Measurement 162 (2020) 107885

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

An efficient Pearson correlation based improved random forest classification for protein structure prediction techniques

B. Kalaiselvi [a,*], M. Thangamani [b]

[a] Department of Computer Science and Engineering, Mahendra Engineering College for Women, Tiruchengode, Namakkal, Tamil Nadu, India
[b] Department of Computer Science and Engineering, Kongu Engineering College,Perundurai,Erode, India.

ARTICLE INFO

Article history:
Received 6 January 2020
Received in revised form 13 March 2020
Accepted 21 April 2020
Available online 7 May 2020

Keywords:
Amino acid features
Improved random forest classification
Protein structure
Weighted covariance
Weighted mean
Weighted pearson correlation

ABSTRACT

In biochemistry, the protein structure prediction from the primary sequence is a significant issue. Few research works are intended for performing protein structure prediction with assist of diverse data mining techniques. However, the existing technique does not provide enhanced performance for protein structure prediction. To resolve this limitation, Weighted Pearson Correlation based Improved Random Forest Classification (WPC-IRFC) Technique is introduced. The WPC-IRFC Technique is developed for enhancing the protein structure prediction performance with higher accuracy and lesser time. The WPC-IRFC uses Weighted Pearson Correlation (WPC) to select relevant amino acid features based on weighted mean and weighted covariance. After selecting the relevant amino acid features, WPC-IRFC Technique designs an Improved Random Forest Classification (IRFC) for predicting the protein structure from a big protein dataset (DS). IRFC significantly lessens the error rate of classification with aid of iteratively reweighted least squares model to accurately identify protein structures.

1. Introduction

Protein Structure Prediction is a procedure for identifying the three dimensional structure of protein from amino acid sequence. Proteins are huge biological molecules that comprise the great amount of amino acid sequence. The prediction of protein structure plays an imperative role in biology for the development of new drug. In Bioinformatics and molecular biology, finding the protein structure from the amino acid sequence has become a demanding issue. Many research works are designed for protein structure prediction using various data mining techniques. The prediction accuracy of existing techniques was not adequate. Therefore, WPC-IRFC Technique is developed.

A novel technique called ProtNN was designed in [1] for fast and accurate protein 3D-structure classification. The accuracy of protein structure prediction was lessened. An enhanced GA framework (GAPlus) was employed in [2] for Ab Initio Protein Structure identification. The FPR of GAPlus was higher.

Bacterial Foraging Optimization (BFO) algorithm was introduced in [3] for protein structure discovery and thereby reducing the free energy level. The time complexity of BFO was very higher.

A Memetic Algorithm (MA) was applied in [4] for protein structure discovery. But, the FPR was minimal.

Machine learning classification model was presented in [5] with aid of six physical and chemical properties to identify the structure of the protein. The precision and recall of this model was not adequate. A neural networks and support vector machine were employed in [6] to execute protein structure identification process.

Genetic Algorithm was applied in [7] for finding the protein structure on a large scale. However, the amount of time taken for protein structure prediction was higher. Machine Learning Techniques was exploited in [8] to increase performance of multiclass protein structure prediction. The misclassification error was not solved.

In [9], three distance based classifiers was introduced to resolve the secondary structure prediction issues. With the assist of estimation of distribution algorithm, A novel method was designed in [10] for accomplishing fragment-based protein structure prediction. The computational complexity of structure prediction was remained open issue.

A combinatorial fusion technique was intended in [11] for feature selection and protein structure classification. Through cascaded bidirectional recurrent neural network (BRNN), a novel prediction system was introduced in [12] for protein secondary structure. The ratio of number of faulty prediction was higher.

* Corresponding author.
E-mail addresses: kalaiselvi.csc1980@gmail.com (B. Kalaiselvi), manithangamani2@gmail.com (M. Thangamani).

https://doi.org/10.1016/j.measurement.2020.107885
0263-2241/

Automatic classification of protein structure was introduced in [13] with help of physicochemical constraints. Two multi-classification strategies applied in [14] to discover protein structural classes according to autocovariance. The protein structure predication time was higher.

2. Related works

In [15], a review of different techniques was analyzed for protein secondary structure identification. Deep Convolutional Neural Fields was developed in [16] to find protein structure and their properties, for example, contact number, disorder regions, and solvent accessibility. Enhanced k-means clustering algorithms were constructed in [17] for finding high-quality protein structural models.

An Artificial Neural Network Classifier was presented in [18] for detection of protein structural classes. For identification of protein structure, a Distributed Tree-based Ensemble Learning Approach was designed in [19]. Protein Structural Class Prediction was executed in [20] with assist of k-separated bigrams and position-specific scoring matrix. NMR spectra analysis was presented in [21] for automated structure identification of proteins.

Sparse Representation based Classification (SRC) technique was introduced in [22] for predicting protein fold using chosen relevant features. A segmented-based feature extraction method was designed in [23] for determining the secondary structure of proteins. A novel technique was designed in [24] for identifying the three-dimensional shape of large proteins. In [25], a survey of diverse machine learning methods introduced for Protein Structure Prediction.

To conquer the existing issues, WPC-IRFC technique is designed. The contributions of WPC-IRFC technique is explained as follows,

- To improve protein structure prediction performance with lesser time as compared to conventional works, WPC-IRFC technique is introduced by combining the WPC and IRFC.
- To enhance the feature selection performance of protein structure prediction with minimal time as compared to conventional works, WPC is applied in WPC-IRFC technique. The WPC estimates the statistical correlation among two amino acid features according to weighted mean and weighted covariance and thereby choose only a more significant amino acid features for identifying protein structures.
- To achieve higher classification accuracy and minimal time for protein structure prediction as compared to traditional works, IRFC is proposed in WPC-IRFC technique. The IRFC finds the majority vote of '*n*'decision tree classifiers and evade the misclassification of protein structure prediction using iteratively reweighted least squares model.

The residual structure of the article is formulated as follows; the related works are depicted in Section 2. The WPC-IRFC technique is described in Section 3 with the architecture diagram. Experimental settings are presented in Section 4 and result analysis is exposed in Section 5. Section 6 provides the conclusion of the article.

3. Weighted pearson correlation based improved random forest classification technique

In bioinformatics and theoretical chemistry, the prediction of protein structure is an imperative process. Since, the prediction of Protein structure is a highly crucial process in medicine (for instance, drug design) and biotechnology. In existing, a lot of research works are designed for finding the protein structure with the application of different classification techniques. But, the performance of existing classification techniques was poor. Toresolve these limitations; WPC-IRFC Technique is developed. The WPC-IRFC technique is designed with application of WPC and IRFC.

On the contrary to traditional techniques, WPC is applied in WPC-IRFC technique for effective feature selection assist computes weighted mean and weighted covariance with intentions of choosing the relevant amino acid features with minimal time. In addition to that, IRFC is used in WPC-IRFC technique on the contrary to conventional techniques too btain higher classification performance for protein structure detection. The IRFC applied in WPC-IRFC technique decreases the error rate of classification for accurately performing protein structure prediction. Thus, WPC-IRFC technique obtains better results in terms of TPR, FPR, PSPA and PSPT as compared to the conventional works.

The overall process of WPC-IRFC technique is portrayed in Fig. 1 for protein structure prediction with greater accuracy and minimal time. As presented in above Fig. 1, WPC-IRFC technique applied two algorithms in order to achieve higher protein structure detection performance. The WPC-IRFC technique initially selects the relevant amino acid features from big protein DS with the application of WPC. Then, the WPC-IRFC technique identifies protein structure with the application of IRFC. This assists for WPC-IRFC technique to improve the protein structure prediction performance and lessen the time for early diagnosis of disease. The exhaustive processes of WPC-IRFC technique are shown in below subsections.

3.1. Weighted Pearson Correlation-based feature selection

The WPC-IRFC Technique exploits WPC with aiming at selecting the relevant amino acid features for protein structure identification. WPC is utilized in WPC-IRFC Technique to improve the feature selection performance for accurate protein structure prediction and improving disease diagnosis efficiency. The WPC determines

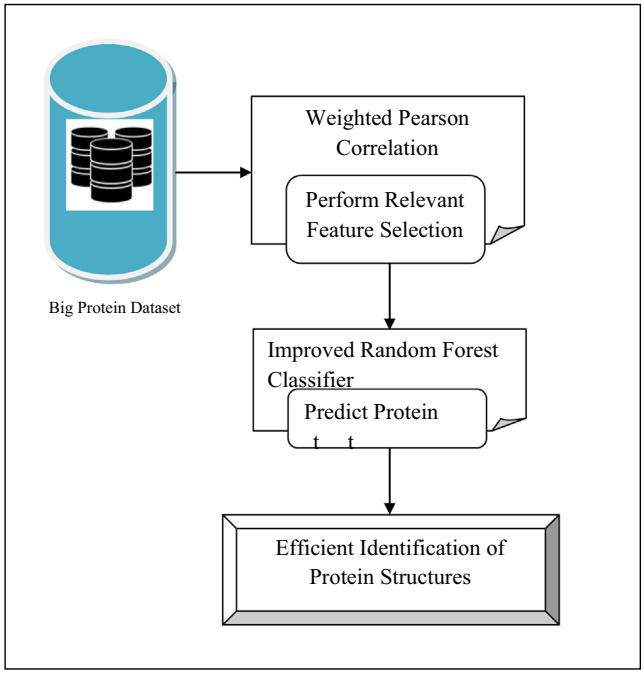

Fig. 1. WPC-IRFC Technique for Protein Structure Prediction.

the statistical correlation, or association, among two amino acid features based on weighted mean and weighted covariance on the contrary to existing Pearson correlation. Therefore, WPC is the best method for selecting relevant amino acid features. With the support of determined degree of correlation, then WPC choose the relevant amino acid features and thereby enhance PSPA. A WPC is a number between −1 and 1 that represents the extent to which two amino acid features are linearly related.

The flow process of WPC is presented in Fig. 2 to carry out feature selection processes. As exposed in the below figure, at first WPC computes weighted mean and weighted covariance between the dependent variable (i.e. disease) and one or more independent variables (i.e. amino acid features).

With help of measured weighted mean and weighted covariance, then WPC evaluates the correlation among amino acid features in input big protein DS. WPC verifies correlation value is 0 to +1.00. If the above condition is true, then WPC chooses the amino acid feature for protein structure prediction. Otherwise, the amino acid feature is not selected. Therefore, WPC-IRFC Technique attains better results in TPR and FPR for feature selection with minimal time and error rate for effectual protein structure detection.

Let us consider a big protein DS represented as '$DS = F_1, F_2, ..F_n$' with a number of amino acid features 'F_i'. The WPC measures the weighted mean using below expression,

$$m(F_i; w) = \frac{\sum_i w_i F_i}{\sum_i w_i} \tag{1}$$

From Eq. (1), 'F_i' represent amino acid feature in the input big protein DS in which 'w_i' indicates the weight vector. By using the above equation, WPC estimates weighted mean between the amino acid feature and disease to find relevant amino acid feature for predicting protein structure. Consequently, WPC determines weighted covariance using below formulation,

$$cov(F_i, x; w) = \frac{\sum_i w_i (F_i - m(F_i, w))(x - m(F_i, w))}{\sum_i w_i} \tag{2}$$

From Eq. (2), 'F_i' denotes amino acid feature in the DS and 'w_i' is a weight vector and 'x'symbolizes corresponding disease. With assist of above equation, WPC determines weighted covariance between the amino acid feature and disease to discover relevant amino acid feature for finding protein structure. Followed by, WPC estimates correlation coefficient value 'Cr' using below mathematical representation,

$$Cr(F_i, x; w) = \frac{cov(F_i, x; w)}{\sqrt{cov(F_i, F_i; w) cov(y, y; w)}} \tag{3}$$

From Eq. (3), results of correlation value 'Cr' is ranges from −1.00 to +1.00. With the estimated correlation value 'Cr', WPC efficiently performs the feature selection processes with higher TPRs. The algorithmic steps of WPC are shown in below.

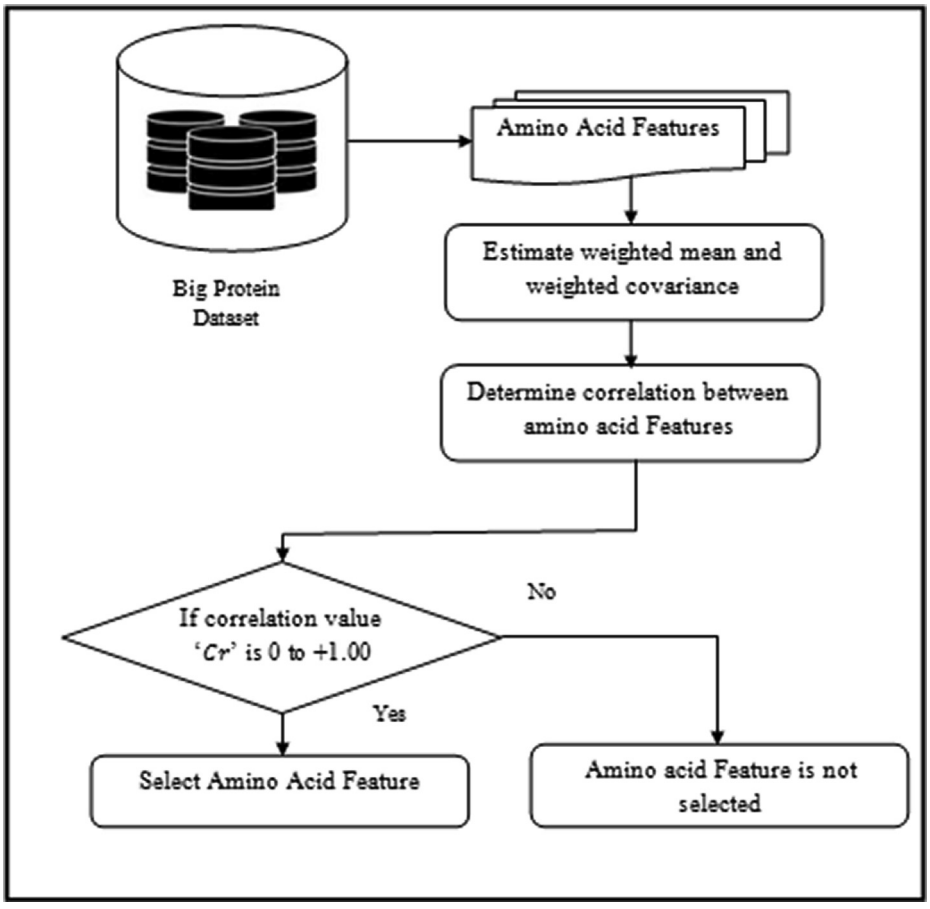

Fig. 2. Flow Processes of WPC for Feature Selection.

Algorithm 1 (*Weighted Pearson Correlation-Based Feature Selection Algorithm*).

//Weighted Pearson Correlation-Based Feature Selection Algorithm

Input: Big Protein Dataset '*DS*', Number of amino acid features '$F_i = F_1, F_2, ..F_n$',
Output: Relevant amino acid features
Step 1: Begin
Step 2: For each amino acid features 'F_i'
Step 3:*Measure weighted mean using* (1)
Step 4: Measure weighted covariance using (2)
Step 5: Determine the correlation coefficient '*Cr*' using (3)
Step 6: If(correlation value '*Cr* value is 0 to +1.00), **then**
Step 7: Choose Amino Acid Feature
Step 8: else
Step 9: Amino Acid Feature is not selected
Step 10: Endif
Step 11: End for
Step 12: End

The algorithmic process of WPC is explained in algorithm 1 for selecting the relevant amino acid features with minimal time. With the help of chosen optimal amino acid features, proposed technique effectively find outs the protein structure. Therefore, WPC-IRFC technique enhances the TPR and minimizes FPR, time of feature selection when compared to existing techniques.

3.2. Improved random forest classification

After selecting the relevant amino acid features, the WPC-IRFC Technique identifies the protein structureusing IRFC. In bioinformatics, the protein structures classification is necessary for their function determination and performs disease diagnosis. The existing random forests classifier takes more time and also affords misclassifications result. In order to solve these limitations, IRFC is developed. On the contrary to existing random forest classifier, iteratively reweighted least squares model is employed in IRFC. The iteratively reweighted least squares model lessens the misclassification error to enhance the classification accuracy for protein structure prediction as compared to conventional works. IRFC is a supervised learning algorithm. IRFC is an ensemble of Decision Trees and trained with "bagging" method. Bagging method enhances the overall result. To acquire accurate and stable prediction, IRFC constructs multiple decision trees and combines them. One big advantage of IRFC is that it exploited for both classification and regression issues. The IRFC builds '*n*' number of decision trees and subsequently applies voting scheme. Then, the majority votes of decision trees are determined to effectively find out the protein structure for disease diagnosis. The processes involved in IRFC are presented in below.

Fig. 3 depicts the block diagram of IRFC to attain higher classification accuracy for protein structure identification with minimal time. From the above figure, IRFC fits '*n*' number of decision tree classifiers to select relevant amino acid features from protein DS and enhance the predictive accuracy and control over-fitting. After that, IRFC combines all the outputs from the individual decision tree through the application of voting scheme and discovers majority votes of classification. Subsequently, IRFC optimizes the error rate of classification with application of iteratively reweighted least squares model to accurately finding protein structure. This assists for WPC-IRFC Technique to enhance the PSPA with lesser time as compared to conventional works.

On the contrary to existing random forest classifier, decision tree is exploited in IRFC that uses linear regression model for predicting protein structures. Every internal node in tree represents

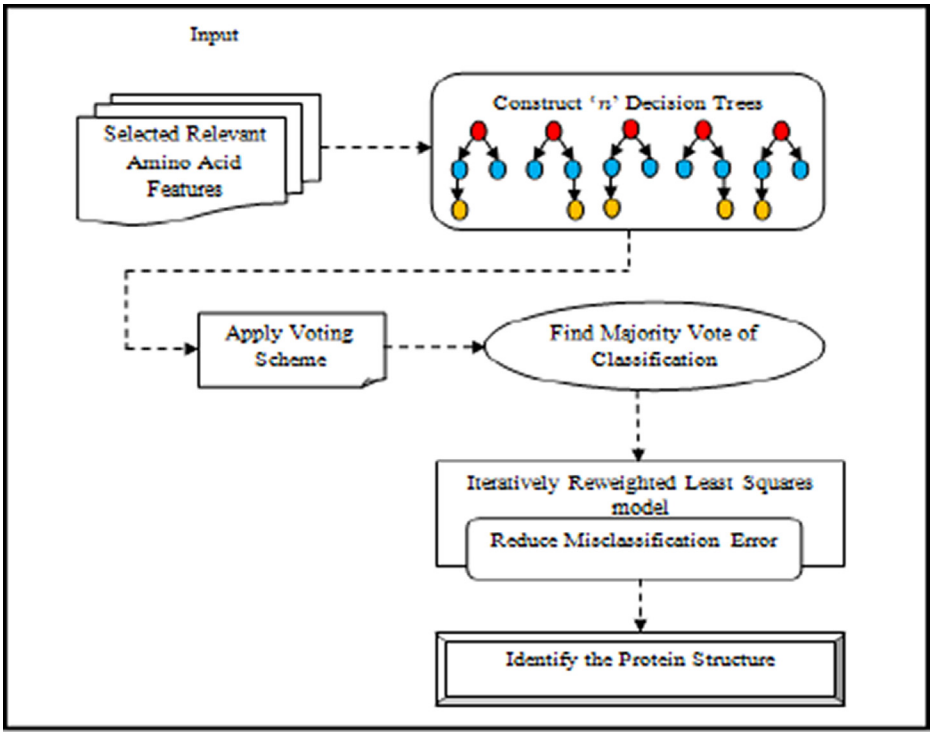

Fig. 3. Flow Process of IRFC for Protein Structure Prediction.

test on training samples. A leaf node contains the different classes and affords prediction results of protein structures. Linear regression model utilized in decision tree estimates relationship between a selected relevant amino acid features using linear predictor functions to find protein structures. From that, the decision tree classification is performed using below expression

$$f(DT_1) = \beta_0 + \beta_1 F_1 + \cdots\cdots + \beta_n F_n \quad (4)$$

From Eq. (4),'$\beta_0, \cdots .., \beta_n$' represents the linear regression coefficients that assist for decision tree classifier to measure the relationship between the selected relevant amino acid features'$F_i = F_1, F_2, ..F_n$' and thereby predicts the protein structures. Then, IRFC combines the result of all 'n' decision trees using below formulation,

$$\gamma = f(DT_1) + f(DT_2) + \cdots + f(DT_n) \quad (5)$$

From Eq. (5), '$f(DT_i)$'represent that result of decision tree classifier. Followed by, IRFC applies voting scheme for build 'n' number of decision tree classifiers using below mathematical expression,

$$\gamma = \vartheta\{f(DT_1) + f(DT_2) + \cdots + f(DT_n)\} \quad (6)$$

From Eq. (6), 'ϑ'indicates a vote applied for results of decision tree classifier. Subsequently, IRFC finds majority votes of 'n' number of decision tree classifiers using below mathematical representation,

$$\gamma = arg\max_{n} \vartheta f(DT_i) \quad \text{Where } 'i \in 1 \ \ to \ \ n' \quad (7)$$

From Eq. (7), 'γ'refers to a classification output with majority votes. Here, the arg max function is exploited to identify majority votes of 'n' number of decision tree classifiers. During the classification process of protein structure prediction, IRFC lessens error rate with aid of iteratively reweighted least squares model on the contrary to conventional random forest classification. The IRFC estimates the error rate as differences among actual output and predicted output. Accordingly, error rate 'ε_i' of classification is computed as,

$$\varepsilon_i = \sum_{i=1}^{n} \gamma_i - p_o \quad (8)$$

From Eq. (8), 'γ' refers to the actual output in which 'p_o' represents the predicted classification output. After computing the error rate, IRFC applied iteratively reweighted least squares model that is employed in the proposed technique to solve misclassification problems with help of objective functions. Here, objective function is to optimize the error rate of protein structure classification to get improved prediction accuracy as compared to conventional works. The IRFC lessens the error rate of protein structure classification by using iteratively reweighted least squares model which is mathematically formulated as,

$$\widehat{\gamma} = \arg min \sum_{i=1}^{n} \gamma_i - p_o \quad (9)$$

From Eq. (9), IRFC increases performance of classification that effectively predicts the protein structures with minimal time. Then,'$\widehat{\gamma}$' is a final classification output for protein structure identification. The IRFC algorithm is presented to enhance classification performance of protein structure identification for disease diagnosis which is illustrated in below.

Algorithm 2 (*Improved Random Forest Classification*).

//Improved Random Forest Classification Algorithm

Input: Relevant amino acid features
Output: improved protein structure prediction accuracy minimum time
Step 1: Begin
Step 2: For selected relevant amino acid features
Step 3: Create '*n*'number of a decision tree using (4)
Step 4: Combines result of all '*n*' decision trees using (5)
Step 5: Apply voting scheme using (6)
Step 6: Find the majority votes of classification using (7)
Step 7: Measure the error rate using (8)
Step 8: Optimize classification error and accurately predicts protein structure using (9)
Step 9: **End For**
Step 10: **End**

From above algorithmic steps, IRFC takes the relevant amino acid features as input. At first, IRFC trains selected relevant amino acid features with decision tree classifier. To choose relevant amino acid features, IRFC formulates the '*n*' number of decision tree classifiers. Then, IRFC combined the output of all'*n*'decision tree classifiers and applies voting scheme. After that, IRFC finds the majority vote of '*n*'decision tree classifiers. During the classification, the error rate is estimated for avoiding the misclassification of protein structure prediction using iteratively reweighted least squares model. From that, IRFC significantly executes protein structure prediction with a minimal amount of time consumption. Thus, the WPC-IRFC Technique improves the PSPA with minimal PSPT as compared to conventional works.

4. Experimental settings

The performance of WPC-IRFC Technique is evaluated in Java Language with two big protein DSs namely VariBench and Protein Data Bank (PDB). The VariBench DS is obtained from http://structure.bmc.lu.se/VariBench/datasets.php. Besides, The PDB DS includes data about the nucleic acid and 3D protein shapes which employed to discover sequences of proteins for disease diagnosis. The PDB DS is taken from https://www.rcsb.org/pdb/home/home.do. The above said two DS comprises information such as Atom serial number, Atom name, Residue name, Chain identifier, Residue sequence number, Code for insertion of residues, etc. For conducting the experimental evaluation, WPC-IRFC Technique employs the dissimilar number of amino acid features in the range of 50–500 from above two DSs. Furthermore, VariBench and PDB are crystallographic databases that contain enormous numbers of biological molecules, such as proteins and nucleic acids.

The result of WPC-IRFC Technique is calculated in TPR, FPR, PSPA and PSPT. The performance result of WPC-IRFC Technique is compared with four existing methods namely, ProtNN [1], enhanced GA framework (GAPlus) [2], Bacterial Foraging Optimization (BFO) algorithm [3] and Memetic Algorithm (MA) [4].

5. Result and discussions

In this section, the WPC-IRFC Technique results are compared with four existing methods namely [1,2,3] and [4]. The effectiveness of WPC-IRFC Technique is estimated with following metrics namely TPR, FPR, PSPA and PSPT with the assist of tables and graphs.

5.1. Experimental result of true positive rate

True Positive Rate '(*TPR*)'is calculated as the ratio of number of amino acid features which are correctly selected as relevant to the total number of features. TPR is calculated in percentage (%) and obtained using below mathematical formulation,

$$TPR = \frac{No.\ of\ features\ correclty\ selected}{Total\ No.\ of\ features} * 100 \quad (10)$$

From Eq. (10), TPR of feature selection is estimated for protein structure prediction. Greater TPR, more efficient the technique is said to be.

WPC-IRFC Technique is executed in Java Language to evaluate the TPR during the feature selection process with number of amino acid features ranges from 50 to 500. The TPR results of WPC-IRFC Technique are compared with existing methods [1–3] and [4] to estimate the WPC-IRFC Technique performance for identifying protein structures from big protein DS. When conducting an experimental work using 50–500 amino acid features from VariBench DS, WPC-IRFC Technique achieves 93% TPR whereas existing [1–3] and [4] obtains 78%, 80%, 83%, and 86%. Therefore, the TPR of WPC-IRFC Technique is higher than the conventional methods. Table 1 illustrates the results analysis of TPR with two DSs (see Table 2).

Table 1
Tabulation for true positive rate.

Datasets Name	True Positive Rate (%)				
	ProtNN	GAPlus	BFO	MA	WPC-IRFC
VariBench Data set	78	80	83	86	93
PDB dataset	80	80.5	84	87	95

Table 2
Tabulation for protein structure prediction accuracy.

Datasets Name	Protein Structure Prediction Accuracy (%)				
	ProtNN	GA Plus	BFO	MA	WPC-IRFC
VariBench Data set	75	78	81	84	90
PDB dataset	77	81	84	87	96

Fig. 4presents the impacts of TPR using five methods namely ProtNN [1], GAPlus [2], BFO [3] and MA [4] and WPC-IRFC Technique and two DS such as VariBench and PDB DS. As exposed in Fig. 4, the proposed WPC-IRFC Technique provides higher TPR when using both big protein DS for obtaining improved protein structures prediction performance when compared to four existing ProtNN [1], GAPlus [2], BFO [3] and MA [4]. This is because of the utilization of WPC in WPC-IRFC Technique on the contrary to conventional works. The WPC utilized in the proposed technique determines the weighted mean and weighted covariance between amino acid features in order to significantly estimate correlation value. The measured correlation value assists for WPC-IRFC Technique to choose amino acid features that are more important for protein structure identification. This supports for IRFC Technique improve the number of amino acid features correctly selected as compared to conventional works. Therefore, WPC-IRFC Technique enhances the TPR of feature selection as compared to conventional works.

5.2. Experimental result of protein structure prediction accuracy

PSPA is defined as the ratio of number of protein structures which are correctly predicted with selected relevant amino acid features to total number of protein structures. PSPA estimated in percentages (%).

$$PSPA = \frac{number\ of\ protein\ structures\ correctly\ predicted}{total\ number\ of\ protein\ structures} * 100 \quad (11)$$

From Eq. (11), the accuracy of Protein structure detection is evaluated. Higher PSPA, more effective the method is said to be.

WPC-IRFC Technique is implemented in Java Language to determine the PSPA with number of amino acid features ranges from 50 to 500. The PSPA result of WPC-IRFC is compared with existing methods to find the performance of WPC-IRFC Technique for classifying protein structures from big protein DS. When accomplishing an experimental process using 50–500 amino acid features from PDB DS, WPC-IRFC Technique attains 96% PSPA whereas existing ProtNN [1], GAPlus [2], BFO [3] and MA [4] gets 77%, 81%, 84%, and 87%. From that, the WPC-IRFC Technique provides higher PSPA as compared to existing methods. The result analysis of PSPA with two DS is depicted in below.

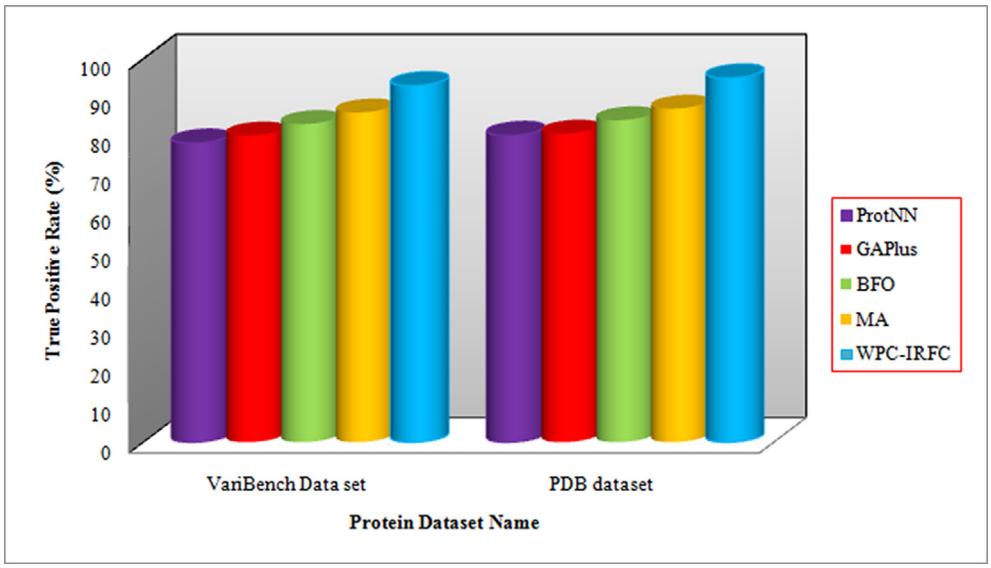

Fig. 4. Comparative Results of True Positive Rate using Two Dataset.

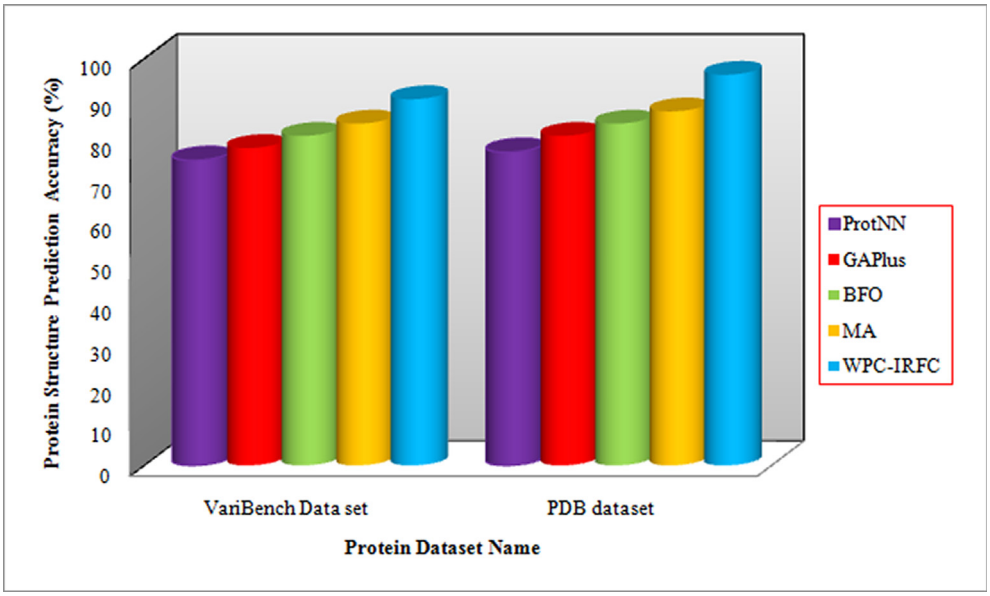

Fig. 5. Comparative Results of Protein Structure Identification Accuracy using Two Dataset.

Fig. 5 demonstrates the impacts of PSPA using five methods namely ProtNN [1], GAPlus [2], BFO [3] and MA [4] and WPC-IRFC Technique and two DSs such as VariBench and PDB DS. As depicted in Fig. 5, the proposed WPC-IRFC Technique affords higher PSPA using both big protein DS for improving disease diagnosis performance ewhen compared to four existing ProtNN [1], GAPlus [2], BFO [3] and MA [4]. This is due to the utilization of WPC and IRFC in WPC-IRFC Technique on the contrary to conventional works. In WPC-IRFC Technique, the algorithmic processes of WPC assist to pick the amino acid features for protein structure prediction. Further, the algorithmic processes of IRFC help for WPC-IRFC Technique to avoid the misclassification error of protein structure detection. Hence, WPC-IRFC Technique improves the number of protein structures which are correctly predicted with relevant amino acid features as compared to conventional works. Therefore, IRFC Technique attains greater PSPA than the existing works.

5.3. Experimental result of False Positive rate

False Positive Rate '(*FPR*)' is calculated as the ratio of number of amino acid features which are incorrectly selected as relevant to total number of features taken as input. FPR is measured in percentage (%) and formalized as below,

$$FPR = \frac{No.\ of\ features\ incorreclty\ selected}{Total\ No.\ of\ features} * 100 \quad (12)$$

From Eq. (12), the FPR of feature selection is measured for protein structure identification. Lesser the FPR, more effective the technique is said to be.

WPC-IRFC Technique is implemented in Java Language to evaluate the FPR of feature selection with number of amino acid features ranges from 50 to 500. The result of FPR using WPC-IRFC Technique is compared with existing methods [1–3] and [4] to estimate the performance of WPC-IRFC Technique for protein structures identification for disease diagnosis. When performing an experimental evaluation with 50–500 amino acid features from VariBench DS, WPC-IRFC Technique obtains 7% FPR whereas existing [1–3] and [4] gets 22%, 20%, 17%, and 14%. Hence, the FPR of WPC-IRFC Technique is lower than the existing methods. Table 3 depicts the result analysis of FPR with two DSs.

Table 3
Tabulation for false positive rate.

Datasets Name	False Positive Rate (%)				
	ProtNN	GAPlus	BFO	MA	WPC-IRFC
VariBench Data set	22	20	17	14	7
PDB dataset	27	19	16	13	5

The impact of FPR is portrayed in Fig. 6 using five methods namely existing [1–3] and [4] and WPC-IRFC Technique and two DSs such as VariBench and PDB DS. As illustrated in Fig. 6, the proposed WPC-IRFC Technique provides lower FPR when using both big protein DS for effective protein structures predictions when compared to four existing ProtNN [1], GAPlus [2], BFO [3] and MA [4]. This is because of the utilization of WPC in WPC-IRFC Technique. With the WPC process, WPC-IRFC Technique measures the correlation between amino acid features in input big protein DS. After that, WPC-IRFC Technique checks if a correlation value is 0 to +1.00. If it is true, then WPC-IRFC Technique selects the amino acid feature for discovering protein structures. Otherwise, the amino acid feature is not elected. From that, WPC-IRFC Technique reduces the ratio of a number of protein structures that are wrongly predicted. Therefore, the WPC-IRFC Technique attains lesser FPR as compared to conventional works.

5.4. Experimental result of protein structure prediction time

In WPC-IRFC Technique, Protein structure prediction time' (*PSPT*)'measures the amount of time taken to identify the protein structures from a big protein DS. PSPT is calculated in milliseconds (ms) and formalized as follows,

$$PSPT = N * time\ \ (predicting\ \ protein\ \ structure) \quad (13)$$

From (13), the amount of time required for protein structure identification is calculated. Here,'*N*' indicates the number of protein structures. Minimal PSPT, more efficient the technique is said to be.

WPC-IRFC Technique is implemented in Java Language for estimating the PSPT with number of amino acid features ranges from

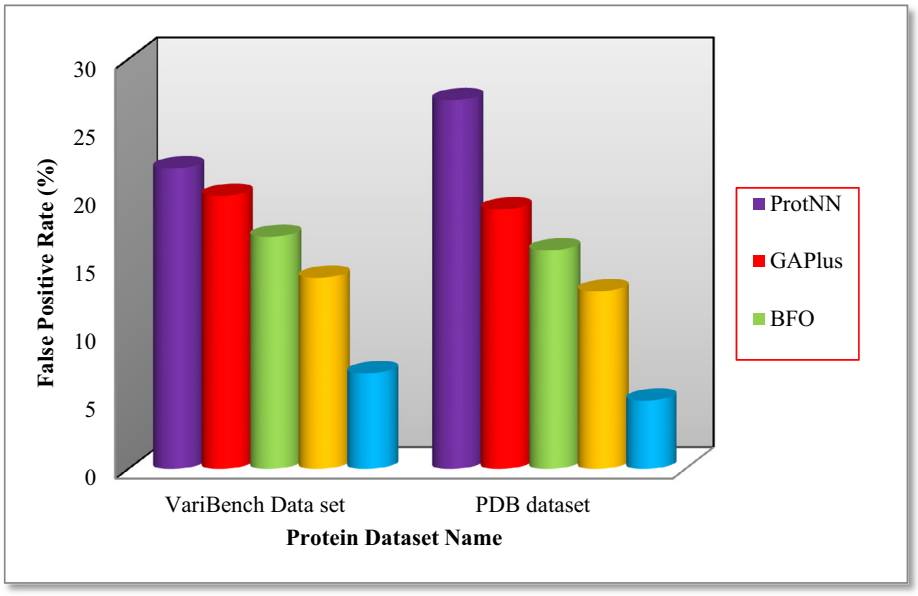

Fig. 6. Comparative Results of False Positive Rate using Two Dataset.

50 to 500. PSPT result of WPC-IRFC Technique is compared with existing [1–3] and [4] to find performance of WPC-IRFC Technique for efficient disease diagnosis. When implementing an experimental process by 50–500 amino acid features from PDB DS, WPC-IRFC Technique takes13 ms time to predict protein structures whereas conventional ProtNN [1], GAPlus [2], BFO [3] and MA [4] obtains 52 ms, 46 ms, 25 ms, and 26 ms respectively. From the results, it is descriptive that PSPT using proposed WPC-IRFC is minimal as compared to conventional methods. Table 4 portrays the experimental result of FPR with two DSs.

Fig. 7 reveals the impacts of PSPT using five methods namely ProtNN [1], GAPlus [2], BFO [3] and MA [4] and WPC-IRFC Technique and two DSs such as VariBench and PDB DS. As shown in Fig. 7, the WPC-IRFC Technique takes minimal PSPT with two big protein DS as compared to existing [1–3] and [4]. This is because of the application of WPC and IRFC in WPC-IRFC Technique on the contrary to conventional works. With the application of WPC, WPC-IRFC Technique chooses the amino acid features that are more relevant for identifying protein structure with the minimal amount of time. In addition, with the application of IRFC, WPC-IRFC Technique accurately predicts protein structures via selected amino acid features with minimal PSPT. This assist for WPC-IRFC Technique to lessen the PSPT than the existing works. Therefore, WPC-IRFC Technique lessens the PSPT as compared to conventional works.

Table 4
Tabulation for protein structure prediction time.

Datasets Name	Protein Structure Prediction Time (ms)				
	ProtNN	GAPlus	BFO	MA	WPC-IRFC
VariBench Dataset	55	49	28	27	16
PDB dataset	52	46	25	26	13

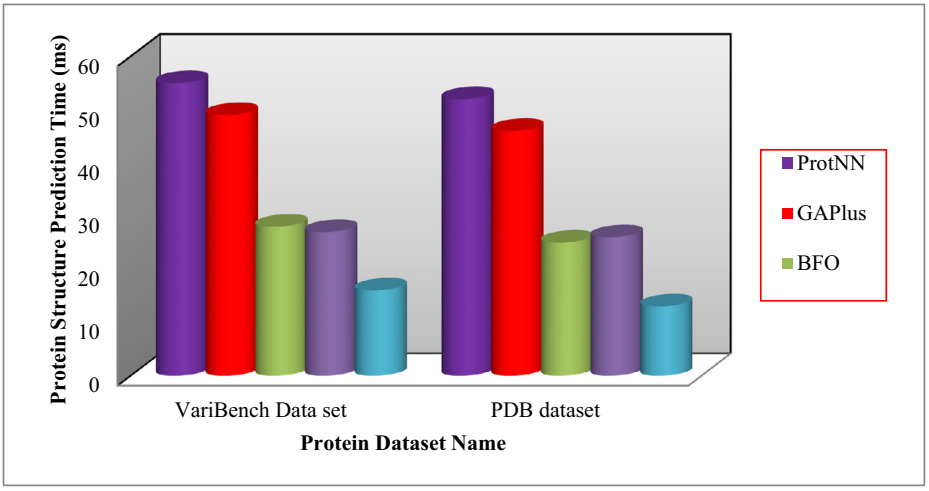

Fig. 7. Comparative Results of Protein Structure Prediction Time using Two Dataset.

6. Conclusion

An efficient WPC-IRFC Technique is introduced for enhancing the protein structure prediction performance with higher accuracy and minimal time. The WPC-IRFC Technique is designed with application of WPC and IRFC. The WPC supports for WPC-IRFC Technique to enhance the ratio of number of correctly selected amino acids features and also to minimize the time complexity involved during feature selection process as compared to traditional techniques. Moreover, IRFC helps for WPC-IRFC Technique to achieve higher classification performances for protein structure identification with minimal amount of time consumption when compared to existing techniques. WPC-IRFC Technique is tested with the metrics namely TPR, FPR, PSPA, PSPT using two big protein DS and compared with conventional works. The result analysis of WPC-IRFC Technique offers better performance in terms of PSPA and PSPT for efficient disease diagnosis as compared to state-of-the-art works.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Wajdi Dhifli, Abdoulaye Baniré Diallo, ProtNN: fast and accurate protein 3D-structure classification in structural and topological space, BioData Mining 9 (30) (2016) 1–17.

[2] Mahmood A. Rashid, Firas Khatib, Md Tamjidul Hoque, Abdul Sattar, An enhanced genetic algorithm for ab initio protein structure prediction, IEEE Trans. Evolut. Comput. 20 (4) (2016) 627–644.

[3] D. Ramyachitra, V. Veeralakshmi, Bacterial Foraging Optimization for protein structure prediction using FCC & HP energy model, Gene Rep. 7 (2017) 43–49, Elsevier.

[4] Leonardo Corrêa, Bruno Borguesan, Camilo Farfán, Mario Inostroza-Ponta, Márcio Dorn, A memetic algorithm for 3D protein structure prediction problem, IEEE/ACM Trans. Comput. Biol. Bioinform. 15 (3) (2018) 1–14.

[5] Sonal Mishra, Yadunath Pathak, Anamika Ahirwar, Classification of protein structure (RMSD <= 6Å) using physicochemical properties, Int. J. Bio-Sci. Bio-Technol. 7 (6) (2015) 141–150.

[6] Badal Bhushan, Manish Kumar Singh, Protein structure prediction using neural networks & support vector machines, Int. J. Eng Sci. Adv. Technol. 3 (3) (2014) 145–156.

[7] Subhendu Bhusan Rout, Satchidananda Dehury, Bhabani Sankar Prasad Mishra, Protein structure prediction using genetic algorithm, Int. J. Comput. Sci. Mobile Comput. 2 (6) (2013) 187–192.

[8] Mayuri Patel, Multi-class protein structure prediction using machine learning techniques, Int. J. Res. Advent Technol. 2 (12) (2014) 54–59.

[9] Ashish Ghosh, Bijnan Parai, Protein secondary structure prediction using distance-based classifiers, Int. J. Approx. Reason. 47 (1) (2008) 37–44, Elsevier.

[10] David Simoncini, Francois Berenger, Rojan Shrestha, Kam Y.J. Zhang, A probabilistic fragment-based protein structure prediction algorithm, PLOS One 7 (10) (2012) 1–11.

[11] Ken-Li Lin, Chun-Yuan Lin, Chuen-Der Huang, Hsiu-Ming Chang, Chiao-Yun Yang, Chin-Teng Lin, Chuan Yi Tang, D. Frank Hsu, Feature selection and combination criteria for improving accuracy in protein structure prediction, IEEE Trans. NanoBiosci. 6 (2) (2007) 186–196.

[12] Jinmiao Chen, Narendra Chaudhari, Cascaded bidirectional recurrent neural networks for protein secondary structure prediction, IEEE/ACM Trans. Comput. Biol. Bioinf. 4 (4) (2007) 572–582.

[13] M. Abhilash Mohan, Divya Rao, Shruthi Sunderrajan, Gautam Pennathur, Automatic classification of protein structures using physicochemical parameters, Interdiscip. Sci.: Comput. Life Sci. 6 (3) (2014) 176–186, Springer.

[14] Wu. Jiang, Yi-Zhou Li, Meng-Long Li, Yu. Le-Zheng, Two multi-classification strategies used on SVM to predict protein structural classes by using autocovariance, Interdiscip. Sci.: Comput. Life Sci. 1 (4) (2009) 315–319, Springer.

[15] Qian Jiang, Xin Jin, Shin-Jye Lee, Shaowen Yao, Protein secondary structure prediction: a survey of the state of the art, J. Mol. Graph. Model. 76 (2017) 379–402, Elsevier.

[16] Sheng Wang, Jian Peng, Jianzhu Ma, Xu. Jinbo, Protein secondary structure prediction using deep convolutional neural fields, Sci. Rep. 6 (2016) 1–11.

[17] Wu Hongjie, Haiou Li, Min Jiang, Cheng Chen, Qiang Lv, Wu Chuang, Identify high-quality protein structural models by enhanced k-means, BioMed Res. Int. (2017) 1–9, Article ID 7294519.

[18] Li Ningbo, Huo Hua, An artificial neural network classifier for the prediction of protein structural classes, Int. J. Curr. Eng. Technol. 7 (3) (2017) 946–952.

[19] Leo Dencelin Xavier, Ramkumar Thirunavukarasu, A distributed tree-based ensemble learning approach for efficient structure prediction of protein, Int. J. Intell. Eng. Syst. 10 (3) (2017) 226–234.

[20] Harsh Saini, Gaurav Raicar, Alok Sharma, Sunil Lal, Abdollah Dehzangi, Rajeshkannan Ananthanarayanan, James Lyons, Neela Biswas, Kuldip K. Paliwal, Protein structural class prediction via k-separated bigrams using position specific scoring matrix, J. Adv. Comput. Intell. Intell. Inform. 18 (4) (2014) 474–479.

[21] Julia M. Würz, Sina Kazemi, Elena Schmidt, Anurag Bagaria, Peter Guntert, NMR-based automated protein structure determination, Arch. Biochem. Biophys. 628 (2017) 24–32, Elsevier.

[22] YongXu KeYan, Xiaozhao Fang, Chunhou Zheng, Bin Liu, Protein fold recognition based on sparse representation based classification, Artif. Intell. Med. 79 (2017) 1–8, Elsevier.

[23] Abdollah Dehzangi, Kuldip Paliwal, James Lyons, Alok Sharma, Abdul Sattar, A segmentation-based method to extract structural and evolutionary features for protein fold recognition, IEEE/ACM Trans. Comput. Biol. Bioinf. 11 (3) (2014) 510–519.

[24] Mario Garza-Fabre, Shaun M. Kandathil, Julia Handl, Joshua Knowles, Simon C. Lovell, Generating, maintaining, and exploiting diversity in a memetic algorithm for protein structure prediction, Evolut. Comput. 24 (4) (2016) 1–30.

[25] Jianlin Cheng, Allison N. Tegge, Pierre Baldi, Machine learning methods for protein structure prediction, IEEE Rev. Biomed. Eng. 1 (2008) 41–49.