Gupta et al. Biology Direct 2013, 8:27
http://www.biologydirect.com/content/8/1/27

BIOLOGY DIRECT

RESEARCH Open Access

Identification of B-cell epitopes in an antigen for inducing specific class of antibodies

Sudheer Gupta[1], Hifzur Rahman Ansari[1], Ankur Gautam[1], Open Source Drug Discovery Consortium[2] and Gajendra PS Raghava[1*]

Abstract

Background: In the past, numerous methods have been developed for predicting antigenic regions or B-cell epitopes that can induce B-cell response. To the best of authors' knowledge, no method has been developed for predicting B-cell epitopes that can induce a specific class of antibody (*e.g.,* IgA, IgG) except allergenic epitopes (IgE). In this study, an attempt has been made to understand the relation between primary sequence of epitopes and the class of antibodies generated.

Results: The dataset used in this study has been derived from Immune Epitope Database and consists of 14725 B-cell epitopes that include 11981 IgG, 2341 IgE, 403 IgA specific epitopes and 22835 non-B-cell epitopes. In order to understand the preference of residues or motifs in these epitopes, we computed and compared amino acid and dipeptide composition of IgG, IgE, IgA inducing epitopes and non-B-cell epitopes. Differences in composition profiles of different classes of epitopes were observed, and few residues were found to be preferred. Based on these observations, we developed models for predicting antibody class-specific B-cell epitopes using various features like amino acid composition, dipeptide composition, and binary profiles. Among these, dipeptide composition-based support vector machine model achieved maximum Matthews correlation coefficient of 0.44, 0.70 and 0.45 for IgG, IgE and IgA specific epitopes respectively. All models were developed on experimentally validated non-redundant dataset and evaluated using five-fold cross validation. In addition, the performance of dipeptide-based model was also evaluated on independent dataset.

Conclusion: Present study utilizes the amino acid sequence information for predicting the tendencies of antigens to induce different classes of antibodies. For the first time, *in silico* models have been developed for predicting B-cell epitopes, which can induce specific class of antibodies. A web service called IgPred has been developed to serve the scientific community. This server will be useful for researchers working in the field of subunit/epitope/peptide-based vaccines and immunotherapy (http://crdd.osdd.net/raghava/igpred/).

Reviewers: This article was reviewed by Dr. M Michael Gromiha, Dr Christopher Langmead (nominated by Dr Robert Murphy) and Dr Lina Ma (nominated by Dr Zhang Zhang).

Keywords: Support vector machine, Prediction, Antibody, Class-specific, B-cell epitope, Isotype

* Correspondence: raghava@imtech.res.in
[1]Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh 160036, India
Full list of author information is available at the end of the article

BioMed Central

© 2013 Gupta et al.; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background

Innate and adaptive immune responses are the two main arms of host immune system to combat invading pathogens. The innate immunity, also known as first line defense, is non-specific and responsible for the immediate action against infection. On the other hand, adaptive immunity is a highly specialized type of defense system, where the immune system first recognizes pathogen and then develops pathogen-specific defense mechanisms. In addition, adaptive immunity also generates memory cells that can handle pathogen effectively and rapidly if the system is attacked later by the same pathogen. The adaptive arm of immunity can be divided broadly into two categories; humoral and cell-mediated, responsible for activating B-cells and T-cells respectively.

Vaccination is an artificial procedure for sensitizing immune response or generating memory cells against a desired pathogen. Over the years, subunit vaccine design has become an integral part of vaccine design in which immunogenic region of protein is used instead of complete pathogen or antigen [1]. Antibodies (Abs) are one of the important components of humoral immunity where B-cells recognize antigenic regions or B-cell epitopes (BCEs) and generate antigen specific Abs. These Abs perform various functions such as phagocytosis [2], cell-mediated cytotoxicity [3], neutralization, compliment activation [4] and mast cell binding [5]. Broadly these Abs can be categorized in five classes or isolates *i.e.,* IgA, IgD, IgE, IgG, and IgM.

It has been observed in the past that particular pathogen/antigen induce defined class or subclass of Abs, for example, infections like schistosomiasis and filariasis induce a mixed response of IgE and IgG [6-8]. In case of protozoan like *Plasmodium falciparum,* Ab response of merozoite surface proteins constitutes mainly IgG1 and IgG3 subclasses [9,10]. On the other hand, viruses like rotavirus, HIV and influenza virus, are well known for inducing IgA type of response [11]. In case of IgE inducing antigens (allergens), the studies showed that the allergens have some features that make them allergenic [12]. These facts together suggest that there are desired effector functions of Abs, which are needed to encounter various types of pathogens. Thus, it is important to understand why the immune system produces different classes of antibodies against different antigens. This understanding will help an experimental biologist to design a better vaccine for the induction of systemic or mucosal immunity as well as immunotherapy. In the past, numerous databases and methods have been developed for maintaining and predicting BCEs in an antigen [13-16]. Till date, limited efforts have been made to develop the method for predicting allergens or BCEs that can induce IgE type of antibodies [17,18]. To the best of authors' knowledge, no comprehensive attempts have been made for predicting BCEs responsible for inducing specific class of Abs or discrimination of epitopes that induce different class of Abs.

In this paper, we have made an attempt to understand the relation between amino acid sequence of epitopes and type of Abs they will induce. First we have collected IgG, IgE and IgA specific BCEs from Immune Epitope Database (IEDB). Subsequently, these three classes of epitopes were analyzed to understand which residues or group of residues are preferred among these sequences. Based on comparative analysis, we developed prediction models using various features like amino acid composition, dipeptide composition and binary profiles. We also developed a user-friendly platform for the scientific community that allows users to predict IgG, IgE and IgA specific BCEs.

Results

Analysis

Composition analysis

In order to ascertain whether certain types of residues are dominated in different classes of BCEs, the percent average amino acid composition of IgG, IgE and IgA specific BCEs and non-B-cell epitopes (non-BCEs) was calculated and compared (Figure 1). The analysis revealed that there are differences in the percent average amino acid composition profiles of four classes (IgG, IgE, IgA, and non-BCEs) of epitopes. As shown in Figure 1, certain types of residues are abundant in each class, for instance Pro and Gln are abundant in IgA inducing epitopes while Cys and Glu are found to be dominated in IgE inducing epitopes. These observations are in accordance to several previous reports, where researchers have shown that there is a propensity of Cys residues in IgE inducing epitopes, and they form stable conformational epitopes through disulphide bonds [19-25]. It has also been observed in the past that IgA binding antigenic regions are Pro/Gly rich [26,27].

Dipeptide composition provides more information than amino acid composition and has been used in the past for developing various classification models [28-30]. We have also computed and compared average dipeptide composition for each class of epitopes (IgG, IgE, IgA and non-BCE). It was observed that each class of epitopes has certain types of dipeptides having significantly higher composition (Welch's *t*-test) than other class of epitopes. Dipeptides AS, GP, WK, YR, *etc.* are prevailing in IgG; IQ LA, NA, NE, *etc.* are frequent in IgE, and ED, FP, PF, PQ, PY, QP, *etc.* are predominant in IgA class of epitopes (Additional file 1).

Residue preference

In order to understand the preference of residues at different positions in epitopes, we analyzed two sample

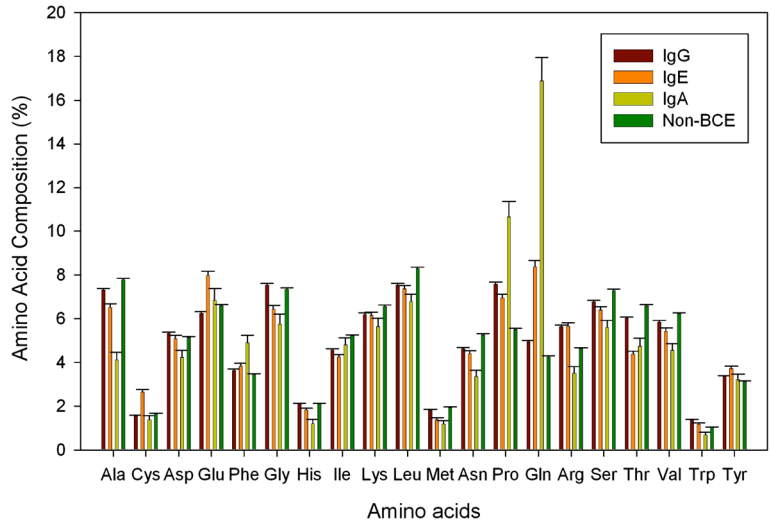


Figure 1 Comparison of average amino acid composition of different class of epitopes.

logos [31] of each class (Figure 2). As shown in Figure 2, certain residues are preferred at specific positions in each class. For example, in IgG inducing epitopes, Pro, Gly and Met are preferred at the first position and Pro and Arg are preferred at second position. Overall, Pro is found to be preferred at almost every position. In IgE inducing epitopes, Gln, Glu and Cys are found to be preferred at various positions. In IgA inducing epitopes, an exclusive preference of Gln and Pro is observed (Figure 2).

Length of epitopes

In order to understand whether the length of epitopes plays any role in inducing specific class of antibodies, we examined the length of different class of epitopes. As shown in Figure 3, most of the epitopes are between 4 and 20 residues in length, only few epitopes having length more than 40 residues. Analysis revealed that more than 55% IgA specific epitopes are less than 10 residues in length while more than 45% IgE specific epitopes are between 11 and 15 residues in length. IgA inducing epitopes have length between 4–10 residues.

Physico-chemical property analysis

We computed and compared the physico-chemical properties of various epitopes to understand their correlation with antibody-class specificity. We did not find any significant differences in physico-chemical properties between

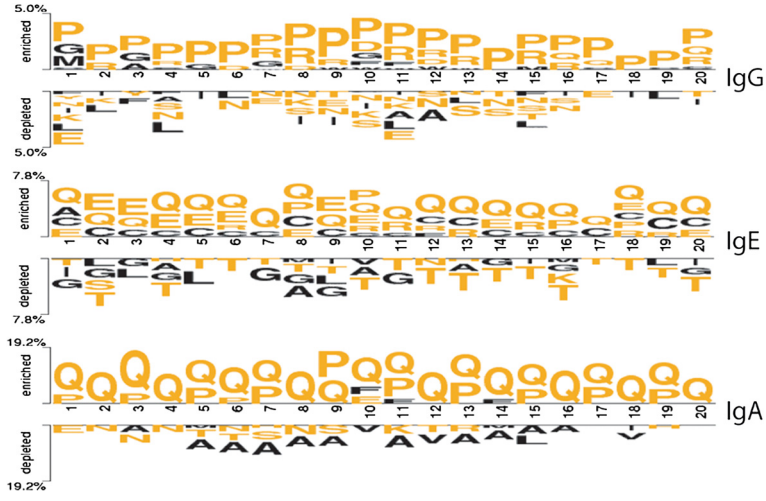


Figure 2 Two sample logos for each class of epitopes where epitopes of a class is taken positive and the rest of peptides as negative examples.

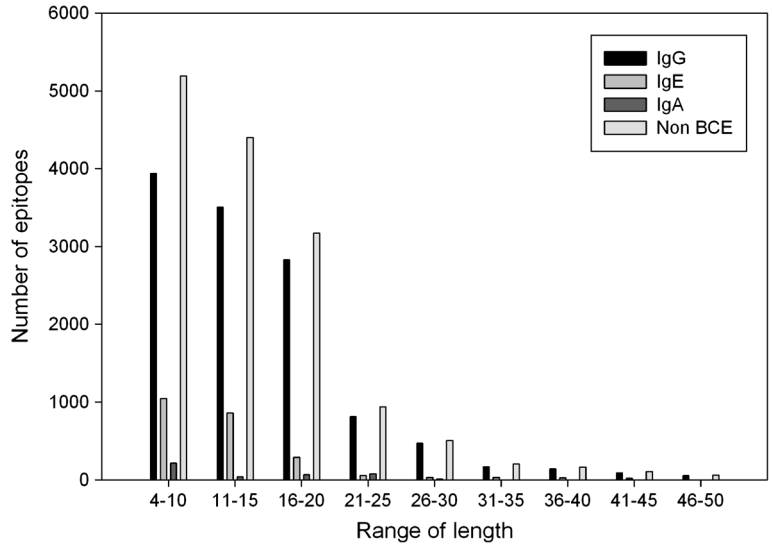


Figure 3 Lengthwise distributions of class-specific epitopes.

the classes of epitopes (Additional file 2: Figure S1), except polar, aliphatic and positively charged residues, which showed differences in their composition both in IgA and IgE classes. In the past, few attempts have been made to differentiate BCEs from non-BCEs based on physico-chemical properties and shown similar observations [32].

Motifs analysis

Since motif-based distinction of IgA epitopes has been reported in the past [11], we extracted motifs from other classes (IgG and IgE) considering the fact that few motifs might be present in these classes, as well. We performed MEME (see Method section) studies for the discovery of motifs in all classes and extracted 20 motifs at default parameters of MEME (Additional file 3: Table S1). We have implemented this motif information at our server where user can scan their epitopes for the presence of specific motifs.

Models for predicting antibody-specific BCEs

In this study, we have developed numerous models for predicting IgG, IgE, and IgA specific BCEs. We built models for each class on all the datasets. The performance of models was optimized, for example, in case of SVM; parameters were tuned for all three types of kernels linear, polynomial and radial bias.

SVM^{light} models developed on BalanceVar dataset

(i) ***Composition-based model.*** Since significant differences were observed in amino acid and dipeptide composition of each class of epitopes, first we have developed SVM models using amino acid composition as input feature and achieved maximum MCC values 0.28, 0.51 and 0.43 for IgG, IgE and IgA respectively. The performances of amino acid composition-based models are summarized in Table 1 and Additional file 3: Table S2. Next, SVM models were developed based on dipeptide composition of epitopes. Dipeptide-based models performed better than the amino acid composition-based model and achieved maximum MCC of 0.41, 0.66 and 0.44 for IgG, IgE and IgA respectively (Table 1 and Additional file 3: Table S2). Detailed performances of dipeptide-based model at different thresholds are summarized in supporting information (Additional file 3: Table S3).

(ii) ***Physico-chemical properties-based model.*** We developed models based on physico-chemical properties (PCP) using 10 physico-chemical properties of BCEs and non-BCEs. These models were further optimized in order to improve the performance. We achieved maximum MCC of 0.32, 0.29 and 0.46 for IgG, IgE and IgA respectively (Table 1 and Additional file 3: Table S2). Further,

Table 1 The performance of SVM models developed for predicting antibody specific BCEs on BalanceVar dataset using various features

Input pattern	IgG epitope			IgE epitope			IgA epitope		
	ACC	MCC	AUC	ACC	MCC	AUC	ACC	MCC	AUC
AAC	63.85	0.28	0.68	75.33	0.51	0.81	71.46	0.43	0.76
AAP	68.30	0.37	0.73	78.3	0.57	0.85	72.93	0.46	0.78
CTD	64.30	0.29	0.69	68.81	0.38	0.71	69.76	0.40	0.74
DPC	**70.42**	**0.41**	**0.76**	**82.7**	**0.66**	**0.88**	**72.07**	**0.44**	**0.78**
PCP	66.18	0.32	0.71	64.31	0.29	0.64	72.8	0.46	0.78

(*ACC* accuracy, *MCC* Matthew's correlation coefficient, *AUC* area under curve).

we developed models using features like composition-transition and distribution (CTD) as input, which has been used in the past for predicting BCEs [33] and achieved MCC of 0.29, 0.38 and 0.40 for IgG, IgE and IgA respectively (Table 1 and Additional file 3: Table S2). The performance of both PCP and CTD-based models on threshold dependent, as well as on threshold independent parameters were poorer than dipeptide-based models. In addition, we developed models using amino acid pairs propensity (AAP) and achieved the performance more or less similar to dipeptide-based model. AAP based model achieved maximum MCC of 0.37, 0.57 and 0.46 for IgG, IgE and IgA respectively (Table 1 and Additional file 3: Table S2).

SVM^{light} models developed on BalanceFix dataset

(i) ***Composition-based model.*** We built models on BalanceFix dataset using amino acid composition as input feature and achieved MCC of 0.33, 0.64 and 0.39 for IgG, IgE and IgA respectively (Table 2 and Additional file 3: Table S4). Similarly, SVM models developed with dipeptide composition profile achieved MCC of 0.44, 0.70 and 0.45 for IgG, IgE and IgA respectively. The AUC values for the models were 0.77, 0.9 and 0.78 for IgG, IgE and IgA respectively. Detailed performances of dipeptide-based model at different thresholds are summarized in supporting information (Additional file 3: Table S5).

(ii) ***Physico-chemical properties-based model.*** Further, SVM models were developed using PCP which achieved MCC of 0.13, 0.16 and 0.27 for IgG, IgE and IgA respectively. The maximum performance of CTD model on this data in term of MCC was 0.15, 0.28 and 0.27 for IgG, IgE and IgA respectively (Table 2 and Additional file 3: Table S4). Furthermore, models on BalanceFix dataset using AAP achieved maximum MCC of 0.39, 0.65 and 0.49 for IgG, IgE and IgA respectively.

(iii) ***Binary profile-based model.*** Since BalanceFix dataset consists of fixed length epitopes, therefore, we developed model using binary profile of epitopes as input features to predict antibody-specific BCEs. The binary based models achieved maximum MCC of 0.08, 0.12, 0.24 for IgG, IgE and IgA respectively (Table 2 and Additional file 3: Table S4).

Table 2 The performance of SVM models developed for predicting antibody specific BCEs on BalanceFix dataset using various features

Input pattern	IgG epitope			IgE epitope			IgA epitope		
	ACC	MCC	AUC	ACC	MCC	AUC	ACC	MCC	AUC
AAC	66.27	0.33	0.70	81.78	0.64	0.86	69.29	0.39	0.75
AAP	69.29	0.39	0.75	82.39	0.65	0.89	74.34	0.49	0.79
CTD	57.41	0.15	0.61	63.99	0.28	0.70	63.3	0.27	0.67
DPC	**71.73**	**0.44**	**0.77**	**84.96**	**0.70**	**0.90**	**72.28**	**0.45**	**0.78**
PCP	56.57	0.13	0.59	58.11	0.16	0.62	63.3	0.27	0.69
BIN	54.02	0.08	0.55	56.17	0.12	0.59	62.17	0.24	0.67

(*ACC* accuracy, *MCC* Matthew's correlation coefficient, *AUC* Area under curve).

In addition, we also developed SVM models for each class on realistic datasets *i.e.* RealVar and RealFix datasets. The dipeptide-based SVM models performed best among the rest of the models. The performance of models developed on realistic datasets is summarized in supporting information (Additional file 3: Table S6 and Additional file 3: TableS7).

Models developed using WEKA

We developed models based on BayesNet, Complement NaiveBayes, NaiveBayes, NaiveBayes Multinomial, SMO, IBk (kNN), J48, and RandomForest using WEKA for predicting antibody-specific BCEs. After tuning different parameters, we found that out of many algorithms of WEKA, three algorithms SMO, kNN and Random Forest performed comparatively better as shown in supporting information (Additional file 3: Table S8 and Additional file 3: TableS9). Balanced set of patterns for both variable (BalanceVar) and fixed (BalanceFix) length were used for all three classes and evaluated using five-fold cross validation technique as shown in the supplementary information (Additional file 3: Table S8 and Additional file 3: Table S9).

The best classifiers of WEKA achieved maximum accuracy 70.07% for IgG [KNN: 0, window size: 0, algorithm: LinearNN search], 81.50% for IgE [SMO, polykernel -C 250007 -O 1.0, c: 1.0, epsilon: 1.0E-12] and 71.16% for IgA [Randomforest, numFeature: 15, numTrees: 10, Seed: 1]. These results were comparable to SVM models implemented using SVM^{light}, where it achieved maximum accuracy of 70.42%, 82.70% and 72.07% for IgG, IgE and IgA respectively.

Performance of model on independent dataset

In order to evaluate the performance of our models on independent data, we first trained our models on 80% of data by ten-fold cross validation and later the performance of the best models were evaluated on remaining 20% independent data. We evaluated dipeptide based model on all the datasets for all classes of epitopes. The performances of these models are summarized in Table 3. On BalanceVar data, model (ten-fold cross validation) developed on training data achieved maximum MCC of 0.42, 0.61 and 0.39 while MCC of 0.37, 0.63 and 0.49

Table 3 The performance of dipeptide composition based SVM models, evaluated using ten-fold cross validation on training data (80%) and independent validation on independent data (20%) on BalanceEval (BalanceFix & BalanceVar) dataset

Dataset		Mode	Data size	ACC	MCC	AUC
BalanceVar	IgG	Training	6063	70.88	0.42	0.76
		Evaluation	1519	68.24	0.37	0.74
	IgE	Training	1873	80.53	0.61	0.87
		Evaluation	468	81.49	0.63	0.88
	IgA	Training	322	69.60	0.39	0.75
		Evaluation	80	74.69	0.49	0.79
BalanceFix	IgG	Training	4893	70.87	0.42	0.76
		Evaluation	1223	71.67	0.43	0.78
	IgE	Training	1524	85.04	0.70	0.90
		Evaluation	381	80.97	0.62	0.86
	IgA	Training	213	73	0.46	0.80
		Evaluation	54	66.67	0.33	0.72

were achieved on evaluation datasets of IgG, IgE and IgA classes respectively. Similarly, for BalanceFix data, model performed reasonably well and achieved maximum maximum MCC of 0.42, 0.70 and 0.46 on training data while MCC of 0.43, 0.62 and 0.33 were achieved on evaluation datasets of IgG, IgE and IgA classes respectively. The performance of every ten fold cross validation model on the evaluation set was comparable to that of five fold cross validation model made on main data. Taken all these results together, it can be speculated that our model performed consistently well, and high accuracy is not due to over optimization.

Implementation and utility of IgPred

We have developed a user-friendly web server 'IgPred' (Figure 4) for predicting antibody-specific BCEs. A number of useful tools have been integrated to IgPred, and their descriptions is as follows:

(i) ***Epitopes in peptides:*** This tool allows users to predict antibody-specific epitopes in their peptide sequences. User can select either variable length, or fixed length peptides option provided at server. In case of variable length peptides, the models were built on variable length data. Here, server allows users to submit multiple peptides in FASTA format for predicting epitopes. In case of fixed length peptides, the models were developed on fixed length data, so on server users may submit multiple peptides of fixed length in FASTA or plain format for predicting epitopes.

(ii) ***Epitopes in proteins:*** This tool allows users to identify antigenic regions (BCEs) in protein sequences, which can induce particular class of Ab. There are two options for users as described above. First option is mapping with variable length, and other is mapping with fixed length window.

(iii) ***Mapping with experimental data:*** This tool allows users to map experimentally verified class-specific BCEs on user's antigen sequence.

(iv) ***MotifScan:*** This module is designed to scan Ab class-specific motifs in an antigen sequence provided by the users. These motifs were derived from experimentally validated BCEs that induce IgG, IgE and IgA Abs using MEME/MAST software as mentioned in method.

(v) ***Similarity search:*** This option allows users to search known (experimentally validated) BCEs in their antigen sequence using Smith-Waterman similarity search algorithm.

In the prediction tools, users can select an SVM threshold for the class of antibody that is going to be generated. We suggest that if high confidence in prediction is needed, user should select high threshold value, but at the same time sensitivity of the prediction will be compromised. The results can be downloaded immediately or can be delivered by email. The common gateway interface (CGI) script for IgPred was written using PERL 5.03. IgPred is freely available at http://crdd.osdd.net/raghava/igpred/.

Discussion

In the past, several methods have been developed for predicting BCEs in an antigen/protein sequence from their primary structure [15,34-37]. To the best of authors' knowledge no comprehensive method has been developed so far for predicting class–specific BCEs, except few methods developed for IgE inducing allergenic epitopes [17,38,39]. The present study is an attempt in the direction to understand differences between the BCEs that induce different classes of antibodies like IgG, IgE and IgA. We hypothesized that induction of different classes of antibodies (*i.e.* IgG, IgE and IgA) could be determined by the sequence of an epitope. Therefore, to understand this, we first extracted BCEs that induce IgG, IgE and IgA types of antibodies from the IEDB database, and then these sequences were systematically analyzed. Amino acid and dipeptide composition analysis revealed that the composition of certain residues/dipeptides is higher in certain antibody-specific epitopes than the others. In addition, few residues are preferred in a particular class suggesting that these residues or dipeptides may play an important role in class switching. For example, Pro and Gln are significantly dominant in

Figure 4 Schematic representation of IgPred webserver.

IgA epitopes. Though Pro is usually not a preferred residue in any of the regular secondary structures, this residue might be contributing in induction of IgA. Based on these observations, it is clear that the composition of sequence (residue or dipeptide) can be used to discriminate epitopes of different class. Therefore, we developed models using amino acid and dipeptide composition as input features. As shown in results section, models based on dipeptide were able to classify the epitope of different classes with reasonably high accuracy. All models were evaluated using both threshold dependent and independent parameters. In addition to SVM models, we also developed models using various modules of WEKA package. It was observed that models based on classifier SMO, kNN and Random Forest perform better than other classifiers of WEKA. It was also observed that overall SVM-based models implemented using SVM^{light} performed better than models developed using WEKA as explained in results section.

We further developed SVM model using binary profiles of patterns as input features. In addition, we also developed models using various other features like PCP, CTD and AAP, but none of the methods achieved accuracy higher than dipeptide-based models. The performance of dipeptide-based model were also evaluated by ten-fold cross validation and performance was almost similar to five-fold. We evaluated the performance of ten-fold cross validation models on independent datasets and achieved reasonable accuracy (as mentioned in the result section). Our results suggest that performance of our models is not due to the over optimization; thus our method will be useful and effective in real life.

In our study, we used one vs. rest approach for creating datasets. It means for developing models for predicting IgA epitopes; we used IgA epitopes as positive set and the rest of the epitopes (IgE, IgG and non-BCE) as negative set. However, we have not used one vs rest approach for prediction, it means our predictions are not exclusive prediction for a single class and our models may predict a peptide inducing for more than one class of antibodies based on prediction score. In a situation where a peptide has equal score for two models then it can be assigned to both the classes if the score for both classes is more than the threshold. The dimensions provided in IgPred webserver enable users to determine the potency of any antigen to induce systemic, allergic or mucosal Ab immune response beforehand.

Conclusion

In the present study, we have made an attempt to establish a relation between an antigenic amino acid sequence and its tendencies to generate systemic (IgG), allergic (IgE) and mucosal (IgA) type of Ab response. For the first time, *in silico* models have been developed for predicting class-specific BCEs. We have implemented our methods in the form of a web server -IgPred. We anticipate that IgPred will be beneficial in designing a better vaccine and immunotherapy, with most appropriate effector function, and several other clinical applications.

Methods

Dataset creation

In this study, datasets were derived from B-cell assays of human and mouse, which were extracted from Immune Epitope Database (IEDB) (http://www.iedb.org/). Of these sequences, B-cell assay positive epitopes were considered as BCEs (positive examples) and B-cell negative epitopes were considered as non-BCEs (negative examples). Since these sequences are of variable length (from 4–100 amino acids), only non- redundant (unique) sequences having a length between 4 to 50 amino acids were taken. These epitopes were divided into three classes on the basis of antibody heavy chain (IgG, IgE and IgA) reported in the IEDB database. Finally, we obtained 11981, 2341, 403 and 22835 sequences for IgG, IgE, IgA specific BCEs and non-BCEs respectively. Non-BCEs are entirely different from the IgG, IgE and IgA specific BCEs and nowhere present in positives. Therefore, we treated it as a separate class. From the above data, we have derived following datasets for developing various models. The approach for developing datasets (BCEs and non-BCEs) and selecting features has been adopted from the previous study by our group [35]. In this study, authors developed a method, which efficiently classify BCEs from non-BCEs. Here, in the present study, we wanted to take it to the next step to classify BCEs into class-specific BCEs. The overview of dataset creation is summarized in Figure 5.

(i) ***Realistic dataset with variable length (RealVar dataset):*** This dataset consists of total 14725 BCEs (*i.e.,* 11981 IgG-specific, 2341 IgE-specific, 403 IgA-specific BCEs) and 22835 non-BCEs of variable lengths (Table 4)

(ii) ***Balanced dataset with variable length (BalanceVar dataset):*** This dataset was created from the above mentioned RealVar dataset. In this dataset, positive examples were balanced by taking equal number of negative examples from total negatives. First, to avoid biases, 4942 negative examples were picked randomly from non-BCEs (22835), which is equivalent to one third of all BCEs (IgG, IgE, and IgA). Subsequently, balanced dataset of each class was developed taking equal number of negative examples randomly from total negatives (Table 4). Since machine learning technique need fixed length pattern and also it is difficult to predict epitopes in an antigen using model trained on variable length datasets as length is not fixed for scanning, we have developed two other datasets of fixed length. These datasets are derived from the above two datasets.

(iii) ***Realistic dataset with fixed length (RealFix dataset):*** In order to utilize the full potential of machine learning techniques, we fixed the length to 20 residues as most epitopes have length up to 20 residues. This has been implemented using extension truncation technique as described previously [40,41]. After removing redundant epitopes, we got 9660, 1905, 267 and 20589 for IgG, IgE, IgA and non-BCEs respectively (Table 4).

(iv) ***Balanced dataset with fixed length epitopes (BalanceFix dataset):*** This dataset was generated from RealFix dataset using the same process as described above for creating BalanceVar dataset. It consists of 9660, 1905, and 267 BCEs for IgG, IgE, and IgA respectively and an equal number of negative examples for the respective class as summarized in Table 4.

(v) ***Independent dataset:*** In order to create an independent dataset, 20% sequences were randomly picked from BalanceFix (for fixed length) and BalanceVar (for variable length) dataset and used as independent dataset. Model was trained on rest 80% sequences using ten-fold cross validation and then performance of the optimized model was validated on independent dataset.

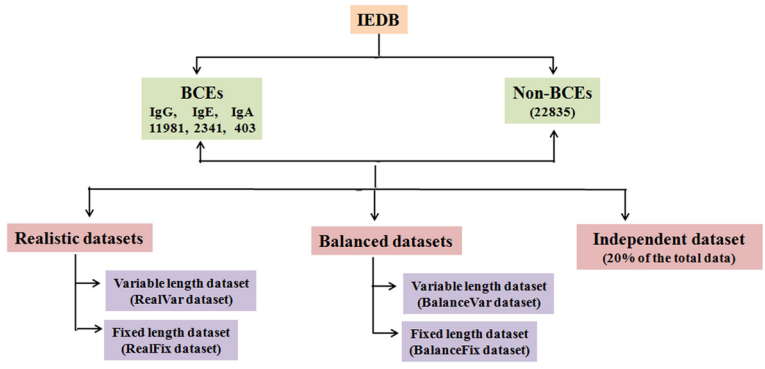


Figure 5 Overview of dataset creation.

Table 4 Various datasets used for developing prediction models in the present study

Dataset type	IgG		IgE		IgA	
	Positive	Negative	Positive	Negative	Positive	Negative
RealVar	11981	25579	2341	35219	403	37157
BalanceVar	7598	7598	2341	2341	403	403
RealFix	9660	22761	1905	30516	267	32154
BalanceFix	6116	6116	1905	1905	267	267

One-vs-rest approach for datasets

We developed models for predicting epitopes inducing different Abs. In order to develop model for epitopes inducing a specific class of Ab (*e.g.* IgA inducing epitopes), we used IgA inducing epitopes as positive examples, and the rest of the epitopes belong to other classes (*e.g.,* IgE, IgG and non-BCE) as negative examples (Additional file 2: Figure S2). Similarly, for developing model for IgG inducing BCEs, IgG-specific BCEs were considered as positive examples and the rest of the epitopes from other classes (IgE specific, IgA specific BCEs and non-BCEs) were considered as negative examples (Additional file 2: Figure S2).

Amino acid composition

Amino acid composition is one of the simplest features, which have been used in the past to encapsulate the global information of protein into a fixed length pattern [42,43]. Amino acid composition of a peptide is proportional to frequency of each type of residue in a peptide. As there are 20 types of residues, so there are 20 types of composition for each peptide. Thus, a vector of 20 is used to represent an epitope as described in previous studies [44].

Dipeptide composition

Dipeptide composition is another important feature, which is used to transform the variable length of peptides to fixed length feature vectors. Dipeptide compositions have been used in earlier studies to classify various classes of peptides [44,45]. Dipeptide composition provides information of the pair of residues in contrast to a single residue and provides a fixed pattern length of 400. We computed percent of occurrence of each type of dipeptide and used this information for developing machine learning models. Following formula has been used for computing dipeptide composition of each epitope

$$Percentage\ of\ Dip(i) = \frac{Total\ number\ of\ Dip(i)}{Total\ number\ of\ all\ possible\ Dips} \times 100 \quad (1)$$

where Dip*(i)* is a dipeptide *i* out of 400 dipeptides.

To analyse the differences between dipeptide composition among all classes of epitopes (IgG-, IgE-, IgA-inducing and Non-BCE), we performed significance test (Welch's *t*-test). For example, we first calculated average of dipeptide compositions (400 dipeptides) for both IgG(+) data and IgG(−) data. Next, we calculated the difference of means in both the datasets for each of the 400 dipeptides for IgG-, IgE-, IgA-inducing BCEs and non-BCEs respectively. We also computed whether the difference in the mean of dipeptide composition is significant or not using *t*-test. Several dipeptides show a significant difference in composition between positive and negative dataset of each class (Additional file 1).

Binary profile of patterns

We used binary profiles of patterns for fixed length datasets where each amino acid was presented by a vector of dimension 20 as described previously [44]. Since the length of epitopes was 20, a pattern of window length 20 was represented by a vector of dimension (20 × 20). In the past, binary profile has been used for developing prediction models [44-47].

Physico-chemical properties

It is well known that function of a peptide is governed by its primary sequence and their physico-chemical properties. Therefore, in the present study, we have selected ten commonly used physico-chemical properties. These properties include hydrophobicity, bulky side chain, net-hydrogen, stearic hindrance, amphipathicity, hydrophilicity, charge, pI value, *etc.* [48,49].

MEME/MAST

MEME/MAST module has been used previously for discovering novel motifs in various classes of proteins and peptides [44]. We have adopted the similar strategy in the present study for discovering various motifs in epitopes using MEME/MAST module. This module comprises of two programs: one is MEME, which is used to discover motifs, and other is MAST, used for searching motifs [50,51]. In the present study, we have used MEME for discovering motifs in IgG, IgE and IgA inducing epitopes and subsequently, these discovered motifs were used further for identifying epitopes using MAST.

Composition-transition-distribution (CTD)

CTD has been used in several machine learning algorithms [33] to get a fixed length vector from variable length data. Each peptide sequence was mapped into a string defined by three symbols. These symbols were resulted from grouping of all amino acids into three groups, on the basis of certain physico-chemical properties. For every physico-chemical property, we got a string of 1, 2 and 3 symbols; three feature given by composition, three feature given by the percent frequency of i followed by j or j followed by i (transition) and five features per symbol. Thus total 15 features representing the fractions of the entire sequence where the first, 25, 50, 75, and 100% of the candidate symbol are contained in string (distribution).

Amino acid pairs propensity scale

It has been reported that some amino acid pairs found more frequently in BCEs than in non-BCEs. Keeping this in mind, we developed an AAP propensity scale as reported previously [33,34]. The frequencies were calculated from positive and negative data sequences [52]. AAP features can be viewed as dipeptide composition features weighted by the amino acid propensity of each dipeptide. The final vector size in AAP is 400 .

Software for extraction of features and implementing machine-learning techniques

The calculations for different features have been carried out using in-house PERL scripts and R package (2.10.1) scripts. Plots were made using SigmaPlot 10.0. We used SVM_Light software (http://svmlight.joachims.org/) for developing SVM based models. SVM is freely available for academic use and has been used in number of research papers [29,53,54]. We also used WEKA 3.2 package (http://www.cs.waikato.ac.nz/ml/WEKA/) to develop various types of models. We used nine algorithms of WEKA package namely BayesNet, Complement NaiveBayes, NaiveBayes, NaiveBayes Multinomial, SMO, IBk, J48, and RandomForest [55].

Evaluating the performance of models

In the present study, we used five-fold cross validation technique to evaluate the performance of our models developed for predicting antibody-specific epitopes. In this technique, one fifth of total data is used for testing and remaining data is used for training the model and this process is repeated till all instances are evaluated. Similarly, we also run ten-fold cross validation, which was preceded by evaluation on independent dataset to validate the performance of the model. In addition, in order to evaluate the performance of models, we included both threshold dependent and threshold independent parameters. In case of threshold dependent parameters, we used standard parameters like sensitivity (Sen), specificity (Spe), Overall accuracy (Acc) and Matthews's correlation coefficient (MCC) using following equations.

$$Sens = \frac{TP}{TP + FN} \times 100 \quad (2)$$

$$Spec = \frac{TN}{TN + FP} \times 100 \quad (3)$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (4)$$

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{[TP + FP][TP + FN][TN + FP][TN + FN]}} \quad (5)$$

[TP = true positive; FN = false negative; TN = true negative; FP = false positive; Sens = Sensitivity; Spec = Specificity; Acc = Accuracy]

To evaluate the performance of models using threshold independent parameters, AUC (Area under curve curve) have been calculated with the help of PERF software.

Response and revision of the manuscript in light of the reviewer comments:

Reviewers' comments on the original manuscript

We are grateful to the reviewers' for their useful thoughts and suggestions. We have incorporated all the suggestions of the reviewers in our manuscript. Here, we are addressing all comments of reviewer's point-by-point.

Reviewer number 1: Dr. M Michael Gromiha

In this work, authors have developed a method for predicting different types of B-cell epitopes. They have utilized several features such as amino acid and dipeptide compositions, physicochemical properties and binary profiles. The method showed a correlation coefficient in the range of 0.44 to 0.70 to various types of epitopes. Further, a web server has been developed for application purposes.

Reviewer comments: The composition analysis shows the preference of Pro in IgA. Pro is usually not a preferred residue in any of the regular secondary structures. The higher occurrence of Pro may be discussed. Further, the abundance of Cys in IgE may be commented.

Authors' response: It has been shown in previous studies [56,57] that most of residues in BCEs fall in non-regular (coil or tight-turns) secondary structure. In revised paper, we have discussed the preference of Pro and Cys in IgA- and IgE-inducing epitopes respectively.

Reviewer comments: The residue pair preference showed the dominance of Pro with other residues only in IgG and IgA and not in IgE. This may be discussed.

Authors' response: This is the first study, where types of residues preferred in the different classes of BCEs have been calculated. We have no idea (biological significance of observation) why Pro with other residues (pair) is more abundant in IgG- and IgA- inducing epitopes and not in IgE-inducing epitopes. This is an interesting point to be studied in the future for understanding the above observation.

Reviewer comments: It has been shown that IgE can be predicted with higher accuracy than other epitopes. The reason may be explained.

Authors' response: It has been observed that IgE-inducing epitopes are more conserved in comparison to other classes of epitopes, which could be responsible for higher accuracy.

Reviewer comments: The expansions for the parameters used in SMO may be given.

Authors' response: As suggested by the reviewer, in revised version of the manuscript, we have described SMO parameters in detail.

Reviewer comments: It is necessary to give the procedure used to remove the redundancy.

Authors' response: In revised manuscript, the procedure to remove the redundant or duplicate peptides has been described.

Reviewer comments: Quality of written English: Acceptable

Reviewer number 2: Dr Christopher Langmead (nominated by Dr Robert Murphy)

The manuscript presents an SVM-based method for predicting antibody-specific epitopes. Three classes were considered: IgA, IgE, IgG. Features included AA composition, dipeptide composition, and physio-chemical properties. Training data were obtained from the IEDB, and machine-learning methods were performed using either SVM^{Light} or WEKA. A website for performing a variety of tasks associated with epitope prediction is also reported.

The study has some flaws that need to be addressed.

Reviewer comments: The results obtained with SVM^{Light} are the result of tuning parameters, whereas the results obtained for the strawman models are the result of using WEKA's default parameters. To be fair, the authors must do parameter searches for these.

Authors' response: We agree with the reviewer, in our revised manuscript, we have reported the performances of strawman models after parameter optimization (see Additional file 3: Table S1).

Reviewer comments: Did the authors try a string kernel for the SVM^{Light} experiments? It would seem appropriate, given the nature of the data.

Authors' Response: In the present study, we have tried only three kernels of SVM^{Light} namely Linear, Polynomial and Radial basis. It is possible that string kernel may perform better than above kernels, but unfortunately, our group does not have expertise in optimization of string kernel parameters. In addition, previous studies have indicated that radial basis is an efficient kernel for discriminating various types of peptide/epitopes.

Reviewer comments: The authors need to explain what they do in the event of a tie in their one-vs-rest approach to multi-class classification.

Authors' response: In this study, we used one verses rest approach for creating datasets only, and not for predicting epitopes, for example, to develop prediction model for IgA-inducing epitopes, we created a dataset containing IgA-inducing epitopes as positive examples and remaining epitopes (IgE-, IgG-inducing epitopes and non-BCEs) as negative examples. We computed performance of models based on the threshold, for example, in IgA model if a peptide having SVM score above the threshold then it is assigned as IgA-inducing epitope. We have not used exclusive prediction for a single class and our predictors may predict a single peptide inducing more than one class of antibodies. Thus, it does not matter if a peptide has equal score for two models, and it may be predicted in two classes if the SVM score is more than the threshold. In our revised manuscript, we have clarified this point.

Reviewer comments: Since the matrics for their method (MCC) are apparently worse than published methods for (antibody-neutral) epitope prediction, the authors should evaluate a two-stage classification process whereby an antibody-neutral classification is performed, and then the positive results are passed to their method. This would simplify the learning task because their method would not have to learn to distinguish non-BCEs from BCEs.

Authors' response: The aim of this study is to predict antibody specific BCEs instead of BCEs. For the first time, we have developed models for predicting antibody class specific-BCEs that may induce three types of antibodies (IgA, IgE and IgG). Thus, it is not possible to compare this method with the previous methods as earlier methods have been developed for predicting BCEs only.

Reviewer comments: The manuscript also has some flaws that need to be addressed. Primarily, they should cite and discuss other SVM-based methods for epitope prediction. Additionally, it is not clear whether the authors understand that SMO is, in fact, an algorithm for learning SVMs. It seems strange to simply list SMO among the non-SVM algorithms.

Typos: There are a number of typos that can be identified by using a spell-checker. The authors meant to say that there are five primary isotypes at the end of the first

paragraph of the Background section, not six (note: they correctly list the five classes).

Authors' response: As suggested by reviewer, we have cited other SVM-based methods published earlier for epitope prediction. We have also edited the manuscript as per reviewer's suggestion. Since there are different versions of Support Vector Machines (like SVM^{light} and SMO), we tried one from SVM^{light} and other from WEKA package (SMO). As the reviewer has advised, we have placed SMO in SVM algorithms from non-SVM algorithms.

Reviewer comments: Quality of written English: Needs some language corrections before being published

Authors' Response: We have tried our best to improve the quality of english in revised version of the manuscript.

Reviewer number: 3

Report form:

OK

Reviewer comments: Quality of written English: Needs some language corrections before being published

Referee 3: Dr Lina Ma (nominated by Dr Zhang Zhang)

This manuscript presented a method to predict B-cell epitopes that can induce a specific class of antibody and attempted to understand the relation between primary sequence of epitopes and the class of antibodies. My comments are listed as follows.

Reviewer comments: It is noticed that one paper published by the authors, entitled "Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence", describes the method of B-cell epitope prediction". Is the method presented here similar with that in the published one? As the authors used non-BCEs as a negative control in both papers, I wonder what is the correlation between this manuscript and the published one and it might be better to describe it clearly or discuss any issue caused.

Authors' response: As indicated by reviewer, recently our group has published a paper [35] describing a method developed for predicting linear B-cell epitopes (in revised version of this paper, we have cited and discussed our recent paper). In the past, other methods have also been developed for predicting B-cell epitopes (including our recent paper [35]). In the present paper, for the first time, we have developed a method for predicting epitopes that can induce specific class of antibodies. In the revised manuscript, we have clarified the difference between IgPred and previous methods.

Reviewer comments: It is highly recommended that the manuscript describe their results in a more clarified and detailed manner.

a) In Figure 1, it is obvious that Pro and Gln are abundant in IgA inducing epitopes, but I do not think that Cys and Glu are dominated in IgE inducing epitopes. In Figure 2, IgA inducing epitopes contain more LP, LQ, PF, PQ, PY, QP, QL and QQ dipeptides while the IgG inducing epitopes and IgE epitopes do not tend to show a significant difference in any dipeptide compositions among the three kinds of epitopes.
b) Some statistical methods should be used to compare the difference between epitopes, and it is better to list results with significant differences.
c) Error bars should be added in the histogram of Figures to show deviations.

Authors' response: We agree with the reviewer in the points (a, b and c) and after getting valuable comments from the reviewer, we performed Welch's *t*-test for each class of epitopes (IgG, IgE, IgA, and non-BCE). As an example, we first calculated average of dipeptide compositions (400 dipeptides) for both IgG(+) data and IgG(−) data. Further with the help of *t*-test, we calculated the difference of means of both the datasets for each of 400 dipeptides. Looking at the large number of dipeptides (400), we removed Figure 2 (showing dipeptide composition as bar graph) and provided the dipeptide composition of all the three classes with their negatives along with p-value in separate Additional file 1. With such analysis, it can be observed (in the Supplementary excel sheet) that significant dipeptides such as AS, GP, WK, YR, *etc*. are found to be dominant in IgG-inducing epitopes; IQ LA, NA, NE, *etc*. are frequent in IgE-inducing epitopes; and ED, FP, PF, PQ, PY, QP, *etc*. are prevailing in IgA-inducing epitopes. We have also discussed these observations in the revised manuscript.

As per reviewer's suggestion, we performed Welch-s *t*-test on dipeptide composition data to look at the significant difference between positive and negative data of each class.

We have also added error bars to the amino acid composition bar graph as per reviewer's recommendation.

d) In Figure 1, non-BCEs were used as a negative control. It is better that negative control is also used in Figure 2. Similar problems also exist in sections of "Residue preference", "Length of epitopes", "Physico-chemical property analysis", "Motifs analysis". It is better that negative control (non-BCEs) should be used consistently with Figure 1.

Authors' response: As suggested by the reviewer, we have added AAC of non-BCE in Figure 1. We replaced Figure 2 with excel sheet showing DPC of 400 dipeptides (Additional file 1).

e) The results of ACC, AUC, and MCC are listed in tables, what about sensitivity and specificity

Authors' response: Since there were about six features for each of the three classes of epitopes, we did not include sensitivity and specificity in the result tables. After getting suggestion of the reviewer, we created additional tables (Table S2 and S4) containing sensitivity and specificity and incorporated as supplementary information (Additional file 3: Table S2 and Additional file 3: Table S4).

f) What does AAP mean? What is the difference between AAP and DPC?

Authors' response: AAP (Amino acid propensity) is a feature, which is derived from DPC of the datasets. This feature has been exploited in previous B-cell epitope prediction algorithms [33,40,41,43]. This represents the DPC of a protein/amino acid sequence weighted by the amino acid propensity of each dipeptide in a matrix made by both positive and negative data. We have edited the manuscript by describing it in detail.

Reviewer comments: Discussion Section- At the end of the first paragraph, it is concluded that SVM performed better than WEKA. This is an important conclusion of this paper, which should be explained in detail.

Authors' response: As we can observe in the supplementary tables S8 and S9, algorithms of WEKA (IBk, Random Forest and SMO) could perform reasonably well with DPC as input feature. At the same time, using SVM^{light} with DPC as input feature, the performances of different models were significantly better than those of WEKA as a whole. As suggested by the reviewer, we have discussed the performances of SVM and WEKA modules in detail in the discussion section.

Reviewer comments: At the end of paragraph 2, please explain "reasonable accuracy" in detail?

Authors' response: As suggested by the reviewer, we have explained the accuracies of our models on independent dataset in detail. For BalanceVar data, model (ten-fold cross validation) developed on training data achieved maximum MCC of 0.42, 0.61 and 0.39 while MCC of 0.37, 0.63 and 0.49 were achieved on evaluation datasets of IgG, IgE and IgA classes respectively. Similarly, for BalanceFix data, model performed well and achieved maximum MCC of 0.42, 0.70 and 0.46 on training data while MCC of 0.43, 0.62 and 0.33 were achieved on evaluation datasets of IgG, IgE and IgA classes respectively. We have also discussed this issue in the revised manuscript as per reviewer's suggestion.

Reviewer comments: IgA inducing epitopes are quite different from IgE or IgG inducing epitopes in AAC and AAP comparison. However, SVM models for predicting IgA inducing epitopes do not seem to perform better than that for predicting IgE and IgG inducing epitopes. This is really confusing for me, or it might be better to provide explanations for this result.

Authors' response: We agree with the reviewer that IgA epitopes are quite different. We examined the performance of prediction of IgA inducing epitopes, as well as the reason for the poor performance of these models. It could be due to the fact that IgA inducing epitopes are very limited in comparison to non-IgA inducing epitopes. It is a well-known fact that machine learning techniques, particularly SVM-based models perform poor especially when positive and negative dataset is unbalanced. We have mentioned these points in the revised manuscript.

Reviewer comments: As there have been methods for predicting IgE inducing epitopes, what is the difference between previous methods and the models described in the manuscript?

Authors' response: We agree with the reviewer that there are methods developed for prediction of IgE inducing epitopes (one is from our own group [17]). The current study can be considered as an extension of the previous studies. This study offers researcher to compare the potential of an epitope to induce systemic type (IgG), mucosal type (IgA) and inflammatory type (IgE) of antibody immune response. Thus in contrast to the predictors of IgE inducing epitopes, our web server is more comprehensive with reference to antibody immune response.

Reviewers' comments on the revised manuscript

Referee 3: Dr Lina Ma (nominated by Dr Zhang Zhang)

Reviewer comments: The authors have answered all the questions seriously, while there are still some mistakes in the revised version. It is better that the authors check the manuscript carefully before submission.

Section "Composition analysis", paragraph 1, line 6, the word "Glu" should be Gln. Section "Discussion", paragraph 1, lines 19–20, this sentence seems incomplete.

The authors say that they have mentioned these points in the revised manuscript in answering my third question of Section "Discussion", while I did not find where they have mentioned.

Authors' response: We are thankful to the reviewer for appreciating our efforts. As suggested by the reviewer, we have edited the manuscript and incorporated all the suggestions. The performance of WEKA has been discussed in result and discussion section after getting the comments of reviewer.

Additional files

Additional file 1: Comparison of dipeptide composition of various types of BCEs and non-BCEs.

Additional file 2: Figure S1. Comparison of physico-chemical properties of various types of BCEs and non-B-cell-epitopes. **Figure S2.** One vs. rest approach used in developing datasets.

Additional file 3: Table S1. List of top 20 motifs discovered in three classes of epitopes. **Table S2.** The performance of SVM models developed for predicting antibody specific BCEs on BalanceVar dataset. **Table S3.** The performance of dipeptide-based model on BalanceVar dataset, evaluated using five-fold cross validation technique. **Table S4.** The performance of SVM models developed for predicting antibody specific BCEs on BalanceFix dataset. **Table S5.** The performance of dipeptide-based model evaluated using five-fold cross validation, performance was evaluated on BalanceFix dataset. **Table S6.** The performance of SVM-based models developed using various features for predicting antibody specific B-cell epitopes on RealFix dataset. **Table S7.** The performance of SVM based models developed using various features for predicting antibody specific B-cell epitopes on RealVar dataset. **Table S8.** Performance of WEKA classifiers developed using various input features for different classes of epitopes on BalanceFix dataset. **Table S9.** The performance of WEKA classifiers developed using various input features for different classes of epitopes on BalanceVar dataset.

Abbreviations

AAC: Amino acid composition; DPC: Dipeptide composition; MCC: Matthews's correlation coefficient; CTD: Composition-transition-distribution; AAP: Amino acid pairs propensity; AUC: Area under the curve; SVM: Support vector machine.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SG and HRA created datasets and developed models. SG created the backend web server and AG contributed in the development of the front end user interface. AG and SG analyzed the data and wrote the manuscript. OSDD consortium provides discussions and feedback on this problem. GPSR conceived the project, coordinated it and refined the final manuscript drafted by AG and SG. All authors have read and approved the final manuscript.

Acknowledgements

Authors are thankful to funding agencies Council of Scientific and Industrial Research (project Open Source Drug Discovery and GENESIS BSC0121) and Department of Biotechnology (project BTISNET), Govt. of India. This report has Institute of Microbial Technology (IMTECH) communication no. 071/2012.

Author details

[1]Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh 160036, India. [2]Open Source Drug Discovery Unit, Council of Scientific and Industrial Research (CSIR) Anusandhan Bhawan, 2 Rafi Marg, New Delhi 110001, India.

Received: 4 June 2013 Accepted: 25 October 2013
Published: 30 October 2013

References

1. Almagro JC: **Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires.** *J Mol Recognit* 2004, **17**(2):132–143.
2. Webster SD, Galvan MD, Ferran E, Garzon-Rodriguez W, Glabe CG, Tenner AJ: **Antibody-mediated phagocytosis of the amyloid beta-peptide in microglia is differentially modulated by C1q.** *J Immunol* 2001, **166**(12):7496–7503.
3. Roos A, Bouwman LH, Van Gijlswijk-Janssen DJ, Faber-Krol MC, Stahl GL, Daha MR: **Human IgA activates the complement system via the mannan-binding lectin pathway.** *J Immunol* 2001, **167**(5):2861–2868.
4. Mitchell TJ, Andrew PW, Saunders FK, Smith AN, Boulnois GJ: **Complement activation and antibody binding by pneumolysin via a region of the toxin homologous to a human acute-phase protein.** *Mol Microbiol* 1991, **5**(8):1883–1888.
5. Galli SJ, Tsai M: **IgE and mast cells in allergic disease.** *Nat Med* 2012, **18**(5):693–704.
6. Lobos E, Nutman TB, Hothersall JS, Moncada S: **Elevated immunoglobulin E against recombinant Brugia malayi gamma-glutamyl transpeptidase in patients with bancroftian filariasis: association with tropical pulmonary eosinophilia or putative immunity.** *Infect Immun* 2003, **71**(2):747–753.
7. Hagan P, Blumenthal UJ, Dunn D, Simpson AJ, Wilkins HA: **Human IgE, IgG4 and resistance to reinfection with Schistosoma haematobium.** *Nature* 1991, **349**(6306):243–245.
8. Ottesen EA, Skvaril F, Tripathy SP, Poindexter RW, Hussain R: **Prominence of IgG4 in the IgG antibody response to human filariasis.** *J Immunol* 1985, **134**(4):2707–2712.
9. Flueck C, Frank G, Smith T, Jafarshad A, Nebie I, Sirima SB, Olugbile S, Alonso P, Tanner M, Druilhe P, *et al*: **Evaluation of two long synthetic merozoite surface protein 2 peptides as malaria vaccine candidates.** *Vaccine* 2009, **27**(20):2653–2661.
10. Garraud O, Mahanty S, Perraut R: **Malaria-specific antibody subclasses in immune individuals: a key source of information for vaccine design.** *Trends Immunol* 2003, **24**(1):30–35.
11. Jain S, Rosenthal KL: **The gp41 epitope, QARVLAVERY, is highly conserved and a potent inducer of IgA that neutralizes HIV-1 and inhibits viral transcytosis.** *Mucosal Immunol* 2011, **4**(5):539–553.
12. Traidl-Hoffmann C, Jakob T, Behrendt H: **Determinants of allergenicity.** *J Allergy Clin Immunol* 2009, **123**(3):558–566.
13. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, Bui HH, Buus S, Frankild S, Greenbaum J, *et al*: **Immune epitope database analysis resource (IEDB-AR).** *Nucleic Acids Res* 2008, **36**(Web Server issue): W513–W518.
14. Saha S, Bhasin M, Raghava GP: **Bcipep: a database of B-cell epitopes.** *BMC Genomics* 2005, **6**:79.
15. Saha S, Raghava GP: **Prediction of continuous B-cell epitopes in an antigen using recurrent neural network.** *Proteins* 2006, **65**(1):40–48.
16. Saha S, Raghava GP: **BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties.** *ICARIS, LNCS* 2004, **3239**:197–204.
17. Saha S, Raghava GP: **AlgPred: prediction of allergenic proteins and mapping of IgE epitopes.** *Nucleic Acids Res* 2006, **34**(Web Server issue): W202–W209.
18. Dall'Antonia F, Gieras A, Devanaboyina SC, Valenta R, Keller W: **Prediction of IgE-binding epitopes by means of allergen surface comparison and correlation to cross-reactivity.** *J Allergy Clin Immunol* 2011, **128**(4):872–879. e878.
19. Ramachandran H, Banerjee B, Greenberger PA, Kelly KJ, Fink JN, Kurup VP: **Role of C-terminal cysteine residues of Aspergillus fumigatus allergen Asp f 4 in immunoglobulin E binding.** *Clin Diagn Lab Immunol* 2004, **11**(2):261–265.
20. Banerjee B, Kurup VP, Greenberger PA, Kelly KJ, Fink JN: **C-terminal cysteine residues determine the IgE binding of Aspergillus fumigatus allergen Asp f 2.** *J Immunol* 2002, **169**(9):5137–5144.
21. Nayak SK, Rathore D, Batra JK: **Role of individual cysteine residues and disulfide bonds in the structure and function of Aspergillus ribonucleolytic toxin restrictocin.** *Biochemistry* 1999, **38**(31):10052–10058.
22. Vrtala S, Hirtenlehner K, Vangelista L, Pastore A, Eichler HG, Sperr WR, Valent P, Ebner C, Kraft D, Valenta R: **Division of the major birch pollen allergen, Bet v 1, into two non-anaphylactic fragments.** *Int Arch Allergy Immunol* 1997, **113**(1–3):246–248.
23. Smith AM, Chapman MD: **Reduction in IgE binding to allergen variants generated by site-directed mutagenesis: contribution of disulfide bonds to the antigenic structure of the major house dust mite allergen Der p 2.** *Mol Immunol* 1996, **33**(4–5):399–405.
24. Vailes LD, Li Y, Bao Y, DeGroot H, Aalberse RC, Chapman MD: **Fine specificity of B-cell epitopes on felis domesticus allergen I (Fel d I): effect of reduction and alkylation or deglycosylation on Fel d I structure and antibody binding.** *J Allergy Clin Immunol* 1994, **93**(1 Pt 1):22–33.

25. Lombardero M, Heymann PW, Platts-Mills TA, Fox JW, Chapman MD: **Conformational stability of B cell epitopes on group I and group II dermatophagoides spp. Allergens. Effect of thermal and chemical denaturation on the binding of murine IgG and human IgE antibodies.** *J Immunol* 1990, **144**(4):1353–1360.
26. Heden LO, Frithz E, Lindahl G: **Molecular characterization of an IgA receptor from group B streptococci: sequence of the gene, identification of a proline-rich region with unique structure and isolation of N-terminal fragments with IgA-binding capacity.** *Eur J Immunol* 1991, **21**(6):1481–1490.
27. Jerlstrom PG, Chhatwal GS, Timmis KN: **The IgA-binding beta antigen of the c protein complex of Group B streptococci: sequence determination of its gene and detection of two binding regions.** *Mol Microbiol* 1991, **5**(4):843–849.
28. Bhasin M, Raghava GP: **ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W414–W419.
29. Bhasin M, Raghava GP: **Prediction of CTL epitopes using QM, SVM and ANN techniques.** *Vaccine* 2004, **22**(23–24):3195–3204.
30. Bhasin M, Garg A, Raghava GP: **PSLpred: prediction of subcellular localization of bacterial proteins.** *Bioinformatics* 2005, **21**(10):2522–2524.
31. Vacic V, Iakoucheva LM, Radivojac P: **Two sample logo: a graphical representation of the differences between two sets of sequence alignments.** *Bioinformatics* 2006, **22**(12):1536–1537.
32. Sun J, Xu T, Wang S, Li G, Wu D, Cao Z: **Does difference exist between epitope and non-epitope residues? analysis of the physicochemical and structural properties on conformational epitopes from B-cell protein antigens.** *Immunome Res* 2011, **7**(3):1–11.
33. El-Manzalawy Y, Dobbs D, Honavar V: **Predicting flexible length linear B-cell epitopes.** *Comput Syst Bioinformatics Conf* 2008, **7**:121–132.
34. Chen J, Liu H, Yang J, Chou KC: **Prediction of linear B-cell epitopes using amino acid pair antigenicity scale.** *Amino Acids* 2007, **33**(3):423–428.
35. Singh H, Ansari HR, Raghava GP: **Improved method for linear B-cell epitope prediction using Antigen's primary sequence.** *PLoS One* 2013, **8**(5):e62216.
36. Larsen JE, Lund O, Nielsen M: **Improved method for predicting linear B-cell epitopes.** *Immunome Res* 2006, **2**:2.
37. Sollner J, Mayer B: **Machine learning approaches for prediction of linear B-cell epitopes on proteins.** *J Mol Recognit* 2006, **19**(3):200–208.
38. Dimitrov I, Flower DR, Doytchinova I: **AllerTOP–a server for in silico prediction of allergens.** *BMC Bioinformatics* 2013, **14**(Suppl 6):S4.
39. 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**(6):e5861.
40. Wang HW, Lin YC, Pai TW, Chang HT: **Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification.** *J Biomed Biotechnol* 2011, **2011**:432830.
41. El-Manzalawy Y, Dobbs D, Honavar V: **Predicting linear B-cell epitopes using string kernels.** *J Mol Recognit* 2008, **21**(4):243–255.
42. Eisenhaber F, Frommel C, Argos P: **Prediction of secondary structural content of proteins from their amino acid composition alone, II: the paradox with secondary structural class.** *Proteins* 1996, **25**(2):169–179.
43. Park KJ, Kanehisa M: **Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs.** *Bioinformatics* 2003, **19**(13):1656–1663.
44. Gautam A, Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A, Raghava GP: **In silico approaches for designing highly effective cell penetrating peptides.** *J Transl Med* 2013, **11**:74.
45. Sharma A, Kapoor P, Gautam A, Chaudhary K, Kumar R, Chauhan JS, Tyagi A, Raghava GP: **Computational approach for designing tumor homing peptides.** *Sci Rep* 2013, **3**:1607.
46. Xiao X, Shao S, Ding Y, Huang Z, Chou KC: **Using cellular automata images and pseudo amino acid composition to predict protein subcellular location.** *Amino Acids* 2006, **30**(1):49–54.
47. Ansari HR, Raghava GP: **Identification of conformational B-cell epitopes in an antigen from its primary sequence.** *Immunome Res* 2010, **6**:6.
48. Eisenberg D: **Three-dimensional structure of membrane and surface proteins.** *Annu Rev Biochem* 1984, **53**:595–623.
49. Hopp TP, Woods KR: **Prediction of protein antigenic determinants from amino acid sequences.** *Proc Natl Acad Sci USA* 1981, **78**(6):3824–3828.
50. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching.** *Nucleic Acids Res* 2009, **37**(Web Server issue):W202–W208.
51. Rashid M, Saha S, Raghava GP: **Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs.** *BMC Bioinforma* 2007, **8**:337.
52. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31**(1):365–370.
53. Wee LJ, Simarmata D, Kam YW, Ng LF, Tong JC: **SVM-based prediction of linear B-cell epitopes using bayes feature extraction.** *BMC Genomics* 2010, **11**(Suppl 4):S21.
54. Bhasin M, Raghava GP: **Analysis and prediction of affinity of TAP binding peptides using cascade SVM.** *Protein Sci* 2004, **13**(3):596–607.
55. Witten R, I.h, Frank E: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco: Morgan Kaufmann; 1999.
56. Brusic V, Petrovsky N, Gendel SM, Millot M, Gigonzac O, Stelman SJ: **Computational tools for the study of allergens.** *Allergy* 2003, **58**(11):1083–1092.
57. Cooper JA, Hayman W, Reed C, Kagawa H, Good MF, Saul A: **Mapping of conformational B cell epitopes within alpha-helical coiled coil proteins.** *Mol Immunol* 1997, **34**(6):433–440.

doi:10.1186/1745-6150-8-27
Cite this article as: Gupta *et al.*: Identification of B-cell epitopes in an antigen for inducing specific class of antibodies. *Biology Direct* 2013 **8**:27.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Additional file 2

Identification of B-cell epitopes in an antigen for inducing specific class of antibodies.

Sudheer Gupta, Hifzur Rahman Ansari, Ankur Gautam, Open Source Drug Discovery Consortium2 and Gajendra P. S. Raghava*

Figure S1. Comparison of physicochemical properties among three classes of epitopes along with non-B-cell epitopes.

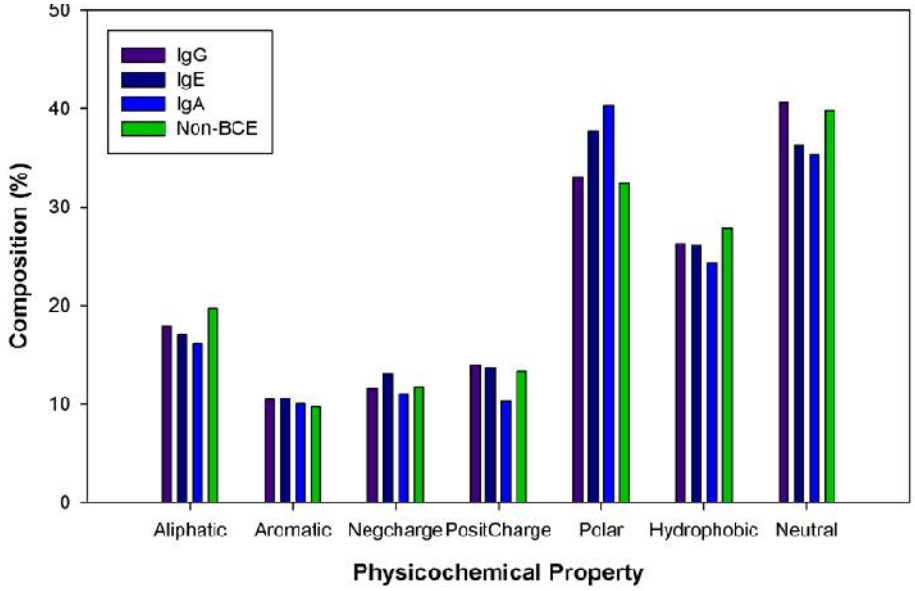
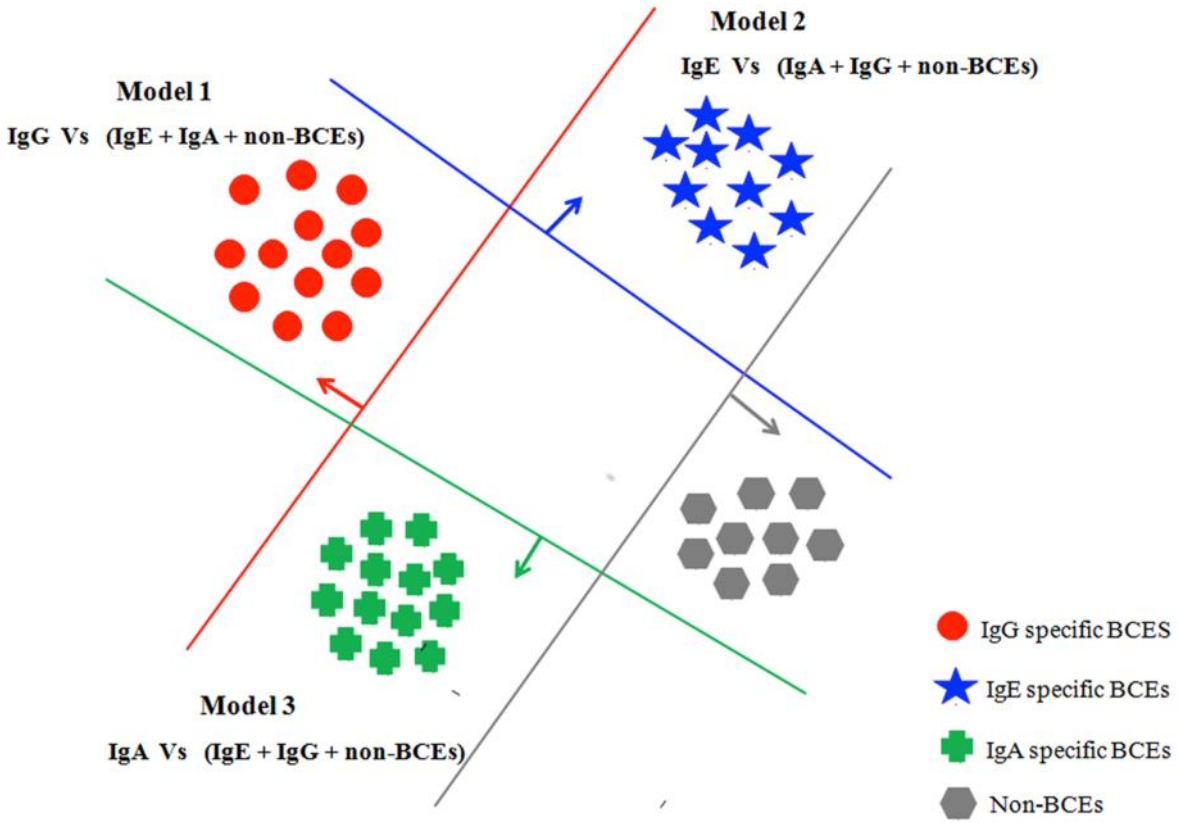


Figure S2. One vs. rest approach used in developing datasets.



Identification of B-cell epitopes in an antigen for inducing specific class of antibodies.

Sudheer Gupta, Hifzur Rahman Ansari, Ankur Gautam, Open Source Drug Discovery Consortium2 and Gajendra P. S. Raghava*

Additional file 3

Table S1. List of top 20 motifs discovered by MEME studies for three classes of epitopes.

	IgG epitopes	IgE epitopes	IgA epitopes
Motif 1	LY[QE][EA]FDEMEEC	QQQQPQQQFPQQQQP	QPY[PQ]Q[QP]QPFPQ[QP]Q[PQ][FQY]
Motif 2	LSG[KR]PAI[IV]PDRE[VA]LY	[CN][SA]QWGW[CT]G[ST]xD	[EK]E[ED]EEDK[EK][EDR][DKE]E
Motif 3	[QFP][PF][QP]Q[PQ]QQPFPQ[QP][QP]Q[PQ]	[FW][YA][PS][EQD][LY][LMF][HR][QR][LF]Y[QH]	[HE]K[SHT][AG]I[VL]T[LV]TY[DH]SE[WET]QR[DQT][QK]F
Motif 4	G[NS]D[YW]EDRYYRENM[HYN]R	[PH][MPY]H[IPF][RHD]LS[FG]	FEEIRNLALQT
Motif 5	[MP]PPPGM[RI]PP[PM][GP]	[LI][LP][FY][VC][MR][GE][NG][ESG][AK][EV]P	CCQ[QH]LWQIPEQS[RQ]C
Motif 6	[KQ][MW][DK]AEF[RH][HR][DW]S[GS]Y[EM]VH	KC[KG][VI]NA[LP]YKIS[AG][SL]TN	G[TD]I[FC][DN][NT][FM][LH][IY]TN
Motif 7	MEEC[AS][SQ][HA]LPYIEQ[GA][MQ]	[YDP][GRPS][RQV][DE]P[YF][ST][PE]S[QP][DYS]	LCLIGCKPPIGEHWG
Motif 8	V[SA]PTHYVPESD	PE[VG]D[DL]E[AI]L[EL][KRQ][FK]	[IN][PYG][CKE][GR][KD][VN][VT]L[IQR][NQI][HKR][ND][IV][RA][CHF]
Motif 9	N[AV][DN][PG]NA[NG][PG]NA[NG][PG]NA[NG]	[RV][HR][AS][KP]A[HQ]IL[QR]W[GQ]	GG[SG][GR]G[GR]A[GR]AGG[RG]GRG
Motif 10	PYIEQGMQLA[EA][QE]FKQ	TY[DGT]N[EKD]C[LN][LF]C[AN][HAY]	VLQESTYQLVQQLCC
Motif 11	H[IV][ST]GHRMAWDMM[LM]NW	[LA][KRA]S[LV]P[QG][QK]CG[VL]R	[VQ][SP][FV][QP]Q[PL]Q[QL]Q[YN]P
Motif 12	[AT]SG[AG][SA][GA][QG][SAP]G[AT][SA]GA[SG]G	[NG][LE]A[RK]TT[PA][DQ]R[QR][AT]ACNC	GGGYVKLFPNS
Motif 13	GSSYGFQYSPG	[PV][SA]G[AT]WY[YS][VL][PA][LM][GA]	[KH]G[SD][PS][CE][TY]N[VI][AM][VF][NG]P[GD][DI]C
Motif 14	AEQFKQK[AV]LGLLQ[TR][AS]	[ILA][NDG][TKN][QDP][YF][SKLT][KDPW][YAR][PLS][SKLY][YF]	[AHNP][LC][SHY][ACRT][TS][FI][EA][PS]F[GS][INV][FKL]G
Motif 15	NPKPQRKTKRNTNRR	E[IV]V[PR]N[SAT][VAN]E[EDQ][KRE][HLY]	[DA]L[YN]CY[EW][QP]LND
Motif 16	NDIRVEESIYQ	[QA][DEK]KT[EK][INE]P[ATC]	AVYFKEQFLDGDGWT

Motif 17	H[PS][NY][IQ]EE[VA]A[LM][GSE][HLNQT][ETK][GV]E[ID]PFY[GW][KRV]	[QM][AT][MA][ES]DI[KS][QLV][ML][ED]A	RDHSYQEEAAM
Motif 18	FGKGGIVTCAMFTCK	[QR][SF][LKF]V[YA]PF[PA]	EEKDKGLQTSQDARF
Motif 19	ILIGVIITWIGMNSR	A[FY]N[HQ]FG[PEK][NGR][LA][GID]QRMPR	[MV]N[FQ][DM]I[PN]E[EA][IK]KQL[QL][QE]F
Motif 20	FLVNTWKSKK[NC]PMGF	Q[GQ][DG]RR[CH]QS	HKSDFGKFVLSSGKF

Table S2. The detailed performance of SVM models developed using various features for predicting antibody specific BCEs on BalanceVar dataset.

Input Feature	IgG Epitope					IgE Epitope					IgA Epitope				
	SEN	**SPEC**	**ACC**	**MCC**	**AUC**	**SEN**	**SPEC**	**ACC**	**MCC**	**AUC**	**SEN**	**SPEC**	**ACC**	**MCC**	**AUC**
AAC	62.45	65.24	63.85	0.28	0.68	73.41	77.23	75.33	0.51	0.81	70.97	71.94	71.46	0.43	0.76
AAP	68.11	68.50	68.30	0.37	0.73	74.26	82.30	78.3	0.57	0.85	73.95	71.94	72.93	0.46	0.78
CTD	63.46	65.12	64.30	0.29	0.69	69.61	68.02	68.81	0.38	0.71	70.22	69.30	69.76	0.40	0.74
DPC	**72.05**	**68.82**	**70.42**	**0.41**	**0.76**	**75.89**	**89.44**	**82.7**	**0.66**	**0.88**	**64.52**	**79.38**	**72.07**	**0.44**	**0.78**
PCP	66.53	65.84	66.18	0.32	0.71	64.62	64.01	64.31	0.29	0.64	72.21	73.38	72.8	0.46	0.78

SEN: Sensitivity; SPEC: Specificity

Table S3. Performance of dipeptide-based model with five fold cross validation on BalanceVar dataset, for all three classes.

IgG class

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	7398	6993	682	200	97.37	8.89	52.90	0.13
-0.9	7274	6594	1081	324	95.74	14.08	54.70	0.17
-0.8	7222	6122	1553	376	95.05	20.23	57.45	0.23
-0.7	7121	5664	2011	477	93.72	26.20	59.79	0.27
-0.6	6996	5120	2555	602	92.08	33.29	62.54	0.31
-0.5	6853	4641	3034	745	90.19	39.53	64.74	0.34
-0.4	6648	4140	3535	950	87.50	46.06	66.67	0.37
-0.3	6411	3667	4008	1187	84.38	52.22	68.22	0.39
-0.2	6143	3245	4430	1455	80.85	57.72	69.23	0.40
-0.1	5831	2808	4867	1767	76.74	63.41	70.05	0.41
0	5474	2393	5282	2124	72.05	68.82	70.42	0.41##
0.1	5095	2028	5647	2503	67.06	73.58	70.33	0.41
0.2	4726	1666	6009	2872	62.20	78.29	70.29	0.41
0.3	4261	1371	6304	3337	56.08	82.14	69.17	0.40
0.4	3786	1134	6541	3812	49.83	85.22	67.62	0.38
0.5	3284	916	6759	4314	43.22	88.07	65.76	0.35
0.6	2804	732	6943	4794	36.90	90.46	63.82	0.32
0.7	2349	576	7099	5249	30.92	92.50	61.86	0.30
0.8	1879	433	7242	5719	24.73	94.36	59.72	0.27

0.9	1448	303	7372	6150	19.06	96.05	57.75	0.24
1	1020	162	7513	6578	13.42	97.89	55.87	0.21

IgE class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	2328	2131	236	15	99.36	9.97	54.44	0.21
-0.9	2292	2028	339	51	97.82	14.32	55.86	0.22
-0.8	2282	1872	495	61	97.40	20.91	58.96	0.28
-0.7	2263	1693	674	80	96.59	28.47	62.36	0.34
-0.6	2240	1530	837	103	95.60	35.36	65.33	0.39
-0.5	2210	1333	1034	133	94.32	43.68	68.87	0.44
-0.4	2160	1099	1268	183	92.19	53.57	72.78	0.50
-0.3	2113	908	1459	230	90.18	61.64	75.84	0.54
-0.2	2053	726	1641	290	87.62	69.33	78.43	0.58
-0.1	1998	562	1805	345	85.28	76.26	80.74	0.62
0	1920	429	1938	423	81.95	81.88	81.91	0.64
0.1	1842	332	2035	501	78.62	85.97	82.31	0.65
0.2	1778	250	2117	565	75.89	89.44	82.70	0.66##
0.3	1683	190	2177	660	71.83	91.97	81.95	0.65
0.4	1564	138	2229	779	66.75	94.17	80.53	0.63
0.5	1436	115	2252	907	61.29	95.14	78.30	0.60
0.6	1268	97	2270	1075	54.12	95.90	75.12	0.55
0.7	1085	81	2286	1258	46.31	96.58	71.57	0.50
0.8	877	72	2295	1466	37.43	96.96	67.35	0.43
0.9	637	61	2306	1706	27.19	97.42	62.48	0.35
1	405	34	2333	1938	17.29	98.56	58.13	0.27

IgA class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	402	416	1	1	99.75	0.24	49.15	-0.00
-0.9	399	410	7	4	99.01	1.68	49.51	0.03
-0.8	398	402	15	5	98.76	3.60	50.37	0.08
-0.7	395	386	31	8	98.01	7.43	51.95	0.13
-0.6	383	368	49	20	95.04	11.75	52.68	0.12
-0.5	379	351	66	24	94.04	15.83	54.27	0.16
-0.4	368	317	100	35	91.32	23.98	57.07	0.21
-0.3	354	271	146	49	87.84	35.01	60.98	0.27
-0.2	339	232	185	64	84.12	44.36	63.90	0.31
-0.1	322	198	219	81	79.90	52.52	65.98	0.34
0	310	171	246	93	76.92	58.99	67.80	0.36
0.1	294	144	273	109	72.95	65.47	69.15	0.39
0.2	278	108	309	125	68.98	74.10	71.59	0.43
0.3	260	86	331	143	64.52	79.38	72.07	0.44##
0.4	241	60	357	162	59.80	85.61	72.93	0.47
0.5	221	35	382	182	54.84	91.61	73.54	0.50
0.6	195	27	390	208	48.39	93.53	71.34	0.47
0.7	177	18	399	226	43.92	95.68	70.24	0.47
0.8	152	16	401	251	37.72	96.16	67.44	0.42
0.9	114	13	404	289	28.29	96.88	63.17	0.35
1	74	6	411	329	18.36	98.56	59.15	0.29

Table S4. The detailed performance of SVM models developed using various features for predicting antibody specific BCEs on BalanceFix dataset.

Input Feature	IgG Epitope					IgE Epitope					IgA Epitope				
	SEN	**SPEC**	**ACC**	**MCC**	**AUC**	**SEN**	**SPEC**	**ACC**	**MCC**	**AUC**	**SEN**	**SPEC**	**ACC**	**MCC**	**AUC**
AAC	67.10	65.43	66.27	0.33	0.70	77.17	86.40	81.78	0.64	0.86	68.54	70.04	69.29	0.39	0.75
AAP	71.42	67.15	69.29	0.39	0.75	81.73	83.04	82.39	0.65	0.89	45.69	92.88	74.34	0.49	0.79
CTD	58.44	56.38	57.41	0.15	0.61	65.72	62.26	63.99	0.28	0.70	55.06	71.54	63.3	0.27	0.67
DPC	**75.28**	**68.18**	**71.73**	**0.44**	**0.77**	**86.04**	**83.88**	**84.96**	**0.70**	**0.90**	**70.41**	**74.16**	**72.28**	**0.45**	**0.78**
PCP	52.98	60.17	56.57	0.13	0.59	59.00	57.22	58.11	0.16	0.62	64.42	62.17	63.3	0.27	0.69

SEN: Sensitivity; SPEC: Specificity

Table S5. Performance of dipeptide based model with five fold cross validation on BalanceFix dataset for all three classes.

IgG class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	5892	5208	908	224	96.34	14.85	55.59	0.19
-0.9	5716	4665	1451	400	93.46	23.72	58.59	0.24
-0.8	5644	4152	1964	472	92.28	32.11	62.20	0.31
-0.7	5558	3725	2391	558	90.88	39.09	64.99	0.35
-0.6	5451	3406	2710	665	89.13	44.31	66.72	0.37
-0.5	5350	3090	3026	766	87.48	49.48	68.48	0.40
-0.4	5208	2778	3338	908	85.15	54.58	69.87	0.42
-0.3	5040	2498	3618	1076	82.41	59.16	70.78	0.43
-0.2	4852	2215	3901	1264	79.33	63.78	71.56	0.44
-0.1	4604	1946	4170	1512	75.28	68.18	71.73	0.44##
0	4316	1695	4421	1800	70.57	72.29	71.43	0.43
0.1	4021	1456	4660	2095	65.75	76.19	70.97	0.42
0.2	3667	1234	4882	2449	59.96	79.82	69.89	0.41
0.3	3351	1026	5090	2765	54.79	83.22	69.01	0.40
0.4	3026	840	5276	3090	49.48	86.27	67.87	0.38
0.5	2683	702	5414	3433	43.87	88.52	66.20	0.36
0.6	2361	568	5548	3755	38.60	90.71	64.66	0.34
0.7	2031	454	5662	4085	33.21	92.58	62.89	0.32
0.8	1686	368	5748	4430	27.57	93.98	60.78	0.29
0.9	1293	278	5838	4823	21.14	95.45	58.30	0.25
1	885	136	5980	5231	14.47	97.78	56.12	0.22

IgE class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	1886	1618	287	19	99.00	15.07	57.03	0.26
-0.9	1864	1502	403	41	97.85	21.15	59.50	0.30
-0.8	1855	1362	543	50	97.38	28.50	62.94	0.36
-0.7	1841	1219	686	64	96.64	36.01	66.33	0.41
-0.6	1823	1068	837	82	95.70	43.94	69.82	0.46
-0.5	1801	932	973	104	94.54	51.08	72.81	0.51
-0.4	1775	777	1128	130	93.18	59.21	76.19	0.56
-0.3	1740	640	1265	165	91.34	66.40	78.87	0.60
-0.2	1711	518	1387	194	89.82	72.81	81.31	0.64
-0.1	1672	410	1495	233	87.77	78.48	83.12	0.67
0	1639	307	1598	266	86.04	83.88	84.96	0.70##
0.1	1595	234	1671	310	83.73	87.72	85.72	0.72
0.2	1557	183	1722	348	81.73	90.39	86.06	0.72
0.3	1506	154	1751	399	79.06	91.92	85.49	0.72
0.4	1457	131	1774	448	76.48	93.12	84.80	0.71
0.5	1391	110	1795	514	73.02	94.23	83.62	0.69
0.6	1304	99	1806	601	68.45	94.80	81.63	0.66
0.7	1199	87	1818	706	62.94	95.43	79.19	0.62
0.8	1029	78	1827	876	54.02	95.91	74.96	0.55
0.9	809	72	1833	1096	42.47	96.22	69.34	0.46
1	563	41	1864	1342	29.55	97.85	63.70	0.38

IgA class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	259	245	22	8	97.00	8.24	52.62	0.11
-0.9	253	233	34	14	94.76	12.73	53.75	0.13
-0.8	248	221	46	19	92.88	17.23	55.06	0.15
-0.7	245	213	54	22	91.76	20.22	55.99	0.17
-0.6	239	199	68	28	89.51	25.47	57.49	0.20
-0.5	234	184	83	33	87.64	31.09	59.36	0.23
-0.4	224	164	103	43	83.90	38.58	61.24	0.25
-0.3	216	130	137	51	80.90	51.31	66.10	0.34
-0.2	207	105	162	60	77.53	60.67	69.10	0.39
-0.1	193	92	175	74	72.28	65.54	68.91	0.38
0	188	69	198	79	70.41	74.16	72.28	0.45##
0.1	178	53	214	89	66.67	80.15	73.41	0.47
0.2	162	40	227	105	60.67	85.02	72.85	0.47
0.3	152	25	242	115	56.93	90.64	73.78	0.51
0.4	139	22	245	128	52.06	91.76	71.91	0.48
0.5	130	16	251	137	48.69	94.01	71.35	0.48
0.6	127	13	254	140	47.57	95.13	71.35	0.49
0.7	117	8	259	150	43.82	97.00	70.41	0.48
0.8	104	6	261	163	38.95	97.75	68.35	0.45
0.9	90	2	265	177	33.71	99.25	66.48	0.44
1	65	1	266	202	24.34	99.63	61.99	0.36

Table S6. The performance of SVM-based models developed using various features for predicting antibody specific B-cell epitopes on RealFix dataset.

Class of Antibodies	Feature	Sen	Spec	Acc	MCC	AUC
IgG	AAC	69.49	71.89	71.18	0.39	0.75
	AAP	79.82	69.76	72.76	0.46	0.79
	CTD	37.03	85.17	70.83	0.25	0.67
	DPC	**74**	**78.74**	**77.33**	**0.5**	**0.83**
	PCP	15.82	93.71	70.5	0.15	0.61
	BIN	61.45	52.01	54.82	0.12	0.60
IgE	AAC	72.91	88.12	87.23	0.4	0.84
	AAP	80.16	89.46	88.92	0.46	0.89
	CTD	19.27	98.01	93.38	0.24	0.74
	DPC	**77.38**	**94**	**93.02**	**0.55**	**0.9**
	PCP	3.88	99.61	93.99	0.11	0.68
	BIN	10.76	97.28	92.20	0.11	0.60
IgA	AAC	41.2	98.85	98.37	0.3	0.77
	AAP	40.07	99.13	98.64	0.33	0.75
	CTD	12.36	99.93	99.21	0.27	0.71
	DPC	**40.82**	**99.67**	**99.18**	**0.45**	**0.76**
	PCP	3	99.98	99.18	0.13	0.67
	BIN	13.86	99.57	98.87	0.17	0.67

Sen: Sensitivity; Spec: Specificity

Table S7. The performance of SVM based models developed using various features for predicting antibody specific B-cell epitopes on RealVar dataset.

Class of Antibodies	Feature	Sen	Spec	Acc	MCC	AUC
IgG	AAC	44.9	84.36	71.77	0.32	0.7
	AAP	**55.30**	**84.41**	**75.74**	**0.41**	**0.79**
	CTD	25.04	92.9	72.68	0.25	0.67
	DPC	56.58	82.19	74.02	0.39	0.73
	PCP	22.94	89.17	69.44	0.16	0.61
IgE	AAC	42.93	94.27	91.07	0.33	0.7
	AAP	**52.81**	**97.12**	**94.51**	**0.5**	**0.89**
	CTD	8.82	99.36	94.04	0.18	0.74
	DPC	54.76	95.91	93.35	0.47	0.86
	PCP	40.21	83.67	81.11	0.15	0.68
IgA	AAC	34.99	99.43	98.74	0.37	0.8
	AAP	30.02	99.52	98.78	0.34	0.73
	CTD	23.6	99.57	98.95	0.27	0.71
	DPC	**42.18**	**99.01**	**98.41**	**0.36**	**0.84**
	PCP	11.61	99.89	99.16	0.23	0.66

Sen: Sensitivity; Spec: Specificity

Table S8. Performance of various WEKA classifiers developed using various features as input for different classes of epitopes on BalanceFix dataset.

Features	Parameter	IgG			IgE			IgA		
		SMO	**IBk**	**RF**	**SMO**	**IBk**	**RF**	**SMO**	**IBk**	**RF**
AAC	**ACC**	63.34	64.51	62.76	76.61	75.59	74.67	71.28	66.85	68.91
	MCC	0.27	0.29	0.26	0.53	0.52	0.5	0.46	0.41	0.41
	AUC	0.63	0.64	0.66	0.76	0.76	0.8	0.72	0.66	0.73
AAP	**ACC**	68.14	70.05	67.27	81.5	79.53	80.81	70.6	60.23	71.1
	MCC	0.35	0.4	0.35	0.64	0.6	0.62	0.49	0.23	0.48
	AUC	0.67	0.7	0.72	0.81	0.8	0.86	0.71	0.59	0.74
DPC	**ACC**	67.40	70.07	67.14	81.34	79.61	80.89	70.6	63.30	71.16
	MCC	0.35	0.4	0.34	0.64	0.6	0.62	0.49	0.27	0.49
	AUC	0.67	0.7	0.72	0.81	0.8	0.86	0.71	0.59	0.76
CTD	**ACC**	55.31	58.09	55.19	61.42	63.02	63.18	62.73	62.92	62.92
	MCC	0.1	0.16	0.11	0.23	0.26	0.27	0.28	0.26	0.34
	AUC	0.55	0.58	0.56	0.61	0.63	0.67	0.62	0.63	0.68
PHY	**ACC**	56.76	56.50	55.96	61.55	63.18	64.07	63.48	66.10	64.79
	MCC	0.14	0.13	0.12	0.23	0.27	0.28	0.28	0.32	0.31
	AUC	0.57	0.56	0.57	0.62	0.63	0.68	0.63	0.66	0.68

Table S9. The performance of various WEKA classifiers developed using various features as input for different classes of epitopes on BalanceVar dataset.

Features	Parameter	IgG			IgE			IgA		
		SMO	**IBk**	**RF**	**SMO**	**IBk**	**RF**	**SMO**	**IBk**	**RF**
AAC	**ACC**	63.53	62.54	64.36	71.00	70.7	70.47	71.68	68.17	70.98
	MCC	0.27	0.25	0.29	0.42	0.42	0.41	0.47	0.37	0.44
	AUC	0.63	0.62	0.68	0.72	0.71	0.76	0.73	0.68	0.76
AAP	**ACC**	67.81	68.27	68.08	77.18	76.94	76.86	71.46	68.66	70.73
	MCC	0.36	0.37	0.36	0.55	0.55	0.54	0.47	0.42	0.42
	AUC	0.69	0.69	0.72	0.77	0.77	0.82	0.68	0.69	0.72
DPC	**ACC**	67.96	68.25	67.71	77.26	76.96	75.99	71.59	70.12	71.83
	MCC	0.36	0.37	0.36	0.55	0.55	0.52	0.49	0.36	0.44
	AUC	0.68	0.69	0.73	0.77	0.77	0.82	0.71	0.70	0.76
CTD	**ACC**	62.5	60.96	59.98	64.78	63.12	62.95	65.61	64.51	66.34
	MCC	0.25	0.22	0.2	0.3	0.26	0.26	0.32	0.29	0.33
	AUC	0.62	0.58	0.63	0.65	0.63	0.67	0.65	0.65	0.71
PHY	**ACC**	63.23	60.41	62.04	62.44	60.72	64.2	66.1	65.98	68.66
	MCC	0.27	0.21	0.24	0.25	0.22	0.31	0.32	0.32	0.37
	AUC	0.63	0.59	0.65	0.62	0.61	0.71	0.66	0.65	0.74

Additional file 2

Identification of B-cell epitopes in an antigen for inducing specific class of antibodies.

Sudheer Gupta, Hifzur Rahman Ansari, Ankur Gautam, Open Source Drug Discovery Consortium2 and Gajendra P. S. Raghava*

Figure S1. Comparison of physicochemical properties among three classes of epitopes along with non-B-cell epitopes.

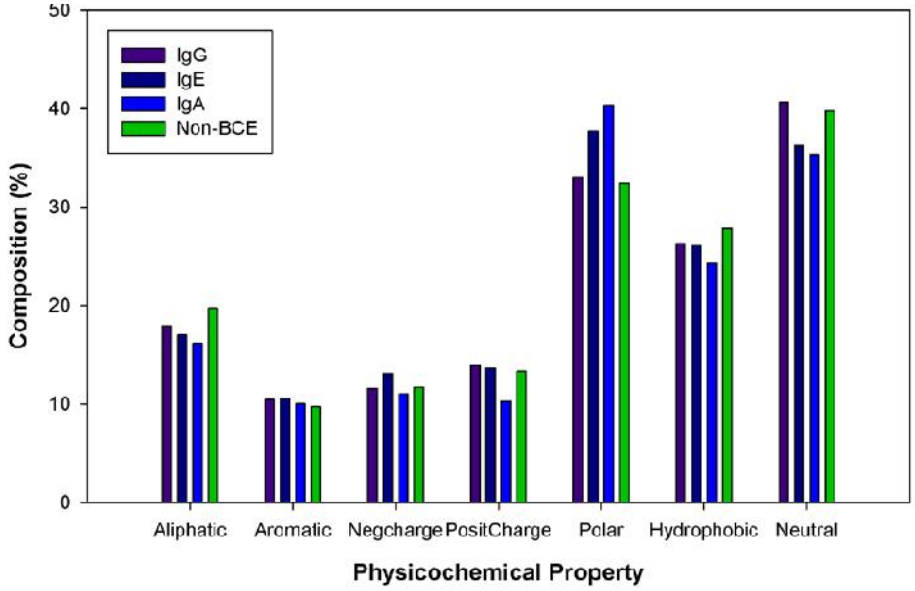
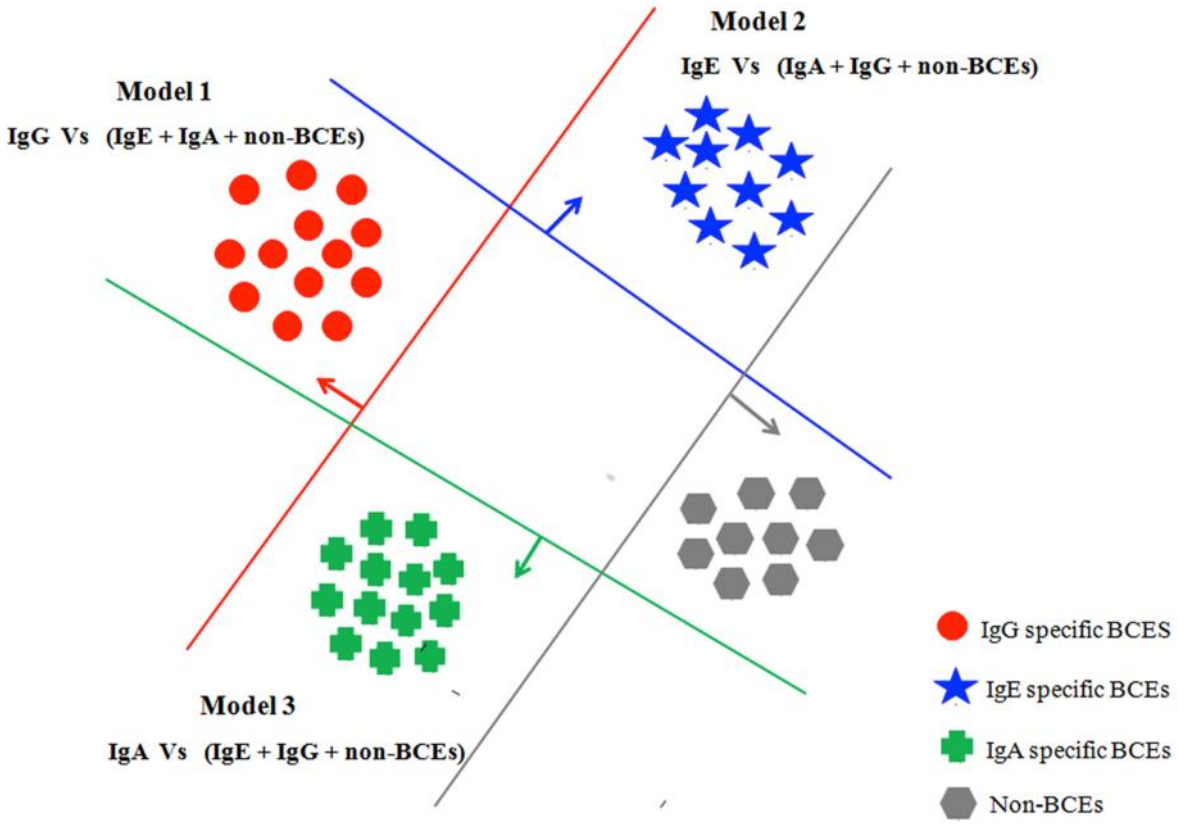


Figure S2. One vs. rest approach used in developing datasets.



Identification of B-cell epitopes in an antigen for inducing specific class of antibodies.

Sudheer Gupta, Hifzur Rahman Ansari, Ankur Gautam, Open Source Drug Discovery Consortium2 and Gajendra P. S. Raghava*

Additional file 3

Table S1. List of top 20 motifs discovered by MEME studies for three classes of epitopes.

	IgG epitopes	IgE epitopes	IgA epitopes
Motif 1	LY[QE][EA]FDEMEEC	QQQQPQQQFPQQQQP	QPY[PQ]Q[QP]QPFPQ[QP]Q[PQ][FQY]
Motif 2	LSG[KR]PAI[IV]PDRE[VA]LY	[CN][SA]QWGW[CT]G[ST]xD	[EK]E[ED]EEDK[EK][EDR][DKE]E
Motif 3	[QFP][PF][QP]Q[PQ]QQPFPQ[QP][QP]Q[PQ]	[FW][YA][PS][EQD][LY][LMF][HR][QR][LF]Y[QH]	[HE]K[SHT][AG]I[VL]T[LV]TY[DH]SE[WET]QR[DQT][QK]F
Motif 4	G[NS]D[YW]EDRYYRENM[HYN]R	[PH][MPY]H[IPF][RHD]LS[FG]	FEEIRNLALQT
Motif 5	[MP]PPPGM[RI]PP[PM][GP]	[LI][LP][FY][VC][MR][GE][NG][ESG][AK][EV]P	CCQ[QH]LWQIPEQS[RQ]C
Motif 6	[KQ][MW][DK]AEF[RH][HR][DW]S[GS]Y[EM]VH	KC[KG][VI]NA[LP]YKIS[AG][SL]TN	G[TD]I[FC][DN][NT][FM][LH][IY]TN
Motif 7	MEEC[AS][SQ][HA]LPYIEQ[GA][MQ]	[YDP][GRPS][RQV][DE]P[YF][ST][PE]S[QP][DYS]	LCLIGCKPPIGEHWG
Motif 8	V[SA]PTHYVPESD	PE[VG]D[DL]E[AI]L[EL][KRQ][FK]	[IN][PYG][CKE][GR][KD][VN][VT]L[IQR][NQI][HKR][ND][IV][RA][CHF]
Motif 9	N[AV][DN][PG]NA[NG][PG]NA[NG][PG]NA[NG]	[RV][HR][AS][KP]A[HQ]IL[QR]W[GQ]	GG[SG][GR]G[GR]A[GR]AGG[RG]GRG
Motif 10	PYIEQGMQLA[EA][QE]FKQ	TY[DGT]N[EKD]C[LN][LF]C[AN][HAY]	VLQESTYQLVQQLCC
Motif 11	H[IV][ST]GHRMAWDMM[LM]NW	[LA][KRA]S[LV]P[QG][QK]CG[VL]R	[VQ][SP][FV][QP]Q[PL]Q[QL]Q[YN]P
Motif 12	[AT]SG[AG][SA][GA][QG][SAP]G[AT][SA]GA[SG]G	[NG][LE]A[RK]TT[PA][DQ]R[QR][AT]ACNC	GGGYVKLFPNS
Motif 13	GSSYGFQYSPG	[PV][SA]G[AT]WY[YS][VL][PA][LM][GA]	[KH]G[SD][PS][CE][TY]N[VI][AM][VF][NG]P[GD][DI]C
Motif 14	AEQFKQK[AV]LGLLQ[TR][AS]	[ILA][NDG][TKN][QDP][YF][SKLT][KDPW][YAR][PLS][SKLY][YF]	[AHNP][LC][SHY][ACRT][TS][FI][EA][PS]F[GS][INV][FKL]G
Motif 15	NPKPQRKTKRNTNRR	E[IV]V[PR]N[SAT][VAN]E[EDQ][KRE][HLY]	[DA]L[YN]CY[EW][QP]LND
Motif 16	NDIRVEESIYQ	[QA][DEK]KT[EK][INE]P[ATC]	AVYFKEQFLDGDGWT

Motif 17	H[PS][NY][IQ]EE[VA]A[LM][GSE][HLNQT][ETK][GV]E[ID]PFY[GW][KRV]	[QM][AT][MA][ES]DI[KS][QLV][ML][ED]A	RDHSYQEEAAM
Motif 18	FGKGGIVTCAMFTCK	[QR][SF][LKF]V[YA]PF[PA]	EEKDKGLQTSQDARF
Motif 19	ILIGVIITWIGMNSR	A[FY]N[HQ]FG[PEK][NGR][LA][GID]QRMPR	[MV]N[FQ][DM]I[PN]E[EA][IK]KQL[QL][QE]F
Motif 20	FLVNTWKSKK[NC]PMGF	Q[GQ][DG]RR[CH]QS	HKSDFGKFVLSSGKF

Table S2. The detailed performance of SVM models developed using various features for predicting antibody specific BCEs on BalanceVar dataset.

Input Feature	IgG Epitope					IgE Epitope					IgA Epitope				
	SEN	**SPEC**	**ACC**	**MCC**	**AUC**	**SEN**	**SPEC**	**ACC**	**MCC**	**AUC**	**SEN**	**SPEC**	**ACC**	**MCC**	**AUC**
AAC	62.45	65.24	63.85	0.28	0.68	73.41	77.23	75.33	0.51	0.81	70.97	71.94	71.46	0.43	0.76
AAP	68.11	68.50	68.30	0.37	0.73	74.26	82.30	78.3	0.57	0.85	73.95	71.94	72.93	0.46	0.78
CTD	63.46	65.12	64.30	0.29	0.69	69.61	68.02	68.81	0.38	0.71	70.22	69.30	69.76	0.40	0.74
DPC	**72.05**	**68.82**	**70.42**	**0.41**	**0.76**	**75.89**	**89.44**	**82.7**	**0.66**	**0.88**	**64.52**	**79.38**	**72.07**	**0.44**	**0.78**
PCP	66.53	65.84	66.18	0.32	0.71	64.62	64.01	64.31	0.29	0.64	72.21	73.38	72.8	0.46	0.78

SEN: Sensitivity; SPEC: Specificity

Table S3. Performance of dipeptide-based model with five fold cross validation on BalanceVar dataset, for all three classes.

IgG class

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	7398	6993	682	200	97.37	8.89	52.90	0.13
-0.9	7274	6594	1081	324	95.74	14.08	54.70	0.17
-0.8	7222	6122	1553	376	95.05	20.23	57.45	0.23
-0.7	7121	5664	2011	477	93.72	26.20	59.79	0.27
-0.6	6996	5120	2555	602	92.08	33.29	62.54	0.31
-0.5	6853	4641	3034	745	90.19	39.53	64.74	0.34
-0.4	6648	4140	3535	950	87.50	46.06	66.67	0.37
-0.3	6411	3667	4008	1187	84.38	52.22	68.22	0.39
-0.2	6143	3245	4430	1455	80.85	57.72	69.23	0.40
-0.1	5831	2808	4867	1767	76.74	63.41	70.05	0.41
0	5474	2393	5282	2124	72.05	68.82	70.42	0.41##
0.1	5095	2028	5647	2503	67.06	73.58	70.33	0.41
0.2	4726	1666	6009	2872	62.20	78.29	70.29	0.41
0.3	4261	1371	6304	3337	56.08	82.14	69.17	0.40
0.4	3786	1134	6541	3812	49.83	85.22	67.62	0.38
0.5	3284	916	6759	4314	43.22	88.07	65.76	0.35
0.6	2804	732	6943	4794	36.90	90.46	63.82	0.32
0.7	2349	576	7099	5249	30.92	92.50	61.86	0.30
0.8	1879	433	7242	5719	24.73	94.36	59.72	0.27

0.9	1448	303	7372	6150	19.06	96.05	57.75	0.24
1	1020	162	7513	6578	13.42	97.89	55.87	0.21

IgE class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	2328	2131	236	15	99.36	9.97	54.44	0.21
-0.9	2292	2028	339	51	97.82	14.32	55.86	0.22
-0.8	2282	1872	495	61	97.40	20.91	58.96	0.28
-0.7	2263	1693	674	80	96.59	28.47	62.36	0.34
-0.6	2240	1530	837	103	95.60	35.36	65.33	0.39
-0.5	2210	1333	1034	133	94.32	43.68	68.87	0.44
-0.4	2160	1099	1268	183	92.19	53.57	72.78	0.50
-0.3	2113	908	1459	230	90.18	61.64	75.84	0.54
-0.2	2053	726	1641	290	87.62	69.33	78.43	0.58
-0.1	1998	562	1805	345	85.28	76.26	80.74	0.62
0	1920	429	1938	423	81.95	81.88	81.91	0.64
0.1	1842	332	2035	501	78.62	85.97	82.31	0.65
0.2	1778	250	2117	565	75.89	89.44	82.70	0.66##
0.3	1683	190	2177	660	71.83	91.97	81.95	0.65
0.4	1564	138	2229	779	66.75	94.17	80.53	0.63
0.5	1436	115	2252	907	61.29	95.14	78.30	0.60
0.6	1268	97	2270	1075	54.12	95.90	75.12	0.55
0.7	1085	81	2286	1258	46.31	96.58	71.57	0.50
0.8	877	72	2295	1466	37.43	96.96	67.35	0.43
0.9	637	61	2306	1706	27.19	97.42	62.48	0.35
1	405	34	2333	1938	17.29	98.56	58.13	0.27

IgA class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	402	416	1	1	99.75	0.24	49.15	-0.00
-0.9	399	410	7	4	99.01	1.68	49.51	0.03
-0.8	398	402	15	5	98.76	3.60	50.37	0.08
-0.7	395	386	31	8	98.01	7.43	51.95	0.13
-0.6	383	368	49	20	95.04	11.75	52.68	0.12
-0.5	379	351	66	24	94.04	15.83	54.27	0.16
-0.4	368	317	100	35	91.32	23.98	57.07	0.21
-0.3	354	271	146	49	87.84	35.01	60.98	0.27
-0.2	339	232	185	64	84.12	44.36	63.90	0.31
-0.1	322	198	219	81	79.90	52.52	65.98	0.34
0	310	171	246	93	76.92	58.99	67.80	0.36
0.1	294	144	273	109	72.95	65.47	69.15	0.39
0.2	278	108	309	125	68.98	74.10	71.59	0.43
0.3	260	86	331	143	64.52	79.38	72.07	0.44##
0.4	241	60	357	162	59.80	85.61	72.93	0.47
0.5	221	35	382	182	54.84	91.61	73.54	0.50
0.6	195	27	390	208	48.39	93.53	71.34	0.47
0.7	177	18	399	226	43.92	95.68	70.24	0.47
0.8	152	16	401	251	37.72	96.16	67.44	0.42
0.9	114	13	404	289	28.29	96.88	63.17	0.35
1	74	6	411	329	18.36	98.56	59.15	0.29

Table S4. The detailed performance of SVM models developed using various features for predicting antibody specific BCEs on BalanceFix dataset.

Input Feature	IgG Epitope					IgE Epitope					IgA Epitope				
	SEN	**SPEC**	**ACC**	**MCC**	**AUC**	**SEN**	**SPEC**	**ACC**	**MCC**	**AUC**	**SEN**	**SPEC**	**ACC**	**MCC**	**AUC**
AAC	67.10	65.43	66.27	0.33	0.70	77.17	86.40	81.78	0.64	0.86	68.54	70.04	69.29	0.39	0.75
AAP	71.42	67.15	69.29	0.39	0.75	81.73	83.04	82.39	0.65	0.89	45.69	92.88	74.34	0.49	0.79
CTD	58.44	56.38	57.41	0.15	0.61	65.72	62.26	63.99	0.28	0.70	55.06	71.54	63.3	0.27	0.67
DPC	**75.28**	**68.18**	**71.73**	**0.44**	**0.77**	**86.04**	**83.88**	**84.96**	**0.70**	**0.90**	**70.41**	**74.16**	**72.28**	**0.45**	**0.78**
PCP	52.98	60.17	56.57	0.13	0.59	59.00	57.22	58.11	0.16	0.62	64.42	62.17	63.3	0.27	0.69

SEN: Sensitivity; SPEC: Specificity

Table S5. Performance of dipeptide based model with five fold cross validation on BalanceFix dataset for all three classes.

IgG class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	5892	5208	908	224	96.34	14.85	55.59	0.19
-0.9	5716	4665	1451	400	93.46	23.72	58.59	0.24
-0.8	5644	4152	1964	472	92.28	32.11	62.20	0.31
-0.7	5558	3725	2391	558	90.88	39.09	64.99	0.35
-0.6	5451	3406	2710	665	89.13	44.31	66.72	0.37
-0.5	5350	3090	3026	766	87.48	49.48	68.48	0.40
-0.4	5208	2778	3338	908	85.15	54.58	69.87	0.42
-0.3	5040	2498	3618	1076	82.41	59.16	70.78	0.43
-0.2	4852	2215	3901	1264	79.33	63.78	71.56	0.44
-0.1	4604	1946	4170	1512	75.28	68.18	71.73	0.44##
0	4316	1695	4421	1800	70.57	72.29	71.43	0.43
0.1	4021	1456	4660	2095	65.75	76.19	70.97	0.42
0.2	3667	1234	4882	2449	59.96	79.82	69.89	0.41
0.3	3351	1026	5090	2765	54.79	83.22	69.01	0.40
0.4	3026	840	5276	3090	49.48	86.27	67.87	0.38
0.5	2683	702	5414	3433	43.87	88.52	66.20	0.36
0.6	2361	568	5548	3755	38.60	90.71	64.66	0.34
0.7	2031	454	5662	4085	33.21	92.58	62.89	0.32
0.8	1686	368	5748	4430	27.57	93.98	60.78	0.29
0.9	1293	278	5838	4823	21.14	95.45	58.30	0.25
1	885	136	5980	5231	14.47	97.78	56.12	0.22

IgE class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	1886	1618	287	19	99.00	15.07	57.03	0.26
-0.9	1864	1502	403	41	97.85	21.15	59.50	0.30
-0.8	1855	1362	543	50	97.38	28.50	62.94	0.36
-0.7	1841	1219	686	64	96.64	36.01	66.33	0.41
-0.6	1823	1068	837	82	95.70	43.94	69.82	0.46
-0.5	1801	932	973	104	94.54	51.08	72.81	0.51
-0.4	1775	777	1128	130	93.18	59.21	76.19	0.56
-0.3	1740	640	1265	165	91.34	66.40	78.87	0.60
-0.2	1711	518	1387	194	89.82	72.81	81.31	0.64
-0.1	1672	410	1495	233	87.77	78.48	83.12	0.67
0	1639	307	1598	266	86.04	83.88	84.96	0.70##
0.1	1595	234	1671	310	83.73	87.72	85.72	0.72
0.2	1557	183	1722	348	81.73	90.39	86.06	0.72
0.3	1506	154	1751	399	79.06	91.92	85.49	0.72
0.4	1457	131	1774	448	76.48	93.12	84.80	0.71
0.5	1391	110	1795	514	73.02	94.23	83.62	0.69
0.6	1304	99	1806	601	68.45	94.80	81.63	0.66
0.7	1199	87	1818	706	62.94	95.43	79.19	0.62
0.8	1029	78	1827	876	54.02	95.91	74.96	0.55
0.9	809	72	1833	1096	42.47	96.22	69.34	0.46
1	563	41	1864	1342	29.55	97.85	63.70	0.38

IgA class-

Thres	TP	FP	TN	FN	Sen	Spec	Accu	MCC
-1	259	245	22	8	97.00	8.24	52.62	0.11
-0.9	253	233	34	14	94.76	12.73	53.75	0.13
-0.8	248	221	46	19	92.88	17.23	55.06	0.15
-0.7	245	213	54	22	91.76	20.22	55.99	0.17
-0.6	239	199	68	28	89.51	25.47	57.49	0.20
-0.5	234	184	83	33	87.64	31.09	59.36	0.23
-0.4	224	164	103	43	83.90	38.58	61.24	0.25
-0.3	216	130	137	51	80.90	51.31	66.10	0.34
-0.2	207	105	162	60	77.53	60.67	69.10	0.39
-0.1	193	92	175	74	72.28	65.54	68.91	0.38
0	188	69	198	79	70.41	74.16	72.28	0.45##
0.1	178	53	214	89	66.67	80.15	73.41	0.47
0.2	162	40	227	105	60.67	85.02	72.85	0.47
0.3	152	25	242	115	56.93	90.64	73.78	0.51
0.4	139	22	245	128	52.06	91.76	71.91	0.48
0.5	130	16	251	137	48.69	94.01	71.35	0.48
0.6	127	13	254	140	47.57	95.13	71.35	0.49
0.7	117	8	259	150	43.82	97.00	70.41	0.48
0.8	104	6	261	163	38.95	97.75	68.35	0.45
0.9	90	2	265	177	33.71	99.25	66.48	0.44
1	65	1	266	202	24.34	99.63	61.99	0.36

Table S6. The performance of SVM-based models developed using various features for predicting antibody specific B-cell epitopes on RealFix dataset.

Class of Antibodies	Feature	Sen	Spec	Acc	MCC	AUC
IgG	AAC	69.49	71.89	71.18	0.39	0.75
	AAP	79.82	69.76	72.76	0.46	0.79
	CTD	37.03	85.17	70.83	0.25	0.67
	DPC	**74**	**78.74**	**77.33**	**0.5**	**0.83**
	PCP	15.82	93.71	70.5	0.15	0.61
	BIN	61.45	52.01	54.82	0.12	0.60
IgE	AAC	72.91	88.12	87.23	0.4	0.84
	AAP	80.16	89.46	88.92	0.46	0.89
	CTD	19.27	98.01	93.38	0.24	0.74
	DPC	**77.38**	**94**	**93.02**	**0.55**	**0.9**
	PCP	3.88	99.61	93.99	0.11	0.68
	BIN	10.76	97.28	92.20	0.11	0.60
IgA	AAC	41.2	98.85	98.37	0.3	0.77
	AAP	40.07	99.13	98.64	0.33	0.75
	CTD	12.36	99.93	99.21	0.27	0.71
	DPC	**40.82**	**99.67**	**99.18**	**0.45**	**0.76**
	PCP	3	99.98	99.18	0.13	0.67
	BIN	13.86	99.57	98.87	0.17	0.67

Sen: Sensitivity; Spec: Specificity

Table S7. The performance of SVM based models developed using various features for predicting antibody specific B-cell epitopes on RealVar dataset.

Class of Antibodies	Feature	Sen	Spec	Acc	MCC	AUC
IgG	AAC	44.9	84.36	71.77	0.32	0.7
	AAP	**55.30**	**84.41**	**75.74**	**0.41**	**0.79**
	CTD	25.04	92.9	72.68	0.25	0.67
	DPC	56.58	82.19	74.02	0.39	0.73
	PCP	22.94	89.17	69.44	0.16	0.61
IgE	AAC	42.93	94.27	91.07	0.33	0.7
	AAP	**52.81**	**97.12**	**94.51**	**0.5**	**0.89**
	CTD	8.82	99.36	94.04	0.18	0.74
	DPC	54.76	95.91	93.35	0.47	0.86
	PCP	40.21	83.67	81.11	0.15	0.68
IgA	AAC	34.99	99.43	98.74	0.37	0.8
	AAP	30.02	99.52	98.78	0.34	0.73
	CTD	23.6	99.57	98.95	0.27	0.71
	DPC	**42.18**	**99.01**	**98.41**	**0.36**	**0.84**
	PCP	11.61	99.89	99.16	0.23	0.66

Sen: Sensitivity; Spec: Specificity

Table S8. Performance of various WEKA classifiers developed using various features as input for different classes of epitopes on BalanceFix dataset.

Features	Parameter	IgG			IgE			IgA		
		SMO	**IBk**	**RF**	**SMO**	**IBk**	**RF**	**SMO**	**IBk**	**RF**
AAC	**ACC**	63.34	64.51	62.76	76.61	75.59	74.67	71.28	66.85	68.91
	MCC	0.27	0.29	0.26	0.53	0.52	0.5	0.46	0.41	0.41
	AUC	0.63	0.64	0.66	0.76	0.76	0.8	0.72	0.66	0.73
AAP	**ACC**	68.14	70.05	67.27	81.5	79.53	80.81	70.6	60.23	71.1
	MCC	0.35	0.4	0.35	0.64	0.6	0.62	0.49	0.23	0.48
	AUC	0.67	0.7	0.72	0.81	0.8	0.86	0.71	0.59	0.74
DPC	**ACC**	67.40	70.07	67.14	81.34	79.61	80.89	70.6	63.30	71.16
	MCC	0.35	0.4	0.34	0.64	0.6	0.62	0.49	0.27	0.49
	AUC	0.67	0.7	0.72	0.81	0.8	0.86	0.71	0.59	0.76
CTD	**ACC**	55.31	58.09	55.19	61.42	63.02	63.18	62.73	62.92	62.92
	MCC	0.1	0.16	0.11	0.23	0.26	0.27	0.28	0.26	0.34
	AUC	0.55	0.58	0.56	0.61	0.63	0.67	0.62	0.63	0.68
PHY	**ACC**	56.76	56.50	55.96	61.55	63.18	64.07	63.48	66.10	64.79
	MCC	0.14	0.13	0.12	0.23	0.27	0.28	0.28	0.32	0.31
	AUC	0.57	0.56	0.57	0.62	0.63	0.68	0.63	0.66	0.68

Table S9. The performance of various WEKA classifiers developed using various features as input for different classes of epitopes on BalanceVar dataset.

Features	Parameter	IgG			IgE			IgA		
		SMO	**IBk**	**RF**	**SMO**	**IBk**	**RF**	**SMO**	**IBk**	**RF**
AAC	**ACC**	63.53	62.54	64.36	71.00	70.7	70.47	71.68	68.17	70.98
	MCC	0.27	0.25	0.29	0.42	0.42	0.41	0.47	0.37	0.44
	AUC	0.63	0.62	0.68	0.72	0.71	0.76	0.73	0.68	0.76
AAP	**ACC**	67.81	68.27	68.08	77.18	76.94	76.86	71.46	68.66	70.73
	MCC	0.36	0.37	0.36	0.55	0.55	0.54	0.47	0.42	0.42
	AUC	0.69	0.69	0.72	0.77	0.77	0.82	0.68	0.69	0.72
DPC	**ACC**	67.96	68.25	67.71	77.26	76.96	75.99	71.59	70.12	71.83
	MCC	0.36	0.37	0.36	0.55	0.55	0.52	0.49	0.36	0.44
	AUC	0.68	0.69	0.73	0.77	0.77	0.82	0.71	0.70	0.76
CTD	**ACC**	62.5	60.96	59.98	64.78	63.12	62.95	65.61	64.51	66.34
	MCC	0.25	0.22	0.2	0.3	0.26	0.26	0.32	0.29	0.33
	AUC	0.62	0.58	0.63	0.65	0.63	0.67	0.65	0.65	0.71
PHY	**ACC**	63.23	60.41	62.04	62.44	60.72	64.2	66.1	65.98	68.66
	MCC	0.27	0.21	0.24	0.25	0.22	0.31	0.32	0.32	0.37
	AUC	0.63	0.59	0.65	0.62	0.61	0.71	0.66	0.65	0.74