frontiers in Immunology

ORIGINAL RESEARCH
published: 30 January 2020
doi: 10.3389/fimmu.2020.00071

A Hybrid Model for Predicting Pattern Recognition Receptors Using Evolutionary Information

Dilraj Kaur†, *Chakit Arora*† *and Gajendra P. S. Raghava**

Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

This study describes a method developed for predicting pattern recognition receptors (PRRs), which are an integral part of the immune system. The models developed here were trained and evaluated on the largest possible non-redundant PRRs, obtained from PRRDB 2.0, and non-pattern recognition receptors (Non-PRRs), obtained from Swiss-Prot. Firstly, a similarity-based approach using BLAST was used to predict PRRs and got limited success due to a large number of no-hits. Secondly, machine learning-based models were developed using sequence composition and achieved a maximum MCC of 0.63. In addition to this, models were developed using evolutionary information in the form of PSSM composition and achieved maximum MCC value of 0.66. Finally, we developed hybrid models that combined a similarity-based approach using BLAST and machine learning-based models. Our best model, which combined BLAST and PSSM based model, achieved a maximum MCC value of 0.82 with an AUROC value of 0.95, utilizing the potential of both similarity-based search and machine learning techniques. In order to facilitate the scientific community, we also developed a web server "PRRpred" based on the best model developed in this study (http://webs.iiitd.edu.in/raghava/prrpred/).

Keywords: pattern recognition receptors, prediction, innate immunity, machine learning, BLAST, toll-like receptors

OPEN ACCESS

Edited by:
Junji Xing,
Houston Methodist Research Institute, United States

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Taruna Madan,
National Institute for Research in Reproductive Health (ICMR), India

***Correspondence:**
Gajendra P. S. Raghava
raghava@iiitd.ac.in

†*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Molecular Innate Immunity, a section of the journal Frontiers in Immunology

Received: *18 November 2019*
Accepted: *13 January 2020*
Published: *30 January 2020*

Citation:
Kaur D, Arora C and Raghava GPS (2020) A Hybrid Model for Predicting Pattern Recognition Receptors Using Evolutionary Information. Front. Immunol. 11:71. doi: 10.3389/fimmu.2020.00071

INTRODUCTION

Pattern Recognition Receptors (PRRs) are germline-encoded proteins that are capable of sensing invading pathogens by binding to the so-called pathogen-associated molecular patterns (PAMPs) found in pathogens. These PRRs also bind to Damage-Associated Molecular Patterns (DAMPs) which are molecules released by damaged cells. This recognition of PAMPs and DAMPs by PRRs initiates a cascade of signaling processes and activates microbicidal and pro-inflammatory responses. It leads to elimination of infectious agents, and at the same time, represents an essential link to the adaptive immune response (1). There are four significant sub-families of PRRs—Toll-like receptors (TLRs), nucleotide-binding oligomerization domain (NOD)-Leucine Rich Repeats (LRR)-containing receptors (NLR), retinoic acid-inducible gene 1 (RIG-1)-like receptors (RLR), and C-type lectin receptors (CLRs). While TLRs and CLRs are transmembrane proteins, NLRs and RLRs are cytoplasmic proteins. These PRRs play essential roles in bacterial, viral, and fungal recognition (2). Several other PRRs such as scavenger receptors, mannose receptors, and β-glucan receptors induce phagocytosis. Other secreted PRRs are complement receptors, collectins, ficolins, pentraxins for instances, serum amyloid, and C-reactive protein, lipid transferases, peptidoglycan recognition proteins (PGRs), XA21D (3).

Various studies in the past have exhibited the importance of PRRs in diseases such as autoimmune disorders (4, 5), atherosclerosis, sepsis, asthma (6), heart failure (4), kidney diseases (7), bacterial meningitis, viral encephalitis, stroke, Alzheimer's disease (AD), Parkinson's disease (PD) (5), immunodeficiency disorders like chronic granulomatous disease (CGD), and X-linked agammaglobulinemia (XLA) (8), Cancer (9–12). Thus, PRRs have emerged as an important area for therapeutic research specifically in adjuvant designing (13–16). Hence, it is vital to have a deep understanding of PRR machinery and their functional roles in innate immunity. Broadly, PAMPs and DAMPs bind to PRRs, which results in signals that prompt leukocyte recruitment (17). Cell types expressing PRRs include innate immune cells such as macrophages, monocytes, dendritic cells, and mast cells but also non-immune cells such as epithelial cells and fibroblasts (18). Pattern recognition receptor-ligand binding and their joint conformational changes elicit a cascade of downstream signaling. This cascade results in transcriptional changes as well as post-translational modifications (17). Traditional methods for identifying PRRs include experimental techniques such as immunofluorescence (19), Quantitative real-time PCR, Cell viability assay, Immunoblot and Immunoprecipitation (20, 21), Microbial Binding and Agglutination Assay (22), PAMP binding assay (22, 23), ELISA (24–26), Growth-inhibition assay (27). These experimental techniques are highly accurate but costly and time-consuming. Recent advances in technology have led to the development of various *in-silico* methods for predicting the function of a protein. Besides being faster and inexpensive, these methods are also reproducible. The data required for building such prediction methods can be obtained from various web-resources/databases/repositories such as InnateDB (28), IEDB (29), IIDB (30), Vaxjo (31), and VIOLIN (32). Due to its crucial role in innate immunity, the prediction of PRRs is a necessity to further aid research and efficient-therapy design. So far, only one prediction method (33) for sub-family classification of PRRs has been developed in the past, based on data obtained from PRRDB (34). This method, however, used a relaxed criterion for dataset preparation (CD hit at 90% cutoff) due to scarce data. This dataset was subsequently used for training and testing of machine learning models. Since their processed dataset contains homologous sequences, the model prediction results could be biased.

In order to complement and overcome the limitations of the existing method, we developed a method using the largest possible dataset, derived from PRRDB 2.0 (3) database, with standard protocols. In this study, we used protocols that divide the data into five data sets in such a way that no two proteins in two different subsets have more than 40% sequence similarity, without reducing the number of sequences in the dataset (35, 36). In order to understand the strength and limitation of the standard similarity-based approach, we evaluate the performance of BLAST on our dataset. In the second step, we developed standard machine-learning based classification models for predicting PRRs using a wide range of descriptors like residue composition and dipeptide composition (37–40). It has been shown in the past that evolutionary information provides more information than single sequence (41, 42). Thus, we developed models using evolutionary information in the form of the composition of the position-specific scoring matrix (PSSM) profile (37). Finally, we developed hybrid models that combine the strength of different approaches used in this study (42, 43). We show that the hybrid model that comprises of BLAST based similarity search and PSSM profile based Random Forest (RF) classifier, is the best *in-silico* classification method for predicting PRRs. This model is freely available for public use in the form of the web-server "PRRpred" (http://webs.iiitd.edu.in/raghava/prrpred/), to assist and aid further research on PRRs.

MATERIALS AND METHODS

Dataset

PRRs sequences (positive data) were obtained from the database PRRDB2.0 (3). Initially, the total PRRs taken were 2,727, which were reduced to 179 unique PRRs after the removal of identical sequences. The negative dataset was created by collecting random sequences from Swiss-Prot (44), which were not PRRs. The negative dataset constituted of 274 Non-PRR sequences. In order to create non-redundant subsets without reducing number of sequences, we used an approach previously described by Bendtsen et al. (35) and Garg et al. (36). Following the same approach, CD-HIT (45) with a cutoff of 40% sequence similarity was used on both positive and negative datasets, to obtain positive and negative clusters respectively. Each cluster is a collection of similar sequences based on the cutoff. A total of 106 positive clusters and 210 negative clusters were obtained. The distribution of the sequences in the clusters is shown in **Figure 1**. For the positive dataset, five subsets were created from the clusters obtained by CD-HIT. All the sequences in the first cluster were assigned to the first subset and the next cluster's sequences to the subsequent subset and so on. We continued this process until all sequences (contained in CD-HIT

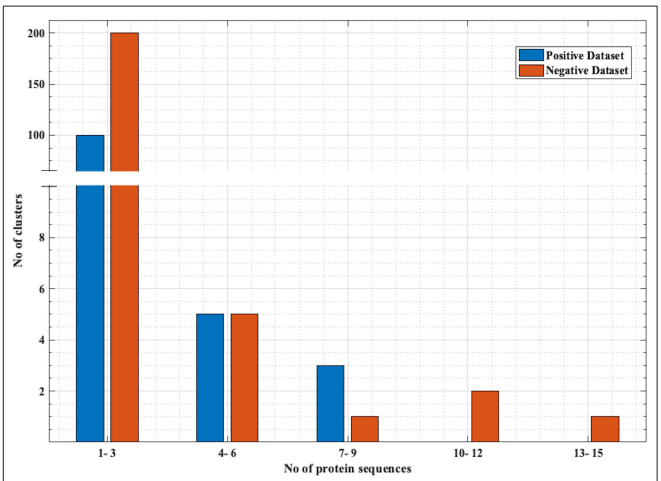

FIGURE 1 | Distribution of the sequences in negative and positive clusters obtained from CD-HIT. x-axis represents the number of sequences and y-axis represents number of clusters that have those number of sequences. Most of the positive and negative clusters have a smaller number of sequences, while there are a few clusters with a comparatively larger number of sequences.

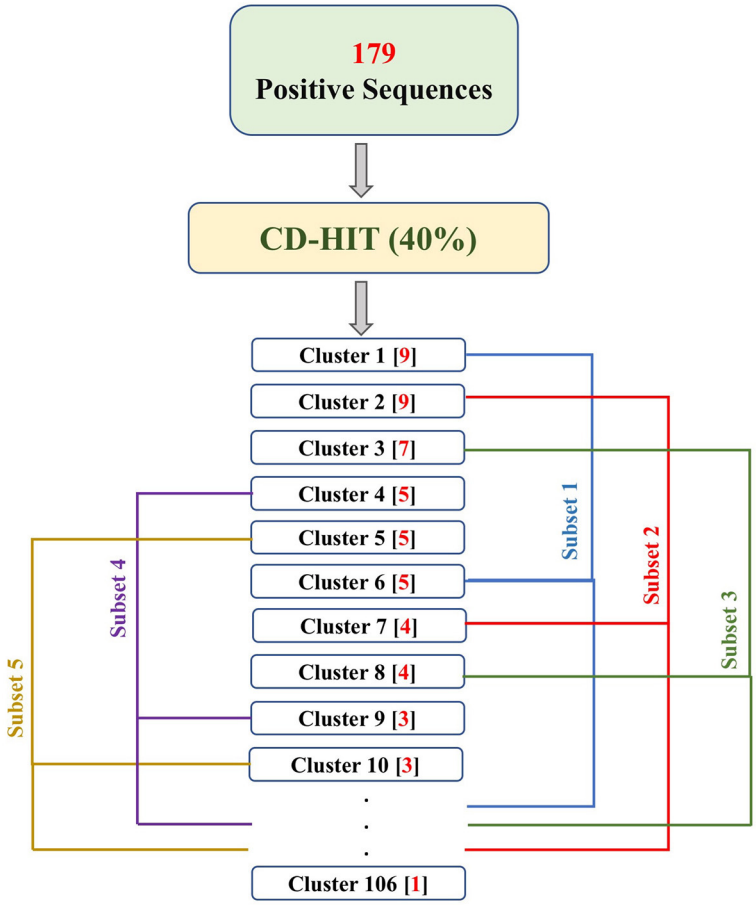

FIGURE 2 | The flowchart explains the process of fractioning positive clusters obtained from CD-HIT into five subsets. The numbers in the parentheses, following the cluster names, represent the number of sequences in that cluster. As a result, Subset 1 contains sequences of clusters 1, 6, 11, …, 106; Subset 2 contains sequences of cluster 2, 7, 12, …, 102; Subset 3 contains sequences of cluster 3, 8, 13, …, 103; Subset 4 contains sequences of cluster 4, 9, 14, …, 104; and Subset 5 contains sequences of cluster 5, 10, 15, …, 105.

generated clusters) were distributed in the five subsets. **Figure 2** explains this procedure diagrammatically. A similar process was implemented for negative dataset to create five negative subsets. This strategy makes sure that the subsets are dissimilar to each other (<40% similarity between sequences in two subsets), which will be beneficial for unbiased training and testing of machine learning models and selection of a better classification model. The aim of this process is to create non-redundant dataset without reducing the number of proteins from the dataset (35, 36).

Five-Fold Cross Validation

The performance of the modules constructed in this report was evaluated using five-fold cross-validation technique. Training and test sets were formed using positive and negative subsets. Four positive and the corresponding four negative subsets were combined to form the training set. The remaining one positive and one corresponding negative subset were combined to form the test set. This process is repeated five times, such that the combination of a positive subset and the corresponding negative subset is used as a test set exactly once. We employed these five training and test sets for performing five-fold cross-validation to select the best machine learning models as well as for developing BLAST similarity search-based module, as explained in the next sections. Five-fold cross-validation is a standard process that has been successfully implemented in several machine learning-based studies in the past (39, 46–50).

BLAST Based Similarity Search

A similarity search based module was designed based on pBLAST (BLAST+ 2.7.1) (51). To evaluate the performance of this module, five-fold cross-validation was implemented. For this, a train set was used to make a local database against which the query sequences (sequences in the test set) were searched at an *e*-value of 0.001. The procedure is repeated five times (for each training and test set), and the evaluation metrics are noted (Results). Finally, in the web-server implementation, the total positive (179 PRRs) and negative dataset (274 Non-PRRs) have been combined to make a database of 453 proteins against which the user's unseen query protein can be searched.

Protein Features

Composition Based Features

Amino acid composition (AAC) and di-peptide composition (DPC) were obtained from Pfeature and used as features that provide residue information of a protein. AAC, for a protein sequence, is a 20-length vector where each element is the fraction of a specific type of residue in the sequence. DPC, on the other hand, is a 400-length vector that gives the composition of the amino-acids present in pairs (e.g., L-M, G-L, and so on) in the protein sequence. The detailed information can be obtained from Pfeature (52).

Evolutionary Information-Based Features

In this study, we obtained evolutionary information for a protein using PSI-BLAST. We implemented evolutionary information in the form of PSSM-400 composition profile as a feature, similar to the previous studies (37, 53–57). PSSM-400 for a protein sequence is a 20 x 20 dimensional vector, which is the composition of occurrences of each type of 20 amino acids corresponding to each type of amino acids in the protein sequence. For each protein sequence, PSSM matrix was created, which was then normalized and converted to a 20 x 20 PSSM composition vector using Pfeature's (52) "Evolutionary Info" module.

Machine Learning Techniques

We used Sci-Kit's sklearn package, consisting of various classifiers, to develop prediction models. Each of these methods requires fixed-length feature vectors. The maximum information about proteins of variable lengths was converted into fixed vectors of equal dimensions (AAC, DPC, PSSM-400), and then these were used as input features. We used Sci-Kit's GridSearch package to tune hyper-parameters in order to get the best performance on the training set. Subsequently, the best-learned model was used for the test. This process was implemented using five-fold cross-validation, and the average performance of five-folds was evaluated. Different Machine Learning based classifiers were then used to develop prediction models. The most basic classifier i.e., Logistic Regression (LR) was used to handle linear data, while to handle non-linear data more advanced classifiers such as Random Forest (RF), Support Vector Machine (SVM), Extra Trees (ET), K-Nearest Neighbor (KNN), and Multi-Layer Perceptron (MLP) were used. All these machine learning methods have been successfully applied in many bioinformatics studies (39, 46, 49, 50, 58).

Performance Evaluation Parameters

Each model used in the study was evaluated using threshold independent and dependent scoring parameters. Threshold dependent parameters used here are Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), and Matthew's correlation coefficient (MCC). "Sens" is defined as true positive rate (TPR) i.e., correctly predicted positives with respect to actual total positives, whereas true negative rate (TNR) is defined by "Spec." "Acc" is the ability of the model to differentiate between true positives and true negatives, while MCC is the correlation coefficient between predicted and actual classes. Following relations were used to calculate these:

$$Sens = \frac{TP}{P} \times 100$$

$$Spec = \frac{TN}{N} \times 100$$

$$Acc = \frac{TP + TN}{P + N} \times 100$$

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

Where TP denotes correctly predicted positive, TN denotes the correct negative predictions, P denotes the total sequences in the positive set, N denotes the total sequences in the negative set, FP denotes actual negative sequences which have been wrongly predicted as positive, and FN represents wrongly predicted positive sequences. These scoring parameters are well-established and have been used in many studies for model's performance evaluation. Area under Receiver Operating Characteristic Curve (AUROC) value is a threshold independent parameter, which is calculated via the plot between True positive rate (TPR or Sens) and False positive rate (FPR or 1-Spec) (59).

Hybrid Models for Classification

In order to improve the accuracy of the machine learning-based models further, hybrid models were constructed that combined the BLAST prediction score with the ML-based scores as done in ALGpred (60). We assigned a score of "+0.5" for positive prediction (PRRs), "−0.5" for negative prediction (Non-PRRs), and "0" for no hits (NH). This score was added to the Machine learning-based model score (i.e., prediction probability of positive class). This is done for each of the sequences in the test set in a five-fold cross-validation process. Then based on this combined score, scoring metrics were evaluated for each ML model at various probability cutoffs.

RESULTS

Prediction Based on Similarity Search

One of the standard software which is commonly used for similarity search is BLAST. Thus, we used BLAST for discriminating PRRs and Non-PRRs. In order to avoid bias, we used five-fold cross-validation, where proteins in the test set were searched against the training set using BLAST at different *e*-value cut-offs (**Table 1**). This process is repeated five times to cover all the proteins in our training sets. The positive dataset consists of 179 PRRs, and a negative dataset consists of 274 Non-PRRs. As shown in **Table 1**, the number of correctly predicted PRRs increased from 74.30 to 82.12% with *e*-value from 10^{-9} to 10^{-0} or 1. Though the performance of correctly predicted PRRs (sensitivity) increased with an increase in *e*-value, the rate of error (% of Non-PRRs) also increased. In the case of Non-PRRs, specificity increased from 32.48 to 49.68%, and the error rate also increased from 1.67 to 10.05% with *e*-value from 10^{-9} to 10^{-0}. The overall accuracy of BLAST was only around 51% at *e*-value 10^{-3}; due to a significant number of no-hits. This poor performance shows that

TABLE 1 | The performance of BLAST on training and testing dataset using five-fold cross validation. PRRs, and non-PRRs were searched at different *e*-values of BLAST.

BLAST (*e*-value)	Positive hits (Searching PRRs)		Negative hits (Searching Non-PRRs)	
	PRRs (Sensitivity)	Non-PRRs (Error)	Non-PRRs (Specificity)	PRRs (Error)
10^{-9}	133 (74.30)	4 (1.45)	89 (32.48)	3 (1.67)
10^{-8}	134 (74.86)	4 (1.45)	90 (32.84)	4 (2.23)
10^{-7}	134 (74.86)	5 (1.82)	90 (32.84)	4 (2.23)
10^{-6}	135 (75.41)	5 (1.82)	93 (33.94)	4 (2.23)
10^{-5}	136 (75.97)	7 (2.55)	98 (35.76)	5 (2.79)
10^{-4}	136 (75.97)	7 (2.55)	99 (36.13)	6 (3.35)
10^{-3}	138 (77.09)	8 (2.92)	101 (36.86)	6 (3.35)
10^{-2}	139 (77.65)	10 (3.64)	102 (37.22)	6 (3.35)
10^{-1}	140 (78.21)	20 (7.29)	107 (39.05)	7 (3.91)
1	147 (82.12)	65 (23.72)	135 (49.27)	18 (10.05)

BLAST is not suitable to discriminate PRRs and Non-PRRs with high precision.

Models Based on Machine Learning Techniques

Composition-Based Features

In order to develop a method for classification of PRRs and Non-PRRs, we used two main sequence composition-based features viz. (i) Amino acid composition and (ii) Dipeptide composition. A wide range of machine learning techniques (e.g., SVM, KNN, RF) were used for developing prediction models. We examined the frequency of the 20 amino acids in both the positive and negative datasets. A comparison of amino acid composition between PRRs and Non-PRRs showed that residues L, N, S, and Q are more abundant in PRRs whereas A, D, E, K, and V are frequent in Non-PRRs (**Figure 3**). The composition of PRRs is different from the composition of Non-PRRs, as shown in **Figure 3**. Thus, amino acid composition (AAC) feature can be used to develop models for discriminating two classes. Following machine learning techniques were used for developing binary classification models; (i) Extra-trees (ET), (ii) Random forest (RF), (iii) Support vector machine (SVM), (iv) K nearest neighbor (KNN), (v) Logistic regression (LR), and (vi) Multi-layer perceptron (MLP). As shown in **Table 2**, ET based models got a maximum AUROC of 0.90 with an MCC value of 0.63 on the training dataset. We achieved AUROC as 0.88 with MCC 0.63 on the test dataset.

Similarly, models were constructed using dipeptide composition and using different machine learning techniques (**Table S1**). The best performance was noted for LR with an average accuracy of 80.25%, MCC value of 0.59, and AUROC of 0.87 at test set, while on the training dataset, an average accuracy of 82.57% was noted with MCC value of 0.64 and AUROC value of 0.88. Overall test accuracy was 83% in the case of LR, with MCC of 0.64 and AUROC of 0.88.

Composition of Sequence Profile

It has been shown in the past that the sequence profile provides more information than single sequence. Thus, in this study, first, we generate sequence profile corresponding to a protein using PSI-BLAST software. In order to generate a fixed number of features, we compute the composition of sequence profile or PSSM profile (see section Materials and Methods). We represent the composition of PSSM profile by PSSM-400, which has a fixed-length vector of 400 elements. We generated the PSSM-400 composition profiles for our dataset and used it as feature vectors for developing classification models. Similar to the AAC and DPC based methods, we use various classifiers such as SVM, RF, ET, MLP, etc. for training and test purposes. As shown in **Table 3**, models based on evolutionary information showed a maximum AUROC of 0.87 with MCC of 0.64 on the training dataset. Similarly, on the test dataset, the maximum AUROC was 0.89, with MCC of 0.66. As compared to composition-based prediction models, PSSM based prediction model showed higher performance in terms of MCC. In terms of AUROC, the performance of both composition and PSSM based methods was nearly the same.

Combination of Sequence and PSSM Composition

The PSSM composition was combined with amino acid composition to generate a 420-length feature vector. Various classifiers were used for training and testing using five-fold cross-validation. As shown in **Table 4**, we got an AUROC value of 0.89 with a MCC value of 0.66 using LR on training sets. Similarly, maximum AUROC was obtained on using MLP with AUROC of 0.90 and of 0.67 on the test dataset. Thus, the performance has been improved as compared to using evolutionary information-based features (PSSM) or composition-based features (AAC or DPC) alone. **Figure 4** shows the ROC curves for different classifiers corresponding to AAC, PSSM, and the combination of AAC and PSSM.

Hybrid Models

It is apparent from the previous results that both similarity-based approach and machine learning-based models have their own pros and cons. Thus, we made an attempt to develop a method that combines the strengths of both approaches. The *e*-value of 10^{-3} was selected for the BLAST-based similarity search method based on the hits against PRRs. Since at this *e*-value, the probability of correct prediction was found to be reasonably high (77.09%), and the rate of error was very low (2.92%). Though the number of no-hits was too high at this cutoff (∼80%), it was compensated by a high prediction accuracy. In order to integrate the two approaches, proteins were first classified using machine learning models. In the second step, the proteins were again classified using BLAST wherein the query proteins which showed similarity with PRRs at *e*-value of 10^{-3} were assigned as PRRs. We gave preference to BLAST over machine learning-based models in predicting PRRs due to the high probability of correct prediction of the BLAST-based similarity search method. In simple words, we used machine learning techniques for classifying proteins as PRRs and Non-PRRs when there is no BLAST hit for query protein at BLAST

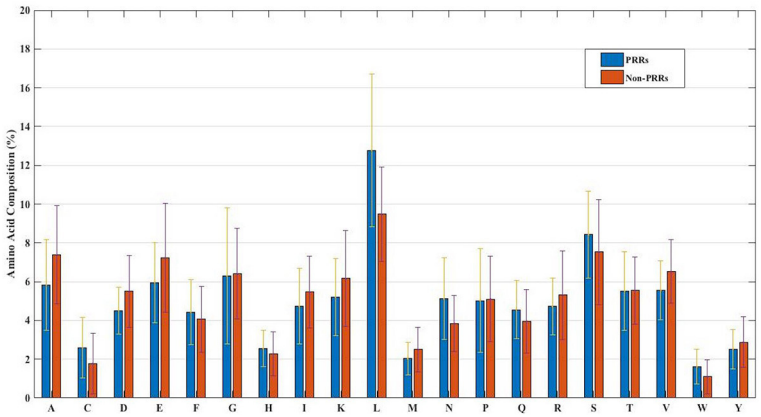

FIGURE 3 | The percent amino acid composition of pattern recognition receptors and non-pattern recognition receptor proteins.

TABLE 2 | The performance of different machine learning techniques-based models on PRR dataset developed using AAC of protein sequences.

	Method	Train dataset (Average)					Test dataset (Average)				
Model	**Hyper-parameters***	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**
ET	ne = 90	80.71	82.56	81.73	0.90	0.63	77.06	84.08	82.46	0.88	0.63
SVM	C = 5, g = 0.01, k = rbf	78.07	83.83	81.62	0.87	0.62	77.95	82.31	81.06	0.88	0.60
RF	ne = 100	77.82	81.46	80.08	0.88	0.59	77.42	80.85	79.97	0.87	0.58
LR	C = 1	77.98	82.50	80.77	0.86	0.60	76.12	81.57	79.57	0.86	0.58
MLP	a = tanh, HL = (19,), m = 200, s = adam	77.02	82.77	80.50	0.86	0.59	78.88	77.94	78.90	0.87	0.57
KNN	al = ball_tree, nn = 20, w = distance	76.17	79.06	77.91	0.85	0.55	77.74	75.00	76.97	0.86	0.53

g, gamma; ne, n_estimators; k, kernel; a, activation; HL, hidden layer size; s, solver; al, algorithm; w, weight; m, max_iter; nn, n_neighbors.

TABLE 3 | The performance of different machine learning techniques-based models on PRR dataset developed using PSSM-400 of protein sequences.

	Method	Train dataset (average)					Test dataset (average)				
Model	**Hyper-parameters***	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**
SVM	C = 10, g = 0.5, k = rbf	77.80	85.89	82.78	0.87	0.64	79.74	85.46	83.64	0.89	0.66
LR	C = 1,000	77.31	86.37	82.84	0.87	0.64	80.80	81.07	81.13	0.89	0.61
KNN	al = ball_tree, nn = 6, w = distance	72.80	83.48	79.36	0.86	0.57	78.40	82.50	81.07	0.87	0.60
RF	ne = 80	75.95	85.01	81.55	0.87	0.61	79.07	81.41	80.74	0.86	0.60
MLP	a = logistic, HL = (14), m = 200, s = adam	75.26	85.09	81.28	0.86	0.61	79.07	81.03	80.26	0.88	0.59
ET	ne = 70	80.33	78.79	79.36	0.88	0.58	83.73	74.97	79.15	0.87	0.59

g, gamma; ne, n_estimators; k, kernel; a, activation; HL, hidden layer size; s, solver; al, algorithm; w, weight; m, max_iter; nn, n_neighbors.

TABLE 4 | The performance of different machine learning techniques-based models on PRR dataset developed using the combination of composition (AAC) and evolutionary information (PSSM-400) based features for protein sequences.

	Method	Train dataset (Average)					Test dataset (Average)				
Model	**Hyper-parameters***	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**
MLP	a = tanh, HL, = (70,), m = 200, s = adam	77.70	86.54	83.20	0.88	0.65	81.23	85.50	84.19	0.90	0.67
LR	C = 1,000	82.97	83.49	83.34	0.89	0.66	83.59	81.49	82.67	0.90	0.64
RF	ne = 60	80.32	83.24	82.16	0.88	0.63	80.43	82.44	82.16	0.87	0.63
ET	ne = 100	77.72	85.25	82.35	0.89	0.63	78.96	83.75	82.13	0.88	0.63
SVC	C = 5, g = 0.01, k = rbf	81.65	83.35	82.73	0.89	0.65	80.62	81.56	81.72	0.88	0.62
KNN	al = ball_tree, nn = 20, w = distance	80.20	76.60	78.12	0.87	0.56	80.41	72.88	76.35	0.86	0.52

g, gamma; ne, n_estimators; k, kernel; a, activation; HL, hidden layer size; s, solver; al, algorithm; w, weight; m, max_iter; nn, n_neighbors.

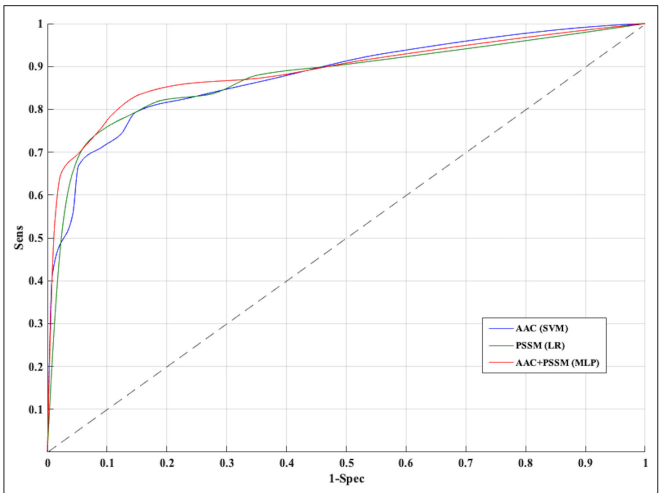

FIGURE 4 | Receiver operating characteristic curves for five-fold cross-validation based on AAC, PSSM, AAC+PSSM using Support vector machine (SVM), Logistic Regression (LR), and Multi-layer perceptron (MLP) based classifier, respectively.

e-value of 10^{-3}. This hybrid strategy improved the coverage, which was earlier missing while using BLAST alone. As shown in **Table 5**, the performance of machine learning techniques improved drastically when BLAST was integrated. Our best hybrid model (RF) based on PSSM achieved an accuracy of 91.39% and AUROC of 0.95 with an MCC of 0.82. In general, the performance of all hybrid models was observed to be better than the BLAST-based similarity search and models based on machine learning techniques.

WEB-SERVER INTERFACE

Providing service to the scientific community is one of the primary goals of this study. We developed a user-friendly web server (http://webs.iiitd.edu.in/raghava/prrpred/), which allows users to predict whether a given protein is a pattern recognition receptor or not. The web interface of the server has two sub-modules under prediction: (i) Composition Based and (ii) Evolutionary Information Based. The "Composition Based" module allows a user to identify a protein sequence based on Amino acid composition. This module further provides the user with the option to choose the non-hybrid method, which is only AAC based and hybrid method, which is AAC+BLAST based. The "Evolutionary Information Based" module facilitates the user to predict PRRs from evolutionary information of a protein sequence. Here, the PSSM-400 composition profile for the entered protein sequence is generated and is used as a feature vector for the prediction. This module also has the facility of non-hybrid and hybrid models, similar to the composition-based module. The web server has been designed by using a responsive HTML template for adjustment to the browsing device. Thus, our web server is compatible with a wide range of devices, including desktops, tablets, and smartphones.

DISCUSSION

Over the past few years, there have been rapid advances in understanding innate immunity, particularly about the mechanisms by which pathogens are recognized and how the signaling molecules respond to them. Innate immunity is gaining more attention than adaptive immunity due to its role in combating the pathogens during the early stages of infection, while adaptive immunity comes later into the picture. Adaptive immunity comprises of receptors which are highly specific to antigens (61). In contrast, innate immunity consists of specialized receptors known as PRRs that recognize infectious pathogens and initiate inflammatory responses for their eradication (62). Several critical implications of PRRs have been reported in the past in the context of adjuvant designing, therapeutic targets, immunomodulator design, cancer immunotherapy, etc. (61, 63, 64). A comprehensive database of pathogen recognizing receptors such as PRRDB (34) is highly essential to understand innate immunity. These kinds of knowledge-based resources can assist researchers working in the area of innate immunity and drug development. In addition to resources, there is a need to develop methods than can annotate newly sequenced PRRs. Recently, a method has been developed using SVM for predicting PRRs and subfamilies (33). This method uses amino acid and pseudo-amino acid composition (PseAAC) for developing models using dataset derived from PRRDB (34). The prediction was based on 332 PRR sequences (containing different families) obtained from 473 sequences (that includes multiple similar UniProt IDs), which were originally present in the database, by employing CD-HIT at 90% cutoff. The model accuracy was reported to be ∼97–98%; however, such a relaxed redundancy reduction process employs sequences that can be similar up to a very high degree. In this paper, we used the dataset obtained from the recently updated version PRRDB2.0 (3) to develop classification models. The positive dataset in our case consists of PRR sequences with unique UniProt IDs, thereby first reducing the redundant data (1,784 sequences) to 179 sequences. Secondly, CD hit at 40% cutoff was applied to divide both the negative (274 random Non-PRR sequences from swiss-prot) and positive datasets to five subsets each. This helped in reducing homology bias amongst the train and test datasets, and thus more precise training of the models during five-fold cross-validation.

Here, we tried various approaches to predict PRRs. We used different protein features such as composition-based features (AAC and DPC) and evolutionary information-based features (PSSM) to develop machine learning-based models in order to distinguish PRRs and Non-PRRs. We also used the combination of composition-based features and evolutionary information-based features for the same. These approaches were used for the first time in the study of predicting PRRs. To do this, we used a variety of classifiers available in Sci-Kit's sklearn such as SVM, RF, ET, MLP. Firstly, we tried BLAST only classification due to its simplicity and wide popularity. Though BLAST resulted in a very high accuracy (*e*-value of 10^{-3}) whenever a hit was found, it was unable to predict around 80% of sequences (No-Hits) during five-fold cross-validation. Thus, we employed a hybrid approach for the problem in hand, which combines ML-based

TABLE 5 | The performance of different machine learning techniques-based models on test dataset when combined with BLAST hits at *e*-value 10^{-3}.

Feature	Model	Hyper-parameters*	Sens	Spec	Acc	AUROC	MCC
PSSM	RF	C = 80	83.24	96.72	91.39	0.95	0.82
AAC	RF	C = 100	82.12	94.53	89.62	0.92	0.78
AAC+PSSM	ET	ne = 100	87.15	89.78	88.74	0.95	0.77
DPC	SVC	C = 2, g = 0.01, k = rbf	79.89	92.34	87.42	0.93	0.73

g, gamma; ne, n_estimators; k, kernel.

methods with BLAST. The major advantage of this strategy is that the proteins which could not be predicted by BLAST alone can be predicted using ML. We tried this approach with each of the protein-features and their combinations, using an extensive range of classifiers. The best performance was achieved in the hybrid case of PSSM and BLAST. The formulation of this hybrid model was implemented in the free web-server. Using the web-server, for an unknown protein sequence, this model will first predict the positive (PRR) or negative (Non-PRR) class, based on BLAST search against the entire database (179 PRRs+274 Non-PRRs). If the result is a "No-Hit," the prediction will then be made by the RF model trained on the complete set. The web-server is freely available and easy to use. We believe that the work done here will be beneficial for the annotation of PRRs and boost the ongoing research in the field of innate immunity.

DATA AVAILABILITY STATEMENT

The datasets used for this study can be found at the PRRpred webserver (https://webs.iiitd.edu.in/raghava/prrpred/dataset.php).

AUTHOR CONTRIBUTIONS

DK and CA generated the dataset, performed data analysis, prepared figures and tables, and developed the web interface. DK, CA, and GR wrote the manuscript. GR conceived the idea and coordinated the project.

FUNDING

This work was supported by J.C. Bose Fellowship (Grant No. SRP076), Department of Science and Technology, India.

ACKNOWLEDGMENTS

Authors are thankful to Indraprastha Institute of Information Technology (IIIT-Delhi) for financial support and fellowships.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.00071/full#supplementary-material

REFERENCES

1. Suresh R, Mosser DM. Pattern recognition receptors in innate immunity, host defense, and immunopathology. *Adv Physiol Educ.* (2013) 37:284–91. doi: 10.1152/advan.00058.2013
2. Kawai T, Akira S. The roles of TLRs, RLRs and NLRs in pathogen recognition. *Int Immunol.* (2009) 21:317–37. doi: 10.1093/intimm/dxp017
3. Kaur D, Patiyal S, Sharma N, Usmani SS, Raghava GPS. PRRDB 2.0: a comprehensive database of pattern-recognition receptors and their ligands. *Database.* (2019) 2019:baz076. doi: 10.1093/database/baz076
4. Farrugia M, Baron B. The role of toll-like receptors in autoimmune diseases through failure of the self-recognition mechanism. *Int J Inflam.* (2017) 2017:8391230. doi: 10.1155/2017/8391230
5. Kumar V. Toll-like receptors in the pathogenesis of neuroinflammation. *J Neuroimmunol.* (2019) 332:16–30. doi: 10.1016/j.jneuroim.2019.03.012
6. Lin Y-T, Verma A, Hodgkinson CP. Toll-like receptors and human disease: lessons from single nucleotide polymorphisms. *Curr Genomics.* (2012) 13:633–45. doi: 10.2174/138920212803759712
7. Komada T, Muruve DA. The role of inflammasomes in kidney disease. *Nat Rev Nephrol.* (2019) 15:501–20. doi: 10.1038/s41581-019-0158-z
8. Mortaz E, Adcock IM, Tabarsi P, Darazam IA, Movassaghi M, Garssen J, et al. Pattern recognitions receptors in immunodeficiency disorders. *Eur J Pharmacol.* (2017) 808:49–56. doi: 10.1016/j.ejphar.2017.01.014
9. O' Donovan DH, Mao Y, Mele DA. The next generation of pattern recognition receptor agonists: improving response rates in cancer immunotherapy. *Curr Med Chem.* (2019) 26:1. doi: 10.2174/0929867326666190620103105
10. do Prado SBR, Castro-Alves VC, Ferreira GF, Fabi JP. Ingestion of non-digestible carbohydrates from plant-source foods and decreased risk of colorectal cancer: a review on the biological effects and the mechanisms of action. *Front Nutr.* (2019) 6:72. doi: 10.3389/fnut.2019.00072
11. Qin S, Dong LP, Bai B, Xue HC. Influence of Toll-like receptor 7 on CD8(+) T lymphocytes in patients with breast cancer. *Zhonghua Yi Xue Za Zhi.* (2019) 99:1562–6. doi: 10.3760/cma.j.issn.0376-2491.2019.20.009
12. Haider T, Tiwari R, Vyas SP, Soni V. Molecular determinants as therapeutic targets in cancer chemotherapy: an update. *Pharmacol Ther.* (2019) 200:85–109. doi: 10.1016/j.pharmthera.2019.04.011
13. Olive C. Pattern recognition receptors: sentinels in innate immunity and targets of new vaccine adjuvants. *Expert Rev Vaccines.* (2012) 11:237–56. doi: 10.1586/erv.11.189
14. Shirota H, Tross D, Klinman DM. CpG oligonucleotides as cancer vaccine adjuvants. *Vaccines.* (2015) 3:390–407. doi: 10.3390/vaccines3020390
15. Dowling JK, Mansell A. Toll-like receptors: the swiss army knife of immunity and vaccine development. *Clin Transl Immunol.* (2016) 5:e85. doi: 10.1038/cti.2016.22
16. Garlapati S, Facci M, Polewicz M, Strom S, Babiuk LA, Mutwiri G, et al. Strategies to link innate and adaptive immunity when designing vaccine adjuvants. *Vet Immunol Immunopathol.* (2009) 128:184–91. doi: 10.1016/j.vetimm.2008.10.298

17. Mogensen TH. Pathogen recognition and inflammatory signaling in innate immune defenses. *Clin Microbiol Rev.* (2009) 22:240–73. doi: 10.1128/CMR.00046-08
18. Tang D, Kang R, Coyne CB, Zeh HJ, Lotze MT. PAMPs and DAMPs: signal 0s that spur autophagy and immunity. *Immunol Rev.* (2012) 249:158–75. doi: 10.1111/j.1600-065X.2012.01146.x
19. D'Souza RS, Bhat KG, Sailaja D, Babji D V, Bandiwadekar TK, Katgalkar RM. Analysis of expression and localization of TLR-2 by immunofluorescent technique in healthy and inflammed oral tissues. *J Clin Diagn Res.* (2013) 7:2683–780. doi: 10.7860/JCDR/2013/6745.3745
20. Kaiser WJ, Sridharan H, Huang C, Mandal P, Upton JW, Gough PJ, et al. Toll-like receptor 3-mediated necrosis via TRIF, RIP3, and MLKL. *J Biol Chem.* (2013) 288:31268–79. doi: 10.1074/jbc.M113.462341
21. Kennedy MN, Mullen GE, Leifer CA, Lee C, Mazzoni A, Dileepan KN, et al. A complex of soluble MD-2 and lipopolysaccharide serves as an activating ligand for Toll-like receptor 4. *J Biol Chem.* (2004) 279:34698–704. doi: 10.1074/jbc.M405444200
22. Jiang S, Wang L, Huang M, Jia Z, Weinert T, Warkentin E, et al. DM9 domain containing protein functions as a pattern recognition receptor with broad microbial recognition spectrum. *Front Immunol.* (2017) 8:1607. doi: 10.3389/fimmu.2017.01607
23. Yang C, Wang L, Jia Z, Yi Q, Xu Q, Wang W, et al. Two short peptidoglycan recognition proteins from *Crassostrea gigas* with similar structure exhibited different PAMP binding activity. *Dev Comp Immunol.* (2017) 70:9–18. doi: 10.1016/j.dci.2016.12.009
24. Yang P, An H, Liu X, Wen M, Zheng Y, Rui Y, et al. The cytosolic nucleic acid sensor LRRFIP1 mediates the production of type I interferon via a beta-catenin-dependent pathway. *Nat Immunol.* (2010) 11:487–94. doi: 10.1038/ni.1876
25. Miao EA, Mao DP, Yudkovsky N, Bonneau R, Lorang CG, Warren SE, et al. Innate immune detection of the type III secretion apparatus through the NLRC4 inflammasome. *Proc Natl Acad Sci USA.* (2010) 107:3076–80. doi: 10.1073/pnas.0913087107
26. Pohlmann S, Zhang J, Baribaud F, Chen Z, Leslie GJ, Lin G, et al. Hepatitis C virus glycoproteins interact with DC-SIGN and DC-SIGNR. *J Virol.* (2003) 77:4070–80. doi: 10.1128/JVI.77.7.4070-4080.2003
27. Krol E, Mentzel T, Chinchilla D, Boller T, Felix G, Kemmerling B, et al. Perception of the Arabidopsis danger signal peptide 1 involves the pattern recognition receptor AtPEPR1 and its close homologue AtPEPR2. *J Biol Chem.* (2010) 285:13471–9. doi: 10.1074/jbc.M109.097394
28. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic Acids Res.* (2013) 41:D1228–33. doi: 10.1093/nar/gks1147
29. Dhanda SK, Mahajan S, Paul S, Yan Z, Kim H, Jespersen MC, et al. IEDB-AR: immune epitope database-analysis resource in 2019. *Nucleic Acids Res.* (2019) 47:W502–6. doi: 10.1093/nar/gkz452
30. Korb M, Rust AG, Thorsson V, Battail C, Li B, Hwang D, et al. The Innate Immune Database (IIDB). *BMC Immunol.* (2008) 9:7. doi: 10.1186/1471-2172-9-7
31. Sayers S, Ulysse G, Xiang Z, He Y. Vaxjo: a web-based vaccine adjuvant database and its application for analysis of vaccine adjuvants and their uses in vaccine development. *J Biomed Biotechnol.* (2012) 2012:831486. doi: 10.1155/2012/831486
32. Xiang Z, Todd T, Ku KP, Kovacic BL, Larson CB, Chen F, et al. VIOLIN: vaccine investigation and online information network. *Nucleic Acids Res.* (2008) 36:D923–8. doi: 10.1093/nar/gkm1039
33. Gao Q-B, Zhao H, Ye X, He J. Prediction of pattern recognition receptor family using pseudo-amino acid composition. *Biochem Biophys Res Commun.* (2012) 417:73–7. doi: 10.1016/j.bbrc.2011.11.057
34. Lata S, Raghava GPS. PRRDB: a comprehensive database of pattern-recognition receptors and their ligands. *BMC Genomics.* (2008) 9:180. doi: 10.1186/1471-2164-9-180
35. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel.* (2004) 17:349–56. doi: 10.1093/protein/gzh037
36. Garg A, Raghava GPS. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biol.* (2008) 8:129–40.
37. Kumar M, Gromiha MM, Raghava GPS. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics.* (2007) 8:463. doi: 10.1186/1471-2105-8-463
38. Usmani SS, Bhalla S, Raghava GPS. Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features. *Front Pharmacol.* (2018) 9:954. doi: 10.3389/fphar.2018.00954
39. Nagpal G, Usmani SS, Dhanda SK, Kaur H, Singh S, Sharma M, et al. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci Rep.* (2017) 7:42851. doi: 10.1038/srep42851
40. Agrawal P, Bhalla S, Chaudhary K, Kumar R, Sharma M, Raghava GPS. *In silico* approach for prediction of antifungal peptides. *Front Microbiol.* (2018) 9:323. doi: 10.3389/fmicb.2018.00323
41. Kaur H, Raghava GPS. Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. *Proteins.* (2004) 55:83–90. doi: 10.1002/prot.10569
42. Garg A, Raghava GPS. ESLpred2: improved method for predicting subcellular localization of eukaryotic proteins. *BMC Bioinformatics.* (2008) 9:503. doi: 10.1186/1471-2105-9-503
43. Bhasin M, Raghava GPS. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* (2004) 32:W414–9. doi: 10.1093/nar/gkh350
44. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* (2017) 45:D158–69. doi: 10.1093/nar/gkw1099
45. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* (2012) 28:3150–2. doi: 10.1093/bioinformatics/bts565
46. Chauhan JS, Mishra NK, Raghava GPS. Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. *BMC Bioinformatics.* (2010) 11:301. doi: 10.1186/1471-2105-11-301
47. Singh H, Kumar R, Singh S, Chaudhary K, Gautam A, Raghava GPS. Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. *BMC Cancer.* (2016) 16:77. doi: 10.1186/s12885-016-2082-y
48. Singh H, Singh S, Singla D, Agarwal SM, Raghava GPS. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol Direct.* (2015) 10:10. doi: 10.1186/s13062-015-0046-9
49. Chaudhary K, Kumar R, Singh S, Tuknait A, Gautam A, Mathur D, et al. A web server and mobile app for computing hemolytic potency of peptides. *Sci Rep.* (2016) 6:22843. doi: 10.1038/srep22843
50. Agrawal P, Kumar S, Singh A, Raghava GPS, Singh IK. NeuroPIpred: a tool to predict, design and scan insect neuropeptides. *Sci Rep.* (2019) 9:5129. doi: 10.1038/s41598-019-41538-x
51. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* (2009) 10:421. doi: 10.1186/1471-2105-10-421
52. Pande A, Patiyal S, Lathwal A, Arora C, Kaur D, Dhall A, et al. Computing wide range of protein/peptide features from their sequence and structure. *bioRxiv [Preprint].* (2019) doi: 10.1101/599126
53. Kaundal R, Raghava GPS. RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. *Proteomics.* (2009) 9:2324–42. doi: 10.1002/pmic.200700597
54. Zhang X, Liu S, Tramontano A. RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics.* (2017) 33:854–62. doi: 10.1093/bioinformatics/btw730
55. Verma R, Varshney GC, Raghava GPS. Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. *Amino Acids.* (2010) 39:101–10. doi: 10.1007/s00726-009-0381-1
56. Verma R, Tiwari A, Kaur S, Varshney GC, Raghava GPS. Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. *BMC Bioinformatics.* (2008) 9:201. doi: 10.1186/1471-2105-9-201
57. Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J Mol Recogn.* (2011) 24:303–13. doi: 10.1002/jmr.1061

58. Laurie SA, Goss GD. Role of epidermal growth factor receptor inhibitors in epidermal growth factor receptor wild-type non-small-cell lung cancer. *J Clin Oncol.* (2013) 31:1061–9. doi: 10.1200/JCO.2012.43.4522
59. Kumar V, Agrawal P, Kumar R, Bhalla S, Usmani SS, Varshney GC, et al. Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues. *Front Microbiol.* (2018) 9:725. doi: 10.3389/fmicb.2018.00725
60. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* (2006) 34:W202–9. doi: 10.1038/nprot.2007.505
61. Zhu G, Xu Y, Cen X, Nandakumar KS, Liu S, Cheng K. Targeting pattern-recognition receptors to discover new small molecule immune modulators. *Eur J Med Chem.* (2018) 144:82–92. doi: 10.1016/j.ejmech.2017.12.026
62. Pahari S, Kaur G, Aqdas M, Negi S, Chatterjee D, Bashir H, et al. Bolstering immunity through pattern recognition receptors: a unique approach to control tuberculosis. *Front Immunol.* (2017) 8:906. doi: 10.3389/fimmu.2017.00906
63. Vasou A, Sultanoglu N, Goodbourn S, Randall RE, Kostrikis LG. Targeting pattern recognition receptors (PRR) for vaccine adjuvantation: from synthetic PRR agonists to the potential of defective interfering particles of viruses. *Viruses.* (2017) 9:186. doi: 10.3390/v9070186
64. Mullen LM, Chamberlain G, Sacre S. Pattern recognition receptors as potential therapeutic targets in inflammatory rheumatic disease. *Arthritis Res Ther.* (2015) 17:122. doi: 10.1186/s13075-015-0645-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.