Briefings in Bioinformatics, 00(00), 2020, 1–12

doi: 10.1093/bib/bbaa294
Problem solving protocol

AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes

Neelam Sharma, Sumeet Patiyal, Anjali Dhall, Akshara Pande, Chakit Arora and Gajendra P. S. Raghava

Corresponding author: Gajendra P. S. Raghava, Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi 110020, India. Tel.: +91-11-26907444; E-mail: raghava@iiitd.ac.in

Abstract

AlgPred 2.0 is a web server developed for predicting allergenic proteins and allergenic regions in a protein. It is an updated version of AlgPred developed in 2006. The dataset used for training, testing and validation consists of 10 075 allergens and 10 075 non-allergens. In addition, 10 451 experimentally validated immunoglobulin E (IgE) epitopes were used to identify antigenic regions in a protein. All models were trained on 80% of data called training dataset, and the performance of models was evaluated using 5-fold cross-validation technique. The performance of the final model trained on the training dataset was evaluated on 20% of data called validation dataset; no two proteins in any two sets have more than 40% similarity. First, a Basic Local Alignment Search Tool (BLAST) search has been performed against the dataset, and allergens were predicted based on the level of similarity with known allergens. Second, IgE epitopes obtained from the IEDB database were searched in the dataset to predict allergens based on their presence in a protein. Third, motif-based approaches like multiple EM for motif elicitation/motif alignment and search tool have been used to predict allergens. Fourth, allergen prediction models have been developed using a wide range of machine learning techniques. Finally, the ensemble approach has been used for predicting allergenic protein by combining prediction scores of different approaches. Our best model achieved maximum performance in terms of area under receiver operating characteristic curve 0.98 with Matthew's correlation coefficient 0.85 on the validation dataset. A web server AlgPred 2.0 has been developed that allows the prediction of allergens, mapping of IgE epitope, motif search and BLAST search (https://webs.iiitd.edu.in/raghava/algpred2/).

Key words: allergens; IgE epitope; prediction; machine learning; MEME/MAST; MERCI; BLAST

Introduction

Allergy is the abnormal behavior of the immune system against foreign substances called allergens. It involves a series of many reactions, which triggers various symptoms like allergic asthma, rhinitis, skin reactions and difficulty in breathing that can lead to death. The rise in the occurrence of allergic diseases in the

Neelam Sharma is currently working as a PhD student in bioinformatics at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Sumeet Patiyal is currently working as a PhD student in bioinformatics at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Anjali Dhall is currently working as a PhD student in Bioinformatics at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Akshara Pande has worked as a research associate in the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India, and is currently working as an associate professor at the Graphic Era Hill University, Dehradun.
Chakit Arora is currently working as a PhD student in bioinformatics at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Gajendra P. S. Raghava is currently working as a professor and head of the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

Submitted: 21 June 2020; **Received (in revised form):** 2 October 2020

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

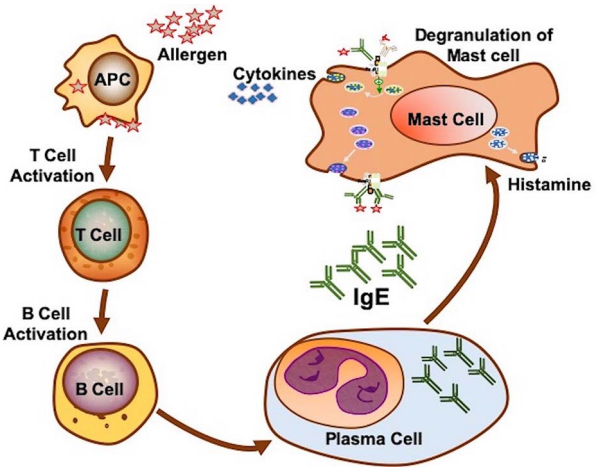

Figure 1. The mechanism of allergy and step-by-step activation of different components of the immune system, from the recognition of allergen by antigen-presenting cells (APCs) to the release of histamine.

last few years has not only enhanced the costs of treatment but also adversely affected the quality of life of a large population [1]. Allergens like dust mites, pollens and many others induce the Type I hypersensitive reactions, which elicit immunoglobulin E (IgE) antibodies. This allergic reaction results in the release of inflammatory mediators, such as histamine, cytokines from mast cells and basophils [2], which affects the population on a large scale, particularly skin sensitization [3, 4].

The first encounter of allergen develops the hypersensitivity, whereas the second encounter of the same allergen leads to the effector response. Type I hypersensitivity is mediated by immunoglobulin E (IgE), which is produced to act against allergens. Allergens induce the Type I hypersensitivity reaction, which sets off the production of allergen-specific IgE epitopes. These epitopes bind to the mast cell and basophils; this is known as the sensitization of mast cells and basophils. Reexposure of the allergens to the sensitized mast cell and basophils (which are already coated with IgE antibodies) leads to the degranulation and release of mediators and inflammatory molecules like histamine, leukotriene, etc., which leads to a mild allergic reaction to sudden death from anaphylactic shock [5].

Overall processing of allergen, activation of IgE antibodies and release of histamine are shown in Figure 1. There is a wide range of molecules that can act as allergens; it includes small chemicals or biological molecules like proteins. In the present study, we aim to develop a method for predicting allergenic proteins. It has a wide range of applications, particularly in biotechnology-derived products, such as genetically modified foods, therapeutics and vaccines [6–8].

The guideline issued by the Food and Agriculture Organization (FAO) fails to identify allergens with high precision due to a large number of false positive predictions [9, 10]. Earlier methods developed before 2005 can be classified in the following categories: (i) similarity search, (ii) supervised learning-based models and (iii) motif-based approaches. In 2006, a hybrid method AlgPred [11] was developed that combines the following approaches for predicting allergenic proteins: (i) support vector machine (SVM)-based model, (ii) mapping of IgE epitopes, (iii) multiple EM for motif elicitation (MEME)/motif alignment and search tool (MAST) motifs [12, 13] and (iv) Basic Local Alignment Search Tool (BLAST)-based similarity search [14]. This method combines the power of different approaches, and it outperformed all methods developed before 2006. Following is a brief description of methods developed in the past 14 years. AllerTool is an SVM-based method developed in 2007; it combines a similarity-based approach for predicting allergenicity and allergic cross-reactivity in proteins [15]. AllerHunter was developed in 2009 on 1356 allergenic proteins, where models were developed using SVM-pairwise sequence similarity [16]. In 2013, AllerTOP was developed on 2210 allergens, and its updated version AllerTOPv2 has been developed on 2427 allergens [17, 18]. In the case of PREAL, SVM models were developed on the 1176 allergenic proteins using biochemical and physicochemical properties [19]. In 2014, AllergenFP was developed on a dataset of 2427 allergens that incorporate descriptor-based fingerprints for developing prediction models [20]. Recently, AllerCatPro has been developed on 4180 allergens for predicting the allergenicity potential of a protein from its sequence and 3D epitope mapping [21].

These allergen prediction methods are heavily used by the scientific community, particularly by experimental researchers in designing proteins with desired allergenicity. These methods have their limitations that include the following: (i) most of the methods have been developed on small datasets, (ii) redundant proteins in the dataset, (iii) no mapping of IgE epitopes and (iv) motif information not incorporated. In order to complement existing methods, we made a systematic attempt to improve our method AlgPred [11]. In this article, we have proposed a method AlgPred 2.0 to classify the allergenic and non-allergenic protein sequences, which is an improved version of AlgPred developed in 2006. Numerous features have been incorporated in AlgPred 2.0 to improve the performance of the method. Models developed in AlgPred 2.0 have been trained and evaluated on the largest possible dataset. In addition, new version uses 10 451 experimentally validated IgE epitopes for mapping.

Materials and Methods

Compilation of dataset

The dataset used in this study was compiled from various databases and repositories. We extracted 2018 allergens from COMPARE (https://comparedatabase.org) and 2078 allergens from Allergen Online [22] repositories. We also obtained datasets used in previous studies and have used it for developing prediction models: (i) 570 allergens and 700 non-allergens from AlgPred [11] and (ii) 2427 allergens and 2427 non-allergens from AllerTOP [17]. Besides, we extracted 1078 allergens from Swiss-Prot release 2019_04 (released on 8 May 2019) using the keyword 'allergen AND reviewed: yes' [23]. All proteins containing non-standard characters (i.e. 'BJOUXZ') or less than 50 amino acids or non-allergen sequences which have similarity with allergen sequences were removed. Finally, we got 10 075 allergen sequences, which we called a positive dataset. It is challenging to obtain a negative dataset, as non-allergens are not readily available like allergens, where experimentally validated data are easily available. Thus, in this study, we carefully extracted and assigned non-allergens from Swiss-Prot release 2019_04. We extracted 545 820 proteins using the query 'NOT allergen NOT cancer NOT allergenic AND reviewed: yes'; these proteins were assigned as non-allergens. In this study, we have only taken the proteins which are reviewed and manually annotated. Thus, chances that some non-allergens used in this study might be allergens are rare as we have only taken proteins that have been annotated manually (only possible if the annotation is incorrect). We got 533 719 non-allergenic

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

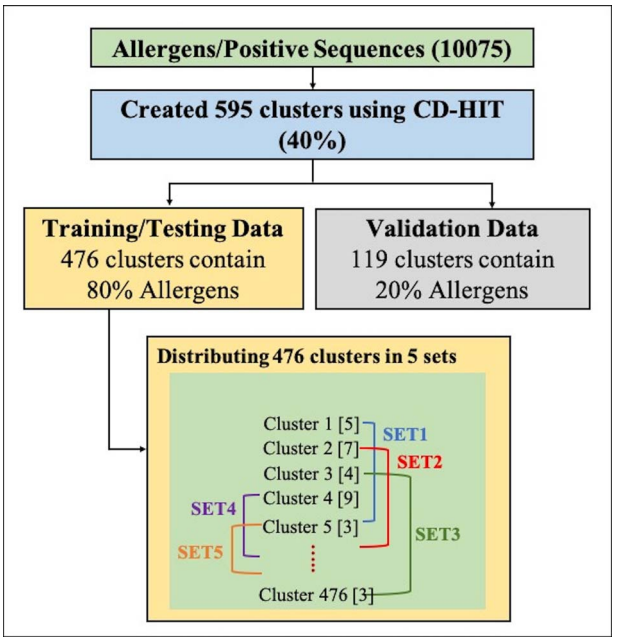

Figure 2. Flowchart shows the process of creating non-redundant training, testing and validation datasets of allergens. First, clusters of all allergens are partitioned in training and validation datasets, and then clusters in the training dataset are partitioned into five sets.

sequences after removing sequences having less than 50 amino acids and containing the non-standard characters. We randomly pick up 10 075 non-allergenic sequences from 533 719 non-allergenic sequences. Finally, we got a dataset that contains 10 075 allergenic and 10 075 non-allergenic sequences.

Creation of non-redundant dataset

One should remove the redundancy among proteins in a dataset to develop a robust method. In the past, researchers have created non-redundant datasets at different levels of similarity from 30% to 100% [11, 24–27]. One of the major reasons to create a non-redundant dataset is to remove similar sequences among proteins in training and testing datasets. Unfortunately, the removal of redundant sequences also reduces the size of the dataset. In a previous study, we introduced the concept of data partitioning to create non-redundant training and testing datasets without reducing the size of the dataset [11]. In this study, we also used the same approach to partition data in training and testing datasets, where no protein in the training set had more than 40% similarity with any protein in the test dataset. First, clusters were created using CD-HIT [28] software at a 40% sequence similarity for the positive dataset (allergens) as well as for the negative dataset (non-allergens). Second, clusters obtained for both allergens and non-allergens dataset were divided into 80% training data and 20% validation data. The clusters in training data (both for positive and negative data) were further fractionated into five sets such that all proteins of a given cluster are kept in one set and sequences in one set do not have any similarity with sequences of other sets. It results in five positive sets and five negative sets. The 20% validation set also consists of both positive and negative clusters that are not present in the training data. The graphical representation of steps used for creating a non-redundant dataset is shown in Figure 2.

Independent dataset

In this study, we already used 20% data for validation of our models using external validation. In addition, we also created an independent dataset of allergens for evaluating the performance of the main model of AlgPred 2.0. These allergens have been recently added to the COMPARE 2020 database and Swiss-Prot after we extracted data for creating a dataset for AlgPred 2.0. We extracted 191 unique allergen sequences from the COMPARE 2020 database; none of these sequences were identical to the AlgPred 2.0 dataset. The protein sequences were searched for unnatural amino acids, 'BJOUZX' and were removed. At last, final 180 allergens were taken from the COMPARE 2020 database. We also extracted proteins from Swiss-Prot release 2020_04 (Released on 12 August 2020) using the keywords 'allergen AND reviewed: yes'. We removed all those proteins which were present in the Swiss-Prot release 2019_04 (release used for creating the main dataset of AlgPred 2.0). We got 117 allergen sequences from Swiss-Prot, which were unique and did not contain any unnatural amino acid. Finally, we created an independent dataset that contains 297 antigen sequences (180 from COMPARE and 117 from Swiss-Prot). In addition, we also created a non-redundant independent dataset of 56 sequences after removing all those sequences having 40% or more similarity with sequences in the main dataset of AlgPred 2.0. This non-redundant independent dataset was created using CD-HIT software at cutoff of 40%. It means no sequence in a non-redundant independent dataset has a similarity of more than 40% with sequences in the main dataset.

Mapping of epitopes

IgE epitopes are responsible for inducing antibodies which in turn induces an allergic reaction in the animal body. It means that a protein would be called as an allergen if it contains an IgE epitope or has high similarity with the sequence of IgE epitopes. Thus, mapping of an IgE epitope can be used to identify any protein as an allergen. Here, IgE epitopes were obtained from following databases/datasets: (i) 15 046 from IEDB [29], (ii) 863 from AllerBase [30] and (iii) 2341 from IgPred [31]. Similarly, we obtained 381 196 and 35 219 non-IgE epitopes from IEDB and IgPred, respectively. Finally, we got 10 451 IgE epitopes and 307 866 non-IgE epitopes after removing redundant epitopes and epitopes having less than 5 and more than 50 amino acids. In this study, we used two techniques for mapping IgE epitopes: (i) BLAST-based searching [14] and (ii) motif-emerging and with classes-identification (MERCI) [32]. In the case of the BLAST-based technique, we searched the proteins against a database of IgE epitopes. A query protein is assigned an allergen if there is a hit against the IgE epitope at a given expectation value (*E*-value) cutoff. We used MERCI software to identify motifs present in experimentally validated IgE epitopes [32]. A protein is assigned as an allergen if it contains an IgE-specific motif identified by MERCI software.

Five-fold cross-validation

In order to train and evaluate our models on the training dataset, we used a standard 5-fold cross-validation technique. A training set was formed by combining four negative and four positive sets, while the corresponding test set was created by combining the remaining one positive and one negative set. To make sure that the combination of a positive set and a negative set is used as a test set only once, this method is repeated five times. These five training and testing sets were used for developing

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

learning-based prediction models. Models were then evaluated by performing prediction on the unseen validation set. It is a standard process that has been successfully implemented in several machine learning (ML)-based studies in the past [11, 33–39].

BLAST for similarity search

BLAST is heavily used in literature to annotate protein sequences [24, 40–42]. We have used it to identify allergens based on the similarity of a protein with allergenic and non-allergenic sequences. The similarity-based search module was developed using the blastp (protein–protein BLAST) suite of BLAST+ version 2.7.1 [14], where the query sequences were hit against the database of allergens and non-allergens. In this study, two strategies were used to identify allergens: (i) top hit of BLAST and (ii) ensemble of top five hits of BLAST. In the case of the top hit of BLAST, a protein is searched against a database of allergens or non-allergens using BLAST at different *E*-value cutoffs. If the top hit of the BLAST is against an allergenic sequence of the database, then the protein is assigned as an allergen and a similar approach is used for assigning protein as non-allergen. For ensemble of top five hits of BLAST, we considered a voting approach to annotate a query protein. If there are at least five or more than five hits corresponding to a query protein sequence, then we consider it as a hit. We assigned the function of query sequence based on maximum hits; if the top five hits have maximum allergens, then the query sequence was assigned as an allergen. Similarly, if the top five hits have maximum non-allergens, then the query was assigned as non-allergen. Various *E*-value cutoffs were used to evaluate the performance of the method.

Motif scanning/motif-based prediction

The motif is a recurring pattern of amino acids or nucleotides that occur in protein or DNA, respectively. MEME/MAST consists of two modules of MEME suite version 5.0.5 [12]. It was downloaded from http://meme-suite.org/index.html. This suite allows the discovering of novel motifs and performing a wide variety of motif-based analyses. The MEME module was used to discover motifs in closely related sequences. It represents motifs as position-dependent letter-probability matrices, which depicts the probability of occurring of each possible letter at each position in the pattern. It takes a file containing primary sequences in FASTA format (training set) as input and gives the output file containing motifs as many as requested using statistical modeling techniques. The MAST [13] module was used to search for matches to a set of motifs (from the output of MEME). It takes an output file of MEME as input and hitting the query file (test set) on the MEME matrix to search for the match for the motifs. In this study, five MEME matrices have been created corresponding to the five training sets, one matrix for each set. Then each matrix was used as an input file for searching motifs in the test sets using MAST. *E*-value cutoffs were also taken into account during MAST analysis.

Protein features

Composition-based features

Residue information of the protein was used in the form of amino acid composition (AAC) and dipeptide composition (DPC). The information in this form was used as a feature for developing ML models. The web server 'Pfeature' was used for this purpose. For a given protein sequence, AAC provides a 20-length vector, where each element is a fraction of a specific type of amino acid residue in the sequence. Whereas, DPC provides the information about the pairwise composition of the amino acids (e.g. A-A, A-C, A-D . . . .Y-W, Y-Y, etc.) present in the protein sequence in the form of a 400-length feature vector. Pfeature can be referred for detailed information on these features [43]. The formula for calculating AAC and DPC for each residue is provided in Equations (1) and (2), respectively.

$$\mathrm{AAC(i)} = \frac{R_i}{N} * 100, \tag{1}$$

where AAC (i) is the amino acid percent composition (i), R_i is the number of residues of type i and *N* represents the length of the peptide sequence.

$$\mathrm{DPC(i)} = \frac{\text{Total number of dipeptide (i)}}{N-1} * 100, \tag{2}$$

where DPC(i) is the percent composition of the dipeptide of type i, and *N* represents the length of a protein.

Evolutionary information-based features

Evolutionary information of the protein in the form of a position-specific scoring matrix (PSSM-400) composition was computed. PSSM-400 uses PSI-BLAST for extracting evolutionary information for a given protein and has been implemented in numerous studies in the past [42, 44–48] PSSM-400 is a 20 × 20 matrix, which encapsulates the composition of occurrences of each type of 20 amino acids with respect to each amino acid present in the given protein sequence. For this study, a PSSM matrix was created for each protein, which is subsequently normalized and converted into a 20 × 20 PSSM-400 vector by utilizing Pfeature web server [43].

ML-based classifiers

Different classification models such as random forest (RF), SVM, decision tree (DT), *K*-nearest neighbor (KNN) and multilayer perceptron (MLP) were implemented using Scikit's sklearn package [49] from Python. GridSearchCV was used for the optimization of hyper-parameters. Protein features such as AAC and PSSM-400 were used as fixed-length vectors for training and testing in classification models. 5-fold cross-validation was used for the evaluation of the models using different performance measures. This method has been implemented in various studies in the past [33, 36, 37, 50, 51]. The complete architecture of AlgPred 2.0 is shown in Figure 3.

Performance evaluation parameters

Threshold-dependent parameters such as sensitivity (Sens), specificity (Spec), accuracy (Acc) and Matthew's correlation coefficient (MCC) were used to measure performance at a different threshold. Besides, threshold-independent parameter, viz., area under receiver operating characteristic curve (AUC) was used to evaluate the performance of our models. 'Sens' is the ratio of true positives and total positives, also known as the true positive rate (TPR), whereas 'Spec' is the true negative rate (TNR). 'Acc' is the number of the total correct predictions (both positive and negative) with respect to total positives and negatives, and MCC is the correlation coefficient between predicted and actual

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

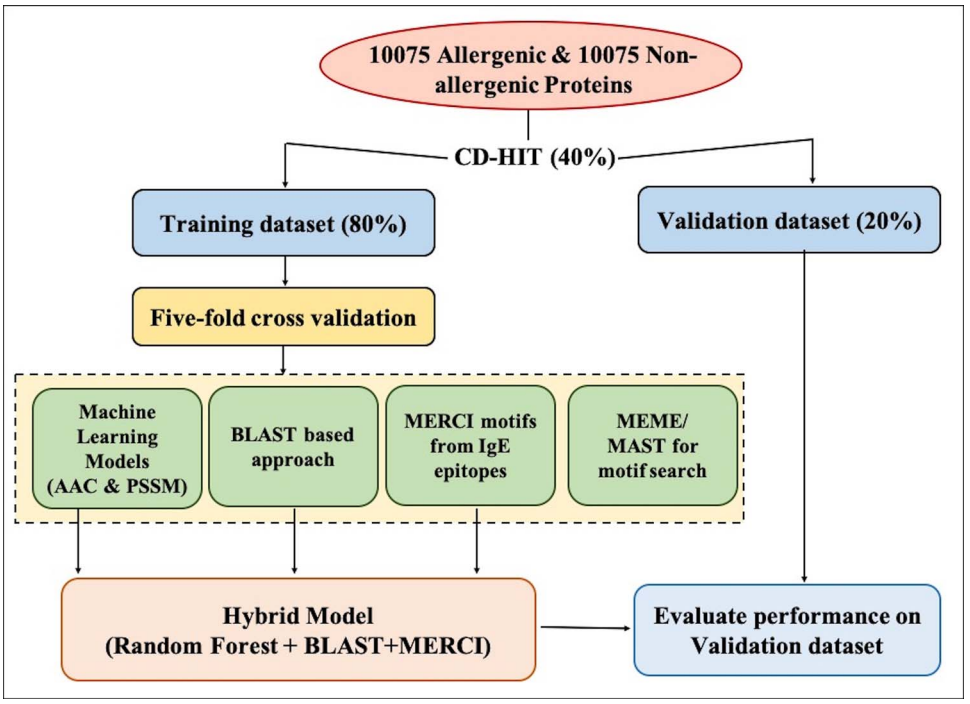

Figure 3. Flowchart shows the overall architecture of AlgPred 2.0, where models are trained and tested using 5-fold cross-validation, further evaluated on validation data. Our hybrid model, which combines RF, BLAST and MERCI, overperforms individual models.

classes. The following standard formulae were used to calculate these parameters:

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (3)$$

$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \quad (4)$$

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100 \quad (5)$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})\,(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})\,(\text{TN} + \text{FN})}} \quad (6)$$

where TP is the number of true positive predictions, TN is the number of true negative predictions, FP denotes the number of sequences falsely predicted as positive sequences and FN represents the number of sequences falsely predicted as negative sequences. These evaluation parameters have been used in many studies for the model's performance evaluation and are well established in the literature [11, 38, 52, 53]. The threshold-independent parameter, area under receiver operating characteristic curve (AUROC), is calculated via the plot between TPR and false positive rate (FPR) [54].

Hybrid approach for classification

In order to improve the accuracy of classifying allergenic and non-allergenic proteins, we implement a hybrid approach as used in AlgPred and other state-of-the-art methods. The studies conducted by Wan *et al.* have proposed the ensemble classifier method using hybrid features like split amino acid composition (SAAC) features and profile alignment (PA) features for predicting subchloroplast localization of multilocation proteins [55]. They have also used ensemble features like pseudo amino acid composition (PseAA) features and PA features for developing method EnTrans-Chlo [56]. In the past, ensemble methods have also been used for predicting protein primary and secondary structures. A study by Han *et al.* [57] has incorporated various features based on amino acid classification as well as physicochemical properties, and they have developed the ensemble method for predicting subnuclear localizations from primary protein structures. Another study by Bouziane *et al.* [58] has developed a voting-based ensemble method for protein secondary structure prediction. In this study, authors have combined different ML algorithms like KNNs, ANNs and multi-class SVMs (M-SVMs) based on PSSM profiles of the proteins. They have used two variants voting concept, that is, simple majority voting (SMV) and weighted majority voting (WMV), to improve the secondary structure prediction.

In AlgPred 2.0, we have also implemented a hybrid or ensemble approach. Here, the following three techniques have been combined: (i) similarity-based approach using BLAST, (ii) motif-based approach using MERCI and (iii) ML-based technique. First, the given protein sequence was classified using BLAST at the *E*-value of 10^{-6}. We assigned the score of '+0.5' for the correct positive predictions (allergenic proteins), '−0.5' for correct negative predictions (non-allergenic proteins) and '0' for no hits. Second, the same protein sequence was classified using MERCI. We assigned the score of '+0.5' if the motifs were found and '0' if the motifs were not found. In the case of a hybrid approach, scores obtained from three methods (i.e. BLAST, MERCI and ML scores) were combined to compute the overall score. This overall score of the hybrid approach was used for assigning the protein as allergenic and non-allergenic protein at different thresholds. These types of hybrid methods using different approaches have been extensively used in many studies in the past [11, 24, 55–58].

Results

Performance on BLAST models

BLAST is a commonly used software in literature for annotating or assigning function to a protein based on similarity search.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

Table 1. The performance of BLAST-based search on training and validation dataset

E-value	Training				Validation			
	Allergens		Non-allergens		Allergens		Non-allergens	
	Chits (Sens)	Whits (Error)	Chits (Spec)	Whits (Error)	Chits (Sens)	Whits (Error)	Chits (Spec)	Whits (Error)
10^{-6}	4618 (57.3%)	164 (2.03%)	912 (11.32%)	107 (1.33%)	878 (43.57%)	48 (2.38%)	264 (13.1%)	24 (1.19%)
10^{-5}	4665 (57.88%)	174 (2.16%)	1001 (12.42%)	111 (1.38%)	883 (43.82%)	48 (2.38%)	284 (14.09%)	24 (1.19%)
10^{-4}	4772 (59.21%)	189 (2.34%)	1093 (13.56%)	120 (1.49%)	887 (44.02%)	50 (2.48%)	311 (15.43%)	24 (1.19%)
10^{-3}	4940 (61.29%)	216 (2.68%)	1201 (14.9%)	127 (1.58%)	899 (44.62%)	52 (2.58%)	342 (16.97%)	24 (1.19%)
10^{-2}	5056 (62.73%)	264 (3.28%)	1349 (16.74%)	135 (1.67%)	913 (45.31%)	55 (2.73%)	383 (19.01%)	28 (1.39%)
10^{-1}	5133 (63.68%)	291 (3.61%)	1552 (19.26%)	152 (1.89%)	947 (47%)	70 (3.47%)	449 (22.28%)	35 (1.74%)

Chits: correct hits; Whits: wrong hits

Thus, we also used BLAST for assigning a protein as allergen or non-allergen. We evaluated the performance of BLAST using a 5-fold cross-validation technique to avoid any biasness. First, a BLAST database is created using sequences in four sets and sequences in the fifth set were searched against the database using BLAST. This process has been repeated five times so that each sequence should be searched at least once. For the evaluation of BLAST on the validation dataset, we created the BLAST database using all sequences in the training dataset, and each sequence in the validation dataset was searched against the database. We used standard BLAST, where the class assignment is based on the top hit; it was observed that BLAST produces a lot of false positives (data not shown). Then we used an ensemble of top five BLAST hits, which reduces the false prediction significantly (Table 1). Though the number of correct hits (sensitivity) increased from 57.3% to 63.68% for the training dataset with the *E*-value ranging from 10^{-6} to 10^{-1}, false predictions or wrong hits also increased proportionally. A similar trend was observed for the validation dataset where sensitivity increases from 43.57% to 47.00%, with the *E*-value ranging from 10^{-6} to 10^{-1}. The overall performance of BLAST is too poor due to the large number of no hits; it means BLAST alone cannot be used for predicting allergenic proteins as represented in Table 1.

Mapping of IgE epitopes

It is known that a protein containing an IgE epitope is an allergen as IgE epitopes are responsible for allergenicity. It has been observed in the case of AlgPred that only a few allergen sequences can be mapped on IgE epitopes. Thus, in this study, we used a similarity-based approach for searching IgE epitopes in a protein sequence. As described in Materials and methods, we used BLAST to search a protein against a database of IgE epitopes. As shown in Figure 4A, the sensitivity increased from 55.3% to 72.99% with *E*-value ranging from 10^{-6} to 10^{-1}, which implies that the allergens have similarities with that to IgE epitopes. It is interesting to note that this technique has a low rate of false positive or error (i.e. 0.03–0.66%). This technique can be used to assign allergens based on the BLAST hit of a protein against IgE epitopes. We integrated this technique in our web server to facilitate users to hit their protein against the database of IgE epitopes. In addition to BLAST-based mapping, we also used MERCI software to map IgE epitopes on a protein. In this case, IgE specific motifs that are exclusively found in the IgE epitope were discovered using MERCI software. Finally, we search these IgE-specific motifs in a query protein. Though this technique was only able to identify 1% allergens, but the false prediction was nearly negligible.

Motif-based prediction

We identified motifs using MEME software from proteins in the training set. Then MAST module was used to search for matches to a set of motifs in the test set. This process is repeated five times to obtain the performance of MEME/MAST on the training dataset. As shown in Figure 4B, sensitivity increases from 21.64% to 41.89% on the training dataset and from 10.52% to 36.97% on the validation dataset, respectively, at the *E*-value ranging from 0.001 to 100. Although the sensitivity increased with an increase in the *E*-value, the percent of the wrong assignment of non-allergens to allergens also increased from 2.07% to 17.95% on the training dataset and 1.49% to 19.9% on the validation dataset, respectively. This depicts that the alone motif-based approach is not sufficient to discriminate allergens and non-allergens (Figure 4B).

ML-based models

First, we compute the AAC of allergen and non-allergen proteins. These features were used for developing models using a wide range of ML techniques (e.g. RF, SVM, KNN, MLP and DT). We optimize these ML models by tuning the different parameters. As shown in Table 2, a model based on RF performs better than other models and achieved maximum AUC 0.93 and 0.92 on the training and validation datasets, respectively. The SVM-based models also achieved reasonable high performance and achieved AUC 0.89 on training and 0.90 on validation. It has been shown in the past that evolutionary information provides more information than a query sequence. In order to capture evolutionary information, we generate PSSM profiles for a protein. These PSSM profiles were used for developing ML-based models. The performance of PSSM-based model is shown in Supplementary Table S1 available online at https://academic.oup.com/bib. In this study, our PSSM-based model does not perform better than composition-based models. We also developed ML techniques based models using DPC of proteins but its performance was not significantly better than AAC-based models. Thus, in this study, the rest of models were developed using AAC which is simple and easy to implement.

ML-based models with motif approach

Composition-based models (AAC) developed using different ML techniques were combined with the MEME/MAST approach. As shown in Supplementary Table S2 available online at https://academic.oup.com/bib, MEME/MAST with RF model performs better than other combinations. It achieves AUC 0.93 with MCC 0.72 on the training dataset and AUC 0.92 with MCC 0.68 on the

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

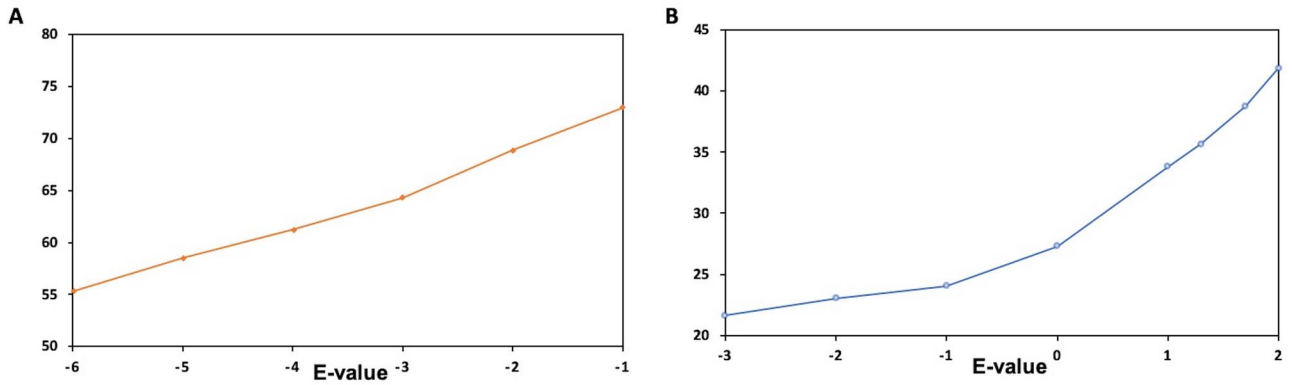

Figure 4. Line graphs show the performance of similarity and motif search methods. (**A**) BLAST; allergens were searched against IgE epitopes (*x*-axis shows BLAST *E*-value; *y*-axis shows the percent of correct hits). (**B**) MAST; motifs were searched in allergens (*x*-axis shows *E*-value of MAST; *y*-axis shows the percent of correct hits).

Table 2. The performance of ML-based models developed using amino acid composition

ML (parameters)	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF (ne =1000, mf = sqrt, md = 100, *c* = gini)	89.16	82.7	85.93	0.93	0.72	84.52	83.62	84.07	0.92	0.68
SVM (*k* = rbf, *g* = 0.05, *C* = 0.1)	85.89	79.9	82.9	0.88	0.66	87.79	77.92	82.85	0.90	0.66
KNN (*m* = minkowski, *w* = distance, nn = 84)	85.57	79.26	82.41	0.90	0.65	85.11	79.95	82.53	0.90	0.65
MLP (*s* = sgd, *a* = relu)	81.81	80.79	81.3	0.87	0.63	85.76	76.67	81.22	0.89	0.63
DT (mf = log2, md = 10 and *c* = entropy)	73.75	78.31	76.03	0.77	0.52	54.24	82.78	68.51	0.74	0.39

Note. ne, *n*_estimators; mf, max_features; md, max_depth; *c*, criterion; *k*, kernel; *g*, gamma; *C*, Regularization parameter; *m*, metric; *w*, weights; nn, *n*_neighbors; *s*, solver; *a*, activation; DT, decision tree.

validation dataset. The input feature combines the probability scores generated by ML after computing AACs and MEME/-MAST scores. Overall, there is no significant improvement in the performance of composition-based models after integration of the MEME/MAST approach. Similarly, the MERCI approach was also combined with composition-based models. As shown in Supplementary Table S3 available online at https://academic.oup.com/bib, there was no significant improvement in ML-based models combined with the MERCI motif approach. Overall, the combination of ML-based models and motif approach is not better than ML-based models.

ML-based models with BLAST search

In order to combine the power of similarity search approach BLAST and ML-based models, we developed a method using these approaches. First, the BLAST search was performed for a query sequence; if we got a BLAST hit, then we assign the query sequence based on the BLAST result. A composition-based model is used to predict allergenic and non-allergenic protein if there is no hit. The performance of our RF-based model improved significantly from AUC 0.93 to 0.99 on the training dataset and AUC 0.92 to 0.98 on the validation dataset. Figure 5 shows the receiver operating characteristic (ROC) curves for different ML classifiers corresponding to the combination of BLAST and AAC, both for training and validation datasets. We also combine the BLAST search with ML-based models developed using the PSSM profile. As shown in Supplementary Table S4 available online at https://academic.oup.com/bib, the performance of ML-based techniques improved significantly.

Hybrid approach

Finally, we combined multiple approaches developed in this study to predict allergens with high precision. In the hybrid approach, we combine two or more methods to overcome the limitation of individual methods. Here, we have combined a composition-based model with BLAST- and MERCI-based approaches. In order to integrate all three approaches, proteins were first classified using BLAST at an *E*-value of 10^{-6}, followed by MERCI. We predicted the allergenicity of protein using a ML-based model if a protein is not predicted using these two approaches (no hits). The hybrid method improved the coverage as well as accuracy, which is not practically possible for a single method or approach. As shown in Table 3, the performance of the hybrid method had improved when all the methods were combined. The best performing model was an RF-based model with AUC 0.99 with MCC 0.88 on the training dataset and AUC 0.98 and MCC of 0.85 on the validation dataset.

Comparison with existing methods

It is important to compare newly developed methods with existing methods to understand the advantage or disadvantages of the newly developed method. As AlgPred 2.0 is a new version of AlgPred, comprehensive comparison is required to understand the merits of the new version. We used evolutionary information as a new feature in AlgPred 2.0 for building prediction models using the PSSM profile. Though evolutionary information is a powerful feature, its performance is lower than the simple AAC in this study. Thus, this feature does not contribute in terms of the method's performances. In order to emphasize our contribution to AlgPred 2.0, we only showed those contributions that improve the performance or service of AlgPred 2.0 over AlgPred (Table 4). Besides the large dataset used for training and

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

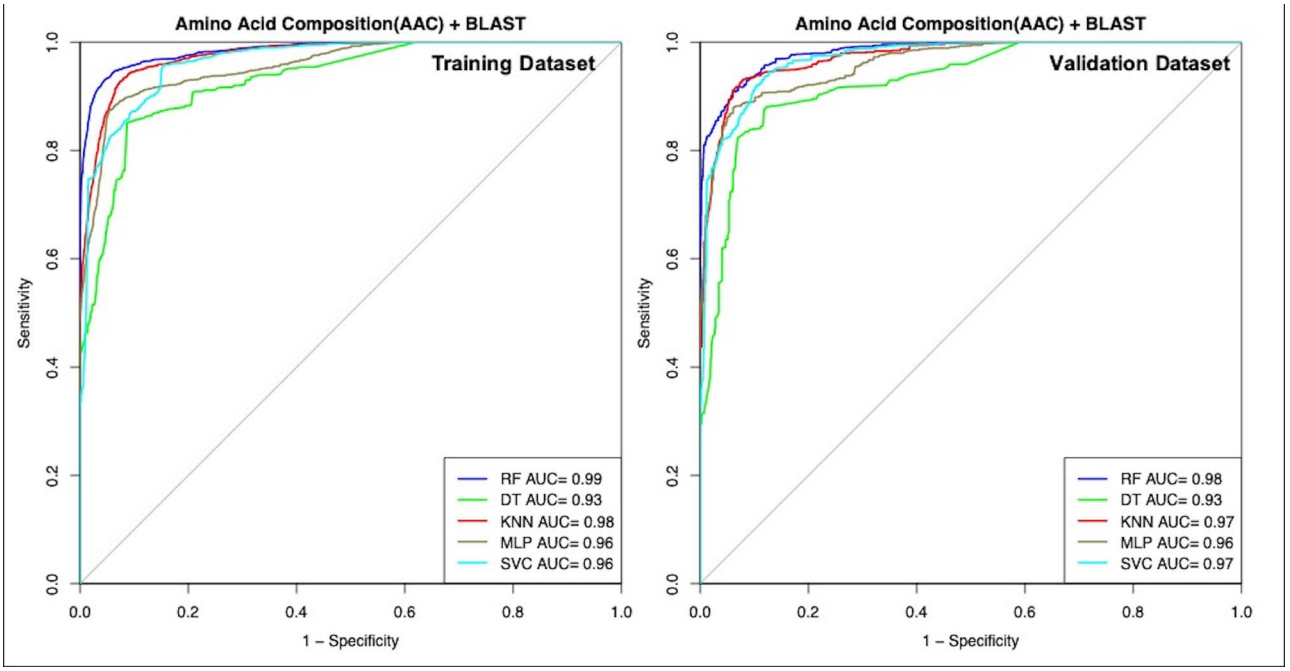

Figure 5. The performance of ML-based models after combining with BLAST is shown by ROC curves on training and validation datasets. These ML-based models were developed using the AAC of proteins.

Table 3. The performance of the hybrid method that combines ML model with BLAST-based search and MERCI-based IgE motifs

ML	Training dataset					Validation dataset				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	93.1 ± 0.02	95.36 ± 0.03	94.23 ± 0.02	0.99 ± 0.01	0.88 ± 0.04	89.43	95.09	92.26	0.98	0.85
KNN	94.47 ± 0.02	90.45 ± 0.01	92.46 ± 0.02	0.98 ± 0.01	0.85 ± 0.03	94.04	91.46	92.75	0.97	0.86
SVM	91.55 ± 0.04	85.22 ± 0.05	88.39 ± 0.06	0.97 ± 0.01	0.77 ± 0.09	96.53	83.97	90.25	0.97	0.81
MLP	90.33 ± 0.03	90.22 ± 0.03	90.28 ± 0.03	0.96 ± 0.03	0.81 ± 0.05	90.82	89.33	90.07	0.96	0.8
DT	86.43 ± 0.05	88.54 ± 0.06	87.48 ± 0.06	0.93 ± 0.02	0.75 ± 0.09	85.06	88.93	87	0.93	0.74

Note. DT, decision tree.

mapping IgE epitopes, first-time motifs were derived from IgE epitopes using MERCI software. As shown above, MERCI-based motif search reduces the false prediction drastically. In previous studies, including AlgPred, the top hit of BLAST was used for the identification of allergens; in this study, for the first time, the top five hits were used for predicting allergens. As shown in the table, this strategy improves the sensitivity and precision power of BLAST. The hybrid method developed in AlgPred 2.0 combines composition-based RF model, BLAST-based search and MERCI motifs, whereas AlgPred combines SVM-based models with IgE epitopes mapping. This is the reason the performance of AlgPred 2.0 is significantly better than that of AlgPred. Several features have been incorporated in the web server to facilitate the scientific community. It includes a provision to submit multiple sequences at a time for the prediction, whereas the old version only allows to submit one sequence at a time. AlgPred 2.0 website has been designed using responsive web design pages to make it suitable for all modern devices, including mobiles, smartphones and iPad. The new version also allows to perform mapping of IgE epitope using MERCI and BLAST, which is not available in AlgPred. In the era of genomics, users wish to predict allergenic proteins in the whole proteome of an organism, which is difficult via web services. Thus, we also developed a stand-alone version of AlgPred 2.0 in Python, which can be run on any operating system having Python. In summary, the new version has many novel features in term of algorithm or services.

In addition to AlgPred, a number of other methods have been developed for predicting allergenic proteins. Thus, it is important to benchmark the existing methods with our AlgPred 2.0. The comparison of AlgPred 2.0 with other existing methods is shown in Table 5. AllerCatPro was developed on 4180 unique allergenic proteins, but the dataset contains the proteins which have 70% or more sequence identity with each other [21]. AllerTOPv2 was developed using 2427 allergens and 2427 non-allergens; in this method, the dataset used for building the model was redundant [18]. The methods mentioned above have not been evaluated on the independent or external dataset. Recently, a state-of-the-art method, AllerHunter, has been developed, which uses both internal (5-fold) and external cross-validation. In this study, they used 1356 allergens and 13 449 non-allergens for developing prediction models [16]. AllerHunter achieved maximum MCC 0.738 on an external dataset that contains 129 allergens and 1314 non-allergens. Our hybrid approach achieved maximum MCC 0.88 on training and 0.85 on validation dataset.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

Table 4. Comprehensive comparison of two methods AlgPred and AlgPred 2.0

S.N.	AlgPred	AlgPred 2.0
Datasets, features and algorithm		
1.	Dataset: 578 allergens, 700 non-allergens and 178 IgE epitopes	Dataset: 10 075 allergens, 10 075 non-allergens and 10 451 IgE epitopes
2.	No MERCI motifs	Identification of MERCI motifs in IgE epitopes
3.	BLAST top hit was used for prediction	BLAST top five hits were used for prediction
4.	SVM model + mapping of IgE epitopes	RF model + BLAST + MERCI motifs
Web service and stand-alone version		
6.	Submission of one sequence at a time	Submission of multiple sequence at a time
7.	Not compatible with smart devices	Compatible with all modern or smart devices
8.	No MERCI motif search	Allow to search MERCI motifs
9.	No BLAST search against IgE epitopes	BLAST search against IgE epitopes
10.	No stand-alone version	Stand-alone software written in Python

Table 5. Comparison of AlgPred 2.0 with existing methods as reported in the literature

Methods	Dataset	Sens	Spec	Acc	AUROC	MCC	Web server working
AlgPred 2.0	20 150	93.1 ± 0.02	95.36 ± 0.03	94.23 ± 0.02	0.99 ± 0.01	0.88 ± 0.04	Yes
AlgPred	1278	88.87	81.86	85.02	NA	0.7053	Yes
AllerCatPro	4180	100	67.00	84.00	NA	NA	Yes
AllerTOPv2	4854	86.70	90.70	88.70	NA	0.775	Yes
AllerTOP	4420	87.60	78.00	82.80	NA	0.671	No
AllerHunter	14 805	83.70	96.40	95.30	0.928 ± 0.004	0.738	No
AllerTool	1274	86.00	86.00	NA	0.90	NA	No
AllergenFP	4854	86.80	89.10	87.90	NA	0.759	Yes

We have also computed the performance of our hybrid model on an independent dataset that contains 297 allergen proteins added recently in the COMPARE and Swiss-Prot database (see Independent dataset section of Materials and methods). No sequence in the independent dataset has similarity with sequences in the main dataset of AlgPred 2.0. Out of the 297 positive proteins, 280 were correctly predicted as allergens by our hybrid model at the default threshold/parameters. Thus, our method achieved high performance (accuracy of 94.28%) on the independent dataset. Though there are no identical sequences between the independent and main dataset, they may still exhibit high similarity. Thus, we also created a non-redundant independent dataset using CD-HIT software at cutoff of 40%. The non-redundant independent dataset contains 56 sequences, where no sequence has more than 40% similarity with sequences in the main dataset. Our hybrid model at the default threshold and parameters correctly predicted 51 out of 56 sequences as allergens, achieving 91.07% accuracy. These results indicate the reliability of our method AlgPred 2.0 on an independent dataset that is not used while training, testing or validating the models.

Web server interface and stand-alone version

A web server AlgPred 2.0 (https://webs.iiitd.edu.in/raghava/algpred2/) has been developed for predicting allergenic proteins. It integrates four major modules: (i) prediction, (ii) IgE epitope mapping, (iii) motif scan and (iv) BLAST search. The 'prediction module' allows users to submit the protein sequences in FASTA format for the prediction of the allergenic and non-allergenic proteins. In this module, the hybrid approach and RF model based on AAC have been integrated. The 'IgE epitope mapping module' facilitates the users to map the IgE epitope on a query protein sequence. The 'motif scan module' allows to scan or map motifs in the protein sequence given by the user. It uses two software, MEME/MAST and MERCI, to derive the motifs. The 'BLAST search module' facilitates the users to perform a similarity-based search using BLAST against allergen and non-allergen database and IgE epitopes database. The web server has been designed by using a responsive HTML template and browser compatibility for different OS systems. In order to facilitate the users to predict allergens at the genome scale, we developed a stand-alone version of AlgPred 2.0, which is available from the download page of our website. We measured the time taken by AlgPred 2.0 for executing 100 proteins; then, the average time for each protein is measured in seconds. We run the stand-alone version on the Linux server as well as on desktop (macOS Catalina version 10.15.7) to compute the execution time. We also submitted the proteins to our web server and calculated the total time taken by the web server from submission to display the result. Table 6 shows the total computation time taken by AlgPred 2.0 for each protein in seconds.

Discussion

In the last five decades, there has been a rise in the prevalence of allergic diseases worldwide. These diseases include allergic rhinitis [59, 60], drug allergy [61, 62], food allergy [63–65], skin allergy [66, 67] and insect allergy [68, 69], among many others [70]. Several *in silico* methods/techniques have been developed in the past that could be used to assess the allergenicity of proteins [16, 17, 20, 21] Each method has its own merits and limitations. In 2006, our group also developed a method called AlgPred, which combines a wide range of approaches that include SVM-based models, BLAST and mapping of IgE epitopes. One of the major

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

Table 6. The total computation time taken by AlgPred 2.0 for each protein in seconds (average)

Method/model	Computation time	Stand-alone version		Web server
		Linux server	Desktop	
RF model	Real time	0.015	0.213	0.736
	User time	0.012	0.010	
	System time	0.003	0.005	
Hybrid model	Real time	0.156	0.209	0.798
	User time	0.149	0.123	
	System time	0.005	0.005	

limitations of AlgPred is that it was trained on limited data (i.e. 578 allergens, 700 non-allergens and 183 IgE epitopes) due to lack of data. In the last 14 years, the numbers of allergens and IgE epitopes have been discovered. Thus, there is a need to update AlgPred using recent advances in the field of immunology. In the present study, we have developed models using 10 075 allergens and 10 075 non-allergens. In addition, 10 451 IgE epitopes were used to identify antigenic regions in proteins.

One of the commonly used techniques to assign the function of a protein is a similarity-based search, where a query sequence is searched in a database of annotated proteins. If the query sequence exhibits high similarity with a protein whose function is known, we assign the same function to the query protein. Though there are several methods available for similarity search, here we used the frequently used method BLAST. As shown in Table 1, BLAST successfully identified the allergens; the probability of correct prediction is more than 50%, or the rate of error is low. This method fails because all unknown proteins do not have similar sequences in the database of allergens and non-allergens. In simple words, BLAST fails due to a large number of no hits.

We observed the same trend in the case of motif search and mapping of IgE epitopes. Though most of these methods successfully identify allergenic protein, they fail due to poor coverage (Figure 4A and B). In order to overcome this limitation, we developed ML-based models using AAC. Overall, the composition-based model has much higher performance in terms of sensitivity as well as specificity. To combine the power of similarity search-based technique and ML-based models, we developed hybrid models. As shown in Table 3, we got the highest performance on both training and testing datasets. We hope our study will be useful for the scientific community working in the areas of protein or peptide therapeutics [71–73]. In the present scenario, there is the utmost need for the development of an efficient prediction tool that can identify the potential allergens from the modified proteins often used in the biotechnology-derived products, such as genetically modified foods, therapeutics and vaccine designing. To facilitate the scientific community and to promote extensive public usage of the proposed prediction method, we have also provided a free web server, AlgPred 2.0. We believe that our method would aid in the more accurate recognition of allergenic proteins and thereby bring a significant improvement in the field of allergy research and therapy.

Data availability

All the datasets generated for this study are available at the 'AlgPred 2.0' web server, https://webs.iiitd.edu.in/raghava/algpred2/stand.html.

Conflict of Interest

Authors declare that they have no conflict of interest.

Authors Contribution

N.S. collected, compiled and processed the datasets. A.P., S.P. and N.S. developed computer programs. N.S. implemented the algorithms and prediction models. A.D. and S.P. created the back end of the web server and front-end user interface. C.A. and N.S. analyzed the results. N.S., A.D., S.P., C.A. and G.P.S.R. wrote the manuscript. G.P.S.R. conceived and coordinated the project and provided overall supervision of the project. All authors have read and approved the final manuscript.

Key Points

- It is an updated version of AlgPred, an allergen prediction method.
- Models have been trained on the largest dataset for developing reliable models.
- Implementation of a wide range of ML techniques for developing models.
- It allows to search motifs in proteins identified using MEME/MAST and MERCI.
- Server integrates facilities like similarity search by BLAST, mapping of IgE epitope.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgement

The authors are thankful to the Department of Science and Technology (DST), Government of India, DST-INSPIRE, DBT for fellowships and financial support, and IIIT-Delhi for providing infrastructure.

References

1. Obermeyer G, Ferreira F. Can we predict or avoid the allergenic potential of genetically modified organisms? *Int Arch Allergy Immunol* 2005;**137**:151–2.
2. Masoli M, Fabian D, Holt S, *et al*. The global burden of asthma: executive summary of the GINA dissemination committee report. *Allergy* 2004;**59**:469–78.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

3. Sutton BJ, Gould HJ. The human IgE network. *Nature* 1993;**366**:421–8.
4. Broadfield E, McKeever TM, Scrivener S, *et al*. Increase in the prevalence of allergen skin sensitization in successive birth cohorts. *J Allergy Clin Immunol* 2002;**109**:969–74.
5. Mak TW, Saunders ME, Jett BD. Immune hypersensitivity. In: *Primer to the Immune Response*. Academic Cell, 2014, 487–516.
6. Goodman RE, Hefle SL, Taylor SL, *et al*. Assessing genetically modified crops to minimize the risk of increased food allergy: a review. *Int Arch Allergy Immunol* 2005;**137**:153–66.
7. Taylor SL. Protein allergenicity assessment of foods produced through agricultural biotechnology. *Annu Rev Pharmacol Toxicol* 2002;**42**:99–112.
8. Dang HX, Lawrence CB. Allerdictor: fast allergen prediction using text classification techniques. *Bioinformatics* 2014;**30**:1120–8.
9. FAO/WHO. *Evaluation of allergenicity of genetically modified foods. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology*, Rome, Italy, 2001. (http://www.fao.org/fileadmin/templates/agns/pdf/topics/ec_jan2001.pdf) (22 October 2020, date last accessed).
10. FAO/WHO. *Joint FAO/WHO Food Standards Programme Codex Alimentarius Commission. Report of the Fourth Session of the Codex Ad Hoc Intergovernmental Task Force on Foods Derived from Biotechnology*, Yokohama, Japan, 2003. (http://www.fao.org/fileadmin/user_upload/gmfp/resources/al0334ae.pdf) (22 October 2020, date last accessed).
11. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res* 2006;**34**(Web Server Issue):W202–9.
12. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994;**2**:28–36.
13. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 1998;**14**:48–54.
14. Camacho C, Coulouris G, Avagyan V, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.
15. Zhang ZH, Koh JL, Zhang GL, *et al*. AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. *Bioinformatics* 2007;**23**:504–6.
16. Muh HC, Tong JC, Tammi MT. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. *PLoS One* 2009;**4**:e5861.
17. Dimitrov I, Flower DR, Doytchinova I. AllerTOP—a server for in silico prediction of allergens. *BMC Bioinformatics* 2013;**14**(Suppl. 6):S4.
18. Dimitrov I, Bangov I, Flower DR, *et al*. AllerTOP v.2—a server for in silico prediction of allergens. *J Mol Model* 2014;**20**:2278.
19. Wang J, Zhang D, Li J. PREAL: prediction of allergenic protein by maximum relevance minimum redundancy (mRMR) feature selection. *BMC Syst Biol* 2013;**7**(Suppl. 5):S9.
20. Dimitrov I, Naneva L, Doytchinova I, *et al*. AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics* 2014;**30**:846–51.
21. Maurer-Stroh S, Krutz NL, Kern PS, *et al*. AllerCatPro—prediction of protein allergenicity potential from the protein sequence. *Bioinformatics* 2019;**35**:3020–7.
22. Goodman RE, Ebisawa M, Ferreira F, *et al*. AllergenOnline: a peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. *Mol Nutr Food Res* 2016;**60**:1183–98.
23. UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2018;**46**:2699.
24. Kaur D, Arora C, Raghava GPS. A hybrid model for predicting pattern recognition receptors using evolutionary information. *Front Immunol* 2020;**11**:71.
25. Bendtsen JD, Jensen LJ, Blom N, *et al*. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel* 2004;**17**:349–56.
26. Singh H, Singh S, Raghava GP. In silico platform for predicting and initiating β-turns in a protein at desired locations. *Proteins* 2015;**83**:910–21.
27. Garg A, Raghava GPS. ESLpred2: improved method for predicting subcellular localization of eukaryotic proteins. *BMC Bioinformatics* 2008;**9**:503.
28. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
29. Vita R, Mahajan S, Overton JA, *et al*. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**(D1):D339–43.
30. Kadam K, Karbhal R, Jayaraman VK, *et al*. AllerBase: a comprehensive allergen knowledgebase. *Database (Oxford)* 2017;**2017**:bax066.
31. Gupta S, Ansari HR, Gautam A, *et al*. Identification of B-cell epitopes in an antigen for inducing specific class of antibodies. *Biol Direct* 2013;**8**:27.
32. Vens C, Rosso MN, Danchin EG. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* 2011;**27**:1231–8.
33. 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.
34. Singh H, Kumar R, Singh S, *et al*. Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. *BMC Cancer* 2016;**16**:77.
35. Singh H, Singh S, Singla D, *et al*. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using random forest. *Biol Direct* 2015;**10**:10.
36. Chaudhary K, Kumar R, Singh S, *et al*. A web server and mobile app for computing hemolytic potency of peptides. *Sci Rep* 2016;**6**:22843.
37. Agrawal P, Kumar S, Singh A, *et al*. NeuroPIpred: a tool to predict, design and scan insect neuropeptides. *Sci Rep* 2019;**9**:5129.
38. Patiyal S, Agrawal P, Kumar V, *et al*. NAGbinder: an approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. *Protein Sci* 2020;**29**:201–10.
39. Dhall A, Patiyal S, Kaur H, *et al*. Computing skin cutaneous melanoma outcome from the HLA-alleles and clinical characteristics. *Front Genet* 2020;**11**:221.
40. Singh H, Raghava GP. BLAST-based structural annotation of protein residues using protein data Bank. *Biol Direct* 2016;**11**:4.
41. Boratyn GM, Schäffer AA, Agarwala R, *et al*. Domain enhanced lookup time accelerated BLAST. *Biol Direct* 2012;**7**:12.
42. Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J Mol Recognit* 2011;**24**:303–13.
43. Pande A, Patiyal S, Lathwal A, *et al*. Computing wide range of protein/peptide features from their sequence and structure. *bioRxiv* 2019;599126.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020

44. Kumar M, Gromiha MM, Raghava GPS. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics* 2007;**8**:463.
45. 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.
46. Zhang X, Liu S. RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics* 2017;**33**:854–62.
47. 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.
48. Verma R, Tiwari A, Kaur S, *et al*. Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. *BMC Bioinformatics* 2008;**9**:201.
49. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
50. Nagpal G, Usmani SS, Dhanda SK, *et al*. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci Rep* 2017;**7**:42851.
51. 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.
52. Agrawal P, Bhagat D, Mahalwal M, *et al*. AntiCP 2.0: an updated model for predicting anticancer peptides [published online ahead of print, 2020 Aug 6]. *Brief Bioinform* 2020; bbaa153.
53. Usmani SS, Bhalla S, Raghava GPS. Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features. *Front Pharmacol* 2018;**9**:954.
54. Kumar V, Agrawal P, Kumar R, *et al*. Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues. *Front Microbiol* 2018;**9**:725.
55. Wan S, Mak MW, Kung SY. Ensemble linear Neighborhood propagation for predicting subchloroplast localization of multi-location proteins. *J Proteome Res* 2016;**15**:4755–62.
56. Wan S, Mak M-W, Kung S-Y. Transductive learning for multi-label protein subchloroplast localization prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:212–24.
57. Han GS, Yu ZG, Anh V, *et al*. An ensemble method for predicting subnuclear localizations from primary protein structures. *PLoS One* 2013;**8**:e57225.
58. Bouziane H, Messabih B, Chouarfia A. Profiles and majority voting-based ensemble method for protein secondary structure prediction. *Evol Bioinform* 2011;**7**:EBO.S7931.
59. Wheatley LM, Togias A. Clinical practice. Allergic rhinitis. *N Engl J Med* 2015;**372**:456–63.
60. Schuler, IV CF, Montejo JM. Allergic rhinitis in children and adolescents. *Pediatr Clin North Am* 2019;**66**:981–93.
61. Waheed A, Hill T, Dhawan N. Drug allergy. *Prim Care* 2016;**43**:393–400.
62. Abrams EM, Khan DA. Diagnosing and managing drug allergy. *CMAJ* 2018;**190**:E532–8.
63. Savage J, Johns CB. Food allergy: epidemiology and natural history. *Immunol Allergy Clin North Am* 2015;**35**:45–59.
64. Iweala OI, Choudhary SK, Commins SP. Food allergy. *Curr Gastroenterol Rep* 2018;**20**:17.
65. Keet CA, Allen KJ. Advances in food allergy in 2017. *J Allergy Clin Immunol* 2018;**142**:1719–29.
66. Kelleher MM, Tran L, Boyle RJ. Prevention of food allergy—skin barrier interventions. *Allergol Int* 2020;**69**:3–10.
67. Roesner LM, Zeitvogel J, Heratizadeh A. Common and different roles of IL-4 and IL-13 in skin allergy and clinical implications. *Curr Opin Allergy Clin Immunol* 2019;**19**: 319–27.
68. Tankersley MS, Ledford DK. Stinging insect allergy: state of the art 2015. *J Allergy Clin Immunol Pract* 2015;**3**:315–23.
69. Tan JW, Campbell DE. Insect allergy in children. *J Paediatr Child Health* 2013;**49**:E381–7.
70. Campbell DE, Mehr S. Fifty years of allergy: 1965–2015. *J Paediatr Child Health* 2015;**51**:91–3.
71. Usmani SS, Bedi G, Samuel JS, *et al*. THPdb: database of FDA-approved peptide and protein therapeutics. *PLoS One* 2017;**12**: e0181748.
72. Usmani SS, Kumar R, Bhalla S, *et al*. In silico tools and databases for designing peptide-based vaccine and drugs. *Adv Protein Chem Struct Biol* 2018;**112**:221–63.
73. Nagpal G, Usmani SS, Raghava GPS. A web resource for designing subunit vaccine against major pathogenic species of bacteria. *Front Immunol* 2018;**9**:2280.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa294/5985292 by guest on 17 November 2020