PROTEINS: Structure, Function, and Bioinformatics 65:40–48 (2006)

SHORT COMMUNICATION

Prediction of Continuous B-Cell Epitopes in an Antigen Using Recurrent Neural Network

Sudipto Saha and G. P. S. Raghava*
Institute of Microbial Technology, Chandigarh, India

ABSTRACT **B-cell epitopes play a vital role in the development of peptide vaccines, in diagnosis of diseases, and also for allergy research. Experimental methods used for characterizing epitopes are time consuming and demand large resources. The availability of epitope prediction method(s) can rapidly aid experimenters in simplifying this problem. The standard feed-forward (FNN) and recurrent neural network (RNN) have been used in this study for predicting B-cell epitopes in an antigenic sequence. The networks have been trained and tested on a clean data set, which consists of 700 non-redundant B-cell epitopes obtained from Bcipep database and equal number of non-epitopes obtained randomly from Swiss-Prot database. The networks have been trained and tested at different input window length and hidden units. Maximum accuracy has been obtained using recurrent neural network (Jordan network) with a single hidden layer of 35 hidden units for window length of 16. The final network yields an overall prediction accuracy of 65.93% when tested by five-fold cross-validation. The corresponding sensitivity, specificity, and positive prediction values are 67.14, 64.71, and 65.61%, respectively. It has been observed that RNN (JE) was more successful than FNN in the prediction of B-cell epitopes. The length of the peptide is also important in the prediction of B-cell epitopes from antigenic sequences. The webserver ABCpred is freely available at www.imtech.res.in/raghava/abcpred/. Proteins 2006; 65:40–48.** **© 2006 Wiley-Liss, Inc.**

Key words: ABCpred; prediction; B-cell epitopes; recurrent neural network; web server

INTRODUCTION

The antigenic regions of a protein that are recognized by the binding sites or paratope of immunoglobulin molecules are called B-cell epitopes. When such specific binding (between epitope of an antigen and paratope of an antibody) is observed experimentally, the particular immunoglobulin establishes the epitope nature of a protein. Epitopes are thus relational entities that can be defined only in a functional sense (i.e. in an immunoassay) by the binding of complementary paratopes.[1] These epitopes play an important role in the designing of peptide-based vaccines and also in the diagnosis of diseases.[2–4] B-cell epitopes are also important for allergy research and in determining the cross-reactivity of IgE-type epitopes of allergens.[5–7] These epitopes may be linear (continuous) or conformational (discontinuous). When linear synthetic peptides are found to cross-react with anti-protein antibodies or when they are able to induce antibodies that cross-react with the parent protein, then these peptides are labeled as linear (continuous) epitopes.[8] The protective linear B-cell epitopes may lead to the synthesis of the efficient peptide vaccine against viral disease.[9] A dominant linear B-cell epitope is used as the target of neutralizing antibody responses in autoimmune diseases.[10] A discontinuous or conformational epitope is composed of several disparate sequences stretches, which are spatially contiguous. These sequences form a compact accessible region when the protein is folded. Deciphering these epitopes is a difficult task, but can give insight into the structural basis of antigen-antibody recognition.[11] Recently, Conformational epitope prediction (CEP) server has been developed for the prediction of conformational epitopes using 3D structural data of protein antigens.[12]

Prediction of immunogenic epitopes remains vital and challenging task using bioinformatic tools. The inherent complexity of antigen recognition complicates epitope prediction.[13] In the past, number of algorithms have been developed for predicting the continuous B-cell epitopes based on physico-chemical properties of amino acids,[14] but their rate of successful prediction is not very high. The commonly used properties for the prediction

Grant sponsors: Council of Scientific and Industrial Research (CSIR); Department of Biotechnology (DBT), Government of India.

*Correspondence to: Dr. G.P.S. Raghava, Scientist, Institute of Microbial Technology, Sector 39A, Chandigarh, India.
E-mail: raghava@imtech.res.in

Received 31 May 2005; Revised 7 March 2006; Accepted 24 April 2006

Published online 7 August 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21078

are hydrophilicity (Parker method),[15] flexibility (Karplus method),[16] accessibility (Emini method),[17] and turns (Pellequer method),[18] which had been correlated with the location of continuous epitopes in a few well-characterized proteins. All the prediction calculations are based on the propensity scales for each of the 20 amino acids and these scales describe the tendency of each residue to be associated with the physico-chemical properties. Based on these properties, few computer programs are developed to assist the user in predicting epitopes in an antigenic sequence. For example, PREDITOP[19] uses 22 normalized scales, corresponding to hydrophilicity, accessibility, flexibility, and secondary structure propensities. Another program, PEOPLE[20] have used the combined prediction methods, taking into account of physico-chemical properties such as β turns, surface accessibility, hydrophilicity, and flexibility. A recent program BEPITOPE[21] aims at predicting the continuous protein epitopes and searching for patterns either in a single protein or on a complete translated genome. An assessment of predictive value of algorithms based on eight physico-chemical parameter scales has been studied for locating of 29 continuous epitopes in four model proteins. The results showed that the percentage of correct prediction varies between 40–68% depending upon the cut-off level of the threshold and the model protein.[22] Van Regenmortel and Pellequer have compared the prediction efficacy of 22 different scales, taking into account both the correct and incorrect predictions, and showed that the prediction accuracy was not >50–60%.[23] Recently, we have studied the performance of various methods on clean and large data set of B-cell epitopes.[24] Based on our observation we also developed a combined method BcePred (www.imtech.res.in/raghava/bcepred/) for predicting the B-cell epitopes using various physico-chemical properties. The performance of the physico-chemical properties varies from 52.9 to 57.5%, whereas combined methods shows 58.7% accuracy.[24] Blythe and Flower found underperformance of the existing 484 amino acid propensity scales while benchmarking B cell epitope prediction.[25]

One of the major problems with existing methods is that they are qualitative rather than quantitative, as most of these methods gave a property plots. In these property plots one can only guess the stretch or region of a protein, which may have B-cell epitope. It is nearly impossible to identify exact region (start and end residue), which can serve as B-cell epitope. To the best of our knowledge, no sophisticated technique like artificial neural network (ANN) has been used for the prediction of B-cell epitopes. The major problem of using machine learning technique is that the input window length has to be fixed, whereas B-cell epitopes sequence vary from 5 to 30 as reported in literature (Bcipep database). This is the reason why machine learning techniques such as ANN were not developed in the past. In this study, an attempt has been made to develop a method using ANN for the prediction of B-cell epitopes. To overcome the problem of varied length of B-cell epitope, we examined all the available B-cell epitopes and observed that most of the epitopes have length of about 20 amino acids or less (as reported in literature), and only few epitopes have length >20 amino acids. Thus in our study, we only considered the epitopes of length 20 amino acids or less for developing our method. It does not mean that B-cell epitope have length 20 amino acids or less. Adding or removing a few residues at the terminals of B-cell epitopes has generated the fixed length patterns. The additional residues were taken from the parent/original antigenic sequences. To train any prediction method, particularly machine learning techniques, one might require both positive (e.g. B-cell epitopes) as well as negative (e.g. non B-cell epitope) datasets. We created the positive B-cell epitope data set from the Bcipep database. In the absence of any proven non-epitopes, we took random peptides generated from proteins as non-epitopes in this study. The creation of negative dataset from random peptides/proteins is a common practice in the literature.[26–28] Though these random peptides may also have B-cell epitopes, we assumed that their probability is low.[29,30]

Both standard feed forward network (FNN) and recurrent neural network (RNN) were applied in the present study for predicting the B-cell epitope in an antigenic sequence. Different window length, that is, 10–20 with two amino acids interval, were used to achieve high accuracy of the B-cell epitope prediction. It was observed that the prediction of B-cell epitopes using RNN was more accurate than FNN.

METHODS

The Data Set

B-cell epitopes have been obtained from Bcipep database[31] that contains 2479 continuous epitopes. To train any machine learning technique one need to have fixed length pattern whereas B-cell epitopes have varying length. We examined the length of B-cell epitopes and observed that large number of epitopes have length less than 20 amino acid (~90%). Thus we discarded all epitopes having length more than 20 residues in order to fix the size of the pattern. We are not justfying that the 20 residues are optimized length for B-cell epitopes but this is the practical aspect to handle the problem of B-cell epitope prediction. We also removed identical epitopes from our dataset to remove any biasness in the prediction. Final dataset consists of 700 unique B-cell epitopes where the maximum length is 20 amino acids. To generate a negative dataset, we created non-epitopes using random peptides of length 20 residues from the proteins in Swiss-Prot.[32] All the random peptides that are identical to B-cell epitopes were removed. Finally, we selected 700 random peptides and used them as non B-cell epitope dataset. Thus our dataset consists of 700 B-cell epitopes and 700 non B-cell epitopes (random peptides).

Creation of Fixed Length Pattern

In this study, we considered the B-cell epitopes of length only 20 amino acids or less to fix the size of the

TABLE I. Creation of Fixed Length Patterns of 20 or Less Than 20 Amino Acids from B-cell Epitopes

Window length/peptide	**AEFPLDIT**[a] (8 amino acid length)	**ACVPTDPNPQEVVLVNVTEN**[b] (20 amino acid length)
20	PKGYVG**AEFPLDIT**AGTEAA	**ACVPTDPNPQEVVLVNVTEN**
18	KGYVG**AEFPLDIT**AGTEA	**CVPTDPNPQEVVLVNVTE**
16	GYVG**AEFPLDIT**AGTE	**VPTDPNPQEVVLVNVT**
14	YVG**AEFPLDIT**AGT	**PTDPNPQEVVLVNV**
12	VG**AEFPLDIT**AG	**TDPNPQEVVLVN**
10	G**AEFPLDIT**A	**DPNPQEVVLV**

[a]Patterns of different length generated from an epitope of eight amino acids.
[b]Patterns of different length generated from an epitope of 20 amino acids.

pattern (upper limit). If the epitope length is less than 20 amino acids, then the length is increased by introducing equal number of residues at both terminals derived from its original antigenic sequence. For example, if any peptide is having length of 8 amino acids, then we added 6 neighbor residues at its both terminals (See Table I). These neighbor residues were obtained from its original antigenic sequence.

Size of the Input Window

After fixing upper limit and creating patterns of length 20, we generated a pattern of different lengths (window length) (10, 12, 14, 16, 18, 20). In this case we removed equal number of residues from both sides of the pattern. Table I shows the different window length obtained for two epitopes of length 8 and 20.

Fivefold Cross-Validation

In this study, a fivefold cross-validation technique has been used, in which the data set is randomly divided into five subsets, each containing an equal number of peptides (280 each). The five subsets have been grouped into training, validation, and testing set. The training set consists of three of these subsets. The network is validated for minimum error on validation set (one set) to avoid overtraining and the network is tested on the remaining set of epitopes called testing set. This process has been repeated five times so that each set was used once for testing. The final prediction results have been the average of five testing sets.

Blind Dataset 1

To evaluate our method on blind or independent dataset, we obtained following four immunogenic proteins from literature. (i) ESAT-6 protein, a low-molecular weight protein secreted by virulent *Mycobacterium tuberculosis*, induced strong antibody response in experimentally infected monkeys. The epitopes were determined using synthesis of overlapping peptides spanning of ESAT-6 protein and by measuring antibody response to ESAT-6 peptides by ELISA in serum samples from monkeys.[33,34] (ii) Ag44 protein is a recombinant antigen expressing the 134 C-terminal RhopH3 residues of *Plasmodium falciparum*. Epitopes was determined using overlapping peptides scanning of the protein and performing ELISA assays.[35] (iii) The nucleocapsid (N) protein of reinderpest virus (RPV) is one of the most abundant and immunogenic viral proteins. Epitope mapping with overlapping peptides revealed three antigenic sites in the regions.[36] (iv) Major surface protein (MSP) 1a of the genus type species *Anaplasma marginale* had been shown to contribute to protective immunity in cattle. Linear B-cell epitopes of MSP1a were mapped using synthetic peptides representing the entire sequence of the protein and the sera from immunized cattle recognized the peptides.[37]

Blind Dataset 2

We also created another independent Blind dataset 2, which consists of total 187 epitopes (128 IgE epitopes obtained from structural database of allergenic proteins (SDAP)[38] and 59 epitopes obtained from Bcipep database[31]), and none of these epitopes were used in the training or testing of ABCpred algorithm. This dataset consists of 109 epitopes having less than 16 residues. To create a pattern of 16 residues, we added equal number of residues on both terminals of these epitopes from its original sequence. We also generated 200 random 16mer peptides from non allergen dataset of Bjorklund et al.[39] and used as non-epitopes. In summary, Blind dataset 2 consists of 187 epitopes and 200 non-epitopes.

Neural Network

In this study, FNN and partial RNN with a single hidden layer have been used. Initially, FNN has been tried, since it is commonly used in the ANN. However, FNN did not yield any satisfactory result and prompted us to try for RNN (Jordan network). Both the networks have been trained using back-propagation algorithm and with various window lengths from 10 to 20 residues. The target output consists of a single binary number and is one or zero (B-cell epitopes or non-epitopes). The final Jordan network has input window of 16 residues and have 35 units in a single hidden layer. For detailed description of Jordan network see supplementary information at www.imtech.res.in/raghava/abcpred/ABC_method.html.

The publicly available free simulation packages SNNS, version 4.2, from Stuttgart University has been used to implement the neural networks.[40] It allows

incorporation of the resulting network into an ANSI C function for use in the stand-alone code. At the start of each simulation, the weights are initialized with random values. The training is carried out by using error back-propagation, with a sum of square error function.[41] The magnitude of the error sum in the test and training set is monitored in each cycle of the training. The ultimate number of cycles is determined when the network converges. During testing, a cut off value is set for each network, and the output produced by the network is compared with the cutoff value. If the output value is greater than the threshold value, then that peptide is predicted as B-cell epitope, otherwise as a non-epitope. For each network, the cutoff value is adjusted so that it yields the highest accuracy for that network. In this study we used uniform/same parameters for learning of five networks on different training sets during fivefold cross validation. It means we have not optimized performance of networks for individual test sets, instead we optimized networks in order to get best average accuracy. We tried different network parameters during the training to get the overall best performance (average accuracy) over five sets. In other words, our best result was achieved by maintaining uniform parameters over the five subsets.

Performance Measure

Threshold-dependent measure

We used commonly used parameter to evaluate the performance of method. The evaluation of performance was at peptide or epitope level and not at residue level. Five parameters have been used in the present work to measure the performance of prediction method. Following is the brief description of the parameters: (1) Q_{sens} (sensitivity) is the percent of epitopes that are correctly predicted as epitopes; (2) Q_{spec} (specificity) is the percent of epitopes correctly predicted as non-epitopes; (3) Q_{acc} (accuracy) is the proportion of correctly predicted peptides; (4) Q_{ppv} (positive prediction value) is the probability that a predicted epitope is infact an epitope; and (5) Matthew's correlation coefficient (MCC) were also calculated. The parameters can be calculated by the following equations.

$$Q_{\text{sens}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%$$

$$Q_{\text{spec}} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\%$$

$$Q_{\text{acc}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\%$$

$$Q_{\text{ppv}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

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

Where TP and FN refer to true positive and false negatives, TN and FP refer to true negatives and false positives.

Threshold-independent measures

One problem with the threshold-dependent measure is that they measure the performance on a given threshold. It is difficult to assess the overall performance of method using these threshold-dependent parameters. The ROC is a threshold-independent measure that was developed as a signal processing technique. For a prediction method, ROC plot is obtained by plotting all sensitivity values (true-positive fraction) on the *y*-axis against their equivalent (1-specificity) values (false-positive fraction) on the *x*-axis. The area under the ROC curve is taken as an important index because it provides a single measure of overall accuracy that is not dependent on a particular threshold.[42] It measures discrimination, the ability of a method to correctly classify B-cell epitopes and non-epitopes.

RESULTS

All the methods have been trained and tested using fivefold cross-validation. The prediction performance measures have been averaged over five sets. First, we trained and tested our method using FNN for different window lengths (input units) like 10, 12, 14, 16, 18, and 20 (See Table I). The performance of FNN at different window lengths with single layer of hidden unit 35 at optimum/default threshold 0.5 is shown in Table II. The

TABLE II. The Performance of Our Neural Network with FNN at Optimum/Default Threshold (0.5)

Window size	Sensitivity (%)	Specificity (%)	PPV (%)	Accuracy (%)	MCC
10	48.14	52.71	50.53	50.43	0.0088
12	53.00 (54.71)[a]	52.14 (54.71)[a]	52.54 (54.72)[a]	52.57 (54.71)[a]	0.0515
14	51.43	55.00	53.17	53.21	0.0645
16	53.29 (55.86)[a]	56.57 (54.29)[a]	55.10 (55.20)[a]	54.93 (55.07)[a]	0.0859
18	51.43	54.57	52.92	53.00	0.0602
20	54.43	59.14	57.28	56.79	0.1374

These results were obtained by FNN using single hidden layer of 35 units.
[a]Indicates maximum percentage at hidden units 10.

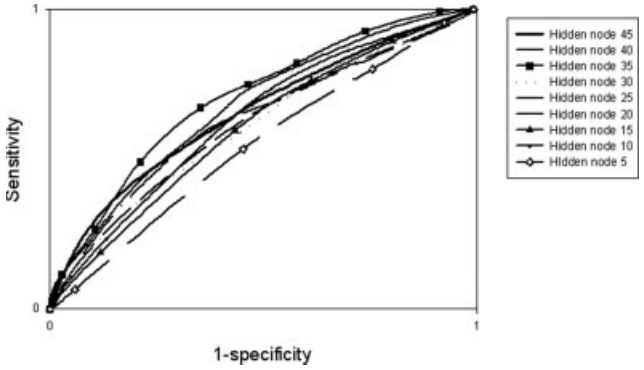

Fig. 1. The overall performance of our method with RNN for window size 16. This ROC plot was obtained between sensitivity (*y*-axis) and 1-specificity (*x*-axis) for RNN at different thresholds from 0.1 to 1.0 at interval of 0.1.

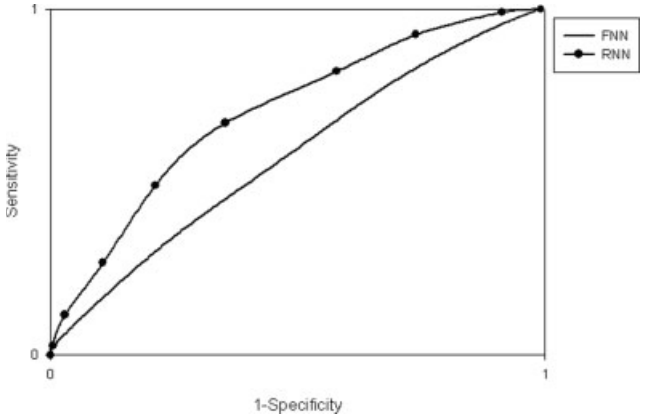

Fig. 2. ROC plot of two neural networks FNN and RNN used in this study at window size 16 and hidden units 35.

maximum performance achieved by FNN varied from 50.43% (nearly random) to 56.79%. We tried number of options, including layers hidden units etc., but the accuracy does not improve further with FNN (data not shown).

The accuracy of the method improved significantly (P value = 0.01732) when we implemented RNN for training and testing (at 0.02 level). The overall performance (ROC plot) of RNN at various thresholds for window length 16 is shown in Figure 1. ROC plot was obtained by plotting all sensitivity values on the *y*-axis against (1-specificity) on the *x*-axis for 0.1–1.0 thresholds at interval of 0.1 (See Methods). The best performance of RNN was at hidden unit 35 with singly hidden layer (see Fig. 1). We also compared the overall performance (ROC plot) of FNN and RNN at hidden unit 35 with window length 16 and observed that RNN was better than FNN for whole range (see Fig. 2). These results clearly indicate the superiority of RNN over FNN in the prediction of B-cell epitopes. We achieved average accuracy, 65.93%; sensitivity, 67.14%; specificity, 64.71%; and MCC, 0.3187 using RNN at threshold 0.5. The learning parameters were same for all five RNN models (e.g., SSE 0.0005, cycles 5000, JE order; hidden nodes 35) in fivefold cross-validation. The accuracy at threshold 0.5 for five test sets was 58.57% (Set 1), 73.57% (Set 2), 72.14% (Set 3), 68.93% (Set 4), and 56.43% (Set 5). We used best RNN model in our server. The sensitivity, specificity, PPV, accuracy, and MCC at different window lengths using RNN are shown in Table III.

Testing of ABCpred server on Blind Dataset 1

To evaluate the performance of ABCpred server, we compute its predictive performance on a blind dataset (Protein sequences not used in the development of ABCpred algorithm). For this purpose, four recently experimentally annotated proteins were obtained from the literature. We predicted the B-cell epitopes in these proteins using ABCpred server at default parameters. The B-cell epitopes (predicted as well as experimentally determined) were mapped on the protein along its amino acid sequence. The B-cell epitopes predicted by ABCpred server in ESAT-6,[33,34] Ag44 protein[35] and MSP1a,[37] Rinderpest virus protein[36] are shown in Figure 3(a,b), respectively. The predicted peptides are displayed rank-wise based on scores obtained by the trained recurrent neural network. All the peptides shown in the figure are at default threshold value (0.5) and window length 16 with overlapping filter. In case of ESAT-6, there was totally four experimentally determined epitopes, our server predicted seven epitopes in this protein. Our four predicted region were in same region in sequence where experimentally determined epitopes (three) were there. Fifth epitope cover nearly half of the fourth B-cell epitope. These results indicate that the server has the ability to detect the potential regions that contain B-cell epitopes, with significant accuracy. However, the server also has lot of over prediction (or false positive) and cannot predict boundary of B-cell epitopes. One of the reasons for the poor prediction is due to the fact that B-cell epitopes do not have any fixed length and we are using a window of fixed length. Also it is not necessary that epitopes determined experimentally have correct boundaries, because in experiment they tried limited peptides (not all peptides of all possible length). The similar trend was observed for other proteins; see Figure 3(a,b) for detail. Overall, the results indicate that the performance of the method is much better than random in real life.

Testing of ABCpred server on Blind Dataset 2

The performance of ABCpred has been evaluated on Blind dataset 2, which consists 187 B-cell epitopes and 200 non-epitopes (16mer random peptides). In case, if the B-cell epitope have more than 16 residues, we examined all overlapping 16mers and if any 16mer have score more than the threshold then whole sequence is predicted as B-cell epitope. As shown in Table IV, we achieved sensitivity of 71.66%, specificity of 61.50%, and

TABLE III. The Performance of Our Method Using RNN at 0.5 Threshold Using Single Hidden Layer of 35 Units

Window Size	Sensitivity (%)	Specificity (%)	PPV (%)	Accuracy (%)	MCC
10	58.71	64.14	61.78	61.43	0.2293
12	53.57	61.71	58.30	57.64	0.1534
14	52.43	65.29	60.12	58.86	0.1786
16	*67.14*	*64.71*	*65.61*	*65.93*	*0.3187*
18	58.70	65.0	62.06	61.86	0.2373
20	57.14	71.57	66.51	64.36	0.2871

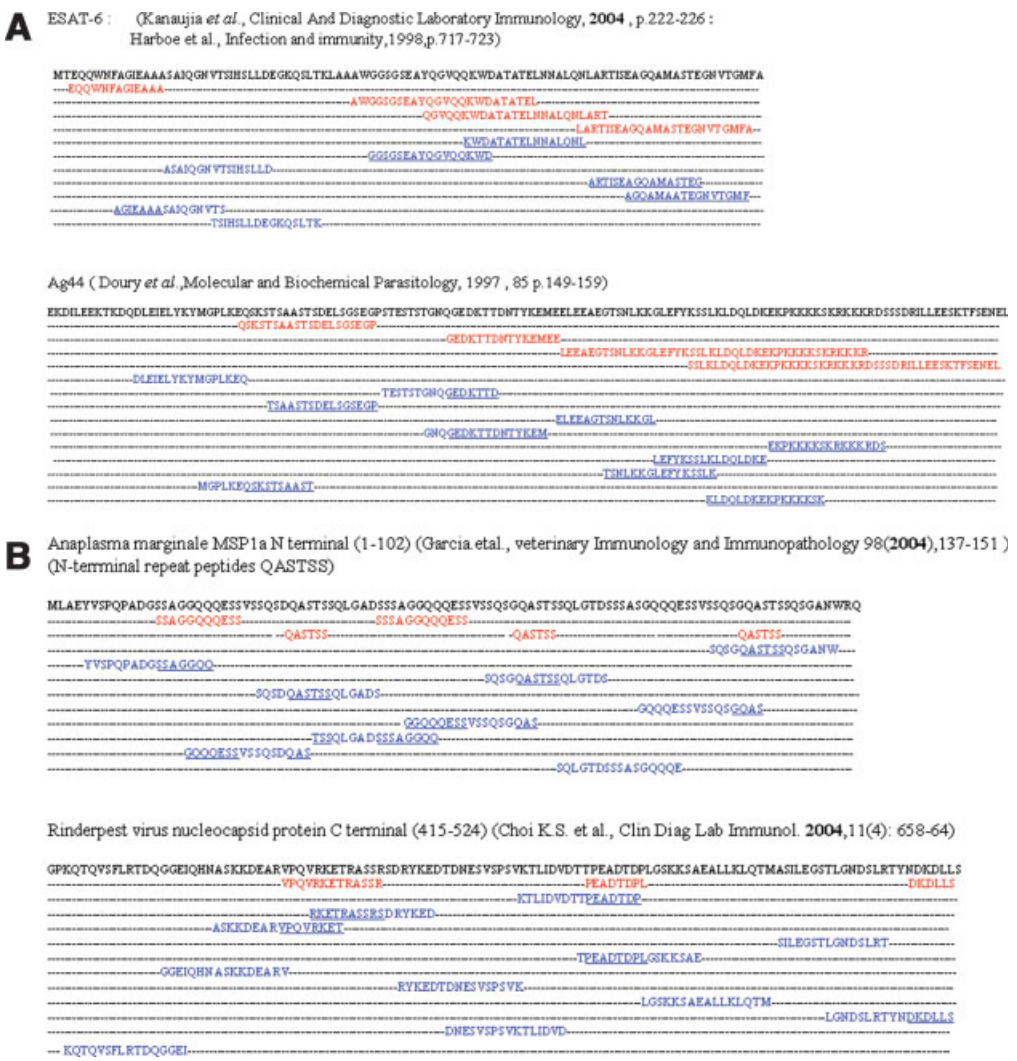

Fig. 3. Comparative epitope mapping of predictions by ABCpred server against experimental data on set of 4 proteins (**a**) *Mycobacterium tuberculosis* ESAT-6 protein and *Plasmodiun falciparum* Ag44; (**b**) Rinderpest nucleocapsid protein C terminal and *Anaplasma marginale* MSP1a N terminal. Red residues are reported in the literature as immunogenic, blue residues are predicted by ABCpred server as epitope, and underlined residues are correctly predicted.

accuracy of 66.41% at default threshold value of 0.5. The maximum accuracy of 69.25% was achieved using ABCpred at threshold value of 0.6. Similarly, we evaluated the performance of Karplus method using flexibility and achieved the maximum accuracy of 59.43% at threshold value of 1.50 (Table V). We achieved maximum accuracy of 61.49% by Parker method using hydrophilicity scale at the threshold value of 2.00 (Table VI). These results demonstrate that the ABCpred can predict B-cell epitopes with reasonably high accuracy.

TABLE IV. The Performance of ABCpred Server on Blind Data Set 2

Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.1	99.47	1.00	48.58
0.2	95.72	7.00	49.87
0.3	92.51	18.00	54.00
0.4	82.89	39.50	60.47
0.5	*71.66*	*61.50*	*66.41*
0.6	60.96	77.00	69.25
0.7	49.73	87.00	68.99
0.8	33.16	95.50	65.37
0.9	4.81	99.50	53.75
1.0	0.00	100.00	51.68

TABLE V. The Performance of Karplus Method Based on Flexibility on Blind Data Set 2

Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.00	100.00	0.00	48.32
0.50	99.47	5.00	50.65
1.00	95.18	20.50	56.59
1.50	*78.60*	*42.00*	*59.43*
2.00	50.27	63.50	57.11
2.50	23.53	79.00	52.19
3.00	4.81	93.50	50.65

TABLE VI. The Performance of Parker Method Based on Hydrophilicity on Blind Data Set 2

Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.00	100.00	0.00	48.32
0.50	99.47	6.00	51.16
1.00	95.72	20.50	56.85
1.50	81.81	41.50	60.98
2.00	*58.82*	*64.00*	*61.49*
2.50	26.74	85.50	57.11
3.00	4.28	98.50	52.97

TABLE VII. The Performance of Various Physico-Chemical Properties in Predicting B-cell Epitope Prediction and ABCpred

Physico-chemical properties/methods	Accuracy	Sensitivity	Specificity
Hydrophilicity[15]	54.47	33.04	76.90
Flexibility[16]	57.53	47.42	67.64
Accessibility[17]	55.49	65.01	45.97
Turns[18]	52.92	17.01	88.82
Antigenic scale[44]	55.59	58.99	52.19
Polarity[24]	54.08	27.50	80.66
Surface[4]	55.73	37.12	74.34
Best combination[24]	58.70	56.07	61.32
ABCpred (window length 16)	65.93	67.14	64.71

Comparison with Existing Methods

It is important to compare the performance of newly developed method with the existing methods. Following is the brief description of major B-cell epitopes prediction methods: (i) Hopps and Woods method is based on analysis of 12 proteins[43]; (ii) Parker et al. method use the modified hydrophilic scale[15]; (iii) Karplus and Schulz developed a method using flexibility scale for predicting the B-cell epitopes[16]; (iv) Emini et al. developed a method using surface accessibility of the amino acids[17]; (v) Kolaskar and Tongaonkar derived their own scale of antigenicity based on the frequency of residues[44]; (vi) Pellequer et al. uses turn scales, which they derived from 87 protein structures[18]; (vii) Pellequer and Westhof developed a program PREDITOP that uses the 22 normalized scales[19]; (viii) PEOPLE uses combination of physico-chemical properties,[20] and (ix) BEPITOPE is a comprehensive program, which allows to combine two or more parameters.[21] It is not practically possible to evaluate all these methods and programs in their original form, because of number of reasons that includes nonavailability of the methods and most of them are qualitative methods. To evaluate the performance of existing methods, we evaluated the performance of various physico-chemical properties of residues, rather than methods itself[24] (http://www.imtech.res.in/raghava/bcepred/). They evaluate major residue properties (hydrophilicity[15]; flexibility[16]; accessibility,[17] etc.), which are used in most of the existing method. As shown in Table VII, the performance of the physico-chemical properties varies from 52.92 to 57.53%. The maximum accuracy of 58.70% has been achieved using the combination of properties.[24] We achieved maximum accuracy of 65.93% using method ABCpred described in this study, which is better than the accuracy achieved using any single property or by combination. We calculated *P*-value to test whether the accuracy of ABCpred is significantly better than accuracy of property based methods. We got *P*-value of 0.012 between accuracies of ABCpred and Karplus[16] method (flexibility) and *P*-value of 0.011 between ABCpred and Parker[15] method (hydrophilicity) accuracies on five test sets at 0.05 level. These results show that the performance of ABCpred is significantly better than the methods based on physico-chemical properties.

Web Server

Based on our observations, a server ABCpred, which allows users to predict continuous B-cell epitopes in a protein sequence, has been developed. Users can submit an amino acid sequence and can select any window length as well as threshold to be used for epitopes prediction. It presents the result in overlap display and tabular frame. In case of tabular frame, the server ranked epitopes based on the score obtained from the trained recurrent neural network. The higher score values of the peptides indicates the higher probability to be predicted for an B-cell epitope. The server is accessible from www.imtech.res.in/raghava/abcpred/.

DISCUSSION

The prediction of B-cell epitopes in an antigen sequence is an important and complex problem. Although, most antigenic determinants of proteins are discontinuous, it is possible to mimic epitopes by synthetic peptides.[8] Many algorithms have been developed to predict the location of continuous epitopes in proteins but their rate of successful prediction is low.[23] One of the major problems faced in developing B-cell epitope prediction is the variable length of the epitope. As all machine learning techniques like SVM, ANN, and PEBLS require fixed length of pattern/peptide, it is not possible to use these techniques for B-cell epitope prediction. Though ANN techniques are used to classify the proteins of variable lengths from their amino acid composition (fixed length pattern of 20), it is not possible in case of epitope/peptide where length is too small to compute the composition. All the existing methods are residue property based where first they generate the property plots (e.g. hydrophilicity, flexibility) and then selects the regions in an antigen, which shows the peaks. These regions are assigned as B-cell epitopes. However, these methods are subjective in nature because one does not know the boundaries of epitopes.

In this study, for the first time a systematic attempt has been made to develop a neural network based method for predicting B-cell epitopes. A major problem in this method is the length of B-cell epitope that varies from 5 to 30 residues. The optimal length of a B-cell epitope is not known, unlike T-cell epitope where MHC molecule core prefer 9 amino acids for binding. On other hand, machine learning method requires a fixed length of window for testing and training. An initial examination of all the B-cell epitopes obtained from Bcipep database reveals that most of the epitopes have 20 or less residues. Therefore, in our study we have only used those epitopes that have 20 or less residues. This way we have fixed the upper limit of size of patterns used in this study. Next problem is how to handle epitopes that have residues <20. For epitopes of length less than 20 amino acids, we have generated patterns of length of 20 amino acids by adding neighboring residues both side of the epitope derived from its original sequence. (See Methods; Table I). This way we get a pattern of fixed length of 20 amino acids corresponding to each epitope. We feel that this is one of the best ways to handle this problem. Another problem we faced in this study was obtaining non B-cell epitopes data. Ideally one should have experimentally proven non B-cell epitopes data. Because of lack of such data in the public domain, we generated random peptides of 20 amino acids from proteins in Swiss-Prot database. We are not justifying that all these random peptides are non B-cell epitopes, and it is possible that these random peptides may also have B-cell epitopes. We adopt this strategy of generating non-epitopes (negative examples) as it has been used in number of investigations in past.[26–28] Final data set contains patterns of length 20, with equal number of positive (B-cell epitope) and negative (non epitope) examples. A machine learning technique (ANN) is used for discriminating B-cell epitopes from non-epitopes. Though FNN is a commonly used network, we obtained poor results using FNN. The percentage of accuracy obtained using FNN is lower than existing methods based on physicochemical properties, and for window lengths of 10 and 12, accuracy of FNN is near random (Table II). It has been observed in the past that RNN performs better than FNN in the prediction of secondary structure of proteins.[45] Therefore, we tried RNN in our study and interestingly the performance of RNN is found to be better than FNN (Table III). The performance of RNN based method described in this study also is significantly better than that reported for any existing B-cell epitope prediction methods. The best performance of our method has been achieved when length of epitope is 16 residues. However, 16 cannot be considered as an ideal length of epitopes as number of epitopes with 15–22 amino acids length have been identified.[46] We also evaluated the performance of our method on blind dataset where we compare the predicted and experimentally determined epitopes in four proteins (not used in testing or training of ABCpred). As shown in Figure 3(a,b), our method was able to predict the experimentally determined epitopes with reasonable accuracy. The performance is much better than random, despite the fact that B-cell epitope prediction is a complex problem. Thus it is worth to use ABCpred server for detecting potential B-cell epitopes in an antigen.

Though we have obtained high prediction accuracy of B-cell epitopes in this study, it has its own limitations. The method described here is not an alternate to existing methods, but will help to complement these methods. A number of assumptions have been made in the algorithm because one cannot directly implement ANN techniques in B-cell epitopes prediction. The aim of this study is to provide an additional quantitative method for B-cell epitopes prediction. The accuracy of method is also not very high, despite our systematic attempts. Users are advised to predict the B-cell epitope in an antigen using all existing methods, including our method, and to find out the regions in antigenic sequences, predicted by most of the methods.

CONCLUSIONS

It was observed that RNN (JE) has been more successful than FNN in prediction of B-cell epitopes. The length of the peptide is also important in prediction of B-cell epitopes from antigenic sequences.

ACKNOWLEDGMENT

We are thankful to Miss Harpreet Kaur for assisting in running SNNS version 4.2.

REFERENCES

1. Van Regenmortel MH. The concept and operational definition of protein epitopes. Philos Trans R Soc Lond B Biol Sci 1989;323: 451–466.
2. Wiesmuller KH, Fleckenstein B, Jung G. Peptide vaccines and peptide libraries. Biol Chem 2001;382:571–579.
3. Zauner W, Lingnau K, Mattner F, Von Gabain A, Buschle M. Defined synthetic vaccines. Biol Chem 2001;382:581–595.
4. Van Regenmortel MH. Pitfalls of reductionism in the design of peptide-cased vaccines. Vaccine 2001;19:2369–2374.
5. Negroni L, Bernard H, Clement G, Chatel JM, Brune P, Frobert Y, Wal JM, Grassi J. Two-site enzyme immunometric assays for

determination of native and denatured β-lactoglobulin. J Immunol Methods 1998;220:25–37.
6. Selo I, Clement G, Bernard H, Chatel J, Creminon C, Peltre G, Wal J. Allergy to bovine β-lactoglobulin: specificity of human IgE to tryptic peptides. Clin Exp Allergy 1999;29:1055–1063.
7. Clement G, Boquet D, Frobert Y, Bernard H, Negroni L, Chatel JM, Adel-Patient K, Creminon C, Wal JM, Grassi J. Epitopic characterization of native bovine β-lactoglobulin. J Immunol Methods 2002;266:67–78.
8. Van Regenmortel MH. Synthetic peptides versus natural antigens in immunoassays. Ann Biol Clin (Paris) 1993;51:39–41.
9. Langeveld JP, Martinez-Torrecuadrada J, Boshuizen RS, Meloen RH, Ignacio CJ. Characterisation of a protective linear B cell epitope against feline parvoviruses. Vaccine 2001;19:2352–2360.
10. Castelletti D, Fracasso G, Righetti S, Tridente G, Schnell R, Engert A, Colombatti M. A dominant linear B-cell epitope of ricin A-chain is the target of a neutralizing antibody response in Hodgkin's lymphoma patients treated with an anti-CD25 immunotoxin. Clin Exp Immunol 2004;136:365–372.
11. Estienne V, Duthoit C, Blanchin S, Montserret R, Durand-Gorde JM, Chartier M, Baty D, Carayon P, Ruf J. Analysis of a conformational B cell epitope of human thyroid peroxidase: identification of a tyrosine residue at a strategic location for immunodominance. Int Immunol 2002;14:359–366.
12. Kulkarni-Kale U, Bhosle S, Kolaskar AS. CEP: a conformational epitope prediction server. Nucleic Acids Res 2005;33:W168–W171. Web server issue.
13. Flower DR. Towards in silico prediction of immunogenic epitopes. Trends Immunol 2003;24:667–674.
14. Pellequer JL, Westhof E, Regenmortel MHV. Predicting location of continuous epitopes in proteins from their primary structures. Methods Enzymol 1991;203:176–201.
15. Parker JMD, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry 1986;25:5425–5432.
16. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen. Naturwissenschaften 1985;72:212,213.
17. Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol 1985;55:836–839.
18. Pellequer J-L, Westhof E, Regenmortel MHV. Correlation between the location of antigenic sites and the prediction of turns in proteins. Immunol Lett 1993;36:83–99.
19. Pellequer JL, Westhof E. PREDITOP: A program for antigenicity prediction. J Mol Graphics 1993;11:204–210.
20. Alix AJ. Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine 1999;18:311–314.
21. Odorico M, Pellequer JL. BEPITOPE: predicting the location of continuous epitope and patterns in proteins. J Mol Recognit 2003;16:20–22.
22. Van Regenmortel MHV, de Marcillac GD. An assessment of prediction methods for locating continuous epitopes in proteins. Immunol Lett 1988;17:95–107.
23. Van Regenmortel MH, Pellequer JL. Predicting antigenic determinants in proteins: looking for unidimensional solutions to a three-dimensional problem? Pept Res 1994;7:224–228.
24. Saha S, Raghava GPS. BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. In: Nicosia G, Cutello V, Bentley PJ, Timis J, editors. ICARIS 2004, LNCS 3239. Berlin: Springer; 2004. pp 197–204.
25. Blythe MJ, Flower DR. Benchmarking B cell epitope prediction: underperformance of existing methods. Prot Sci 2005;14:246–248.
26. Brazma A, Jonassen I, Eidhammer I, Gilbert D. Approaches to the automatic discovery of patterns in biosequences. J Comput Biol 1998;5:279–305.
27. Singh H, Raghava GPS. ProPred1: prediction of promiscuous MHC class-I binding sites. Bioinformatics 2003;19:1009–1014.
28. Singh H, Raghava GPS. PropPred: prediction of HLA-DR binding sites. Bioinformatics 2001;17:1236,1237.
29. Lesenechal M, Becquart L, Lacoux X, Ladaviere L, Baida RC, Paranhos-Baccala G, da Silveira JF. Mapping of B-cell epitopes in a *Trypanosoma cruzi* immunodominant antigen expressed in natural infections. Clin Diagn Lab Immunol 2005;12:329–333.
30. Choi KS, Nah JJ, Ko YJ, Kang SY, Yoon KJ, Jo NI. Antigenic and immunogenic investigation of B-cell epitopes in the nucleocapsid protein of peste des petits ruminants virus. Clin Diagn Lab Immunol 2005;12:114–121.
31. Saha S, Bhasin M, Raghava GPS. Bcipep: A database of B-cell epitopes. BMC Genom 2005;6:79.
32. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
33. Kanaujia GV, Motzel S, Garcia MA, Andersen P, Gennaro ML. Recognition of ESAT-6 sequences by antibodies in sera of tuberculous nonhuman primates. Clin Diagn Lab Immunol 2004;11:222–226.
34. Harboe M, Malin AS, Dockrell HS, Wiker HG, Ulvund G, Holm A, Jorgensen MC, Andersen P. B-cell epitopes and quantification of the ESAT-6 protein of *Mycobacterium tuberculosis*. Infect Immun 1998;66:717–723.
35. Doury JC, Goasdoue JL, Tolou H, Martelloni M, Bonnefoy S, Mercereau-Puijalon O. Characterisation of the binding sites of monoclonal antibodies reacting with the *Plasmodium falciparum* rhoptry protein RhopH3. Mol Biochem Parasitol 1997;85:149–159.
36. Choi KS, Nah JJ, Ko YJ, Kang SY, Yoon KJ, Joo YS. Characterization of immunodominant linear B-cell epitopes on the carboxy terminus of the rinderpest virus nucleocapsid protein. Clin Diagn Lab Immunol 2004;11:658–664.
37. Garcia-Garcia JC, de la Fuente J, Kocan KM, Blouin EF, Halbur T, Onet VC, Saliki JT. Mapping of B-cell epitopes in the N-terminal repeated peptides of *Anaplasma marginale* major surface protein 1a and characterization of the humoral immune response of cattle immunized with recombinant and whole organism antigens. Vet Immunol Immunopathol 2004;98:137–151.
38. Ivanciuc O, Schein CH, Braun W. SDAP: database and computational tools for allergenic proteins. Nucleic Acids Res 2003;31:359–362.
39. Bjorklund AK, Soeria-Atmadja D, Zorzet A, Hammerling U, Gustafsson MG. Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. Bioinformatics 2005;21:39–50.
40. Zell A, Mamier G. Stuttgart neural network simulator, version 4.2. University of Stuttgart, Stuttgart, 1997.
41. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagation errors. Nature 1986;323:533–563.
42. Deleo JM. In: Proceedings of the second international symposium on uncertainity modelling and analysis, IEEE 1993. College Park, MD: Computer Society Press; 1993. pp 318–325.
43. Hopp TP, Woods RK. Predictions of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 1981;78:3824–3828.
44. Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett 1990;276:172–174.
45. Baldi P, Brunak S. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999;15:937–946.
46. Colman PM, Laver WG, Varghese JN, Baker AT, Tulloch PA, Air GM, Webster RG. Three-dimensional structure of a complex of antibody with influenza virus neuraminidase. Nature 1987;326:358.