498 Dataset

Models	All–α	All–β	α/β	$\alpha+\beta$	Overall Accuracy
IGA-SVM (2008) [60]	96.30	93.60	97.80	89.20	94.20
IB1 (2008) [61]	94.95	95.83	97.81	94.16	95.74
CWT-PCA-SVM (2009) [23]	94.40	96.80	97.00	92.30	95.20
NN-CDM (2010) [16]	96.30	93.70	95.60	89.90	93.80
Information Theoretical (2010) [62]	95.30	93.70	97.80	88.30	93.80
AAC-PSSM-AC (2012) [35]	94.40	96.80	97.80	93.80	95.80
LZ-BMKL (2013) [63]	96.30	94.40	96.30	93.80	95.20
COMSPA (2013) [17]	95.20	97.60	98.50	90.50	95.40
Dehzangi *et al.* (2013) [18]	95.30	97.60	96.10	97.80	96.80
PSSM-LPC (2015) [38]	99.10	96.80	97.80	93.80	96.70
Kavianpour *et al.* (2017) [19]	96.58	98.49	97.67	96.50	97.31
PMCI-RFE (2018) [64]	-	-	-	-	93.84
Proposed Model (this study)	96.27	98.42	99.26	96.13	**97.58**

TABLE 7
The Performance Comparison (in Percentage) of the Proposed Model Against State-of-the-Art Methods for 25PDB Dataset

Models	All–α	All–β	α/β	$\alpha+\beta$	Overall Accuracy
Specific Tri-Peptides (2009) [65]	60.60	60.70	67.90	44.30	58.60
AAD-CGR (2009) [66]	64.30	65.00	65.00	61.70	64.00
CWT-PCA-SVM (2009) [23]	76.50	67.30	66.80	45.80	64.00
AADP-PSSM (2010) [67]	83.30	78.10	76.30	54.40	72.90
AATP (2012) [33]	81.90	74.70	75.10	55.80	71.70
AAC-PSSM-AC (2012) [35]	85.30	81.70	73.70	55.30	74.10
Xia *et al.* (2012) [34]	92.60	72.50	71.70	71.00	77.20
Dehzangi *et al.* (2013) [18]	86.10	80.80	60.10	80.60	76.70
MEDP (2014) [36]	87.81	78.33	76.01	57.37	74.84
EEDP (2014) [36]	88.04	78.56	78.03	57.14	75.31
LCC-PSSM (2014) [37]	91.70	80.80	79.80	64.00	79.00
PSSM-LPC (2015) [38]	87.40	81.70	75.10	57.60	75.50
MBMGAC-PSSM (2015) [68]	86.70	81.50	79.50	61.70	77.20
Proposed Model (this study)	91.43	82.40	78.62	74.15	**81.82**

TABLE 8
The Performance Comparison (in Percentage) of the Proposed Model Against State-of-the-Art Methods for 1189 Dataset

Models	All–α	All–β	α/β	$\alpha+\beta$	Overall Accuracy
IB1 (2008) [61]	65.30	67.73	79.93	40.68	64.65
Specific Tri-Peptides (2009) [65]	-	-	-	-	59.90
AAD-CGR (2009) [66]	62.30	67.70	66.50	63.10	65.20
AADP-PSSM (2010) [67]	69.10	83.70	85.60	35.70	70.70
AATP (2012) [33]	72.70	85.40	82.90	42.70	72.60
AAC-PSSM-AC (2012) [35]	80.70	86.40	81.40	45.20	74.60
COMSPA (2013) [17]	73.25	77.26	76.34	54.91	72.53
Dehzangi *et al.* (2013) [18]	80.20	83.60	44.60	85.40	75.82
EEDP (2014) [36]	84.75	81.97	82.04	47.72	75.00
MEDP (2014) [36]	85.20	84.01	84.43	45.23	75.80
LCC-PSSM (2014) [37]	89.20	88.80	85.60	58.50	**81.20**
PSSM-LPC (2015) [38]	82.10	86.30	82.60	43.70	74.90
MBMGAC-PSSM (2015) [68]	79.80	850	84.70	50.60	76.30
PMCI-RFE (2018) [64]	-	-	-	-	62.37
Proposed Model (this study)	88.79	87.41	82.34	64.73	81.12

TABLE 9
The Performance Comparison (in Percentage) of the Proposed Model Against State-of-the-Art Methods for FC699 Dataset

Models	All–α	All–β	α/β	$\alpha+\beta$	Overall Accuracy
SCPRED (2008) [45]	-	-	-	-	87.50
Kong *et al.* (2014) [31]	96.20	90.70	96.30	69.50	92.00
PMCI-RFE (2018) [64]	-	-	-	-	82.58
SVM-RFE (2018) [64]	-	-	-	-	83.06
Proposed Model (this study)	98.46	94.80	95.76	75.61	**93.93**

datasets. Thus, the proposed feature modeling is efficient and effective in the prediction of the PSSC.

For low similarity datasets, the proposed model reported promising results with an overall accuracy of 81.82, 81.12 and 93.93 percent for 25PDB, 1189, and FC699 datasets respectively and it is shown in Tables 7, 8, and 9.

The proposed model outperformed all the state-of-the-art models by a minimum margin of around 3 percent and the maximum margin of around 22 percent for the 25PDB dataset as shown in Table 7. The best model from the literature, i.e., LCC-PSSM [37] reported the overall accuracy of 79 percent where 3600 features were extracted using the linear correlation coefficient approach on PSI-BLAST profiles and 278 features were selected to predict PSSC problem. It is worth noting that the proposed model extracts 50 percent lesser number of features than the LCC-PSSM [37] and outperforms the LCC-PSSM [37] model by a factor of around 3 percent. Moreover, the proposed model's accuracy in predicting $\alpha+\beta$ class has been improved by more than 10 percent compared to the LCC-PSSM.

For 1189 dataset, the proposed model outperforms all the state-of-the-art models except the LCC-PSSM [37] as shown in Table 8. However, the proposed model's accuracy in predicting $\alpha+\beta$ class has been improved by more than 6 percent compared to the LCC-PSSM [37]. It is worth noting that many efforts have been carried out to improve the prediction accuracy for $\alpha+\beta$ class prediction and it remains a challenge for the low sequence similarity datasets.

For FC699 dataset, the performance of the proposed model outperforms the state-of-the-art models by a minimum factor of around 2 percent and a maximum of around 10 percent as shown in Table 9. The best performing model from the literature on FC699 dataset was reported by Kong *et al.* [31] in 2014. It can be observed that the proposed

TABLE 10
The Various Performance Metrics Result of the Proposed Model on Large-Scale Updated Dataset Using 10-Fold Cross-Validation

Dataset	Class	Sens (%)	Spec (%)	MCC	Overall Accuracy (%)
SCOPe_2.07 (Similarity $\leq 25\%$)	All–α	92.44	91.97	0.8941	**81.11**
	All–β	80.07	88.25	0.7892	
	α/β	82.47	79.83	0.7287	
	$\alpha+\beta$	71.48	68.80	0.5810	

feature modeling is effective by a factor of 6 percent in predicting α+β class when compared to Kong *et al.* [31].

The accuracy improvement in predicting α+β class for low similarity datasets is mainly due to the discriminating features which are extracted using SkipXGram and fastText embedding techniques. Thus, we can state that the proposed model is effective in predicting PSSC for both low and high similarity datasets and achieves promising results.

3.5 Performance Analysis of the Proposed Model on Large-Scale Updated Dataset

In the previous subsection, the performance of the proposed model was evaluated on the benchmark datasets which consisted of fewer volume sequences and did not include newly discovered sequences. To evaluate the robustness of the proposed model, we have carried out experiments on the SCOPe_2.07 dataset. The characteristics of this dataset are available under Large-Scale Updated Dataset in Section 2.1.B.

The prediction results of the proposed model on SCOPe_2.07 dataset using 10-fold cross-validation is tabulated in Table 10. The SCOPe_2.07 dataset consists of a high volume of protein sequences (i.e., 7906) and exhibits a low sequence similarity of $\leq$25%. The overall accuracy of the proposed model on SCOPe_2.07 dataset reported 81.11 percent, and the results are consistently in par with the results of the 25PDB dataset which is also of $\leq$25% similarity. By this, we can say that the proposed model performance is consistent and robust even for the large-scale updated dataset.

3.6 Statistical Significance Analysis

To analyze the statistical significance of the proposed model, we performed paired *t*-test on the overall prediction accuracy among the proposed model and state-of-the-art models. Since no state-of-the-art models were evaluated on all the five benchmark datasets, we have considered six state-of-the-art models (AAC-PSSM-AC [35], Dehzangi *et al.* [18], PSSM-LPC [38], PMCI-RFE [64], CWT-PCA-SVM [23] and COMSPA [17]) which have reported overall prediction accuracy on a minimum of any three benchmark datasets out of five.

The results of paired *t*-test among the proposed and each of these six state-of-the-art models with a significance level of 5 percent (i.e., 0.05) are as follows:

t value is -7.602135 and $p \leq 0.00001$ for AAC-PSSM-AC [35], t value is -7.184642 and $p \leq 0.00001$ for Dehzangi *et al.* [18], t value is -8.016546 and $p \leq 0.00001$ for PSSM-LPC [38], t value is -10.935406 and $p \leq 0.00001$ for PMCI-RFE [64], t value is -7.181012 and $p \leq 0.00001$ for CWT-PCA-SVM [23], and for COMSPA [17] t value is -7.45969 and $p \leq 0.00001$.

For the above mentioned *t*-test results, the null hypothesis of all the six cases are rejected. Hence, the overall prediction accuracy of the proposed model is statistically significant than that of the state-of-the-art models.

4 Conclusion and Future Work

The protein secondary structural class prediction plays an important role in analyzing and identifying protein folds, protein tertiary structures, and protein functions. To address the PSSC prediction problem, we have proposed a generic approach which predicts the PSSC effectively for both high and low similarity datasets. The proposed model consists of an enhanced feature modeling with ensemble of three classifiers. The proposed feature modeling consists of three feature extraction techniques such as Embedding (E), SkipXGram (SXG), and General Statistical (GS) based feature extraction technique. As a part of feature modeling, various sets of features were extracted using the proposed feature extraction techniques and finally, six effective sets of features, constituting a total of 1618 features, were shortlisted. The prediction performance of these extracted sets of features was analyzed in detail using an ensemble of three classifiers (i.e., SVM, RF, and GBM). The proposed model reported 93.55 and 97.58 percent overall accuracy for high similarity datasets namely, z277 and z498 respectively. For low sequence similarity datasets, the proposed model attained 81.82, 81.12 and 93.93 percent on 25PDB, 1189, and FC699 datasets respectively. The performance of the proposed model reported the highest overall accuracy across various benchmark datasets and outperformed all the state-of-the-art models for both low and high similarity datasets. Further, the assessment of the proposed model on the large-scale updated dataset, i.e., SCOPe_2.07 showed that the performance of the proposed model is consistent and robust even for the high volumes of protein sequences. From statistical paired t-test results, it was observed that the overall accuracy of the proposed model significantly outperformed state-of-the-art models. Hence, we conclude that the proposed model is effective and robust in solving the PSSC problem. In the future, we would like to investigate the feature reduction approaches to minimize the training time and to enhance the prediction accuracy by pruning conflicting, redundant, and irrelevant features. We would also like to explore the suitability of the proposed model by consuming the predicted classification output as one of the input features for protein fold recognition problem.

Acknowledgments

This work was funded [KSTePS/VGST-RGS/F/GRD No.727/2017-18] by Vision Group on Science and Technology, Department of Information Technology, Biotechnology and Science & Technology, Govt. of Karnataka, India.

References

[1] M. Levitt and C. Chothia, “Structural patterns in globular proteins,” *Nature*, vol. 261, no. 5561, 1976, Art. no. 552.

[2] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, “SCOP: A structural classification of proteins database for the investigation of sequences and structures,” *J. Mol. Biol.*, vol. 247, no. 4, pp. 536–540, 1995.

[3] A. Andreeva, D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin, "Scop database in 2004: Refinements integrate structure and sequence family data," *Nucleic Acids Res.*, vol. 32, no. suppl 1, pp. D226–D229, 2004.

[4] A. Andreeva *et al.*, "Data growth and its impact on the scop database: New developments," *Nucleic Acids Res.*, vol. 36, no. suppl_1, pp. D419–D425, 2007.

[5] I. B. Kuznetsov, Z. Gou, R. Li, and S. Hwang, "Using evolutionary and structural information to predict dna-binding sites on dna-binding proteins," *PROTEINS: Struct. Function Bioinf.*, vol. 64, no. 1, pp. 19–27, 2006.

[6] I. Rahal and J. Walz, "Secondary protein structure prediction combining protein structural class, relative surface accessibility, and contact number," *Int. J. Data Sci.*, vol. 3, no. 1, pp. 68–85, 2018.

[7] R. Z. Aram and N. M. Charkari, "A two-layer classification framework for protein fold recognition," *J. Theor. Biol.*, vol. 365, pp. 32–39, 2015.

[8] G. Raicar, H. Saini, A. Dehzangi, S. Lal, and A. Sharma, "Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids," *J. Theor. Biol.*, vol. 402, pp. 117–128, 2016.

[9] W. Ibrahim and M. S. Abadeh, "Extracting features from protein sequences to improve deep extreme learning machine for protein fold recognition," *J. Theor. Biol.*, vol. 421, pp. 1–15, 2017.

[10] W. Ibrahim and M. S. Abadeh, "Protein fold recognition using deep kernelized extreme learning machine and linear discriminant analysis," *Neural Comput. Appl.*, vol. 31, pp. 4201–4214, 2018.

[11] M. M. Gromiha, "A statistical model for predicting protein folding rates from amino acid sequence with structural class information," *J. Chem. Inf. Model.*, vol. 45, no. 2, pp. 494–501, 2005.

[12] L. Carlacci, K. C. Chou, and G. M. Maggiora, "A heuristic approach to predicting the tertiary structure of bovine somatotropin," *Biochemistry*, vol. 30, no. 18, pp. 4389–4398, 1991.

[13] K.-C. Chou, D.-Q. Wei, Q.-S. Du, S. Sirois, and W.-Z. Zhong, "Progress in computational approach to drug development against SARS," *Current Med. Chem.*, vol. 13, no. 27, pp. 3263–3270, 2006.

[14] S. W. Provencher and J. Gloeckner, "Estimation of globular protein secondary structure from circular dichroism," *Biochemistry*, vol. 20, no. 1, pp. 33–37, 1981.

[15] P. Klein and C. Delisi, "Prediction of protein structural class from the amino acid sequence," *Biopolymers*, vol. 25, no. 9, pp. 1659–1672, 1986.

[16] T. Liu, X. Zheng, and J. Wang, "Prediction of protein structural class using a complexity-based distance measure," *Amino Acids*, vol. 38, no. 3, pp. 721–728, 2010.

[17] D.-J. Yu *et al.*, "Learning protein multi-view features in complex space," *Amino Acids*, vol. 44, no. 5, pp. 1365–1379, 2013.

[18] A. Dehzangi, K. Paliwal, A. Sharma, O. Dehzangi, and A. Sattar, "A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 3, pp. 564–575, May 2013.

[19] H. Kavianpour and M. Vasighi, "Structural classification of proteins using texture descriptors extracted from the cellular automata image," *Amino Acids*, vol. 49, no. 2, pp. 261–271, 2017.

[20] T.-L. Zhang, Y.-S. Ding, and K.-C. Chou, "Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern," *J. Theor. Biol.*, vol. 250, no. 1, pp. 186–193, 2008.

[21] E. Contreras-Torres, "Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chou's PseAAC," *J. Theor. Biol.*, vol. 454, pp. 139–145, 2018.

[22] C. Chen, Y.-X. Tian, X.-Y. Zou, P.-X. Cai, and J.-Y. Mo, "Using pseudo-amino acid composition and support vector machine to predict protein structural class," *J. Theor. Biol.*, vol. 243, no. 3, pp. 444–448, 2006.

[23] Z.-C. Li, X.-B. Zhou, Z. Dai, and X.-Y. Zou, "Prediction of protein structural classes by Chou's pseudo amino acid composition: Approached using continuous wavelet transform and principal component analysis," *Amino Acids*, vol. 37, no. 2, 2009, Art. no. 415.

[24] S. S. Sahu and G. Panda, "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," *Comput. Biol. Chem.*, vol. 34, no. 5, pp. 320–327, 2010.

[25] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, and L. Hao, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowl.-Based Syst.*, vol. 163, pp. 787–793, 2019.

[26] L. Kurgan and L. Homaeian, "Prediction of secondary protein structure content from primary sequence alone–A feature selection based approach," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.*, 2005, pp. 334–345.

[27] L. Kurgan and K. Chen, "Prediction of protein structural class for the twilight zone sequences," *Biochem. Biophys. Res. Commun.*, vol. 357, no. 2, pp. 453–460, 2007.

[28] M. J. Mizianty and L. Kurgan, "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences," *BMC Bioinf.*, vol. 10, no. 1, 2009, Art. no. 414.

[29] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000.

[30] T. Liu and C. Jia, "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information," *J. Theor. Biol.*, vol. 267, no. 3, pp. 272–275, 2010.

[31] 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," *J. Theor. Biol.*, vol. 344, pp. 12–18, 2014.

[32] S. F. Altschul *et al.*, "Gapped blast and psi-blast: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[33] S. Zhang, F. Ye, and X. Yuan, "Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM," *J. Biomol. Struct. Dyn.*, vol. 29, no. 6, pp. 1138–1146, 2012.

[34] X.-Y. Xia, M. Ge, Z.-X. Wang, and X.-M. Pan, "Accurate prediction of protein structural class," *PLoS One*, vol. 7, no. 6, 2012, Art. no. e37653.

[35] T. Liu, X. Geng, X. Zheng, R. Li, and J. Wang, "Accurate prediction of protein structural class using auto covariance transformation of psi-blast profiles," *Amino Acids*, vol. 42, no. 6, pp. 2243–2249, 2012.

[36] L. Zhang, X. Zhao, and L. Kong, "Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 355, pp. 105–110, 2014.

[37] S. Ding, S. Yan, S. Qi, Y. Li, and Y. Yao, "A protein structural classes prediction method based on PSI-BLAST profile," *J. Theor. Biol.*, vol. 353, pp. 19–23, 2014.

[38] Y. Qin, X. Zheng, J. Wang, M. Chen, and C. Zhou, "Prediction of protein structural class based on linear predictive coding of PSI-BLAST profiles," *Open Life Sci.*, vol. 10, no. 1, pp. 529–536, 2015.

[39] Z.-X. Wang and Z. Yuan, "How good is prediction of protein structural class by the component-coupled method?" *Proteins: Struct. Function Bioinf.*, vol. 38, no. 2, pp. 165–175, 2000.

[40] 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 Recognit.*, vol. 39, no. 12, pp. 2323–2343, 2006.

[41] Y.-D. Cai and G.-P. Zhou, "Prediction of protein structural classes by neural network," *Biochimie*, vol. 82, no. 8, pp. 783–785, 2000.

[42] L. Ningbo and H. Hua, "An artificial neural network classifier for the prediction of protein structural classes," *Int. J. Current Eng. Technol.*, vol. 7, no. 3, pp. 946–952, 2017.

[43] K. D. Kedarisetti, L. Kurgan, and S. Dick, "Classifier ensembles for protein structural class prediction with varying homology," *Biochem. Biophys. Res. Commun.*, vol. 348, no. 3, pp. 981–988, 2006.

[44] G.-P. Zhou, "An intriguing controversy over protein structural class prediction," *J. Protein Chem.*, vol. 17, no. 8, pp. 729–738, 1998.

[45] L. Kurgan, K. Cios, and K. Chen, "SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences," *BMC Bioinf.*, vol. 9, no. 1, 2008, Art. no. 226.

[46] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, "SCOPe: Structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D304–D309, 2013.

[47] T. Nugent, S. Ward, and D. T. Jones, "The MEMPACK alpha-helical transmembrane protein structure prediction server," *Bioinformatics*, vol. 27, no. 10, pp. 1438–1439, 2011.

[48] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, 1999.

[49] N. Q. K. Le, E. K. Y. Yapp, Q.-T. Ho, N. Nagasundaram, Y.-Y. Ou, and H.-Y. Yeh, "iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding," *Analyt. Biochem.*, vol. 571, pp. 53–61, 2019.

[50] G. S. Krishnan *et al.*, "Evaluating the quality of word representation models for unstructured clinical text based ICU mortality prediction," in *Proc. 20th Int. Conf. Distrib. Comput. Netw.*, 2019, pp. 480–485.

[51] X. He, L. Li, Y. Liu, X. Yu, and J. Meng, "A two-stage biomedical event trigger detection method integrating feature selection and word embeddings," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 4, pp. 1325–1332, Jul./Aug. 2018.

[52] S. Bankapur and N. Patil, "Protein secondary structural class prediction using effective feature modeling and machine learning techniques," in *Proc. IEEE 18th Int. Conf. Bioinf. Bioeng.*, 2018, pp. 18–21.

[53] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arXiv:1301.3781

[54] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[55] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016. [Online]. Available: https://arxiv.org/abs/1607.01759

[56] A. Dehzangi, S. P. Amnuaisuk, K. H. Ng, and E. Mohandesi, "Protein fold prediction problem using ensemble of classifiers," in *Proc. Int. Conf. Neural Inf. Process.*, 2009, pp. 503–511.

[57] A. Dehzangi, S. Phon-Amnuaisuk, and O. Dehzangi, "Enhancing protein fold prediction accuracy by using ensemble of different classifiers," *Australian J. Intell. Inf. Process. Syst.*, vol. 26, no. 4, pp. 32–40, 2010.

[58] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[59] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *Amer. Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

[60] Z. Li, X. Zhou, Y. Lin, and X. Zou, "Prediction of protein structure class by coupling improved genetic algorithm and support vector machine," *Amino Acids*, vol. 35, no. 3, pp. 581–590, 2008.

[61] K. Chen, L. A. Kurgan, and J. Ruan, "Prediction of protein structural class using novel evolutionary collocation-based sequence representation," *J. Comput. Chem.*, vol. 29, no. 10, pp. 1596–1604, 2008.

[62] X. Zheng, C. Li, and J. Wang, "An information-theoretic approach to the prediction of protein structural class," *J. Comput. Chem.*, vol. 31, no. 6, pp. 1201–1206, 2010.

[63] Z. Mao, G.-S. Han, and T.-T. Wang, "Effects of amino acid classification on prediction of protein structural classes," in *Proc. 10th Int. Conf. Fuzzy Syst. Knowl. Discov.*, 2013, pp. 718–723.

[64] M. Yuan, Z. Yang, G. Huang, and G. Ji, "A novel feature selection method to predict protein structural class," *Comput. Biol. Chem.*, vol. 76, pp. 118–129, 2018.

[65] S. Costantini and A. M. Facchiano, "Prediction of the protein structural class by specific peptide frequencies," *Biochimie*, vol. 91, no. 2, pp. 226–229, 2009.

[66] J.-Y. Yang, Z.-L. Peng, Z.-G. Yu, R.-J. Zhang, V. Anh, and D. Wang, "Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation," *J. Theor. Biol.*, vol. 257, no. 4, pp. 618–626, 2009.

[67] T. Liu, X. Zheng, and J. Wang, "Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile," *Biochimie*, vol. 92, no. 10, pp. 1330–1334, 2010.

[68] Y. Liang, S. Liu, and S. Zhang, "Prediction of protein structural class based on different autocorrelation descriptors of position-specific scoring matrix," *MATCH: Commun. Math. Comput. Chem.*, vol. 73, no. 3, pp. 765–784, 2015.

Sanjay Bankapur (Student Member, IEEE) received the bachelors of engineering degree in computer science from Visvesvaraya Technological University, Karnataka, India, in 2005, and the master of technology degree in computer science and engineering from IIIT Hyderabad, Hyderabad, India, in 2013. He has several years of work experience in software development industry. Currently, he is a doctoral research scholar with the National Institute of Technology Karnataka, Surathkal, India. His research interests include algorithms, data mining, soft computing, machine learning, optimization, statistics, computational biology.

Nagamma Patil received the PhD degree in computer science and engineering from the Indian Institute of Technology Roorkee, Roorkee, India. She is currently with the Department of Information Technology, National Institute of Technology Karnataka, Surathkal, India. Her research interests include data mining, soft computing, big data analytics, machine learning, and bioinformatics. She received a Grant from the Vision Group on Science and Technology, Government of Karnataka, in 2018 for her work on protein structure prediction. She has more than 35 research publications in reputed and peer-reviewed International Journals and Conference Proceedings.