BcePred: Prediction of Continuous B-cell epitopes in antigenic sequences using physico-chemical properties

Sudipto Saha[1] and G. P. S. Raghava [1]*

[1]Institute of Microbial Technology, Bioinformatics centre
Sector 39A, Chandigarh, INDIA
*Corresponding author: raghava@imtech.res.in

Abstract. A crucial step in designing of peptide vaccines involves the identification of B-cell epitopes. In past, numerous methods have been developed for predicting continuous B-cell epitopes, most of these methods are based on physico-chemical properties of amino acids. Presently, its difficult to say which residue property or method is better than the others because there is no independent evaluation or benchmarking of existing methods. In this study the performance of various residue properties commonly used in B-cell epitope prediction has been evaluated on a clean dataset. The dataset used in this study consists of 1029 non-redundant B cell epitopes obtained from Bcipep database and equally number of non-epitopes obtained randomly from SWISS-PROT database. The performance of each residue property used in existing methods has been computed at various thresholds on above dataset. The accuracy of prediction based on properties varies between 52.92% and 57.53%. We have also evaluated the combination of two or more properties as combination of parameters enhance the accuracy of prediction. Based on our analysis we have developed a method for predicting B cell epitopes, which combines four residue properties. The accuracy of this method is 58.70%, which is slightly better than any single residue property. A web server has been developed to predict B cell epitopes in an antigen sequence. The server is accessible from http://www.imtech.res.in/raghava/bcepred/

Introduction

The antigenic regions of protein that are recognized by the binding sites or paratopes of immunoglobulin molecules are called B-cell epitopes. These epitopes play a vital role in designing peptide-vaccines and in disease diagnosis. These epitopes provide information for the synthesis of peptides that induces cross-reacting antibodies, thereby promoting in the development of synthetic peptide vaccines [1]. The Bioinformatics approach of prediction of immunogenic epitopes remains challenging but vital. The inherent complexity of immune presentation and recognition processes complicates epitope prediction [2]. Number of methods has been developed for predicting B cell epitopes, which are based on physico-chemical properties of the amino acids [3]. Hopps and Woods [4] used hydrophilic analysis (on twelve proteins) to investigate the possibility that at least some antigenic determinants might be associated with stretches of amino acids sequence that contain charged and polar

residue and lack large hydrophobic residue. Parker et al., [5] use the modified hydrophilic scale based on peptide retention times during high-performance liquid chromatography (HPLC) on a reversed-phase column. Karplus and Schulz [6] suggested a link between antigenicity and segmental mobility and developed a method for predicting mobility of protein segments on the basis of the known temperature B factors of the a-carbons of 31 proteins of known structure. They utilize the flexibility scale for predicting the B-cell epitopes. Emini et al., [7] developed method for predicting epitopes based on surface accessibility of the amino acids. Kolaskar and Tongaonkar [8] derived their own scale of antigenicity based on frequency of residues in 169 experimentally known epitopes. Pellequer et al., [9] derived turn scales based on the occurrence of amino acids at each of the four positions
of a turn using a structural database comprised of 87 proteins. The turn scales correctly predicted 70% of the known epitopes.

A number of computer programs have also been developed to assist the users in predicting epitopes in an antigen sequence. Pellequer and Westhof [10] developed program PREDITOP that utilize the 22 normalized scales, corresponding to hydrophilicity, accessibility, flexibility and secondary structure propensities. Another program PEOPLE [11] use combined prediction methods, taking into account physico-chemical properties like b turns, surface accessibility, hydrophilicity and flexibility. PEOPLE have been applied for prediction of only two proteins, tropoelastin and antigen protein P30 of Toxoplasma gondii. The BEPITOPE [12] program aims at predicting continuous protein epitopes and searching for patterns in either a single protein or a complete translated genome. This program provide various options like i) selecting any residue property (e.g. hydrophilicity, flexibility, protein accessibility, turns scale); ii) graphical interface so that users can decide the antigenic region; and iii) combining two and more parameters.

It is not practically possible to evaluate all these methods and programs in their original form because many of these programs are not freely available, while some provides qualitative information (visualization etc.) rather than quantitative. Most of these programs are not automatic where you can give the query sequence and get the predicted epitopes. The selection of threshold is another problem. Thus, we have evaluated the various residue properties, which are commonly used in these existing methods rather than methods as such. As far as authors know, no study has been carried out in the past to evaluate these residues properties on large and uniform dataset of experimentally determined B cell epitopes. In this study, we have also evaluated two new properties, polarity and exposed surface area. The effect of combination of two or more properties on accuracy of predicting of B cell epitopes has also been determined in addition to individual properties.

It has been observed that the combination of two or more properties gives better accuracy than individual property, which agree with previous observations [12]. Based on these observations, a web server has been developed for predicting B cell epitopes in an antigenic sequence. This is almost similar to stand alone computer program BEPITOPE except that this is a web server, which allows on-line computation over Internet.

MATERIALS AND METHODS

Data set: B-cell epitopes have been obtained from Bcipep database [13; See http://www.imtech.res.in/raghava/bcipep/ or http://bioinformatics.uams.edu/mirror/bcipep/], which contains 2479 continuous epitopes, including 654 immunodominant, 1617 immunogenic epitopes. All the identical epitopes and non-immunogenic peptides were removed, yielding 1029 unique experimentally proved continuous B cell epitopes. The dataset covers a wide range of pathogenic group like virus, bacteria, protozoa and fungi. The final dataset consists of 1029 B-cell epitopes and 1029 non-epitopes or random peptides (equal length and same frequency generated from SWISS-PROT [14].

Measure of prediction accuracy: Both threshold dependent and independent measures have been used to evaluate the prediction performance. The threshold dependent measures include standard parameters such as sensitivity, specificity and accuracy. The parameter ROC has been used as threshold independent measure.

Brief description of existing methods:
Parker Method: In this method, hydrophilic scale based on peptide retention times during high-performance liquid chromatography (HPLC) on a reversed-phase column was constructed [5]. A window of seven residues was used for analyzing epitope region. The corresponding value of the scale was introduced for each of the seven residues and the arithmetical mean of the seven residue value was assigned to the fourth, (i+3), residue in the segment.

Karplus Method: In this method, flexibility scale based on mobility of protein segments on the basis of the known temperature B factors of the a-carbons of 31 proteins of known structure was constructed [6]. The calculation based on a flexibility scale is similar to classical calculation, except that the center is the first amino acid of the six amino acids window length, and there were three scales for describing flexibility instead of a single one. In the present study, 3Karplus scale has been used for prediction of the epitope region.

Emini Method: The calculation was based on surface accessibility scale on a product instead of an addition within the window. The accessibility profile was obtained using the formulae

$$\mathrm{Sn} = \left(\prod_{i=1}^{6} \delta_{\mathrm{n+4+i}} \right) (0.37)^{-6} \qquad \textbf{(1)}$$

Where Sn is the surface probability, dn is the fractional surface probability value, and i vary from 1 to 6. A hexapeptide sequence with Sn equal to unity and probability greater than 1.0 indicates an increased chance for being found on the surface [7].

Pellequer Method: This method is based on incidence of b turns [10]. The calculation was based on a turn scale and there were three scales for describing turns instead of a single one. A window of seven residues is used for analyzing epitope region. The corresponding value of the scale was introduced for each of the seven residues and the arithmetical mean of the seven residue value is assigned to the fourth, (i+3), residue in the segment. Gaussian smoothing curve was used, which assigns the residue weights in a window of seven residues (the weights were 0.05/0.11/0.19/0.22/0.19/0.11/0.05).

Kolaskar Method: In this method, 156 antigenic determinants (< 20 amino acids) in 34 different proteins were analyzed [8] to calculate the antigenic propensity (Ap) of residues. This antigenic scale was used to predict the epitopes in sequence.

Exposed surface scale and Polarity scale: The physico-chemical properties like exposed surface (15) and polarity [16] has also been evaluated in this study. A window of seven residues has been used for analyzing the epitope region. The corresponding value of the scale has been introduced for each of the seven residues and the arithmetical mean of the seven residue value is assigned to the fourth, (i+3), residue in the segment.

Normalization procedure

Each property scale consists of 20 values assigned to each of the amino acid types on the basis of their relative propensity as described by the scale. In order to compare the profiles obtained by different methods, normalization of the various scales has been done. We have calculated the average of seven maximum and seven minimum values of a given physico-chemical scale and then calculated the difference between the two. The original values of the each scale are set between +3 to –3 by using the formulae

$$\text{Normalization Score} = \frac{AMS}{DS} * 6 \tag{2}$$

Where AMS refer to Average of seven maximum/minimum values from the physico-chemical scale and DS refer to difference between the maximum and minimum score. Normalization score are set to +3 (Maximum) and –3 (Minimum) by subtracting or adding additional values.

Table 1. The performance of various residue properties in B-cell epitope prediction.

Physico-chemical Properties	**Threshold**	**Sensitivity**	**Specificity**	**Accuracy % (Max)**
Hydrophilicity [1]## (Parker et al., 1986)**	2.00	33	76	54.47
Accessibility[2] (Emini et al., 1985)	2.00	65	46	55.49
Flexibility [3] **(**Karplus and Schulz, 1985)	**1.90**	**47**	**68**	**57.53**
Surface [4] (Janin and Wodak, 1978)	2.40	37	74	55.73
Polarity [5] (Ponnuswamy et al., 1980)	2.30	2.8	81	54.08
Turns [6] (Pellequer et al., 199)	1.90	17	89	52.92
Antigenic Scale [7] (Kolaskar and Tongaonkar, 1990)	1.80	59	52	55.59
[3]+[1]	2.00	53	64	58.31
[3]+[1]+[5]	2.30	50	68	58.70
[3]+[1]+[5]+[4]	**2.38**	**56**	**61**	**58.70**
[3]+[1]+[5]+[4]+[6]	2.38	59	58	58.41
[3]+[1]+[5]+[4]+[6]+[2]	2.38	60	56	57.97

Residue property number, for each property a number is assigned. [3]+[1] means combination of Flexibility and Hydrophilicity.** Reference, which describes property scale used .

RESULTS AND DISCUSSION

We have evaluated seven different physico-chemical scales as implemented in existing epitope prediction methods on the B-cell epitope dataset. The performance of all these methods are threshold dependent, so we select threshold value for each scale at which sensitivity and specificity are nearly equal. The performance of various property scales is shown in Table 1. As shown in Table 1 the performance of all the methods is poor and accuracy varies between 52.92% and 57.53%. It has been observed that flexibility as implemented by Karplus and Schulz [6], relatively perform better than any other property scale used in the past. We observe that some methods have higher sensitivity but lower specificity value or vice-versa (Table 1). This fact makes it difficult to compare the methods objectively. Therefore, we use a single threshold independent measure of performance called the Receiver Operating Characteristics (ROC), to assess the performance of the methods [17]. ROC plot, 1-specificity vs sensitivity from threshold –1.5 to 3 has been computed. It is clear from the ROC plot (Figure 1) that the flexibility property based method performs better in comparison to other methods.

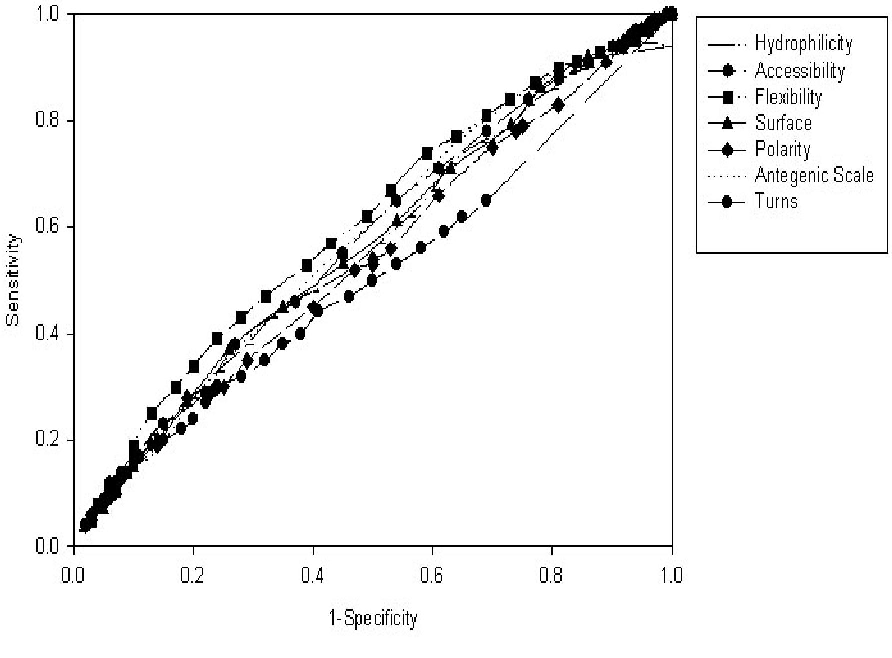
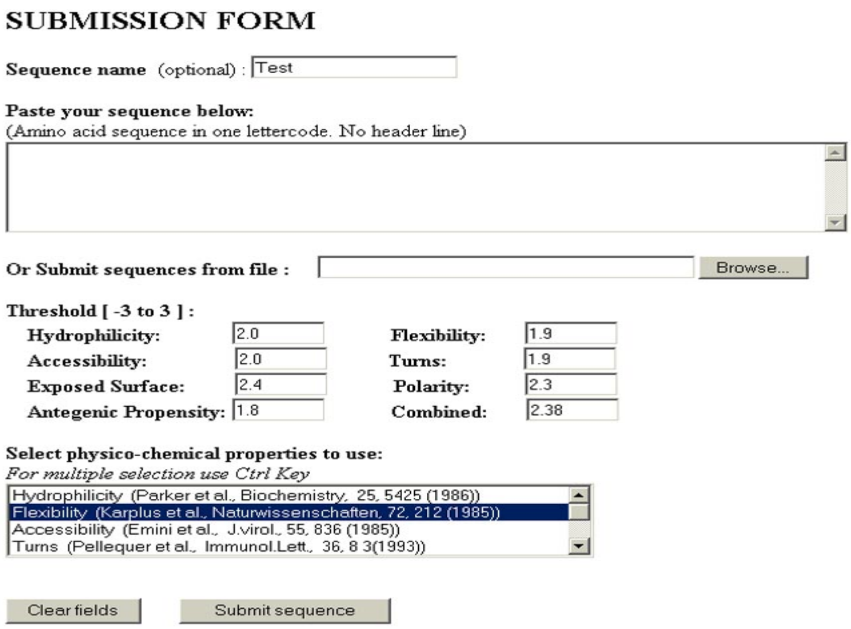


Fig. 1. ROC plot of various residue properties.

In order to see the effect of combination of properties, we have combined the best parameter, flexibility, with other properties one-by-one. It is found that on combination of hydrophilicity and flexibility, the algorithm performs marginally better (accuracy 58.31 %) than any single property or combination of any other two properties. After trying various combinations it is found that combination of properties hydrophilicity, flexibility, polarity and exposed surface performs better than any other combination at a threshold of 2.38. Though combination achieved accuracy 58.70%, sensitivity 56 % and specificity 61% but overall performance is quite poor. The results suggest that for a large dataset the performance of all methods or properties is much below than that is claimed. In past most methods were examined on small set of epitopes and property scales were derived from same epitopes (i.e., same training and testing dataset). It is important to have non-epitopes along with true epitopes to evaluate any threshold dependent method. Earlier non-epitopes were not used in any evaluation. Therefore, most methods never considered the possibility of over prediction. In the present study, we have considered random peptides obtained from SWISS-PROT as non-epitopes. We felt this as necessary since there is no existing database of B-cell non-epitopes.

Web Server

The server BcePred allows user to predict B cell epitopes in protein sequences. As shown in Figure 2a one can submit and can select any residue property or combination of two or more properties as well as threshold to be used for epitope prediction. It presents the results in graphical and tabular frame. An example of graphical output of BcePred is shown in Figure 2b. In case of graphical frame, server plots the residue properties along protein backbone, which assist the users in rapid visualization of B-cell epitope on protein. The peak of the amino acid residue segment above the threshold value (default is 2.38) is considered as predicted B-cell epitope. The tabular output is in the form of a table, which will give the normalized score of the selected properties with the corresponding amino acid residue of a protein along with the maximum, minimum and average values of the combined methods, selected.



SUBMISSION FORM

Sequence name (optional) : Test

Paste your sequence below:
(Amino acid sequence in one lettercode. No header line)

Or Submit sequences from file :

Threshold [-3 to 3] :

Hydrophilicity:	2.0	**Flexibility:**	1.9
Accessibility:	2.0	**Turns:**	1.9
Exposed Surface:	2.4	**Polarity:**	2.3
Antegenic Propensity:	1.8	**Combined:**	2.38

Select physico-chemical properties to use:
For multiple selection use Ctrl Key

Hydrophilicity (Parker et al., Biochemistry, 25, 5425 (1986))
Flexibility (Karplus et al., Naturwissenschaften, 72, 212 (1985))
Accessibility (Emini et al., J.virol., 55, 836 (1985))
Turns (Pellequer et al., Immunol.Lett., 36, 8 3(1993))

Fig. 2 (a). The display of BcePred server submission form

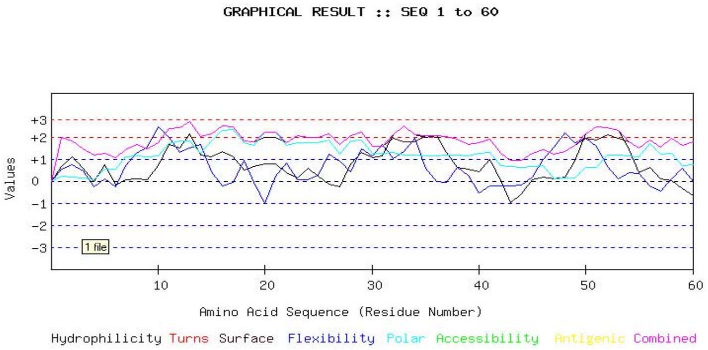


Fig. 2(b). The display of BcePred server graphical output

Acknowledgment

We are thankful to Mr Biju Issac for assisting in graphic display of the output format of BcePred server. We also thank Mr Manoj Bhasin and Miss Harpreet Kaur for their help in writing web programs. We gratefully acknowledge the financial support from the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Govt. of India.

References

1. Nicholson.B.H.: Experimental determination of immunogenic sites. In "Synthetic Vaccines", Published by Blackwell scientific publications, London. (1994) 137-168.

2.Flower,D.R.: Towards in silico prediction of immunogenic epitopes. TRENDS in immunology, 24 (2003) 667-674

3. Pellequer,J.L., Westhof,E. and Regenmortel, M.H.V.: Predicting location of continuous epitopes in proteins from their primary structures. Methods in enzymology, 203 (1991) 176-201.

4. Hopp,T.P. and Woods,R.K.: Predictions of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA., 78 (1981) 3824-3828.

5. Parker,J.M.D., Guo,D. and Hodges,R.S.: 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, 25 (1986) 5425-5432.

6. Karplus,P.A. and Schulz,G.E.: Prediction of chain flexibility in proteins: a
tool for the selection of peptide antigen. Naturwissenschaften, 72 (1985) 212-213.

7. Emini,E.A., Hughes,J.V., Perlow,D.S. and Boger,J.: Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J.Virol., 55 (1985) 836-839.

8. Kolaskar, A.S. and Tongaonkar,P.C.: A semi-emperical method for prediction of antigenic determinants on protein antigens. FEBS, 276 (1990) 172-174.

9. Pellequer,J-L., Westhof,E. and Regenmortel M.H.V.: Correlation between the location of antigenic sites and the prediction of turns in proteins Immunol.Lett., 36, (1993) 83-99.

10. Pellequer.J.L and Wasthof.: PREDITOP: A program for antigenicity prediction. J. Mol. Graphics., 11 (1993) 204-210.

11. Alix AJ. (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine, 18, 311-314.

12. Odorico, M. and Pellequer, J.L. (2003) BEPITOPE: predicting the location of continuous epitope and patterns in proteins. J Mol Recognit., 16, 20-22.

13. Saha,S., Bhasin,M. and Raghava,G.P.S.: Bcipep: A database of B cell epitopes. (2004) (Submitted)

14. Bairoch,A. and Apweiler,R.: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res., 28 (2000) 45-48.

15. Janin,J. and Wodak,S.: Conformation of amino acid side-chains in proteins. J.Mol.Biol., 125 (1978) 357-86.

16. Ponnuswamy,P.K., Prabhakaran,M. and Manavalan,P.: Hydrophobic packing and spatial arrangements of amino acid residues in globular proteins. Biochim.Biophys.Acta., 623, (1980) 301-316.

17. Deleo,J.M.: Proceedings of the Second International Symposium on Uncertainity Modelling and Analusis. IEEE. Computer Society Press,College Park, MD. (1993) 318-325 .