2018 IEEE 18th International Conference on Bioinformatics and Bioengineering

Protein Secondary Structural Class Prediction using Effective Feature Modeling and Machine Learning Techniques

Sanjay Bankapur
Department of Information Technology
National Institute of Technology Karnataka, Surathkal
Mangalore, India
Email: sanjaybankapur.mit@gmail.com

Nagamma Patil
Department of Information Technology
National Institute of Technology Karnataka, Surathkal
Mangalore, India
Email: nagammapatil@nitk.ac.in

***Abstract*—Protein Secondary Structural Class (PSSC) prediction is an important step to find its further folds, tertiary structure and functions, which in turn have potential applications in drug discovery. Various computational methods have been developed to predict the PSSC, however, predicting PSSC on the basis of protein sequences is still a challenging task. In this study, we propose an effective approach to extract features using two techniques (i) SkipXGram bi-gram: in which skipped bi-gram features are extracted and (ii) Character embedded features: in which features are extracted using word embedding approach. The combined feature sets from the proposed feature modeling approach are explored using various machine learning classifiers. The best performing classifier (i.e. Random Forest) is benchmarked against state-of-the-art PSSC prediction models. The proposed model was assessed on two low sequence similarity benchmark datasets i.e. 25PDB and FC699. The performance analysis demonstrates that the proposed model consistently outperformed state-of-the-art models by a factor of 3% to 23% and 4% to 6% for 25PDB and FC699 datasets respectively.**

***Keywords*-amino acid sequence; bi-gram; character embedding; machine learning; protein secondary structural sequence; skip-gram;**

I. Introduction

Indentifying PSSC is a primary step to decode further activities like protein fold recognition, tertiary structure prediction and analysis of protein functions for drug discovery. In 1976, Levitt et al. [1] proposed the concept of structural classes on the basis of visual inspection on the topologies of polypeptide chain from 31 globular protein dataset. These protein structural classes are mainly of four i.e. All-α, All-β, α/β and α+β.

Earlier investigations on identification of PSSC was carried out by experimental methods [2]. However, it is well known that these methods are expensive and also time consuming. To overcome the limitations of experimental methods, numerous computational methods have been proposed over the decades [3], [4], [5], [6]. These computational methods are categorized under multiclass classification problems which mainly involves two activities such as (i) feature modeling and (ii) classification.

For the former activity, indentifying the relevant features which contribute to predict PSSC accurately is termed as feature modeling. The process of feature extraction and selection from protein sequences are broadly categorized into two groups, namely, sequence based and structure based features. Sequence based features are those which are extracted on composition and their distribution, such as, amino acid composition (AAC) [3], pseudo amino acid composition (PseAAC) [7] and physicochemical properties of protein residues [5]. These sequence based features provide significant discriminatory information to predict PSSC when the protein sequences exhibit high similarity among them. However, these features fail to predict for low sequence similarity datasets i.e. for twilight zone [8]. To overcome this, recent works focused extracting features based on the possible structural information [9]). Several previous studies [8], [10], [11], [12] have revealed that the sequence and structure information provide a promising way to enhance the effectiveness of PSSC prediction. However, we still feel that both sequence based and structure based features are not explored completely.

For the classification activity, various classifiers such as logistic regression [8], Bayesian classifier [13], artificial neural network (ANN) [14], [7], support vector machine (SVM) [15], [12] and ensemble classifiers [16], [11] have been developed to tackle the PSSC prediction. SVM and ensemble based classifiers, which have been widely used in the literature, have been proven to be better in comparison to other classifiers for solving the PSSC prediction.

From the literature, it is evident that the feature modeling plays a crucial role in finding the discriminating features to address PSSC prediction problem effectively. Therefore, in this study, we propose a novel and effective feature modeling approach to extract high discriminating features to predict PSSC effectively.

Rest of this paper is organized as follows: Section 2 highlights the datasets used and the data preparation part. Further, we introduce and discuss the proposed feature modeling along with the various state-of-the-art machine learning classifiers. Section 3 shows the performance analysis of the

2471-7819/18/$31.00 ©2018 IEEE
DOI 10.1109/BIBE.2018.00012

proposed model in detail. Finally, Section 4 concludes the paper with a prospect of future work.

II. Materials and Methodology

A. Benchmark Datasets

In this study, to assess the performance of the proposed model, we have considered two low sequence similarity benchmark datasets, namely, 25PDB dataset [16] and FC699 [12]. Frequency characteristics with respect to each class in respective datasets are as shown in Table I.

Table I
NUMBER OF PROTEIN SEQUENCES BELOGING TO DIFFERENT STRUCTURAL CLASSES IN THE DATASETS

Dataset	Sequence Similarity	All-α	All-β	α/β	α+β	Total
25PDB	$< 25\%$	443	443	346	441	1673
FC699	$< 40\%$	130	269	377	82	858

B. Data Preparation

Input protein sequence is converted to possible secondary structural sequence by converting every amino acid residue to either one of the structure elements such as Helix (H), Sheets (E) or Coil (C). In this study, we adopt PSIPRED method [9] to predict possible secondary structural sequence since PSIPRED prediction accuracy outperforms other existing methods. For both the datasets, we prepared the possible protein secondary structural sequence from the given input protein sequences.

C. Feature Modeling

Quality features play a major role in predicting PSSC accurately. From the given amino acid sequences and secondary structural sequences, the proposed feature modeling extracts features using two techniques i.e. SkipXGram bi-gram (SXGbg) and Character Embedding.

1) SkipXGram bi-gram (SXGbg): Protein secondary structure is mainly due to the hydrogen bonds among two amino acid molecules. Hence the proposed model concentrates and extracts on bi-gram features. Moreover, one turn of α-helix is observed to be exhibited on an average 3.6 amino acid residues [17], therefore to mimic the α-helix nature, the proposed model extracts all possible bi-grams by skipping X-grams between the two residues, where, X value varies from 0 to 5 in our experiment. By this, six sets of bi-gram features are generated from protein sequences where, each set consists of 400 features. Six more sets of bi-gram features are extracted from secondary structural sequences in which each set consists of 9 features, as secondary structural sequence is represented using three elements - H (Helix), E (Sheets) and C (Coil). Let S be a protein sequence, which is made up of amino acid residues, of length L i.e. r_0, r_1 .. , r_{L-1} where r_0 is the residue at first position and r_{L-1} is the residue at L^{th} position. From the protein sequence S, the bi-gram features are extracted and added to the SXGbg feature set and it is as shown in equation 1.

$$SXGbg = \sum_{i=X}^{L-1} r_{(i-X)} r_{(i+1)} \tag{1}$$

Where, X indicates the number of skipped grams and the values are varied from 0 to 5 to obtain six set of SXGbg features.

Algorithm to generate SXGbg sets of features is as shown in Algorithm 1.

Algorithm 1 : Proposed Algorithm to Extract SkipXGram bi-gram Features

Input: List of n protein sequences of variable length and a SkipGram value, X
Output: SXGbg feature set with the occurrence count for all n protein sequences

for each protein sequence *j* do
  L be a j^{th} sequence length
  for i=leave to L do
    bg=char at *[i-X]* and *[i+1]* from the j^{th} sequence
    if($SXGbg_j$ in [*bg*]) then
      Increment the *bg* count by 1 in $SXGbg_j$ set
    else
      Add the *bg* to $SXGbg_j$ set with count=1
  end for
end for

2) Character Embedded Features (CE): We adopted and modified the Word2Vec word embeddings technique [18] such that it generates character embeddings, where each character is a protein residue. The embedding model represents each protein sequence in to a vector of size 400. As similar to the ability of Word2Vec model to map words belonging to same domain in close proximity in the vector space, the character embedding model works in such a way that the residues which share similar characteristics in protein sequence are placed in close vicinity in the vector space.

D. Classification Techniques

The protein secondary structure class prediction is a multiclass classification problem. The quality features are extracted using the proposed feature modeling framework and from these extracted features, classification is performed using various machine learning techniques. In this study, we have considered most popular state-of-the-art machine learning classification techniques such as logistic regression, k nearest neighbor classifier, multi layer perceptron, support vector machine, gradient boosting machines and random forest to classify the given protein sequences into its respective secondary classes.

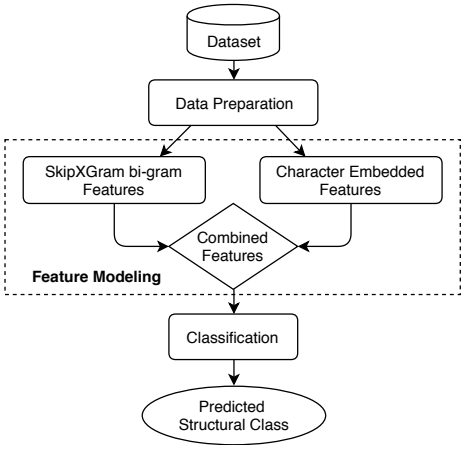

Figure 1. Architecture of the proposed model.

III. Results and Analysis

The overall architecture of the proposed model is as shown in the Fig. 1. Various sets of features are extracted using the proposed feature modeling i.e. SXGbg and CE techniques. These sets of features are evaluated with various state-of-the-art machine learning classifiers with ten-fold cross validation. The detail analysis is as follows:

A. Analysis on SkipXGram bigram (SXGbg) Features

From SXGbg technique, a total of twelve sets of features were extracted in which six sets of features (obtained by varying X value from 0 to 5 and each set consisting of 400 features) are from protein sequences and the other six sets (each set consisting of 9 features) are from possible secondary structural sequences. These twelve sets of features are extracted and evaluated individually using the various machine learning techniques. On both datasets, Random Forest (RF) classifier recorded the highest accuracy for S3Gbg sets of features (i.e. S3Gbg-Seq & S3Gbg-Str) when compared to other classifiers. Hence, these two feature sets (i.e. S3Gbg-Seq & S3Gbg-Str) extracted from SXGbg technique are shortlisted and considered for further analysis.

B. Analysis on Character Embedded (CE) Features

Using CE technique, two sets of features, say *CE-Seq* and *CE-Str* are extracted from protein sequences and possible secondary structural sequences respectively, in which each set is constituting 400 features. These extracted CE features are evaluated using state-of-the-art classifiers. Random Forest (RF) classifier recorded the highest accuracy for *CE-Seq* and *CE-Str* sets of features when compared to other classifiers on both the datasets.

C. Analysis on the Proposed Feature Model using State-of-the-art Classifiers

From all the extracted sets of features, four effective feature sets are shortlisted from SXGbg and CE techniques.

Table II
Evaluation (in %) of State-of-the-art Classifiers' using the Proposed Feature Model on Benchmark Datasets

Datasets	**25PDB**	**FC699**
Classifiers	**Proposed Feature Modeling**	**Proposed Feature Modeling**
LR	76.14	86.94
KNN	74.24	87.99
MLP	76.81	89.86
SVM	77.11	90.21
GBM	77.59	90.09
RF	**79.79**	**91.61**

These four shortlisted sets of features (a total of 1209 features) are combined (as shown in Fig. 1) and evaluated using the state-of-the-art classifiers. The performance comparison using various classifiers on both datasets is as shown in Table II. The combination of the proposed feature modeling with Random Forest classifier reported better prediction accuracy consistently for both the datasets and the same is evident from the Table II and hence, this combination is considered as the proposed model for this study.

D. Performance Analysis of the Proposed Model against State-of-the-art Models

The proposed model consists of four effective sets of features and Random Forest as the classifier. The performance of the proposed model is evaluated against state-of-the-art models on two low sequence similarity benchmark datasets and the respective results are shown in the Tables III and IV. From the Table III, it is clearly evident that the performance of the proposed model outperforms other state-of-the-art models by a factor of 3% to 23% on 25PDB dataset. Similarly, for FC699 dataset, the performance of the proposed model outperforms other state-of-the-art models by a factor of 4% to 6% (as shown in Table IV).

IV. Conclusion and Future Work

The PSSC prediction plays a vital role in identifying and analyzing protein folds and its functions further. In this study, we have proposed an effective feature modeling consisting of two feature extraction techniques such as SXGbg and CE. Various sets of features are extracted using both the proposed techniques and the prediction performance of these sets of features is analyzed using various state-of-the-art machine learning classifiers. From these sets of features, four effective sets of features constituting a total of 1209 features are shortlisted as part of the proposed feature modeling. The performance of the shortlisted features is assessed with various state-of-the-art supervised methods and Random Forest (RF) classifier consistently reported higher prediction accuracy when compared to other classifiers. Further, the proposed model is evaluated against state-of-the-art models on two low sequence similarity benchmark datasets i.e. 25PDB and FC699. The performance of the proposed model outperformed state-of-the-art models by a factor of 3% to 23% for 25PDB dataset and 4% to 6% for FC699 dataset.

Hence, we conclude that the proposed model is effective in predicting PSSC. In future, we would like to explore optimized feature selection to improve the prediction accuracy further by removing irrelevant, redundant and conflicting features.

Table III
PERFORMANCE COMPARISON OF THE PROPOSED MODEL (IN %) AGAINST STATE-OF-THE-ART MODELS FOR 25PDB DATASET

Models	**All-α**	**All-β**	**α+β**	**α/β**	**Overall**
Stacking Ensemble[16]	-	-	-	-	59.90
LLSC-PRED [8]	75.20	67.50	62.10	44.00	62.20
AAD-CGR [6]	64.30	65.00	65.00	61.70	64.00
AADP-PSSM [10]	83.30	78.10	76.30	54.40	72.90
AAC-PSSM-AC [19]	85.20	81.30	73.70	55.20	73.90
Ensemble Model [11]	86.10	80.80	80.60	60.10	76.70
Proposed Model	92.70	78.90	71.90	74.50	**79.79**

Table IV
PERFORMANCE COMPARISON OF THE PROPOSED MODEL (IN %) AGAINST STATE-OF-THE-ART MODELS FOR FC699 DATASET

Models	**All-α**	**All-β**	**α+β**	**α/β**	**Overall**
CBF-PSSE [12]	84.62	91.45	93.90	34.50	86.01
PBF-PSSE [12]	88.46	81.41	88.86	80.49	85.66
SCPRED [4]	-	-	-	-	87.50
Proposed Model	96.40	92.50	95.10	65.10	**91.61**

V. ACKNOWLEDGEMENT

This research work is supported by Vision Group on Science and Technology, Dept. of Science and Technology, Govt. of Karnataka, India, to the second author.

REFERENCES

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

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

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

[4] 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, p. 226, 2008.

[5] T.-L. Zhang, Y.-S. Ding, and K.-C. Chou, "Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern," *Journal of theoretical biology*, vol. 250, no. 1, pp. 186–193, 2008.

[6] 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," *Journal of Theoretical Biology*, vol. 257, no. 4, pp. 618–626, 2009.

[7] S. S. Sahu and G. Panda, "A novel feature representation method based on chou's pseudo amino acid composition for protein structural class prediction," *Computational biology and chemistry*, vol. 34, no. 5, pp. 320–327, 2010.

[8] L. Kurgan and K. Chen, "Prediction of protein structural class for the twilight zone sequences," *Biochemical and Biophysical Research Communications*, vol. 357, no. 2, pp. 453–460, 2007.

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

[10] 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.

[11] 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," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 10, no. 3, pp. 564–575, 2013.

[12] Q. Dai, Y. Li, X. Liu, Y. Yao, Y. Cao, and P. He, "Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position," *BMC bioinformatics*, vol. 14, no. 1, p. 1, 2013.

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

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

[15] C. Chen, Z.-B. Shen, and X.-Y. Zou, "Dual-layer wavelet svm for predicting protein structural class via the general form of chou's pseudo amino acid composition," *Protein and peptide letters*, vol. 19, no. 4, pp. 422–429, 2012.

[16] K. D. Kedarisetti, L. Kurgan, and S. Dick, "Classifier ensembles for protein structural class prediction with varying homology," *Biochemical and Biophysical Research Communications*, vol. 348, no. 3, pp. 981–988, 2006.

[17] J. P. Segrest, M. K. Jones, A. E. Klon, C. J. Sheldahl, M. Hellinger, H. De Loof, and S. C. Harvey, "A detailed molecular belt model for apolipoprotein ai in discoidal high density lipoprotein," *Journal of Biological Chemistry*, vol. 274, no. 45, pp. 31 755–31 758, 1999.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[19] 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.