ISSN 0026-8933, Molecular Biology, 2018, Vol. 52, No. 2, pp. 285–293. © Pleiades Publishing, Inc., 2018.
Published in Russian in Molekulyarnaya Biologiya, 2018, Vol. 52, No. 2, pp. 333–343.

BIOINFORMATICS

UDC 512.017.1:616_-97:57.083.3

BCIgEPRED—a Dual-Layer Approach for Predicting Linear IgE Epitopes[1]

Vijayakumar Saravanan* and Namasivayam Gautham

*Center for Advanced Study in Crystallography and Biophysics, University of Madras,
Guindy Campus, Chennai, Tamil Nadu, 600025 India*

**e-mail: brsaran@gmail.com*

Received December 2, 2016; in final form, March 13, 2017

Abstract—Allergy is a common health problem worldwide, especially food allergy. Since B cell epitopes that are recognized by the IgE antibodies act as antigenic determinants for allergy, they play a vital role in diagnostics. Hence, knowledge of an IgE binding epitope in a protein is of particular interest for identifying allergenic proteins. Though IgE epitopes may be conformational or linear, identification of the later is useful especially in food allergens that undergo processing or digestion. Very few computational tools are available for the prediction of linear IgE epitopes. Here we report a prediction system that predicts the exact linear IgE epitope. Since our earlier study on linear B-cell epitope prediction demonstrated the effectiveness of using an exact epitope dataset (in contrast to epitope containing region datasets), the dataset in this study uses only experimentally verified exact IgE, IgG, IgM and IgA epitopes. Models for Support Vector Machine (SVM) and Random Forest (RF) were constructed adopting Dipeptide Deviation from the Expected mean (DDE) feature vector. Extensive validation procedures including five-fold cross validation and two different independent dataset tests have been performed to validate the proposed method, which achieved a balanced accuracy ranging from 74 to 78% with area under receiver operator curve greater than 0.8. Performance of the proposed method was observed to be better (accuracy difference of 16–28%) in comparison to the existing available method. The proposed method is developed as a standalone tool that could be used for predicting IgE epitopes as well as to be incorporated into any allergen prediction toolhttps://github.com/brsaran/BCIgePred.

Keywords: epitopes, immunoglobulin E, food allergy, B-cell epitope, dipeptide deviation from expected mean, BCIgEPred

DOI: 10.1134/S0026893318020127

INTRODUCTION

Incidences of food allergy are increasing worldwide, especially among infants and children [1]. For most of the food allergic responses, a portion of the allergen protein (epitope) has to be recognized by the immunoglobulin E antibodies (IgE). These epitopes are termed as IgE epitopes [2]. Since the IgE antibodies are the primary effectors in developing an immunological response to food allergens, knowledge of the IgE epitope in an allergen will aid in the design of immunotherapeutic agents [3]. Two forms of B cell epitopes, namely linear and conformational, are recognized by the IgE antibodies. A linear epitope is a sequential epitope that contains a continuous stretch of amino acids capable of being recognized by the antibodies. A conformational epitope is the one which contains generally non-continuous amino acids close to each other in space. In other words, the three dimensional structure of the allergen is recognized by the antibodies [4]. Though the majority of IgE epitopes are conformational in nature, due to denaturation/digestion during food processing, in diagnosis as well as production of hypoallergenic foods linear epitopes play the key role [5, 6]. Methods that have been used for the identification of B cell (including IgE) epitopes include enzymatic cleavage, peptide arrays, phage display techniques, peptide microarray-based immunoassay, Nuclear magnetic resonance spectroscopy and X-crystallography [7]. However, these methods are considered expensive and time consuming [8].

Several algorithms for computational identification of B cell epitopes (both liner and conformational) have been reported [9–12]. However, to our knowledge there is only one tool reported in the literature for the prediction of immunoglobulin E (IgE) class-specific linear B-cell epitopes. This lack may be due to the lack of experimentally verified class-specific B cell epitopes. The tool AlgPred [13], which is an allergen prediction server, includes a module for predicting allergens based on the presence of linear IgE epitopes. The presence of a linear IgE epitope in AlgPred is predicted based on a similarity search against 183 epi-

[1] The article is published in the original.

topes. AlgPred is designed to predict allergens and not IgE epitopes, and the IgE similarity module in AlgPred cannot be directly used to predict IgE epitopes. AlgPred [13] attempted to predict linear IgE epitopes, but achieved poor performance due to the lack of fixed length experimentally verified IgE epitopes at the time of development. The AlgPred server supports only proteinsfor input (i.e. allows only proteinswith lengths >10 to be submitted to the server) and not peptides. The majority of independent datasets used in the present study contain peptides of length <10. Thus, the tool proposed in this paper cannot be directly compared with AlgPred. The creation of experimentally verified epitope databases, such as Immune Epitope Database (IEDB) [14] allowed the development of class-specific B cell epitope prediction tools. IgPred [15] is one of such tools, which is a support vector machine (SVM) based type-specific B cell epitope predictor. It utilizes experimentally verified IgE, IgA, IgG, and non-B-cell epitopes from IEDB for training the model. It adopts a dipeptide composition based feature vector for classification. Though the dataset used in IgPred is experimentally verified, the training dataset consisted of whole peptide regions that contain the epitope, rather than the specifically identified epitope alone andour earlier study [16] on linear B-cell epitope prediction clearly indicated the underperformance of methods for exact epitope prediction that used epitope containing peptide regions rather than exact epitopes, in their training procedures. Also, we found that Dipeptide Deviation from Expected Mean (DDE) feature vector was better at classifying the exact epitopes in comparison to other feature vectors including amino acid composition, dipeptide composition, physiochemical parameters, and amino acid pair propensity [16]. In the present study the main objective was set to develop a tool that is capable of predicting linear IgE-specific B cell epitopes (predominant for food allergy responses) focusing on use of an exact epitope dataset and DDE feature vector for enhancing the prediction accuracy. The proposed method is a two-layer prediction system, where the first layer predicts whether the peptide's/protein's regions of interest are exact linear B-cell epitopes or not via our previously developed LBEEP method [16] and in the second layer, the regions predicted to be exact linear epitopes are subjected to the prediction model developed in this study for IgE specificity. The validation results show an improvement in the prediction performance in comparison to the existing method. The proposed method is developed as a standalone tool that could be used for predicting IgE epitopes as well as to be incorporated into any allergen prediction tool: https://github.com/brsaran/BCIgePred.

MATERIALS AND METHODS

Dataset and pre-processing. IEDB V3.0 [14], an updated database on experimentally verified epitopes and non-epitopes, was utilized to construct the dataset used in this study. The assay data (as of January, 2016) was accessed from the website (http://www.iedb.org/learn_more_v3.php). Unlike IgPred [15], where all experimentally verified class-specific epitopes are used, in this study only experimentally verified exact epitopes were used. (The reasons for using only exact epitopes are elaborated in the "Discussion" section). This was achieved by specifying "Exact epitopes" in the "Structure defines" column of the database. From such exact epitopes, the data was further refined to positive assay by filtering the "Qualitative measurement" column to positive, from which the IgE, immunoglobulin G (IgG), immunoglobulin A (IgA), and immunoglobulin M (IgM) epitopes were partitioned based on the "Assayed antibody heavy chain" column of the data. In this study, the epitopes that are not assigned IgE in the assayed antibody heavy chain column are grouped and collectively named as non-IgE epitopes, which include IgG, IgA, and IgM heavy chain types. Following the above mentioned procedures, a total of 2020 exact IgE epitopes and 12.094 non-IgE epitopes were obtained. To remove redundancy in the dataset, CD-HIT suite [17] was used with an identity cut-off set to 0.6 (60%). The final dataset contained 1414 IgE and 4695 non-IgE epitopes and was named Nr_Dataset. Ninety percent of the Nr_Dataset, that is 1273 IgE and 4226 non-IgE epitopes, were used for training the model and named as Tr_Data. The remaining10%of the data (141 IgE and 469 non-IgE) were used for the independent set test and named as Ind_Set_1. The data in the Ind_Set_1 was never exposed to the model at any stage of the training process. In order to compare the performance of proposed method with the existing tool IgPred, Ind_Set_1A was created from the Ind_Set_1, which contained no data included for training the IgPred, as well as the proposed method. To achieve this, peptides in both Ind_Set_1 and all of the training data of IgPred were compared and those that were identical were excluded from the Ind_Set_1, which resulted in 60 IgE and 256 non-IgE epitopes. Another Independent dataset, Ind_Set_2, was obtained from the updated version of Allergen database for food safety (February, 2016 updated version) [18], making sure no data in the set had greater than 60% similarity with the Tr_Data. The distribution of the datasets used in this study is listed in Table 1. The datasets constructed and used in this study are available as Supplementary text 1 ((see Supplementary on the web-site http://www.molecbio.ru/downloads/2018/2/supp_VijayakumarSaravanan_engl.pdf).

Feature vector and algorithm. Sequence derived feature vectors like amino acid pair propensity [19], dipeptide composition [20], amino acid string kernels [21] and tri-peptide similarity score [22] have been used in earlier studies for the prediction of linear B-cell epitopes. However, our earlier study demonstrated the efficiency of using the DDE (Dipeptide Deviation

Table 1. Distribution of dataset constructed in this study*

Dataset	Class	No. of instances
Original Dataset	IgE epitopes	2020
	Non-IgE epitopes**	12094
Nr_Data	IgE epitopes	1414
	Non-IgE epitopes**	4695
Tr_Data	IgE epitopes	1273
	Non-IgE epitopes	4226
Ind_Set_1	IgE epitopes	141
	Non-IgE epitopes**	469
Ind_Set_1A	IgE epitopes	60
	Non-IgE epitopes**	256
Ind_Set_2	IgE epitopes	102

* Nr—non-redundant; Tr—training; Ind—independent.
** Epitopes belonging to the class IgG, IgA, and IgM.

from Expected Mean) feature vector for the prediction of linear B-cell epitopes [16]. We have used this feature vector here for linear IgE specific B-cell epitope prediction. DDE is a 60 dimensional feature vector constructed from three parameters viz. dipeptide composition measure, theoretical mean of dipeptides, and theoretical variance of dipeptides. A detailed description of the procedure for the construction of DDE from the amino acid sequence has been given in our earlier paper [16].

The training data (Tr_Data) contains 1273 IgE epitopes and 4226 non-IgE epitopes. Since there is three times less IgE epitope data than non-IgE data, the dataset is imbalanced. In order to make the training dataset balanced we adopted an under-sampling technique [23], in which the class with the larger amount of data is reduced to approximately match the class with the less data. This will result in loss of non-IgE information due to under-sampling. To avoid this, in this study we developed three different training data sets. In each of the three sets, the IgE class contains the 1273 IgE epitopes referred to above and the non-IgE class contains 1409 (set 1), 1409 (set 2), and 1408 (set 3) non-IgE epitopes. The non-IgE epitopes were partitioned into three sets by random sampling while ensuring that the non-IgE data in each set is non-redundant. With these three sets, three different models were created. The final prediction score was calculated as the average of the probability scores given by each model. The models were trained using two widely used machine learning classifiers—Support Vector Machine (SVM) [24] and Random forest (RF) [25]. The number of trees for RF were set to 100, while for SVM, radial basis function kernel was adopted with the tuning parameters C and gamma set to 32 and 0.0078125, respectively. The architecture of the method is shown in Fig. 1.

Validation measures. An independent dataset test was used to validate the proposed method. In the independent dataset test, the dataset that was never used in any of the training procedures is tested against the developed models. Two such independent datasets, Ind_Set_1 and Ind_Set_2, were used to validate the proposed method. The independent dataset test was performed on the models trained through *k*-fold cross-validation. For the k-fold cross-validation, $k = 5$ was chosen (5-CV). Tr_Data was used to perform the 5-CV, where the dataset was divided into 5-subsets and in each mode of evaluation one subset was validated against the model constructed using the remaining subsets, ensuring that each subset was validated at least once. Sensitivity (Sn), Specificity (Sp), Precision (P), Mathew's correlation coefficient (MCC), receiver operator characteristic curve (ROC), area under ROC (AU_ROC), overall accuracy (OA), and balanced accuracy (BA) were computed as follows for validation.

Sensitivity (Sn):

$$S_n = \frac{TP}{(TP + FN)} \times 100.$$

Specificity (Sp):

$$S_p = \frac{TN}{(TN + FP)} \times 100.$$

Precision (P):

$$P = \frac{TP}{(TP + FP)} \times 100.$$

Mathew's correlation coefficient (MCC):

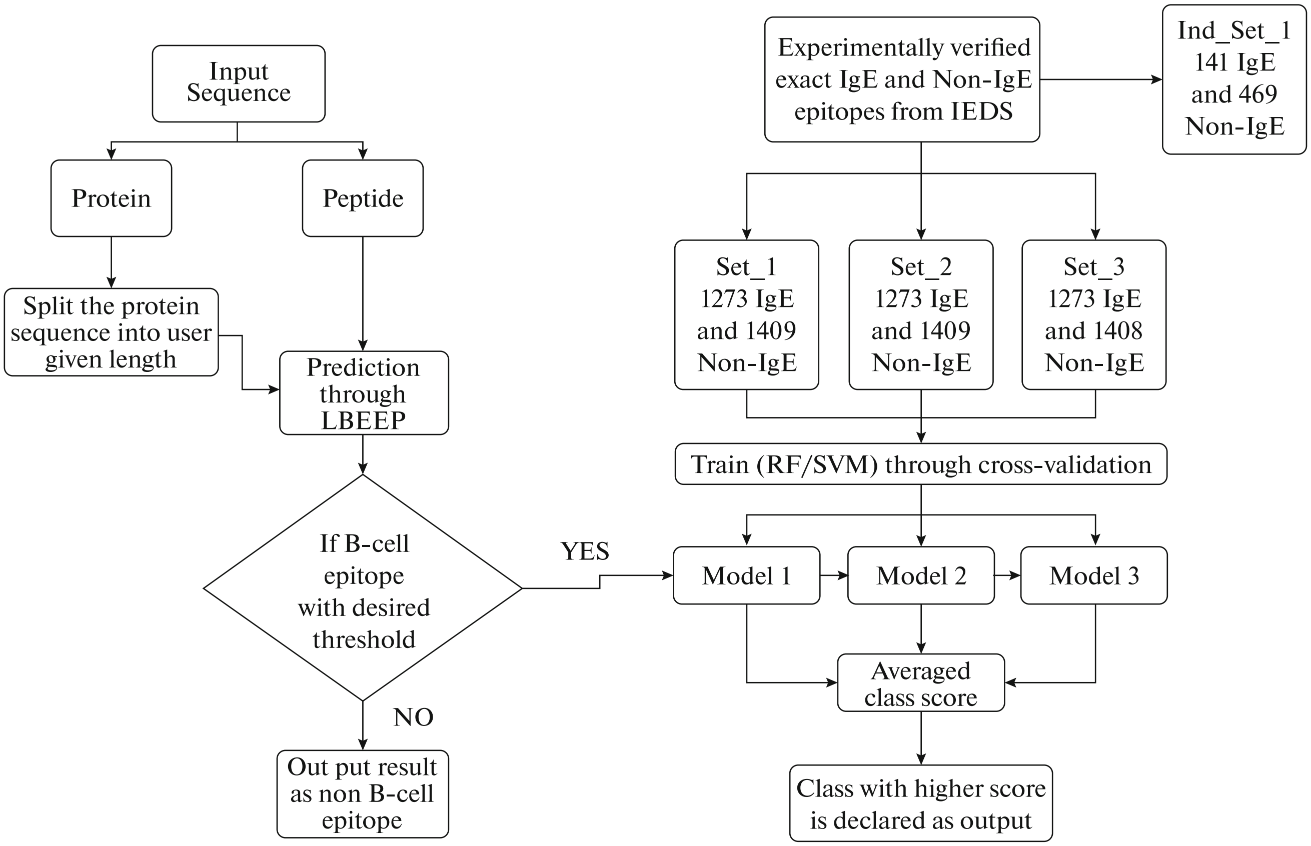
Input Sequence
Protein
Peptide
Split the protein sequence into user given length
Prediction through LBEEP
If B-cell epitope with desired threshold
YES
NO
Out put result as non B-cell epitope
Experimentally verified exact IgE and Non-IgE epitopes from IEDS
Ind_Set_1 141 IgE and 469 Non-IgE
Set_1 1273 IgE and 1409 Non-IgE
Set_2 1273 IgE and 1409 Non-IgE
Set_3 1273 IgE and 1408 Non-IgE
Train (RF/SVM) through cross-validation
Model 1
Model 2
Model 3
Averaged class score
Class with higher score is declared as output

Fig. 1. Architecture of BCIgEPred.

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

Overall accuracy (OA):

$$\mathrm{OA} = \frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})}\times 100.$$

Balanced accuracy (BA):

$$\mathrm{BA} = \frac{\mathrm{S_n}+\mathrm{S_p}}{2}\times 100.$$

Where, TP is true positives (IgE epitopes predicted as IgE epitopes); TN is true negatives (non-IgE epitopes predicted as non-IgE epitopes); FP is false positives (non-IgE epitopes predicted as IgE epitopes); FN is false negatives (IgE epitopes predicted as non-IgE epitopes).

RESULTS

Differentiation between IgE and non-IgE DDE Composition

DDE was computed for both IgE and non-IgE epitopes in Nr_Data. Discrepancy ratios between IgE and non-IgE dipeptides were computed and those with values greater than 10 were listed in Table 2. A total of 24 dipeptides were found to have discrepancy a ratio greater than 10. The highest discrepancy ratio (94.93) was found to be with the dipeptide containing residues Histidine (H) and Phenylalanine (F). Amino acids Histidine (H), Aspartic acid (D), Glutamine (Q), Alanine (A), and Leucine (L) were found to appear more frequently (4 times) in the dipeptides with high discrepancy ratios. The absolute discrepancy ratios for all of the possible dipeptides of 20 amino acids (20 × 20 = 400) were listed as a matrix in Supplementary Table 1 (see Supplementary on the web-site http://www.molecbio.ru/downloads/2018/2/supp_VijayakumarSaravanan_engl.pdf).

Cross-Validation of DDE Discriminating Ability on Partitioned Dataset

Since the training set was partitioned into three different sets (set 1, set 2, and set 3), a cross-validation test was performed on all three sets for inspecting biases, if any, in the data. Cross-validation was performed for both SVM and RF methods, the results are listed in Table 3. In RF, set 1 achieved highest OA and BA, while in SVM, set 3 achieved highest OA and BA. However, the margin of difference among the three sets with respect to all validation measures seems to be less (<2), in both RF and SVM. In comparison to RF, SVM obtained marginally better averaged OA and BA. The averaged sensitivity, specificity, and precision among RF and SVM varied with a high percentage difference (≈>10%). The sensitivity of RF was better than SVM, while the specificity of SVM was better than of RF. Both RF and SVM had a positive MCC, with marginal difference. On varied threshold, the area under the receiver operator characteristic curve (AU_ROC) of both RF and SVM was close to 1, depicting that the predictions of both models were significantly better than a random guess (Fig. 2). Overall, only a mild variation (margin of difference <2) was observed among validation measures within the three sets on both RF and SVM, indicating no biases in the partitioned set.

Independent Dataset Test

Procedures including sub-sampling, jack-knife, and independent dataset test are generally performed for validating a prediction model [26]. However, independent dataset test was considered to be optimal method to validate the generalization of the proposed model [27]. In an independent dataset test, the data that was never exposed to the model during the training procedures will be tested. In this study, three different independent dataset were created (Table 1). Ind_Set_1 is completely independent for the proposed method, whereas Ind_Set_1A and Ind_Set_2 are independent to both the proposed method and IgPred. Independent test results for Ind_Set_1 on proposed method (both SVM and RF) are listed in Table 4. Set_1, Set_2, and Set_3 are the outputs from partitioned model and the "Merged" is the combined score of partitioned models as explained in the "Feature Vector and Algorithm" section of this article. The results of the independent set test on Ind_Set_1A and Ind_Set_2 are listed in Table 5 and 6 respectively. Since the Ind_Set_2 was obtained from the Allergen database for food safety, which contains only the IgE epitopes, non-IgE epitopes from other sources are not included.

Performance Comparison of RF and SVM on Ind_Set_1

It is observed that there is no high variations (<4%) in the overall accuracy between the three sets on both RF and SVM. Also, no other validation measures exceeded a difference of >4% between the three sets on both RF and SVM. However, sensitivity of Set_2 on RF had a difference of 6.4% in comparison to Set_3. Similarly, precision of Set_2 on SVM had a difference of 8% in comparison to Set_3. Except sensitivity, all other score-merged validation measures (Row tagged "Merged" under RF and SVM of Table 4) of RF was found to be enhanced with respect to the individual validation measures of Set_1, Set_2, and Set_3 models. Similarly, in SVM, except BA and AU_ROC, all other validation measures were found to be enhanced. Beside this, the score-merged results on both SVM and RF was better in most of the validation aspects (considered in this study) than the majority voting scheme (Row tagged "Majority" under RF and SVM of Table 4). This indicates that the employed score merging strategy of partitioned balanced models was effective. Hence, in this section the comparison of RF and SVM was made with respect to the "Merged" results. Also, for further independent test result validation (Ind_Set_1A and Ind_Set_2 dataset) and for the proposed tool, BCIgEPred, only the score-merge strategy was adopted.

The sensitivity of RF was ≈10% greater than of SVM, while the specificity of SVM was ≈14% greater than that of RF. The precision was observed to be bet-

Table 2. TheDDE ratio between IgE exact epitopes and non-IgE epitopes (other classes) that possess ratios greater than 10

Dipeptide[a]	DDE Ratio[b]
HF	94.93
DS	70.18
DL	57.23
EH	36.16
TD	29.71
QL	29.08
MT	23.09
WP	20.71
WL	19.06
FP	16.86
TM	15.92
DM	15.68
AF	15.02
SM	14.61
FY	14.54
QQ	13.74
GF	12.70
LQ	12.27
PT	12.06
HY	11.87
AM	11.14
HW	11.04
AT	10.41
AG	10.11

[a] Standard single letter amino acid code. [b] Absolute ratio between DDE IgE epitopes and DDE non-IgE epitopes from Nr_dataset.

Table 3. Five-fold cross-validation results on Tr_Dataset*

Model	Sn, %	Sp, %	P, %	MCC	AU_ROC	OA, %	BA, %
RF							
Set_1	72.9	76.1	73.4	0.49	0.833	74.5	74.5
Set_2	72.9	75.3	72.7	0.48	0.827	74.1	74.1
Set_3	72.0	76.2	73.2	0.48	0.837	74.3	74.1
Average	72.6	75.8	73.1	0.48	0.833	74.3	74.2
SVM							
Set_1	59.2	90.2	84.5	0.52	0.791	75.5	74.7
Set_2	59.3	88.0	82.7	0.50	0.780	74.7	73.6
Set_3	60.6	89.3	83.6	0.52	0.793	75.6	74.9
Average	59.7	89.1	83.6	0.51	0.788	75.2	74.4

* Hereinafter: Sn—sensitivity; Sp—specificity; P—precision; MCC—Mathews correlation coefficient; AU_ROC—area under receiver operator characteristic curve; and OA—overall accuracy; BA—balanced accuracy.

ter in SVM with a percentage difference of ≈22% from RF. SVM outperformed RF in MCC and OA, however AU_ROC of RF was marginally greater than SVM. AU_ROC of both RF and SVM was close to 1, signifying the near perfect prediction, and the ROC graph (Fig. 3) of both RF and SVM indicate that the predictions made are far better than a random guess by varying threshold. Though the overall accuracy of SVM was higher than the RF, balanced accuracy is the measure that validates the balance between both the sensitivity and specificity, which was almost equal in both cases. Hence, in order to compare the performance, both models were subjected to the Ind_Set_1A and Ind_Set_2 independent dataset test along with the existing method IgPred.

Performance Comparison of Proposed Method with Existing Methods

IgPred [15] is the only existing tool, to the best of the authors' knowledge, that predicts class specific linear B-cell epitopes and hence, the proposed method could be directly compared only with IgPred. Since only the IgE data has been used in the Ind_Set_1A and Ind_Set_2 and IgPred is capable of reporting class specific types, all results reported by IgPred (for the two datasets Ind_Set_1A and Ind_Set_2) as IgG, IgA, and non- B-cell epitope were considered as false predictions. All settings of IgPred were set to default values.

On Ind_Set_1A (Table 5), RF outperformed SVM and IgPred in sensitivity, AU_ROC and BA. The specificity of IgPred was perfect (100%), however, the sensitivity was lower than that of RF and SVM. This was also reflected in the balanced accuracy, where the accuracy ofRF was more than 6% greater than that of IgPred. The ROC graph (Fig. 4) for Ind_Set_1A indicates that on varied threshold the RF performance was better than SVM, IgPred, and random guess. In addition, the area under ROC (Table 5) of RF was significantly greater than that of IgPred and SVM, indicating

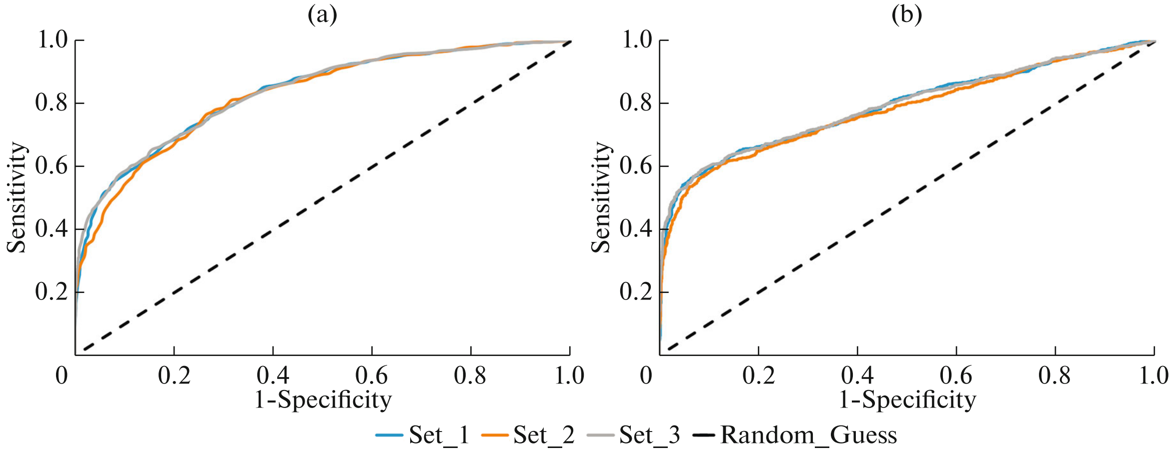

Fig. 2. Receiver operator graph of cross-validation results. A—Random Forest Models. B—Support Vector machine Models.

Table 4. Independent dataset test result on Ind_Set_1*

Model	Sn, %	Sp, %	P, %	MCC	AU_ROC	OA, %	BA, %
RF							
Set_1	74.5	78.0	50.5	0.46	0.857	77.21	76.25
Set_2	78.0	78.3	51.9	0.49	0.863	78.19	78.15
Set_3	71.6	76.8	48.1	0.42	0.842	75.57	74.20
Majority	73.0	80.8	53.3	0.48	0.860	79.01	76.92
Merged	75.8	80.8	54.3	0.51	0.868	79.67	78.34
SVM							
Set_1	63.1	91.3	68.5	0.56	0.822	84.75	77.20
Set_2	64.5	94.0	76.5	0.62	0.830	87.21	79.25
Set_3	64.5	92.5	72.2	0.59	0.835	86.06	78.50
Majority	62.4	93.8	75.2	0.60	0.830	86.55	78.11
Merged	64.5	94.4	77.0	0.63	0.829	87.54	78.31

Table 5. Independent dataset test result on Ind_Set_1A

Model	Sn, %	Sp, %	MCC	AU_ROC	OA, %	BA, %
RF*	68.33	84.76	0.51	0.85	81.64	76.54
SVM*	43.33	94.14	0.43	0.75	84.49	68.73
IgPred	40.00	100.00	0.59	0.78	88.60	70.00

* Models developed in this study.

that RF performance was balanced and better on different threshold in comparison to SVM and IgPred.

Similarly on Ind_Set_2, RF outperformed SVM and IgPred, by having an accuracy difference of ≈15% with IgPred and over 50% with SVM. Comparison of SVM and IgPred results suggest that SVM produced only a marginally better result on Ind_Set_1A and a poorer result on Ind_Set_2. These results suggest that the proposed method (RF model) was significantly better (>10% accuracy difference on different independent datasets) than the existing method. The poor performance of SVM suggests that though SVM performance was equally well (and in some case better) than RF on cross-validation and independent dataset tests (Ind_Set_1) it suffers to maintain balance in prediction on Ind_Set_1A and Ind_Set_2. In contrast, RF performance was consistent on all validation procedures and performed better than IgPred, indicating that the proposed RF model may be preferred for exact linear IgE epitope prediction.

BCIgEPred usage and Features

The standalone version of BCIgEPred was developed using PERL V5.18.4 (Practical Extract Report Language). The complete source code and dependencies are available at https://github.com/brsaran/BCIgePred and free to download. The details of command line options and their detailed usage may be found in a readme file available at https://github.com/brsaran/BCIgePred.

Table 6. Independent dataset results comparison of existing and proposed method on Ind_Set_2

Model	No. of IgE	No. Correctly predicted	Acc*, %
RF**	102	80	78.43
SVM**	102	28	27.45
IgPred	102	64	62.74

* Acc = (No. correctly predicted/No. of IgE) × 100.
** Models developed in this study.

DISCUSSION

In the presented linear IgE epitope prediction tool, a new dataset containing experimentally verified exact IgE and non-IgE (IgA, IgG and IgM) epitopes was created and utilized for creating models. Initial analysis of IgE and non-IgE with a DDE feature vector revealed that amino acids Histidine (H), Aspartic acid (D), Glutamine (Q), Alanine (A), and Leucine (L) were more frequently present in dipeptides with high discrepancy ratios. This was in accordance to an earlier study [15] where in IgE epitope dipeptides containing amino acids Q, L and A were highly preferred. A total of 24 dipeptides had high (>10) DDE discrep-

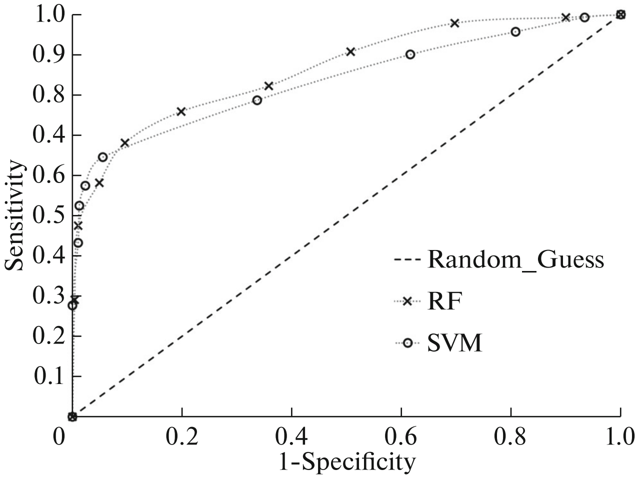

Fig. 3. Receiver operator graph of Independent dataset test results.

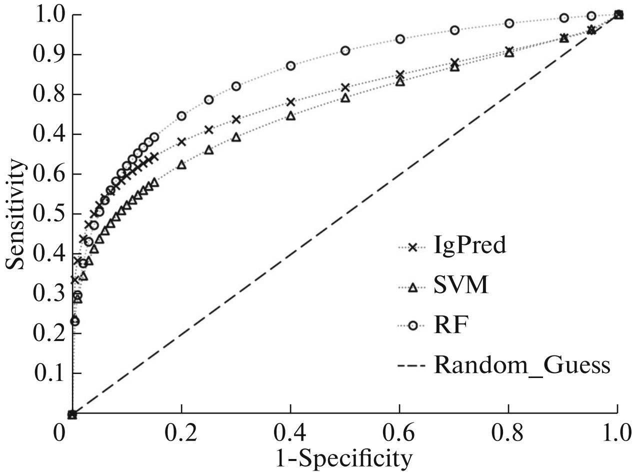

Fig. 4. Receiver operator graph of Independent dataset (Ind_Set_1A) test results.

ancy ratios between the IgE and non-IgE epitopes, suggesting the feature vector's discriminating ability between the epitope classes. To avoid misbalance between the classes and loss of information in the data, the dataset was partitioned into three subsets and a model was developed using each set; and predictions of each models were taken into consideration for this study. Only mild variations in the cross-validation results on different sets indicate no bias in the partitioned datasets and the same models could be utilized for prediction. The independent dataset test (Ind_Set_1) results (Table 4) suggest that both RF and SVM models, developed with the DDE feature vector, were equally well in predicting IgE and non-IgE epitopes. Also, the same results suggest that use of score merge strategy (refer to the methodology section) improved the overall prediction results in contrast to the majority voting scheme (which is often used in prediction systems with multiple models). Independent dataset test (Ind_Set_1A and Ind_Set_2) between the proposed method and IgPred clearly indicates the superior performance of the proposed method (RF model) over the existing method IgPred. Though both RF and SVM models developed in this study performed equally well on different validation tests, independent dataset test (Ind_Set_1A and Ind_Set_2) indicates that the RF model is more stable and balanced in predicting linear IgE epitopes than the SVM model. However, the developed tool includes the feature of selecting the model by the user for their prediction. The proposed method is a two-layer prediction system, which includes our previously developed method LBEEP [16] for the first layer prediction and the model developed in this study for the second layer prediction that predicts whether the peptide possesses IgE specificity or not. The results of our earlier study [16] and those from this one suggest that the DDE feature vector and the use of exact epitopes (rather than epitope contacting regions) are more effective in predicting linear B-cell and IgE epitopes. Since both peptide and protein may be provided as an input to the developed tool, this prediction system could be practically used as a module in a hybrid allergen prediction system.

CONCLUSION

This study proposed a two-layer prediction system for predicting linear IgE epitopes, by utilizing the DDE feature vector. An exact experimentally verified class specific epitope dataset was constructed. Extensive validation procedures were carried out to validate the proposed method. The results suggest that with the DDE feature vector the RF model developed in this study was efficient in predicting the linear exact IgE epitopes. The performance of the proposed method was also compared with the existing tool and found to be better in predicting linear IgE epitopes. B-cell epitopes that are recognized by the IgE antibodies act as an antigenic determinants for allergy, especially in food allergy. Hence, we believe that the proposed tool could be effectively utilized to predict IgE binding epitopes in proteins or to predict a peptide's efficiency as a linear IgE epitope, which on further validation could be used as a diagnostic tool as well as in production of hypoallergenic foods.

ACKNOWLEDGMENTS

Vijayakumar Saravanan is supported by University Grants Commission Dr. S.Kothari post-doctoral fellowship (BSR/BL/14-15/0103). The authors thank the Center for Advanced Study in Crystallography and Biophysics, University of Madras for providing advanced computational facility to carry out this study.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Rueter K., Prescott S. 2014. Hot topics in paediatric immunology: IgE mediated food allergy and allergic rhinitis. *Aust. Fam. Physician.* **43**, 680–685.
2. Lin J., Sampson H.A. 2009. The role of immunoglobulin E-binding epitopes in the characterization of food allergy. *Curr. Opin. Allergy Clin. Immunol.* **9**, 357–363.
3. Tanabe S. 2008. Analysis of food allergen structures and development of foods for allergic patients. *Biosci. Biotechnol. Biochem.***72**, 649–659.
4. Chen X., Negi S.S., Liao S., et al. 2016. Conformational IgE epitopes of peanut allergens Ara h 2 and Ara h 6. *Clin. Exp. Allergy.* **46** (8), 1120–1128. doi 10.1111/cea.12764
5. Matsuo H., Yokooji T., Taogoshi T. 2015. Common food allergens and their IgE-binding epitopes. *Allergol. Int.* **64**, 332–343.
6. Pomés A. 2009. Relevant B cell epitopes in allergic disease. *Int. Arch. Allergy Immunol.* **152,** 1–11.
7. Zhong-Shan G., Hua-Hao S., Min Z., (Eds.). 2012. *Multidisciplinary Approaches to Allergies.* Berlin: Springer, **vol.** 7, pp. 113–126.
8. Costa J.G., Faccendini P.L., Sferco S.J., et al. 2013. Evaluation and comparison of the ability of online available prediction programs to predict true linear B-cell epitopes. *Protein Pept. Lett.* **20**, 724–730.
9. El-Manzalawy Y., Honavar V. 2010. Recent advances in B-cell epitope prediction methods. *Immunome Res.* **6** (Suppl 2), S2.
10. Soria-Guerra R.E., Nieto-Gomez R., Govea-Alonso D.O., Rosales-Mendoza S. 2015. An overview of bioinformatics tools for epitope prediction: Implications on vaccine development. *J. Biomed. Inf.* **53,** 405–414.
11. Dhanda S.K., Usmani S.S., Agrawal P., et al. 2017. Novel in silico tools for designing peptide-based subunit vaccines and immunotherapeutics. *BriefingsBioinf.* **18** (3), 467–478. doi 10.1093/bib/bbw025
12. Davydov Y.I., Tonevitsky A.G. 2009. Prediction of linear B-cell epitopes. *Mol. Biol.* (Moscow). **43** (1), 150–158.
13. Saha S., Raghava G. 2006. AlgPred: Prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* **34**, W202–W209.
14. Vita R., Overton J.A., Greenbaum J.A., Ponomarenko J., Clark J.D., Cantrell J.R., Wheeler D.K., Gabbard J.L., Hix D., Sette A., Peters B. 2015. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405–D412.
15. Gupta S., Ansari H.R., Gautam A.; Open Source Drug Discovery Consortium, Raghava G.P. 2013. Identification of B-cell epitopes in an antigen for inducing specific class of antibodies. *Biol. Direct.* **8**, 27. doi 10.1186/1745-6150-8-27
16. Saravanan V., Gautham N. 2015. Harnessing computational biology for exact linear B-cell epitope prediction: A novel amino acid composition-based feature descriptor. *OMICS.* **19**, 648–658.
17. Huang Y., Niu B., Gao Y., Fu L., Li W. 2010. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics.* **26**, 680–682.
18. Nakamura R., Teshima R., Takagi K., Sawada J. 2004. Development of Allergen Database for Food Safety (ADFS): An integrated database to search allergens and predict allergenicity. *Kokuritsu Iyakuhin Shokuhin Eisei Kenkyusho Hokoku.* **123**, 32–36.
19. Chen J., Liu H., Yang J., Chou K.C. 2007. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids.* **33**, 423–428.
20. Singh H., Ansari H.R., Raghava G.P. 2013. Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PloS One.* **8**, e62216.
21. El-Manzalawy Y., Dobbs D., Honavar V. 2008. Predicting linear B-cell epitopes using string kernels. *J. Mol. Recognit.* **21**, 243–255. doi 10.1002/jmr.893
22. Yao B., Zhang L., Liang S., Zhang C. 2012. VMTriP: A method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. *PloS One.* **7**, e45152. doi 10.1371/journal.pone.0045152
23. Chawla N.V. 2005. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook.* New York: Springer-Verlag, pp. 853–867.
24. Cortes C., VapnikV. 1995. Support-vector networks. *Mach. Learn.* **20,** 273–297.
25. Breiman L. 2001. Random forests. *Mach. Learn.* **45**, 5–32.
26. Lin W.-Z., Fang J.A., Xiao X., Chou K.C. 2011. iDNA-Prot: Identification of DNA binding proteins using random forest with grey model. *PloS One.* **6**, e24756. doi 10.1371/journal.pone.0024756
27. Štambuk N., Konjevoda P. 2011. The role of independent test set in modeling of protein folding kinetics. In: *Software Tools and Algorithms for Biological Systems.* Eds. Arabnia, H.R.R., Tran, Q.N. Advances in Experimental Medicine and Biology, vol. 696. New York: Springer, pp. 279–284.