Artificial Intelligence in Medicine 78 (2017) 61–71

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Machine learning based identification of protein–protein interactions using derived features of physiochemical properties and evolutionary profiles

Muhammad Tahir, Maqsood Hayat*

Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan

ARTICLE INFO

Article history:
Received 15 March 2017
Received in revised form 9 June 2017
Accepted 11 June 2017

Keywords:
Physiochemical properties
PPIs
PSSM
PNN

ABSTRACT

Proteins are the central constitute of a cell or biological system. Proteins execute their functions by interacting with other molecules such as RNA, DNA and other proteins. The major functionality of protein–protein interactions (PPIs) is the execution of biochemical activities in living species. Therefore, an accurate identification of PPIs becomes a challenging and demanding task for investigators from last few decades. Various traditional and computational methods have been applied but they have not achieved quite encouraging results. In order to extend the concept of computational model by incorporating intelligent, contemporary machine learning algorithms have been utilized for identification of PPIs. In this prediction model, protein sequences are expressed by using two distinct feature extraction methods namely: physiochemical properties of amino acids and evolutionary profiles method position specific scoring matrix (PSSM). Jackknife test and numerous performance parameters namely: specificity, recall, accuracy, MCC, precision, and F-measure were employed to compute the predictive quality of proposed model. After empirical analysis, it is determined that the proposed prediction model yielded encouraging predictive outcomes compared to existing state-of-the-art models. This achievement is ascribed with PSSM because it has clearly discerned a motif of PPIs. It is realized that the proposed prediction model will lead to be a practical and very useful tool for research community.

1. Introduction

Proteins are vital molecules of a biological system or cell that play significant roles in biological processes i.e., DNA synthesis, translation, transcription and splicing. Proteins carry out their functions by interacting with other molecules i.e., RNA, DNA and other proteins, binding into temporary or stable complex [1]. These processes refer to protein–protein interactions (PPIs), which are responsible for carrying out biochemical activities in living cells [2,3]. In the life cycles of cells, a PPIs play key roles i.e., regulation of gene expression, cell apoptosis, organism growth and reproduction, changing the specificity of a protein, genetic material duplication, cell signal transduction, and cell necrosis [4–8]. Therefore, the scientists and researchers have realized that eukaryotic and prokaryotic organisms have not only huge numbers of genes but also have widely PPIs networks [9–12]. The human PPIs network is illustrated in Fig. 1. The entire entra and extracellular processes of living organism depend on PPIs [13]. Therefore, it has great biological importance in constructing intermolecular regulatory networks, containing signal transduction pathway, metabolism, and genetic regulatory pathways [14–19]. This study can help to analyze a targeted new drugs and investigates the techniques for the designing of new drugs [20–22]. Transformation and cell growth are disturbed in signaling events, and the competitions and changes in PPIs provide a multipurpose mechanism for pathway regulation [23,24].

In the initial study, the determination of protein–protein interaction was commonly based on biological experimental techniques [25]. However, for every living cells these wet laboratory techniques were inefficient because they were costly, time consuming and susceptible to errors [26,27]. Recently, the researchers have launched a systematic identification of PPI pairs in the budding yeast and examined all possibility by applying computational methods to accurately and rapidly predict PPIs on huge amount of protein datasets [28]. During the past, for the PPIs prediction several machine learning algorithms have been successfully applied

* Corresponding author.
E-mail addresses: Maqsood.hayat@gmail.com, m.hayat@awkum.edu.pk (M. Hayat).

http://dx.doi.org/10.1016/j.artmed.2017.06.006
0933-3657/© 2017 Elsevier B.V. All rights reserved.

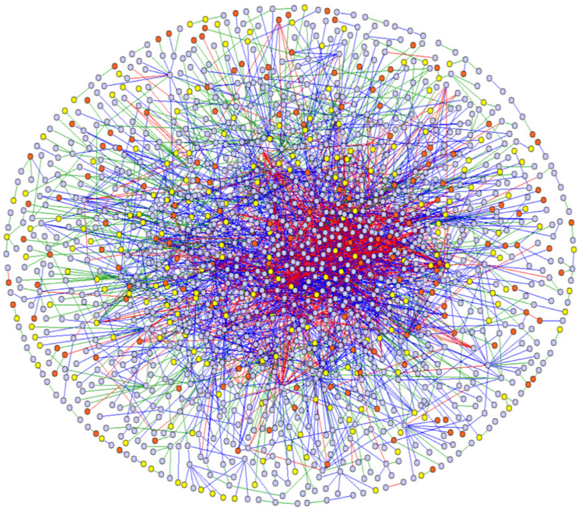

Fig. 1. The human PPIs network. From: www.mdc-berlin.de.

i.e., support vector machine (SVM) [29], random forest (RF) [30], neural network (NN) [31], and so on.

In Machine learning based approaches, the input of prediction models is in the form of structural features, sequential features or both [32–34]. Bradford et al. have developed SVM based classifier on surface patch analysis [29] and then enhanced the performance of prediction model using a Bayesian network [35]. Similarly, Jones and Thornton correctly examined series of residue patches on the surface of protein 3D structures using six parameters i.e., hydrophobicity, protrusion, residue interface propensity, solvation potential, accessible surface area and planarity, and developed a technique for evaluating the relative combined score of a surface patch for developing PPIs [36–39]. Likewise, Chen et al., have developed a prediction method for 3D probability distribution interacting atoms on protein surfaces and after that applied classification techniques to learn the characteristic patterns of the probability density maps specific to PPIs [40]. Ofran and Rost have established a NN model is known as ISIS [41] for the prediction of protein–protein interactions based on the predicted evolutionary information and structural features evaluated from the sub-sequences of nine serial residues. Alike, Porollo and Meller have established a statistical model is known as SPPIDER model [42] using NN and SVMs in combination with relative solvent accessibility (RSA). The RSA features have produced good performance compared to other PPI feature predictors. Mizuguchi and Murakami have developed a naive Bayesian classifier is called PSIVER [43] with predicted accessibility and position-specific scoring matrices (PSSM) as feature sources. In recent time, Dhole et al. have implemented two methods for identification of PPIs such as LORIS [44] and SPRINGS [45] by applying L1-regularized logistic regression and artificial NN, respectively, based on predicted relative averaged cumulative hydropathy, solvent accessibility, and evolutionary conservation features. In a sequel, Liu et al., have developed a machine learning based method named DC-RF-RUS-PF [5], which has shown considerable performance.

Series of publications have demonstrated that Chou's five step rules are the foundation for developing a statistical predictor, which are [5,46–48], i) select or construct a benchmark dataset for training and testing the predictor ii) formulate samples into a discrete vector that truly reflects the effectiveness of target classes. iii) develop/select machine learning algorithm iv) the predictor is examined by cross-validation test v) finally, to established a user-friendly web server.

In this research, we proposed a powerful and an accurate sequence based prediction model for identification of protein–protein interactions. In this proposed prediction model, two feature extraction methods such as physiochemical properties and position specific scoring matrix are used for extracting salient features. Jackknife test is utilized to examine the prediction performance of various classification algorithms.

The rest of paper is structured as follows: Section 2 illustrates materials and methods, Section 3 presents evaluation methodology, Section 4 demonstrates results and discussion. Finally, Section 5 conclusion has been drawn.

2. Materials and methods

2.1. Datasets

In this attempt, we have selected three various benchmark datasets namely: Dtestset72, Dset186, and PDBtestset164 for training. The first dataset Dtestset72 contains 72 non-redundant sequences [5,43]. It was constructed based on the protein–protein docking benchmark set version 3.0 using a homology reduction procedure [5,47,49]. The Dset186 dataset was established by Mizuguchi and Murakami contains 186 non-redundant (sequence identity <25%), non-transmembrane, heterodimeric, and transient protein sequences, which have structurally resolved by X-ray crystallography with a resolution of $\leqq 3.0\,\text{A}\ ^{\circ}$ [5,47]. The third dataset PDBtestset164 constructed by Singh et al., this dataset was achieved, using newly annotated proteins [45] from June, 2010 to November, 2013. It contains non-redundant 164 protein sequences derived from PDB (Protein Data Bank) [45] with the similar filters that were employed to create Dtestset72 and Dset186 [5].

2.2. Feature extraction methods

Two distinct feature extraction methods namely: physiochemical properties and Position Specific Scoring Matrix (PSSM) are used for extracting salient, useful and pertinent discrete information from the protein sequences which are further used for training and testing the predictor efficiently.

2.2.1. Physiochemical properties

Suppose a protein sequence P with L amino acid residues i-e.

$$P = R_1R_2R_3R_4R_5R_6R_7\ldots R_L \tag{1}$$

where R_1 denotes the amino acid residue at the sequence position-1 of the protein P, R_2 the amino acid residue at position-2 and so forth. In Eq. (1) various kinds of amino acids have various physiochemical properties [47]. In this study, we have used the following ten physicochemical properties: polarity, secondary structure, molecular volume, codon diversity, electrostatic charge, hydrophobicity, hydrophilicity, side-chain volume, polarizability and solvent-accessible surface area (SASA), these physicochemical properties are shortly represented by $\Phi^{(1)}$, $\Phi^{(2)}$, $\Phi^{(3)}$, $\Phi^{(4)}$, $\Phi^{(5)}$, $\Phi^{(6)}$, $\Phi^{(7)}$, $\Phi^{(8)}$, $\Phi^{(9)}$ and $\Phi^{(10)}$, respectively [47]. Table 1 shows the numerical values of the ten physicochemical properties for the 20

Table 1
The original values of the ten physiochemical properties for all amino acids.

Amino acid	$\Phi^{(1)}$	$\Phi^{(2)}$	$\Phi^{(3)}$	$\Phi^{(4)}$	$\Phi^{(5)}$	$\Phi^{(6)}$	$\Phi^{(7)}$	$\Phi^{(8)}$	$\Phi^{(9)}$	$\Phi^{(10)}$
A	8.100	−1.302	−0.733	1.57	−0.146	0.620	−0.500	27.500	0.046	1.181
C	5.500	0.465	−0.862	−1.02	−0.255	0.290	−1.000	44.600	0.128	1.461
D	13.000	0.302	−3.656	−0.259	−3.242	−0.900	3.000	40.000	0.105	1.587
E	12.300	−1.453	1.477	0.113	−0.837	−0.740	3.000	62.000	0.151	1.862
F	5.200	−0.59	1.891	−0.397	0.412	1.190	−2.500	115.500	0.290	2.228
G	9.000	1.652	1.33	1.045	2.064	0.480	0.000	0.000	0.000	0.881
H	10.400	−0.417	−1.673	−1.474	−0.078	−0.400	−0.500	79.000	0.230	2.025
I	5.200	−0.547	2.131	0.393	0.816	1.380	−1.800	93.500	0.186	1.810
K	11.300	−0.561	0.533	−0.277	1.648	−1.500	3.000	100.000	0.219	2.258
L	4.900	−0.987	−1.505	1.266	−0.912	1.060	−1.800	93.500	0.186	1.931
M	5.700	−1.524	2.219	−1.005	1.212	0.640	−1.300	94.100	0.221	2.034
N	11.600	0.828	1.299	−0.169	0.933	−0.780	2.000	58.700	0.134	1.655
P	8.000	2.081	−1.628	0.421	−1.392	0.120	0.000	41.900	0.131	1.468
Q	10.500	−0.179	−3.005	−0.503	−1.853	−0.850	0.200	80.700	0.180	1.932
R	10.500	−0.055	1.502	0.44	2.897	−2.530	3.000	105.000	0.291	2.560
S	9.200	1.399	−4.76	0.67	−2.647	−0.180	0.300	29.300	0.062	1.298
T	8.000	0.326	2.213	0.908	1.313	−0.050	−0.400	51.300	0.108	1.525
V	5.900	−0.279	−0.544	1.242	−1.262	1.080	−1.500	71.500	0.140	1.645
W	5.400	0.009	0.672	−2.128	−0.184	0.810	−3.400	145.500	0.409	2.663
Y	6.200	0.83	3.097	−0.838	1.512	0.260	−2.300	117.300	0.298	2.368

amino acids. Therefore, the P protein of Eq. (1) can be expressed into ten various mathematical series, as formulated by

$$\mathrm{P} = \begin{cases} \Phi_1^{(1)}\Phi_2^{(1)}\Phi_3^{(1)}\Phi_4^{(1)}\Phi_5^{(1)}\Phi_6^{(1)}\Phi_7^{(1)}\ldots\Phi_L^{(1)} \\ \Phi_1^{(2)}\Phi_2^{(2)}\Phi_3^{(2)}\Phi_4^{(2)}\Phi_5^{(2)}\Phi_6^{(2)}\Phi_7^{(2)}\ldots\Phi_L^{(2)} \\ . \\ . \\ . \\ . \\ . \\ . \\ \Phi_1^{(10)}\Phi_2^{(10)}\Phi_3^{(10)}\Phi_4^{(10)}\Phi_5^{(10)}\Phi_6^{(10)}\Phi_7^{(10)}\ldots\Phi_L^{(10)} \end{cases} \tag{2}$$

where $\Phi^{(1)}$ is the polarity value of R_1 in Eq. (1), $\Phi^{(2)}$ the secondary structure of R_2 and so on.

2.2.2. Position specific scoring matrix (PSSM)

PSSM is generally used for protein sequence pattern representation [50]. For an input novel protein sequence, set parameter with 0.001 as E-value cutoff for several protein sequences alignment against the sequence [50,51], The obtained $\mathrm{P}_{\mathrm{PSSM}}$ matrix against the protein sequence is composed of L × 20 entries using PSI-BLAST [51] to search the Swiss Port database through three iterations, where *L* represents the number of amino acids in a protein sequence and 20 represents the natïve amino acids [52–54]. The PSSM of a protein sequence P with L residues of amino acids can be formulated as follows:

$$\mathrm{P}_{\mathrm{PSSM}} = \begin{bmatrix} \mathrm{E}_{1,1} & \mathrm{E}_{1,2} & \cdots & \mathrm{E}_{1,j} & \cdots & \mathrm{E}_{1,20} \\ \mathrm{E}_{2,1} & \mathrm{E}_{2,2} & \cdots & \mathrm{E}_{2,j} & \cdots & \mathrm{E}_{2,20} \\ . & . & \cdots & . & \cdots & . \\ .. & . & \cdots & . & \cdots & . \\ .. & . & \cdots & . & \cdots & . \\ \mathrm{E}_{i,1} & \mathrm{E}_{i,2} & \cdots & \mathrm{E}_{i,j} & \cdots & \mathrm{E}_{i,20} \\ . & . & \cdots & . & \cdots & . \\ . & . & \cdots & . & \cdots & . \\ . & . & \cdots & . & \cdots & . \\ \mathrm{E}_{L,1} & \mathrm{E}_{L,2} & \cdots & \mathrm{E}_{L,j} & \cdots & \mathrm{E}_{L,20} \end{bmatrix} \tag{3}$$

where $\mathrm{E}_{i,j}$ denotes the value of amino acid residue in the i^{th} location of the protein sequence is being replaced by amino acid residue type j during the biological evolution processes. The values of j = 1, 2, 3. . . 20 denote the twenty amino acids according to their alphabetical order [50]. In PSSM matrix, each amino acid residue has 20 values that reflect the frequencies of mutation detected at the particular location in the protein family. The PSSM matrix contains both positive and negative values, positive values indicate that more mutation has occurred in the alignment, while negative shows less substitution has taken place in the alignment [55]. After obtaining the original value in each location is normalized by using the following logistic function:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

where ‘x’ represents the original PSSM values. In order to extract features vector from the PSSM, the concept of a sliding window of size *W* was employed. The window size determines the number of residues in a sub-sequence, which is used for the study of neighborhood outcome. In this study, we have used sliding window size W = 11, 13, 15, 17, 19, 21 and 23. According to the analysis [56,57], it is observed that PSSM features have several limitations. Generation of PSSM for protein sequence is mostly depended on the searching sequences. PSSM does not reflect the exact information in case of no homologous sequence find out in the executing sequences, consequently, leads toward wrong prediction.

2.3. Classification algorithms

Classification is the sub-part of pattern recognition, machine learning, and data mining where the statistical data is classified into recognized classes on the basis of instances. Classification and regression are the two categories of the supervised learning problems. In this research, we have used three classification algorithms. Fig. 2 shows the framework of our proposed model.

2.3.1. K-Nearest neighbor (KNN)

KNN is widely used classification algorithm among supervised classifiers, which can be used for classification, regression, and pattern recognition purposes. It is generally applied due to its simplicity, good performance, and smooth comprehension. It has no prior information about the distribution of the data. Owing to nonparametric nature, it is used for regression and classification [58]. Besides of its simplicity, as compared to other learning algorithms, it has incredible and competitive performance. There is no training phase in KNN while keeping all the training data in testing phase.

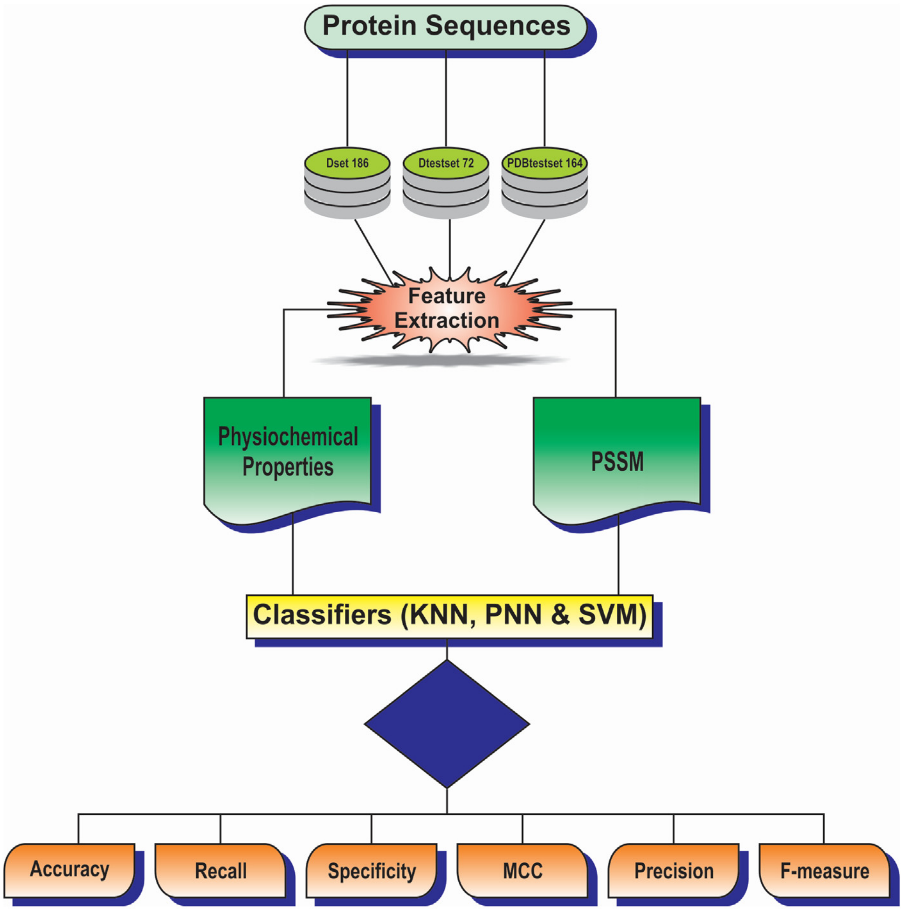

Fig. 2. Framework of proposed Model.

Therefore, this algorithm is called lazy learning or instant based leaning algorithm. It is efficient for that data, which changes and updates dynamically [59]. Distance is measured with the help of Euclidean Distance formula. Fig. 3 shows the general structure of KNN.

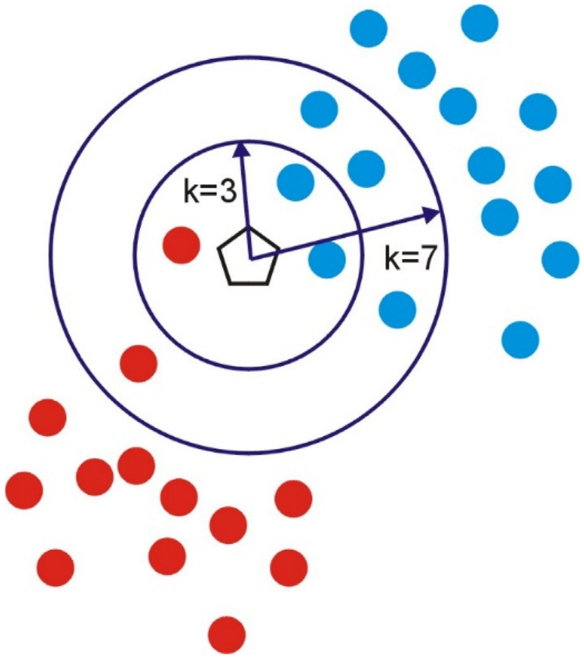

Fig. 3. Structure of KNN.

$$Edis\,(x1, x2) = \sum_{i=1}^{n} \sqrt{(xi1 - xi2)^2} \tag{5}$$

2.3.2. Probabilistic neural network (PNN)

PNN was first time introduced by Specht in 1990 [60], which is based upon Bayes theory. It provides an interactive approach on the basis of probability density function for interpreting the structure of network [61]. The most effective and interesting property of PNN is the ability to represent any number of inputs/output complex relationships [62]. The computational power of PNN is much closer to back-propagation neural network. The simplicity and transparency of PNN is also similar to conventional statistical classification approaches. It operates in completely parallel way, due to which it does not need feedback from the individual neurons for the input [63]. This peculiar and remarkable advantage has made it effective compare to other neural networks.

PNN training method is easy and instantaneous. It has four different operational layers illustrated in Fig. 4. The input layer is directly attached to the input neurons; it is further passed to the pattern layer, which is the second layer of PNN. The numbers of presented samples in the network are equals to the dimensions of pattern layer. For training data sample first and second layers (input layer, pattern layer), there is one-to-one corresponding connectivity between neurons. The third layer is summation layer. It has the same dimensions as the number of classes in data sample. The final layer is called decision or output layer. It categorizes the total given samples into predefined classes.

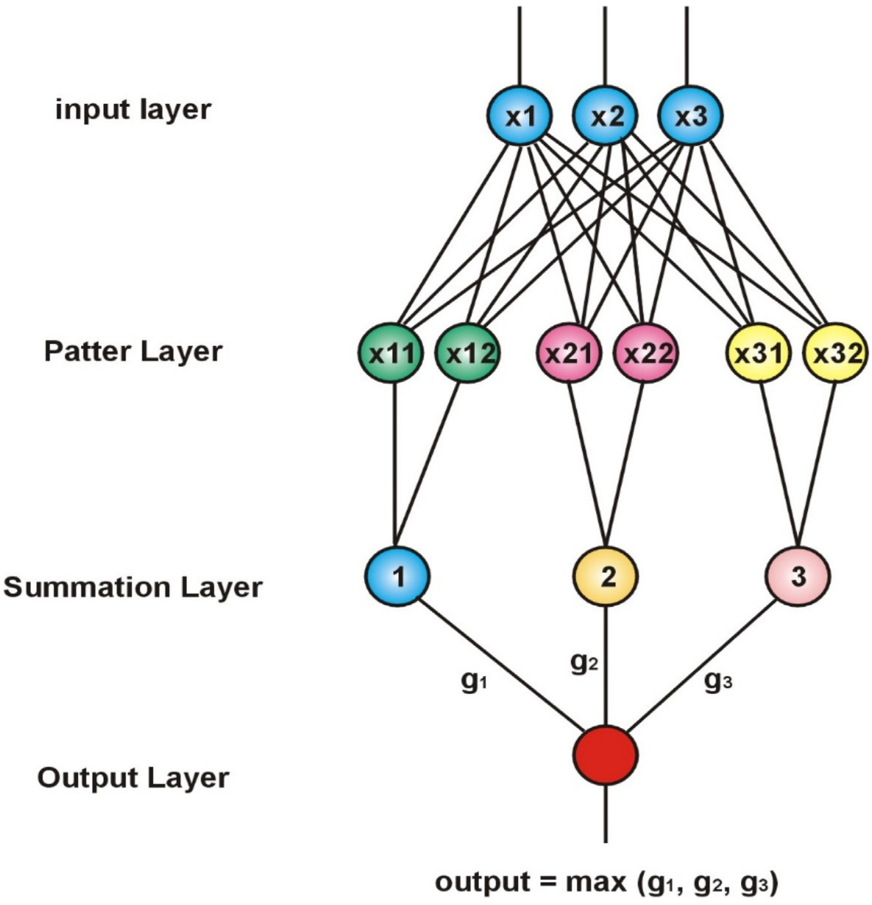

Fig. 4. Architecture of PNN.

2.3.3. Support vector machine (SVM)

SVM is mostly used in the area of pattern recognition and bioinformatics. It was developed by Cortes and Vapnik for supervised classification and regression [64,65]. The main goal of SVM is to determine an optimal separating hyperplane that maximizes the distance from the hyperplane to the instances closest to it on both sides. The margin is defined as the distance between the hyperplane and the samples of the two classes that are closed to the hyperplane [66].

3. Evaluation methodology

Various statistical tests namely: Jackknife cross-validation, subsampling, n-fold cross-validation, and independent dataset tests have been applied to measure the performance of a predictor [67–70]. Owing to unique outcome of Jackknife test [71,72], it has been widely used in bioinformatics [69,73–78]. In this research, the jackknife cross-validation test was utilized to examine the success rate of the predictor. In this test, each protein sequence is in turn singled out as testing sample and the training is performed on the remaining samples. The performance parameters i.e., accuracy, precision, recall, specificity, F-measure, and Mathew's correlation coefficient (MCC) are calculated on the basis of the following formulas:

$$Accuracy = \frac{TN + TP}{TP + FN + FP + TN} \times 100\% \tag{6}$$

$$Recall = \frac{TP}{FN + TP} \times 100\% \tag{7}$$

$$Specticity = \frac{TN}{FP + TN} \times 100\% \tag{8}$$

$$MCC = \frac{TN \times TP - FP \times FN}{\sqrt{[TN + FP]\,[TN + FN]\,[TP + FN]\,[TP + FP]}} \tag{9}$$

$$Precision = \frac{TN}{TP + FP} \times 100\% \tag{10}$$

$$F - measure = 2 \times \frac{Recall \times Precision}{Recall + Precision} \times 100\% \tag{11}$$

where FP, TN,TP, and FN denotes the false positive, true negative, true positive, and false negative, respectively.

4. Results and discussion

4.1. Performance analysis of classification algorithms using derived feature space of physiochemical property

Table 2 describes the success rates of KNN, PNN and SVM classifiers in junction with physiochemical property based feature space using dataset Dtestset72. Various sizes of window are empirically examined such as W of 11, 13, 15, 17, 19, 21 and 23. After executing KNN on different window sizes, the better results are obtained on the window size 23, which are 86.09% accuracy, 91.96% recall, 48.75% specificity, 0.407 MCC, 91.95% precision and 91.95% F-measure. Similarly, the accuracy, recall, specificity, MCC, precision, and F-measure of PNN are 83.43%, 90.62%, 37.86%, 0.288, 90.23%, and 92.58%, respectively, on the window size 15. In the same way, the accuracy, recall, specificity, MCC, precision, and F-measure of SVM are 86.07%, 97.84%, 8.85%, 0.132, 87.56%, and 92.41%, respectively, on the window size 23. Table 3 shows the

Table 2
Success rates of classification algorithms on physiochemical properties using Dataset72.

Window Size	Hypothesis	ACC (%)	Recall (%)	Sp (%)	MCC	Precision (%)	F-Measure (%)
11	KNN	84.05	90.59	42.93	0.332	90.89	90.74
13		84.19	90.49	44.39	0.342	91.12	90.81
15		84.59	90.67	46.05	0.359	91.41	91.04
17		84.81	90.96	45.94	0.364	91.41	91.18
19		85.75	91.60	48.73	0.400	91.87	91.74
21		85.46	91.32	48.24	0.390	91.81	91.56
23		**86.09**	**91.96**	**48.75**	**0.407**	**91.95**	**91.95**
11	PNN	82.95	90.46	35.75	0.267	89.85	90.15
13		83.31	90.79	36.12	0.276	89.96	90.37
15		**83.43**	**90.62**	**37.86**	**0.288**	**90.23**	**90.42**
17		83.03	90.52	35.62	0.266	89.89	90.21
19		85.68	93.91	33.54	0.316	89.94	91.89
21		83.06	90.58	35.32	0.264	89.89	90.29
23		83.03	90.51	35.38	0.263	89.92	90.21
11	SVM	84.03	91.21	26.58	0.180	90.85	91.03
13		84.54	94.02	24.91	0.232	88.73	91.30
15		84.97	94.13	26.90	0.255	89.08	91.54
17		85.16	97.44	6.16	0.071	86.97	91.93
19		85.20	94.14	28.64	0.273	89.30	91.66
21		85.38	96.19	17.35	0.201	87.98	91.90
23		**86.07**	**97.84**	**8.85**	**0.132**	**87.56**	**92.41**

Bold values signifies the highest values in the Table.

Table 3
Success rates of classification algorithms on physiochemical properties using Dataset186.

Window Size	Hypothesis	ACC (%)	Recall (%)	Sp (%)	MCC	Precision (%)	F-Measure (%)
11	KNN	81.60	89.48	19.17	0.085	89.77	89.62
13		82.27	89.95	21.37	0.112	90.07	90.01
15		82.98	90.28	24.97	0.151	90.53	90.41
17		83.07	90.49	24.00	0.145	90.46	90.47
19		83.35	90.65	25.22	0.159	90.61	90.63
21		84.03	91.21	26.58	0.180	90.85	91.03
23		**84.19**	**91.21**	**27.78**	**0.191**	**91.02**	**91.12**
11	PNN	81.97	90.31	15.87	0.064	89.48	89.89
13		82.10	90.31	16.92	0.074	89.61	89.96
15		82.35	90.35	18.71	0.092	89.83	90.09
17		82.15	90.10	18.85	0.090	89.84	89.97
19		82.35	90.31	18.89	0.093	89.86	90.09
21		82.35	90.32	18.60	0.090	89.87	90.09
23		**82.50**	**90.37**	**19.35**	**0.098**	**89.89**	**90.18**
11	SVM	83.26	90.82	35.43	0.270	89.90	90.35
13		82.28	89.90	32.12	0.221	89.70	89.80
15		82.44	89.90	32.95	0.228	89.89	89.89
17		82.60	89.76	37.56	0.271	90.04	89.90
19		83.50	94.73	11.33	0.086	87.29	90.86
21		83.59	91.17	35.94	0.281	89.95	90.55
23		**83.76**	**90.80**	**27.79**	**0.185**	**90.90**	**90.85**

Bold values signifies the highest values in the Table.

predicted results of classification algorithms on dataset Dset186 on various window sizes. KNN again achieved the better results on window size 23, which are 84.19% accuracy, 91.21% recall, 27.78% specificity, 0.191 MCC, 91.02% precision, and 91.12% F-measure. Likewise, PNN has obtained accuracy, recall, specificity, MCC, precision, and F-measure are 82.50%, 90.37%, 19.35%, 0.098, 89.89%, and 90.18%, respectively. Similarly, SVM has yielded the accuracy, recall, specificity, MCC, precision, and F-measure are 83.76%, 90.80%, 27.79%, 0.185, 90.90%, and 90.85%, respectively. Table 4 reports the experimental outcomes of classification algorithms on dataset PDBtestset164. The obtained results of KNN are 82.44% accuracy, 89.90% recall, 32.95% specificity, 0.228 MCC, 89.89% precision and 89.89% F-measure. Similarly, PNN has achieved the accuracy, recall, specificity, MCC, precision, and F-measure are 80.39%, 88.96%, 23.58%, 0.127 MCC, 88.53% and 88.74%, respectively. Likewise, SVM has obtained the accuracy, recall, specificity, MCC, precision, and F-measure are 81.97%, 90.31%, 15.87%, 0.064, 89.48%, and 89.89%, respectively. The better results were reported on window size 23.

4.2. Performance analysis of classification algorithms using PSSM feature space

The predicted outcomes of KNN, PNN and SVM classifiers in combination with PSSM based feature space using dataset Dtestset72 are reported in Table 5. KNN has achieved the highest results on the window size 17, which are 87.20% accuracy, 92.63% recall, 53.08% specificity, 0.458 MCC, 92.53% precision and 92.58% F-measure. Similarly, the accuracy, recall, specificity, MCC, precision, and F-measure of PNN are 87.20%, 92.64%, 53.04%, 0.458, 92.52% and 92.58%, respectively. Likewise, the accuracy, recall, specificity, MCC, precision, and F-measure of SVM are 87.07%, 99.18%, 11.12%, 0.240, 87.49%, and 92.97%, respectively. Table 6 illustrates the success rates of classification algorithms on dataset Dset186. All classification algorithms have obtained the highest results on the window size 15. The results of KNN are 88.17% accuracy, 93.28% recall, 48.25% specificity, 0.414 MCC, 93.37% precision and 93.32% F-measure, whereas the accuracy, recall, specificity, MCC, precision, and F-measure of PNN are 88.17%, 93.28%, 48.23%, 0.413, 93.37%

Table 4
Success rates of classification algorithms on physiochemical properties using PDBtestset164.

Window Size	Hypothesis	ACC (%)		Recall (%)	Sp (%)	MCC	Precision	F-Measure
11		79.40	88.13	23.24	0.114	88.07	88.10	
13		80.23	88.90	24.31	0.134	88.34	88.62	
15		80.84	88.84	27.50	0.163	88.83	88.83	
17	KNN		80.91	89.00	28.17	0.171	88.99	88.99
19		81.41	89.12	30.89	0.198	89.42	89.27	
21		82.28	89.90	32.12	0.221	89.70	89.80	
23		**82.44**	**89.90**	**32.95**	**0.228**	**89.89**	**89.89**	
11		79.64	88.75	21.06	0.101	87.84	88.28	
13		79.83	88.83	21.74	0.108	87.99	88.41	
15		79.84	88.61	22.92	0.116	88.18	88.39	
17	PNN		79.88	88.66	22.55	0.114	88.19	88.43
19		80.41	89.10	23.46	0.128	88.41	88.75	
21		80.30	89.06	22.64	0.119	88.34	88.70	
23		**80.39**	**88.96**	**23.58**	**0.127**	**88.53**	**88.74**	
11		80.13	89.04	22.02	0.113	88.16	88.60	
13		80.35	89.38	21.41	0.112	88.12	88.74	
15		80.65	88.84	27.50	0.163	88.83	88.83	
17	SVM		81.23	90.64	19.90	0.115	88.07	89.33
19		80.76	89.96	20.70	0.113	88.09	89.09	
21		81.55	91.05	19.54	0.118	88.07	89.53	
23		**81.97**	**90.31**	**15.87**	**0.064**	**89.48**	**89.89**	

Bold values signifies the highest values in the Table.

Table 5
Success rates of classification algorithms on PSSM using Dtestset72.

Window Size	Hypothesis	Acc (%)		Recall (%)	Sp (%)	MCC	Precision (%)	F-Measure (%)
11		87.00	92.58	51.95	0.448	92.36	92.47	
13		86.73	92.18	52.54	0.444	92.41	92.30	
15		86.69	92.32	51.40	0.438	92.26	92.29	
17	KNN		**87.20**	**92.63**	**53.08**	**0.458**	**92.53**	**92.58**
19		86.67	92.20	51.99	0.440	92.33	92.27	
21		86.64	92.20	51.77	0.438	92.30	92.25	
23		86.60	92.21	51.40	0.435	92.24	92.23	
11		87.01	92.62	51.81	0.448	92.34	92.48	
13		86.80	92.25	52.58	0.446	92.43	92.34	
15		86.71	92.35	51.31	0.438	92.25	92.30	
17	PNN		**87.20**	**92.64**	**53.04**	**0.458**	**92.52**	**92.58**
19		86.70	92.23	51.95	0.440	92.33	92.28	
21		86.65	92.21	51.77	0.438	92.30	92.26	
23		86.62	92.23	51.45	0.436	92.25	92.24	
11		86.74	92.19	52.54	0.444	92.42	92.30	
13		86.77	92.22	52.54	0.445	92.42	92.32	
15		86.78	92.78	50.68	0.436	92.16	92.35	
17	SVM		**87.07**	**99.18**	**11.12**	**0.240**	**87.49**	**92.97**
19		87.01	92.62	51.77	0.447	92.34	92.48	
21		86.89	99.18	9.76	0.218	87.88	92.88	
23		86.77	92.22	52.54	0.445	92.42	92.32	

Bold values signifies the highest values in the Table.

and 93.32%, respectively. Similarly, the accuracy, recall, specificity, MCC, precision, and F-measure of SVM are 88.12%, 95.27%, 43.23%, 0.441, 91.32%, and 93.26%, respectively. Table 7 describes the predicted results of classification algorithms on dataset PDBtestset164 on various window sizes. The highest results of the classifiers are reported on 21 window size. The obtained results of KNN are 86.85% accuracy, 92.37% recall, 51.87% specificity, 0.442 MCC, 92.41% precision and 92.39% F-measure. Similarly, PNN has achieved the accuracy, recall, specificity, MCC, precision, and F-measure are 86.87%, 92.38%, 51.89%, 0.442 MCC, 92.41% and 92.39%, respectively. Likewise, SVM has yielded the accuracy, recall, specificity, MCC, precision, and F-measure are 86.83%, 97.41%, 20.10%, 0.276, 88.49%, and 92.74%, respectively. After analyzing the experimental results, it is observed that the performance of all the three used classifiers is outstanding in conjunction with PSSM feature space. PSSM is used for representation of amino acids structure/motif in a biological sequence which provides evolutionary information for the given protein sequence.

4.3. *Comparison of proposed model with existing models*

The performance comparison of proposed prediction model has been carried out with recently published ‘DC-RF-RUS-PF’ model using the same datasets in Table 8. The predicted results of ‘DC-RF-RUS-PF’ model on dataset Dset186 were 65.1% accuracy, 61.2% recall, 66.6% specificity, 0.229 MCC, 31.7% precision and 38.2% F-measure. Similarly, the success rates of ‘DC-RF-RUS-PF’ model for dataset Dtestset72, accuracy, recall, specificity, MCC, precision, and F-measure were 63.3%, 62.2%, 63.8%, 0.193, 24.7% and 32.4%, respectively. The anticipated results of ‘DC-RF-RUS-PF’ model for dataset PDBtestset164 were 61.1% accuracy, 52.6% recall, 65.3% specificity, 0.148 MCC, 32.4% precision and 36.0% F-measure. In contrast, our proposed prediction model on dataset Dset186 has achieved 88.17% accuracy, 93.28% recall, 48.25% specificity, 0.414 MCC, 93.37% precision and 93.32% F-measure. Similarly, in case of dataset Dtestset72 the accuracy, recall, specificity, MCC, precision, and F-measure of proposed model are 87.20%, 92.63%, 53.08%, 0.458, 92.53% and 92.58%, respectively. The predicted outcomes of

Table 6
Success rates of classification algorithms on PSSM using Dset186.

Window Size	Hypothesis	Acc (%)		Recall (%)	Sp (%)	MCC	Precision (%)	F-Measure (%)
11		87.63	93.12	44.70	0.380	92.93	93.03	
13		87.78	93.06	46.48	0.394	93.14	93.17	
15		**88.17**	**93.28**	**48.25**	**0.414**	**93.37**	**93.32**	
17	KNN		88.01	93.21	47.35	0.405	93.26	93.23
19		88.10	93.30	47.43	0.407	93.27	93.29	
21		88.08	93.27	47.55	0.408	93.28	93.27	
23		87.83	93.15	46.21	0.394	93.11	93.13	
11		87.62	93.12	44.63	0.380	92.93	93.02	
13		87.79	93.08	46.45	0.394	93.14	93.11	
15		**88.17**	**93.28**	**48.23**	**0.413**	**93.37**	**93.32**	
17	PNN		88.1	93.22	47.26	0.404	93.25	93.23
19		88.11	93.31	47.45	0.408	93.27	93.29	
21		88.17	93.37	47.55	0.410	93.29	93.33	
23		87.83	93.16	46.16	0.394	93.11	93.14	
11		87.58	98.44	19.43	0.313	88.46	93.18	
13		87.39	94.23	44.45	0.424	91.41	92.80	
15		**88.12**	**95.27**	**43.23**	**0.441**	**91.32**	**93.26**	
17	SVM		88.00	97.22	30.19	0.381	89.73	93.32
19		87.96	95.38	41.41	0.428	91.08	93.18	
21		87.32	92.81	52.86	0.460	92.51	92.66	
23		87.72	96.03	35.60	0.394	90.34	93.10	

Bold values signifies the highest values in the Table.

Table 7
Success rates of classification algorithms on PSSM using Dtestset164.

Window Size	Hypothesis	Acc (%)		Recall (%)	Sp (%)	MCC	Precision (%)	F-Measure (%)
11		86.04	91.85	49.11	0.408	91.97	91.97	
13		86.75	92.30	51.52	0.437	92.35	92.33	
15		86.80	92.45	50.99	0.436	92.29	92.37	
17	KNN		86.57	92.14	51.23	0.431	92.30	92.22
19		86.57	92.19	50.95	0.430	92.26	92.22	
21		**86.85**	**92.37**	**51.87**	**0.442**	**92.41**	**92.39**	
23		86.64	92.31	50.66	0.430	92.23	92.27	
11		86.05	91.87	49.11	0.408	91.97	91.92	
13		86.76	92.31	51.52	0.437	92.35	92.33	
15		86.82	92.47	50.99	0.437	92.29	92.38	
17	PNN		86.57	92.14	51.25	0.432	92.30	92.22
19		86.58	92.20	50.90	0.430	92.25	92.23	
21		**86.87**	**92.38**	**51.89**	**0.442**	**92.41**	**92.39**	
23		86.64	92.31	50.68	0.431	92.27	92.27	
11		85.20	94.14	28.64	0.273	89.30	91.66	
13		85.38	96.19	17.35	0.201	87.98	91.90	
15		86.38	97.84	14.31	0.218	87.78	92.54	
17	SVM		86.55	92.25	50.36	0.427	92.18	92.21
19		86.76	97.55	18.87	0.267	88.32	92.71	
21		**86.83**	**97.41**	**20.10**	**0.276**	**88.49**	**92.74**	
23		86.43	92.10	50.42	0.424	92.17	92.14	

Bold values signifies the highest values in the Table.

Table 8
Performance Comparison of the proposed model with existing methods.

Dataset	Model	Acc (%)	Recall (%)	Sp (%)	MCC	Precision (%)	F-Measure (%)
Dset186	DC-RF-RUS-PF [5]	65.1	61.2	66.6	0.229	31.7	38.2
	Proposed Model	**88.17**	**93.28**	**48.25**	**0.414**	**93.37**	**93.32**
Dtestset72	DC-RF-RUS-PF [5]	63.3	62.2	63.8	0.193	24.7	32.4
	Proposed Model	**87.20**	**92.63**	**53.08**	**0.458**	**92.53**	**92.58**
PDBtestset164	DC-RF-RUS-PF [5]	61.1	52.6	65.3	0.148	32.4	36.0
	Proposed Model	**86.87**	**92.38**	**51.89**	**0.442**	**92.41**	**9**

Bold values signifies the highest values in the Table.

proposed model for dataset PDBtestset164 were 86.87% accuracy, 92.38% recall, 51.89% specificity, 0.442 MCC, 92.41% precision and 92.39% F-measure. The experimental results have shown that the proposed model has obtained better results compared to the existing models so far and graphically illustrated in Figs. 5–7. These significant outcomes are obtained due to the highly efficient features of PSSM because it has clearly discerned the pattern/motif of PPIs in protein sequences; as a result, the discrimination power of KNN, PNN, and SVM is enhanced.

5. Conclusion

In this attempt a precise and accurate intelligent computational model for the identification of PPIs is developed. Two different feature extraction methods namely: physiochemical properties and

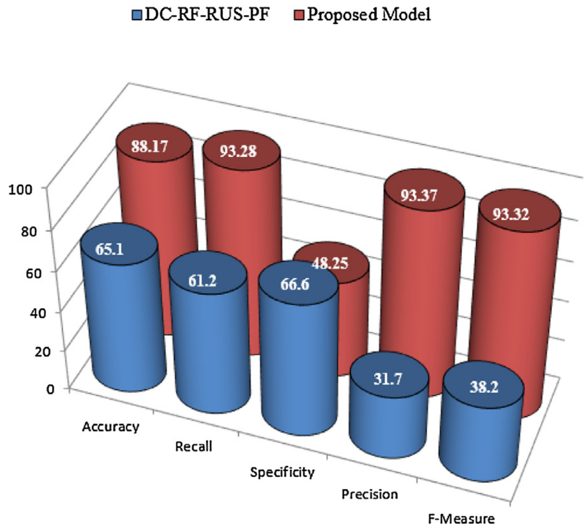

Fig. 5. Shows the performance of Dset186 dataset.

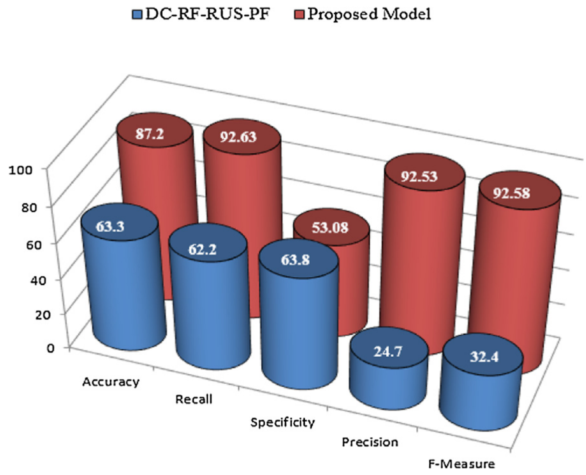

Fig. 6. Shows the performance of Dtestset72 dataset.

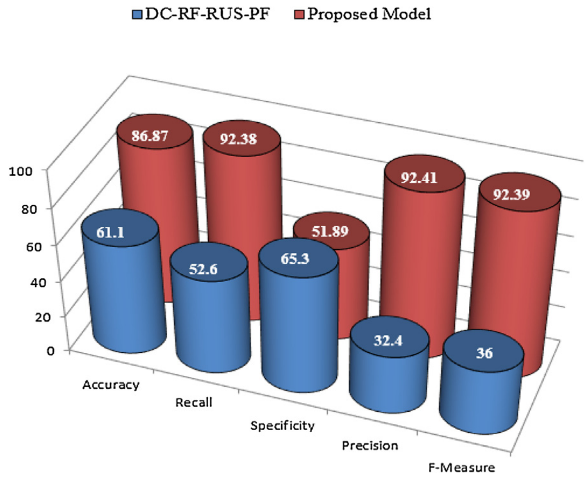

Fig. 7. Shows the performance of PDBtestset164 dataset.

PSSM are employed to extract nominal and prominent features from the protein sequences. Three numerous classification algorithms such as SVM, KNN and PNN are used for classification; jackknife test was employed to measure the performance rate of the proposed predicator. In the final analysis of the classification algorithms and feature spaces, it is observed that the performance outcome of our proposed prediction model is very efficient on all the three datasets than the existing methods in the literature so far. It is anticipated that the proposed prediction model will become a useful and predictive tool for basic academia and research. As demonstrated in many recent publications [5,77,79–81], a user friendly web-servers represent the future direction for developing practically more useful models. Therefore, we shall make effort to provide a web-server in future work for the proposed method presented in this research that will provide practical aid for future researchers in this area. In addition, Pseudo amino acid composition (PseAAC) can not only include the residues composition, but also contain the long-range correlation of the physicochemical properties between two residues. Thus, PseAAC has been extensively applied in protein classification [82–85]. In future, we will add this feature into our model in order to investigate the performance of hypothesis.

Conflict of interest

Authors have no conflict of interests.

References

[1] Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002;12:368–73.

[2] Ahmed Z, Tetlow IJ, Ahmed R, Morell MK, Emes MJ. Protein–protein interactions among enzymes of starch biosynthesis in high-amylose barley genotypes reveal differential roles of heteromeric enzyme complexes in the synthesis of A and B granules. Plant Sci 2015;233:95–106.

[3] Marceau AH, Bernstein DA, Walsh BW, Shapiro W, Simmons LA, Keck JL. Protein interactions in genome maintenance as novel antibacterial targets. PloS One 2013;8:e58765.

[4] De Las Rivas J, Fontanillo C. Protein–protein interaction networks: unraveling the wiring of molecular machines within the cell. Briefings in functional genomicse 2012:ls036.

[5] Liu G-H, Shen H-B, Yu D-J. Prediction of protein–protein interaction sites with machine-learning-based data-cleaning and post-Filtering procedures. J Membr Biol 2016;249:141–53.

[6] Hayat M, Khan A. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J Theor Biol 2011;271:10–7.

[7] Hayat M, Khan A. WRF-TMH: predicting transmembrane helix by fusing composition index and physicochemical properties of amino acids. Amino Acids 2013;44:1317–28.

[8] Hayat M, Khan A. Prediction of membrane protein types using pseudo-amino acid composition and ensemble classification. Int J Comput Electr Eng 2013;5:456.

[9] Pitre S, Alamgir M, Green JR, Dumontier M, Dehne F, Golshani A. Computational methods for predicting protein–protein interactions. In: Protein–Protein Interaction. Springer; 2008. p. 247–67.

[10] Li Z-W, You Z-H, Chen X, Li L-P, Huang D-S, Yan G-Y, et al. Accurate prediction of protein–protein interactions by integrating potential evolutionary information embedded in PSSM profile and discriminative vector machine classifier. Oncotarget 2017;8:23638–49.

[11] Dias R, Kolaczkowski B. Improving the accuracy of high-throughput protein–protein affinity prediction may require better training data. BMC Bioinf 2017;18:102.

[12] Yugandhar K, Gromiha MM. Computational approaches for predicting binding partners, interface residues, and binding affinity of protein–protein complexes. Prediction Protein Secondary Struct 2017:237–53.

[13] Chua HN, Wong L. Increasing the reliability of protein interactomes. Drug Discov Today 2008;13:652–8.

[14] Betel D, Breitkreuz KE, Isserlin R, Dewar-Darch D, Tyers M, Hogue CW. Structure-templated predictions of novel protein interactions from sequence information. PLoS Comput Biol 2007;3:e182.

[15] Hall DA, Ptacek J, Snyder M. Protein microarray technology. Mech Ageing Dev 2007;128:161–7.

[16] Hu L, Huang T, Shi X, Lu W-C, Cai Y-D, Chou K-C. Predicting functions of proteins in mouse based on weighted protein–protein interaction network and protein hybrid properties. PloS One 2011;6:e14556.

[17] Jia J, Liu Z, Xiao X, Liu B, Chou K- C. Identification of protein–protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. J Biomol Struct Dyn 2015:1–16.

[18] Skrabanek L, Saini HK, Bader GD, Enright AJ. Computational prediction of protein–protein interactions. Mol Biotechnol 2008;38:1–17.

[19] Wei L, Xing P, Zeng J, Chen J, Su R, Guo F. Improved prediction of protein?protein interactions using novel negative samples, features, and an ensemble classifier. Artif Intell Med 2017;16:30569–73.
[20] Ako-Adjei D, Fu W, Wallin C, Katz KS, Song G, Darji D, et al. HIV-1, human interaction database: current status and new features. Nucleic Acids Res 2015;43:D566–70.
[21] Burgoyne NJ, Jackson RM. Predicting protein interaction sites: binding hot-spots in protein–protein and protein–ligand interfaces. Bioinformatics 2006;22:1335–42.
[22] Russell RB, Aloy P. Targeting and tinkering with interaction networks. Nat Chem Biol 2008;4:666–73.
[23] Couzens AL, Knight JD, Kean MJ, Teo G, Weiss A, Dunham WH, et al. Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. Sci Signal 2013;6, rs15-rs.
[24] Romano D, Nguyen LK, Matallanas D, Halasz M, Doherty C, Kholodenko BN, et al. Protein interaction switches coordinate Raf-1 and MST2/Hippo signalling. Nat Cell Biol 2014;16:673–84.
[25] Drewes G, Bouwmeester T. Global approaches to protein–protein interactions. Curr Opin Cell Biol 2003;15:199–205.
[26] Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M. Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 2002;18:529–36.
[27] Friedrich T, Pils B, Dandekar T, Schultz J, Müller T. Modelling interaction sites in protein domains with interaction profile hidden Markov models. Bioinformatics 2006;22:2851–7.
[28] Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, et al. Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci 2000;97:1143–7.
[29] Bradford JR, Westhead DR. Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics 2005;21:1487–94.
[30] Jia J, Xiao X, Liu B. Prediction of protein–protein interactions with physicochemical descriptors and wavelet transform via random forests. J Lab. Automat 2016;21:368–77.
[31] Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–61.
[32] Sudha G, Nussinov R, Srinivasan N. An overview of recent advances in structural bioinformatics of protein–protein interactions and a guide to their principles. Prog Biophys Mol Biol 2014;116:141–50.
[33] Agrawal NJ, Helk B, Trout BL. A computational tool to predict the evolutionarily conserved protein–protein interaction hot-spot residues from the structure of the unbound protein. FEBS Lett 2014;588:326–33.
[34] Cukuroglu E, Gursoy A, Nussinov R, Keskin O. Non-redundant unique interface structures as templates for modeling protein interactions. PloS One 2014;9:e86738.
[35] Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR. Insights into protein–protein interfaces using a Bayesian network prediction method. J Mol Biol 2006;362:365–86.
[36] Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272:121–32.
[37] Jones S, Thornton JM. Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 1997;272:133–43.
[38] Garcia-Garcia J, Valls-Comamala V, Guney E, Andreu D, Muñoz FJ, Fernandez-Fuentes N, et al. iFraG: a protein–protein interface prediction server based on sequence fragments. J Mol Biol 2017;429:382–9.
[39] Taghipour S, Zarrineh P, Ganjtabesh M, Nowzari-Dalini A. Improving protein complex prediction by reconstructing a high-confidence protein–protein interaction network of Escherichia coli from different physical interaction data sources. BMC Bioinf 2017;18:10.
[40] Chen C-T, Peng H-P, Jian J-W, Tsai K-C, Chang J-Y, Yang E-W, et al. Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces. PloS One 2012;7:e37706.
[41] Ofran Y, Rost B. ISIS: interaction sites identified from sequence. Bioinformatics 2007;23:e13–6.
[42] Porollo A, Meller J. Prediction-based fingerprints of protein–protein interactions. Proteins 2007;66:630–45.
[43] Murakami Y, Mizuguchi K. Applying the Naïve bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. Bioinformatics 2010;26:1841–8.
[44] Dhole K, Singh G, Pai PP, Mondal S. Sequence-based prediction of protein–protein interaction sites with L1-logreg classifier. J Theor Biol 2014;348:47–54.
[45] Singh G, Dhole K, Pai PP, Mondal S. SPRINGS: prediction of protein–protein interaction sites using artificial neural networks. Peer J PrePrints 2014.
[46] Chen W, Feng P-M, Deng E-Z, Lin H, Chou K-C. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Analytical biochemistry 2014;462:76–83.
[47] Jia J, Liu Z, Xiao X, Liu B, Chou K-C. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol 2015;377:47–56.
[48] Tahir M, Hayat M. iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC. Mol Biosyst 2016.
[49] Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein–protein docking benchmark version 3.0. Proteins 2008;73:705–9.
[50] Hayat M, Khan A. MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM. J Theor Biol 2012;292:93–102.
[51] Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
[52] He X, Han K, Hu J, Yan H, Yang J-Y, Shen H-B, et al. TargetFreeze: identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition. J Membr Biol 2015;248:1005–14.
[53] Wang L, You Z-H, Xia S-X, Liu F, Chen X, Yan X, et al. Advancing the prediction accuracy of protein-Protein interactions by utilizing evolutionary information from position-Specific scoring matrix and ensemble classifier. J Theor Biol 2017.
[54] Ahmad J, Javed F, Hayat M. Intelligent computational model for classification of sub-golgi protein using oversampling and fisher feature selection methods. Artif Intell Med 2017;78:14–22.
[55] Hayat M, Tahir M. PSOFuzzySVM-TMH: identification of transmembrane helix segments using ensemble feature space by incorporated fuzzy support vector machine. Mol BioSyst 2015;11:2255–62.
[56] Wuyun Q, Zheng W, Zhang Y, Ruan J, Hu G. Improved species-specific lysine acetylation site prediction based on a large variety of features set. PloS One 2016;11:e0155370.
[57] Lin H, Chen W, Ding H. AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PloS One 2013;8:e75726.
[58] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 1992;46:175–85.
[59] Han J, Kamber M, Pei J. Data mining, Southeast Asia edition: concepts and Techniques: Morgan kaufmann; 2006.
[60] Specht DF. Probabilistic neural networks. Neural Netw 1990;3:109–18.
[61] Santhanam T, Radhika S. Probabilistic Neural Network–A better solution for noise classification. J Theor Appl Inf Technol 2011;27:39–42.
[62] Khan ZU, Hayat M, Khan MA. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J Theor Biol 2015;365:197–203.
[63] Devi CJ, Reddy BSP, Kumar KV, Reddy BM, Nayak NR. ANN approach for weather prediction using back propagation. Int J Eng Trends Technol 2012:2012.
[64] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
[65] Tahir M, Hayat M, Kabir M. Sequence based predictor for discrimination of Enhancer and their Types by applying general form of Chou's trinucleotide composition. Comput Methods Programs Biomed 2017;146:69–75.
[66] Tahir M, Hayat M. iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC. Mol Biosyst 2016;12:2587–93.
[67] Yang H, Tang H, Chen X-X, Zhang C-J, Zhu P-P, Ding H, et al. Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition. BioMed Res Int 2016;2016.
[68] Zhang C-J, Tang H, Li W-C, Lin H, Chen W, Chou K-C. iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. Oncotarget 2016;7:69783–93.
[69] Ding H, Li D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. Amino Acids 2015;47:329–33.
[70] Che Y, Ju Y, Xuan P, Long R, Xing F. Identification of multi-functional enzyme with multi-label classifier. PloS One 2016;11:e0153503.
[71] Lin H, Ding C, Song Q, Yang P, Ding H, Deng K-J, et al. The prediction of protein structural class using averaged chemical shifts. J Biomol Struct Dyn 2012;29:1147–53.
[72] Chou K-C, Zhang C-T. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
[73] Ding H, Liang Z-Y, Guo F-B, Huang J, Chen W, Lin H. Predicting bacteriophage proteins located in host cell with feature selection technique. Comput Biol Med 2016;71:156–61.
[74] Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. J Microbiol Methods 2011;84:67–70.
[75] Yuan L-F, Ding C, Guo S-H, Ding H, Chen W, Lin H. Prediction of the types of ion channel-targeted conotoxins based on radial basis function network. Toxicol In Vitro 2013;27:852–6.
[76] Ding H, Feng P-M, Chen W, Lin H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Mol Biosyst 2014;10:2229–35.
[77] Chen X-X, Tang H, Li W-C, Wu H, Chen W, Ding H, et al. Identification of bacterial cell wall lyases via pseudo amino acid composition. BioMed Res Int 2016;2016.
[78] Ding H, Luo L, Lin H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Pept Lett 2009;16:351–5.
[79] Cai Y-D, Liu X-J, Xu X-b Chou K-C. Prediction of protein structural classes by support vector machines. Comput Chem 2002;26:293–6.
[80] Lin H, Liang Z-Y, Tang H, Chen W. Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM Trans Comput Biol Bioinform 2017.

[81] Zhang T, Tan P, Wang L, Jin N, Li Y, Zhang L, et al. RNALocate: a resource for RNA subcellular localizations. Nucleic Acids Res 2017;45:D135–8.

[82] Lin H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 2008;252(2):350–6.

[83] Che Y, Ju Y, Xuan P, Long R, Xing F. Identification of multi-functional enzyme with multi-label classifier, PLoS ONE 11(4): e0153503.

[84] Zuo Y, Lv Y, Wei Z, Yang L, Li G, Fan G. iDPF-PseRAAAC: a web-server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition, PLoS ONE 10(12): e0145541.

[85] Wuyun Q, Zheng W, Zhang Y, Ruan J, Hu G. Improved Species-Specific Lysine Acetylation Site Prediction Based on a Large Variety of Features Set, PLoS ONE 11(5): e0155370.